



Robert Bosch GmbH



Bosch Automotive Electrics and Automotive Electronics

Systems and Components,
Networking and Hybrid Drive

5th Edition



Springer Vieweg

Bosch Professional Automotive Information

Robert Bosch GmbH (Ed.)

Bosch Automotive Electrics and Automotive Electronics

Systems and Components,
Networking and Hybrid Drive

5th Edition



Springer Vieweg

Editor:

Robert Bosch GmbH
Automotive Aftermarket (AA/COM3)
Robert Bosch GmbH
Plochingen, Germany

Published by:

© Robert Bosch GmbH, 2007
Postfach 11 29
D-73201 Plochingen
Automotive Aftermarket Division, Business Unit Diagnostics Marketing – Test Equipment
(AA-DG/MKT)
3rd Edition updated and extended, pub. 1999
4th Edition, completely revised and extended, January 2004
5th Edition, completely revised and extended, July 2007
Straight reprint of the 5th edition, published by John Wiley & Sons. Inc. and Bentley Publishers until
2007.

ISBN 978-3-658-01783-5
DOI 10.1007/978-3-658-01784-2

ISBN 978-3-658-01784-2 (eBook)

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Library of Congress Control Number: 2013938481

Springer Vieweg

© Springer Fachmedien Wiesbaden 1999, 2004, 2007, 2013, 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer Vieweg is a brand of Springer DE.
Springer DE is part of Springer Science+Business Media.
www.springer-vieweg.de



In recent decades, the development of the motor vehicle has been marked by the introduction of electronics. At first, electronic systems were used to control the engine (electronic fuel-injection systems), then electronic components entered the domain of driving safety (e.g. antilock brake system, ABS). More recently, completely new fields of application have emerged in the areas of driving assistance, infotainment and communication as a result of continuous advancements in semiconductor technology. Consequently, the proportion of electrics and electronics in the motor vehicle has continuously increased.

A typical feature of many of these new systems is that they no longer perform their function as stand-alone systems but operate in interaction with other systems. If the flow of information between these systems is to be maintained, the electronic control units must be networked with each other. Various bus systems have been developed for this purpose. Networking in the motor vehicle is a topic that receives comprehensive coverage in this book.

Powerful electronic systems not only require information about operating states, but also data from the vehicle's surroundings. Sensors therefore play an important role in the area of automotive electronics. The number of sensors used in the motor vehicle will continue to rise.

The complexity of the vehicle system is set to increase still further in the near future. To guarantee operational reliability in view of this complexity, new methods of electronics development are called for. The objective is to create a standardized architecture for the electrical system/electronics that also offers short development times in addition to high reliability for the electronic systems.

Besides the innovations in the areas of comfort/convenience, safety and infotainment, there is a topic that stands out in view of high fuel prices and demands for cutting CO₂ emissions: fuel consumption. In the hybrid drive, there is great potential for lowering fuel consumption and reducing exhaust-gas emissions. The combination of internal-combustion engine and electric motor enables the use of smaller engines that can be operated in a more economically efficient range. Further consumption-cutting measures are start/stop operation and the recuperation of brake energy (recuperative braking). This book addresses the fundamental hybrid concepts.

The traditional subject areas of automotive electrical systems are the vehicle electrical system, including starter battery, alternator and starter. These topics have been revised for the new edition. New to this edition is the subject of electrical energy management (EEM), which coordinates the interaction of the alternator, battery and electrical consumers during vehicle operation and controls the entire electrical energy balance.

The new edition of the "Automotive Electric/Automotive Electronics" technical manual equips the reader with a powerful tool of reference for information about the level of today's technology in the field of vehicle electrical systems and electronics. Many topics are addressed in detail, while others – particularly the electronic systems – are only presented in overview form. These topics receive in-depth coverage in other books in our series.

 Contents

10 Electrical and electronic systems in the vehicle	168 Electronic components in the vehicle
10 Overview	168 Basic principles of semiconductor technology
13 Motronic-engine management system	172 Passive components
24 Electronic diesel control (EDC)	176 Semiconductor components
32 Lighting technology	186 Manufacture of semiconductor components and circuits
46 Electronic stability program (ESP)	
54 Adaptive cruise control (ACC)	
62 Occupant-protection systems	
70 Basic principles of networking	196 Control units
70 Network topology	196 Operating conditions
74 Network organization	196 Design
76 OSI reference model	196 Data processing
78 Control mechanisms	200 Digital modules in the control unit
	204 Control unit software
82 Automotive networking	208 Automotive sensors
82 Cross-system functions	208 Basics and overview
83 Requirements for bus systems	211 Automotive applications
85 Classification of bus systems	214 Details of the sensor market
85 Applications in the vehicle	215 Features of vehicle sensors
87 Coupling of networks	216 Sensor classification
87 Examples of networked vehicles	218 Error types and tolerance requirements
92 Bus systems	219 Reliability
92 CAN bus	222 Main requirements, trends
106 LIN bus	229 Overview of the physical effects for sensors
112 MOST bus	231 Overview and selection of sensor technologies
122 Bluetooth	
132 FlexRay	
144 Diagnosis interfaces	
152 Architecture of electronic systems	232 Sensor measuring principles
152 Overview	232 Position sensors
155 Vehicle system architecture	259 Speed and rpm sensors
162 Mechatronics	271 Acceleration sensors
162 Mechatronic systems and components	276 Pressure sensors
164 Development methods	279 Force and torque sensors
166 Outlook	288 Flowmeters
	294 Gas sensors and concentration sensors
	298 Temperature sensors
	308 Imaging sensors (video)

310 Sensor types	384 Vehicle electrical systems	486 Electromagnetic compatibility (EMC) and interference suppression
310 Engine-speed sensors	384 Electrical energy supply in the passenger car	486 EMC ranges
312 Hall phase sensors	388 Electrical energy management	487 EMC between different systems in the vehicle
313 Speed sensors for transmission control	390 Two-battery vehicle electrical system	494 EMC between the vehicle and its surroundings
316 Wheel-speed sensors	391 Vehicle electrical systems for commercial vehicles	498 Guarantee of immunity and interference suppression
320 Micromechanical yaw-rate sensors	394 Wiring harnesses	
323 Piezoelectric "tuning-fork" yaw-rate sensor	396 Plug-in connections	
324 Micromechanical pressure sensors		
326 High-pressure sensors	400 Starter batteries	500 Symbols and circuit diagrams
327 Temperature sensors	400 Function and requirements	500 Circuit symbols
328 Accelerator-pedal sensors	402 Design	508 Circuit diagrams
330 Steering-angle sensors	407 Operating principle	519 Designations for electrical devices
332 Position sensors for transmission control	411 Battery designs	521 Terminal designations
335 Axle sensors	418 Battery characteristics	
336 Hot-film air-mass meters	422 Type designations	524 Index of technical terms
339 Piezoelectric knock sensors	423 Practical and laboratory battery testing	Technical terms
340 SMM acceleration sensors	427 Battery maintenance	Abbreviations
342 Micromechanical bulk silicon acceleration sensors	434 Alternators	Background Information
343 Piezoelectric acceleration sensors	434 Electrical power generation in the vehicle	52 ABS versions
344 iBolt™ force sensor	435 Operating principle of the alternator	53 History of radar
346 Torque sensor	443 Voltage regulation	69 Micromechanics
347 Rain/light sensor	448 Overvoltage protection	81 Comparison of bus systems
348 Two-step Lambda oxygen sensors	451 Characteristic curves	175 Miniaturization
352 LSU4 planar wide-band lambda oxygen sensor	453 Power losses	199 Performance of electronic control units
354 Actuators	453 Alternator circuits	297 Piezoelectric effect
354 Electromechanical actuators	455 Alternator designs	383 Greenhouse effect
359 Fluid-mechanical actuators		399 History of the alternator
360 Electrical machines	462 Starting systems	426 History of the battery
366 Hybrid drives	462 Overview	
366 Drive concepts	462 Starter	
370 Operating strategies for electric hybrid vehicles	472 Other types of starter motor	
376 Recuperative brake system	476 Starting systems	
380 Electrical energy accumulators	481 Design	
	484 Overview of the types of starters	

► Authors

Electrical and electronic systems in the vehicle

Dipl.-Ing. Bernhard Mencher;
Dipl.-Ing. (BA) Ferdinand Reiter;
Dipl.-Ing. Andreas Glaser;
Dipl.-Ing. Walter Gollin;
Dipl.-Ing. (FH) Klaus Lerchenmüller;
Dipl.-Ing. Felix Landhäußer;
Dipl.-Ing. Doris Boebel,
Automotive Lighting Reutlingen GmbH;
Dr.-Ing. Michael Hamm,
Automotive Lighting Reutlingen GmbH;
Dipl.-Ing. Tilman Spangler,
Automotive Lighting Reutlingen GmbH;
Dr.-Ing. Frank Niewels;
Dipl.-Ing. Thomas Ehret;
Dr.-Ing. Gero Nenninger;
Prof. Dr.-Ing. Peter Knoll;
Dr. rer. nat. Alfred Kuttenberger.

Networking

Dipl.-Inform. Jörn Stuphorn,
Universität Bielefeld;
Dr. Rainer Constapel,
DaimlerChrysler AG Sindelfingen;
Dipl.-Ing. (FH) Stefan Powolny;
Dipl.-Ing. Peter Häußermann,
DaimlerChrysler AG, Sindelfingen;
Dr. rer. nat. Alexander Leonhardi,
DaimlerChrysler AG, Sindelfingen;
Dipl.-Inform. Heiko Holtkamp,
Universität Bielefeld;
Dipl.-Ing. (FH) Norbert Löchel.

Architecture of electronic systems

Dr. phil. nat. Dieter Kraft;
Dipl.-Ing. Stefan Mischo.

Mechatronics

Dipl.-Ing. Hans-Martin Heinkel;
Dr.-Ing. Klaus-Georg Bürger.

Electronic components

Dr. rer. nat. Ulrich Schaefer.

Control units

Dipl.-Ing. Martin Kaiser;
Dr. rer. nat. Ulrich Schaefer;
Dipl.-Ing. (FH) Gerhard Haaf.

Sensors

Dr.-Ing. Erich Zabler;
Dr. rer. nat. Stefan Finkbeiner;
Dr. rer. nat. Wolfgang Welsch;
Dr. rer. nat. Hartmut Kittel;
Dr. rer. nat. Christian Bauer;
Dipl.-Ing. Günter Noetzel;
Dr.-Ing. Harald Emmerich;
Dipl.-Ing. (FH) Gerald Hopf;
Dr.-Ing. Uwe Konzelmann;
Dr. rer. nat. Thomas Wahl;
Dr.-Ing. Reinhard Neul;
Dr.-Ing. Wolfgang-Michael Müller;
Dr.-Ing. Claus Bischoff;
Dr. Christian Pfahler;
Dipl.-Ing. Peter Weiberle;
Dipl.-Ing. (FH) Ulrich Papert;
Dipl.-Ing. Christian Gerhardt;
Dipl.-Ing. Klaus Miekley;
Dipl.-Ing. Roger Frehoff;
Dipl.-Ing. Martin Mast;
Dipl.-Ing. (FH) Bernhard Bauer;
Dr. Michael Harder;
Dr.-Ing. Klaus Kasten;
Dipl.-Ing. Peter Brenner,
ZF Lenksysteme GmbH, Schwäbisch Gmünd;
Dipl.-Ing. Frank Wolf;
Dr.-Ing. Johann Riegel.

Actuators

Dr.-Ing. Rudolf Heinz;
Dr.-Ing. Robert Schenk.

Hybrid drives

Dipl.-Ing. Michael Bildstein;
Dipl.-Ing. Boyke Richter;
Dr. rer. nat Richard Aumayer;
Dr.-Ing. Karsten Mann;
Dipl.-Ing. Tim Fronzek,
Toyota Deutschland GmbH;
Dipl.-Ing. Hans-Peter Wandt,
Toyota Deutschland GmbH.

Vehicle electrical systems

Dipl.-Ing. Clemens Schmucker;
Dipl.-Ing. (FH) Hartmut Wanner;
Dipl.-Ing. (FH) Wolfgang Kircher;
Dipl.-Ing. (FH) Werner Hofmeister;
Dipl.-Ing. Andreas Simmel.

Starter batteries

Dipl.-Ing. Ingo Koch,
VB Autobatterie GmbH & Co. KGaA, Hannover;
Dipl.-Ing. Peter Etzold;
Dipl.-Kaufm. techn. Torben Fingerle.

Alternators

Dipl.-Ing Reinhard Meyer.

Starting systems

Dipl.-Ing. Roman Pirsch;
Dipl.-Ing. Hartmut Wanner.

Electromagnetic compatibility

Dr.-Ing. Wolfgang Pfaff

and the editorial team in cooperation with the
responsible technical departments at Bosch.

Unless otherwise specified, the above are
all employees of Robert Bosch GmbH.

Electrical and electronic systems in the vehicle

The amount of electronics in the vehicle has risen dramatically in recent years and is set to increase yet further in the future. Technical developments in semiconductor technology support ever more complex functions with the increasing integration density. The functionality of electronic systems in motor vehicles has now surpassed even the capabilities of the Apollo 11 space module that orbited the Moon in 1969.

Overview

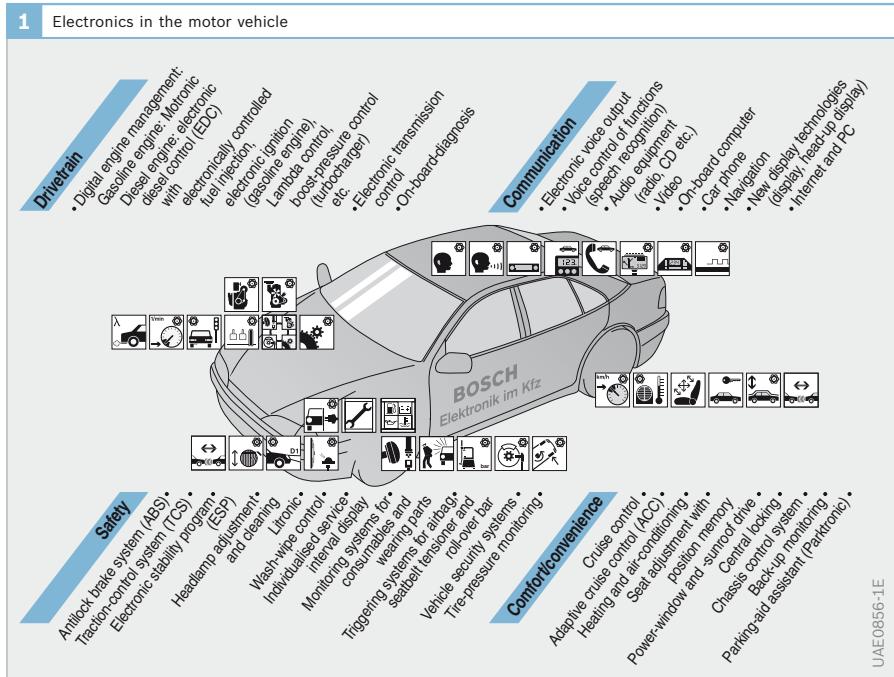
Development of electronic systems

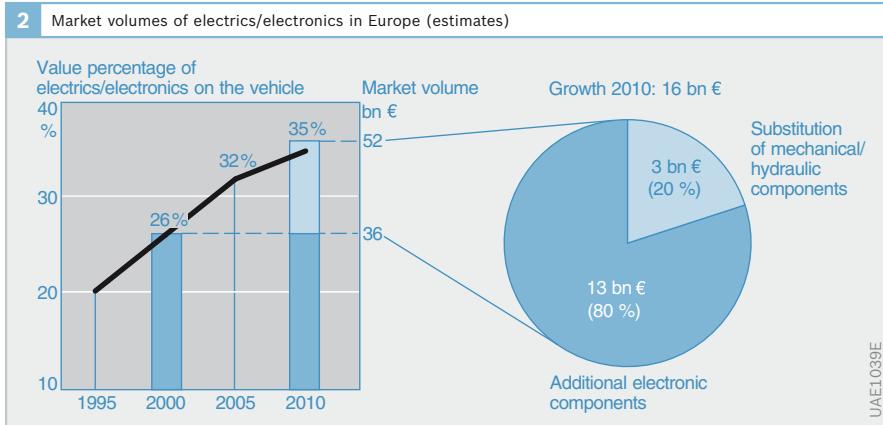
Not least in contributing to the success of the vehicle has been the continuous string of innovations which have found their way into vehicles. Even as far back as the 1970s, the aim was to make use of new technologies to help in the development of safe, clean and economical cars. The pursuit of economic efficiency and cleanliness was closely linked to other customer benefits

such as driving pleasure. This was characterized by the European diesel boom, upon which Bosch had such a considerable influence. At the same time, the development of the gasoline engine with gasoline direct injection, which would reduce fuel consumption by comparison with intake-manifold injection, experienced further advancements.

An improvement in driving safety was achieved with electronic brake-control systems. In 1978, the antilock brake system (ABS) was introduced and underwent continual development to such an extent that it is now fitted as standard on every vehicle in Europe. It was along this same line of development that the electronic stability program (ESP), in which ABS is integrated, would debut in 1995.

The latest developments also take comfort into account. These include the hill hold control (HHC) function, for example, which makes it easier to pull away on uphill gradients. This function is integrated in ESP.





Many kinds of new functions appear in conjunction with driver-assistance systems. Their scope extends far beyond today's standard features such as Parkpilot or electronic navigation systems. The aim is to produce the "sensitive vehicle" that uses sensors and electronics to detect and interpret its surroundings. Tapping into ultrasound, radar and video sensor technologies has led to solutions that play an important role in assisting the driver, e.g. through improved night vision or distance control.

Value creation structure for the future

The latest studies show that the production costs of an average car will increase only slightly by 2010 despite further innovations. No significant value growth for existing systems is expected in the mechanics/hydraulics domain despite the expected volume growth. One reason here being the electrification of functions that have conventionally been realized mechanically or hydraulically. Brake control systems are an impressive example of this change. While the conventional brake system was characterized more or less completely by mechanical components, the introduction of the ABS brake-control system was accompanied by a greater proportion of electronic components in

the form of sensor technology and an electronic control unit. With the more recent developments of ESP, the additional functions, such as HHC, are almost exclusively realized by electronics.

Even though significant economies of scale are seen with the established solutions, the value of the electrics and electronics will increase overall (Fig. 1). By 2010, this will amount to a good third of the production costs of an average vehicle. This assumption is based not least on the fact that the majority of future functions will also be regulated by electrics and electronics.

The increase in electrics and electronics is associated with a growth in software. Even today, software development costs are no longer negligible by comparison with hardware costs. Software authoring is faced with two challenges arising from the resulting increase in complexity of a vehicle's overall system: coping with the volume and a clearly structured architecture. The Autosar initiative (Automotive Open Systems Architecture), in which various motor vehicle manufacturers and suppliers participate, is working towards a standardization of electronics architecture with the aim of reducing complexity through increased reusability and interchangeability of software modules.

Task of an electronic system

Open-loop and closed-loop control

The nerve center of an electronic system is the control unit. Figure 3 shows the system blocks of a Motronic engine-management system. All the open-loop and closed-loop algorithms of the electronic system run inside the control unit. The heart of the control unit is a microcontroller with the program memory (flash EPROM) in which is stored the program code for all functions that the control unit is designed to execute.

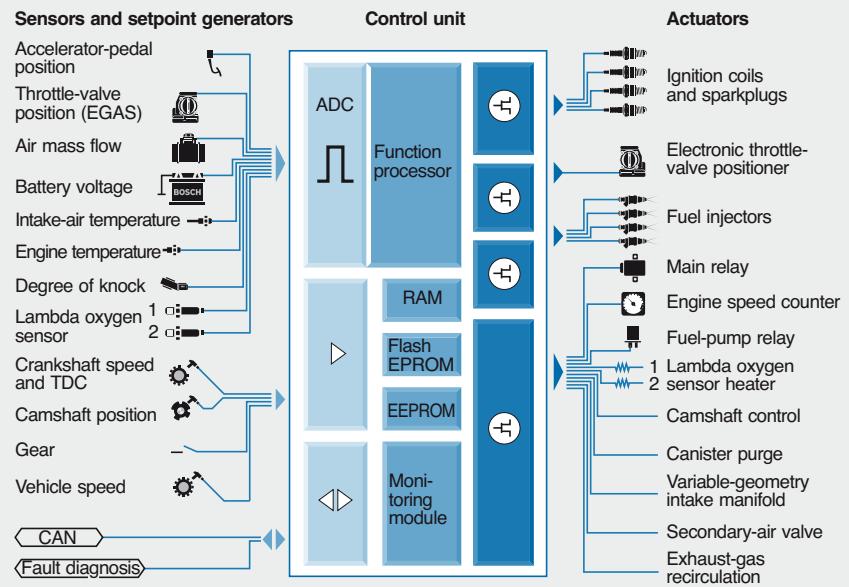
The input variables for the sequence control are derived from the signals from sensors and setpoint generators. They influence the calculations in the algorithms, and thus the triggering signals for the actuators. These convert into mechanical variables the electrical signals that are output by the microcontroller and amplified in the output stage modules. This could be mechanical energy generated by a servomotor (power-window unit), for example, or thermal energy generated by a sheathed-element glow plug.

Communication

Many systems have a mutual influence on each other. For example, it may sometimes be necessary to not only have the electronic stability program carry out a braking intervention in the event wheel spin but also to request that the engine-management system reduce torque and thus counteract wheel spin. Similarly, the control unit for the automatic transmission outputs a request to the engine-management system to reduce torque during a gearshift and thereby promote a soft gear change. To this end, the systems are networked with each other, i.e. they are able to communicate with each other on data buses (e.g. CAN, LIN).

In a premium-class vehicle, there may be up to 80 control units performing their duties. The examples below are intended to give you an insight into the operating principle of these systems.

3 Function modules of an electronic system



Motronic engine-management system

“Motronic” is the name of an engine-management system that facilitates open- and closed-loop control of gasoline engines within a single control unit.

There are Motronic variants for engines with intake-manifold injection (ME Motronic) and for gasoline direct injection (DI Motronic). Another variant is the Bifuel Motronic, which also controls the engine for operation with natural gas.

System description

Functions

The primary task of the Motronic engine-management system is:

- ▶ To adjust the torque desired and input by the driver depressing the accelerator pedal
- ▶ To operate the engine in such a way as to comply with the requirements of ever more stringent emission-control legislation
- ▶ To ensure the lowest possible fuel consumption but at the same time
- ▶ To guarantee high levels of driving comfort and driving pleasure

Components

Motronic comprises all the components which control and regulate the gasoline engine (Fig. 1, next page). The torque requested by the driver is adjusted by means of actuators or converters. The main individual components are:

- ▶ The electrically actuated throttle valve (air system): this regulates the air-mass flow to the cylinders and thus the cylinder charge
- ▶ The fuel injectors (fuel system): these meter the correct amount of fuel for the cylinder charge
- ▶ The ignition coils and spark plugs (ignition system): these provide for correctly timed ignition of the air-fuel mixture present in the cylinder

Depending on the vehicle, different measures may be required to fulfill the requirements demanded of the engine-management system (e.g. in respect of emission characteristics, power output and fuel consumption). Examples of system components able to be controlled by Motronic are:

- ▶ Variable camshaft control: it is possible to use the variability of valve timing and valve lifts to influence the ratio of fresh gas to residual exhaust gas and the mixture formation
- ▶ External exhaust-gas recirculation: adjustment of the residual gas content by means of a precise and deliberate return of exhaust gas from the exhaust train (adjustment by the exhaust-gas recirculation valve)
- ▶ Exhaust-gas turbocharging: regulated supercharging of the combustion air (i.e. increase in the fresh air mass in the combustion chamber) to increase torque
- ▶ Evaporative emission control system: for the return of fuel vapors that escape from the fuel tank and are collected in an activated charcoal canister

Operating variable acquisition

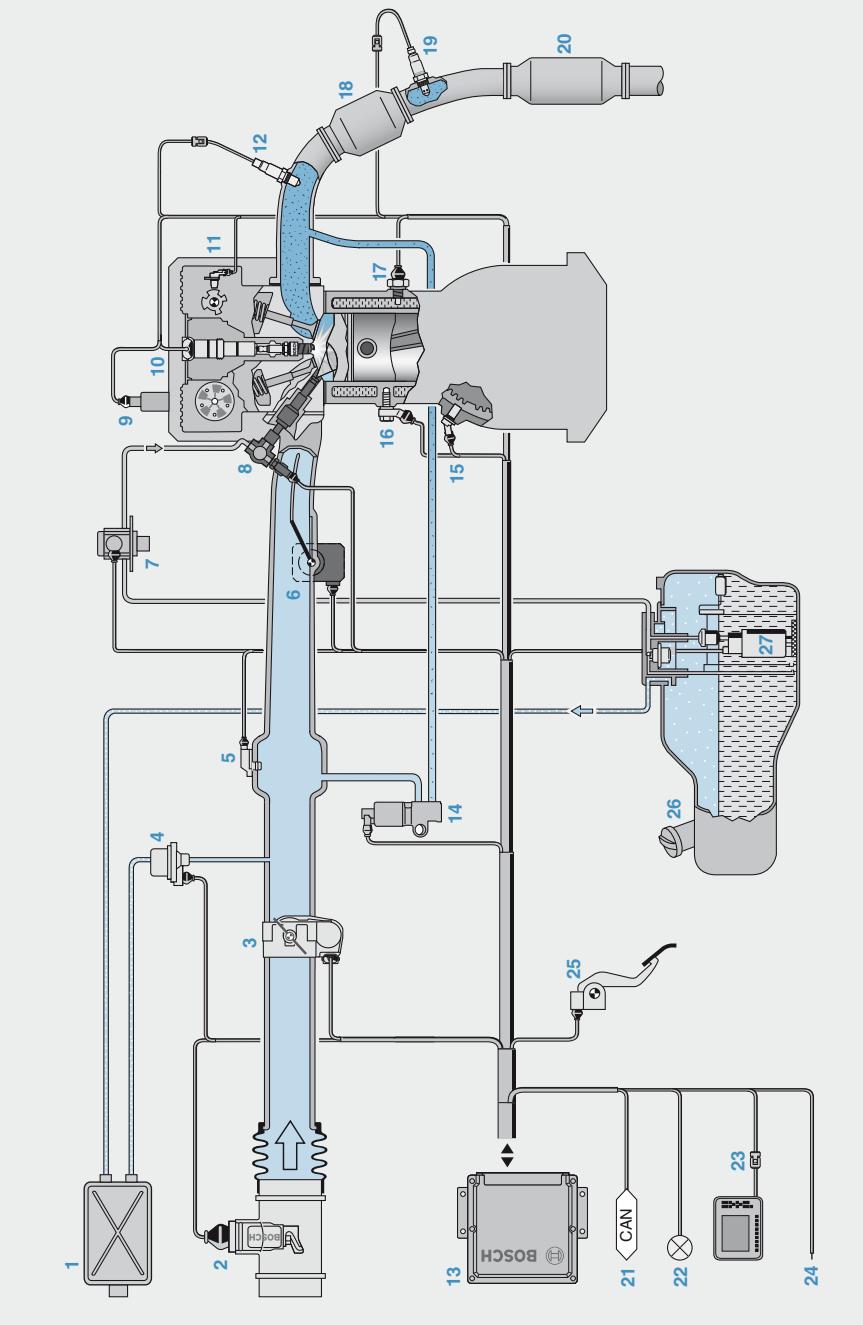
Motronic uses sensors to record the operating variables required for the open and closed-loop control of the engine (e.g. engine speed, engine temperature, battery voltage, intake air mass, intake-manifold pressure, Lambda value of the exhaust gas).

Setpoint generators (e.g. switches) record the adjustments made by the driver (e.g. position of the ignition key, cruise control).

Operating variable processing

From the input signals, the engine ECU detects the current operating status of the engine and uses this information in conjunction with requests from auxiliary systems and from the driver (accelerator-pedal sensor and operating switches) to calculate the command signals for the actuators.

1 Components used for open-loop electronic control of a DI-Motronic system
(example of a naturally aspirated engine, $\lambda = 1$)

**Fig. 1**

- 1 Activated charcoal canister
- 2 Hot-film air-mass meter
- 3 Throttle device (ETC)
- 4 Canister-purge valve
- 5 Intake-manifold pressure sensor
- 6 Swirl control valve
- 7 High-pressure pump
- 8 Rail with high-pressure fuel injector
- 9 Camshaft adjuster
- 10 Ignition coil with spark plug
- 11 Camshaft phase sensor
- 12 Lambda oxygen sensor (LSU)
- 13 Motronic ECU
- 14 EGR valve
- 15 Speed sensor
- 16 Knock sensor
- 17 Engine-temperature sensor
- 18 Primary catalytic converter
- 19 Lambda oxygen sensor
- 20 Primary catalytic converter
- 21 CAN interface
- 22 Diagnosis lamp
- 23 Diagnosis interface
- 24 Interface with immobilizer control unit
- 25 Accelerator-pedal module
- 26 Fuel tank
- 27 Fuel delivery module with electric fuel-supply pump

Air system

A specific air-fuel mixture is required to achieve the desired torque. For this purpose, the throttle valve (Fig. 1, Item 3) regulates the air necessary for the mixture formation by adjusting the metering orifice in the intake port for the fresh air taken in by the cylinders. This is effected by a DC motor (Fig. 2) integrated in the throttle device that is controlled by the Motronic control unit. The position of the throttle valve is fed back to the control unit by a position sensor to make position control possible. This sensor may be in the form of a potentiometer, for example. Since the throttle device is a component relevant to safety, the sensor is designed with redundancy.

The intake air mass (air charge) is recorded by sensors (e.g. hot-film air-mass meter, intake-manifold pressure sensor).

Fuel system

The control unit (Fig. 1, Item 13) calculates the fuel volume required from the intake air mass and the current operating status of the engine (e.g. intake-manifold pressure, engine speed), and also the time at which fuel injection should take place.

In gasoline injection systems with intake manifold injection, the fuel is introduced into the intake duct upstream of the intake valves. To this end, the electric fuel-supply pump (27) delivers fuel (primary pressure up to approximately 450 kPa) to the fuel injectors. Each cylinder is assigned a fuel injector that injects the fuel at intermittent intervals. The air-fuel mixture in the intake passage flows into the cylinder during the induction stroke. Corrections are made to the injected fuel quantity, e.g. by the Lambda control (Lambda oxygen sensor, 12) and the canister purge (evaporative-emissions control system, 1, 4).

With gasoline direct injection, fresh air flows into the cylinder. The fuel is injected directly into the combustion chamber by high-pressure fuel injectors (8) where it forms an air-fuel mixture with the intake air. This requires a higher fuel pressure, which is generated by additional high-pressure pump (7). The pressure can be variably adjusted (up to 20 MPa) in line with the operating point by an integrated fuel-supply control valve.

2 Throttle device with potentiometric position feedback

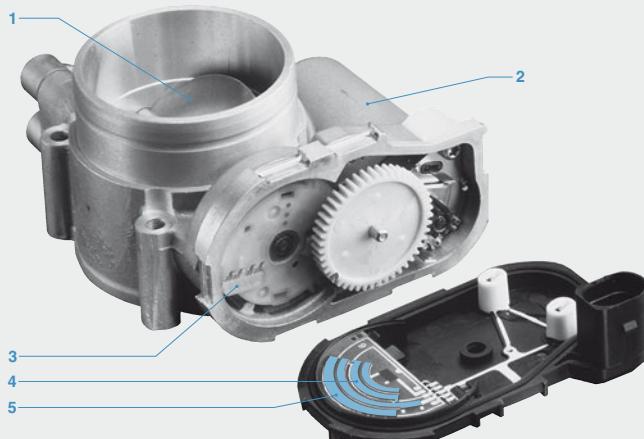


Fig. 2

- 1 Throttle valve
- 2 DC motor
- 3 Wiper
- 4 Resistance track 1
- 5 Resistance track 2

Fuel injector for intake-manifold injection

Function

The electromagnetic (solenoid-controlled) fuel injectors spray the fuel into the intake manifold at primary pressure. They allow fuel to be metered in the precise quantity required by the engine. They are actuated by driver stages which are integrated in the engine ECU with the signal calculated by the engine-management system.

3 EV14 electromagnetic fuel injector

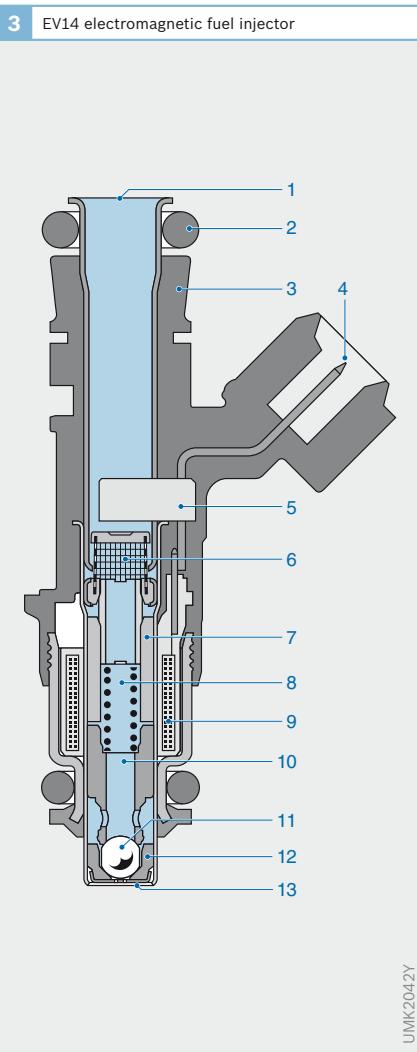


Fig. 3

- 1 Hydraulic port
- 2 O-ring
- 3 Valve housing
- 4 Electrical connection
- 5 Plastic clip with injected pins
- 6 Filter strainer
- 7 Internal pole
- 8 Valve spring
- 9 Solenoid coil
- 10 Valve needle with armature
- 11 Valve ball
- 12 Valve seat
- 13 Injection-orifice plate

Design and operating principle

Essentially, electromagnetic fuel injectors (Fig. 3) are comprised of the following components:

- Valve housing (3) with electrical connection (4) and hydraulic port (1)
- Solenoid coil (9)
- Moving valve needle (10) with solenoid armature and valve ball (11)
- Valve seat (12) with injection-orifice plate (13) and
- Valve spring (8)

In order to ensure trouble-free operation, stainless steel is used for the parts of the fuel injector which come into contact with fuel. The fuel injector is protected against dirt by a filter strainer (6) at the fuel inlet.

Connections

On the fuel injectors presently in use, fuel supply to the fuel injector is in the axial direction, i.e. from top to bottom ("top feed"). The fuel line is secured to the hydraulic port by means of a clamping fixture. Retaining clips ensure reliable fastening. The sealing ring (O-ring) on the hydraulic port (2) seals off the fuel injector at the fuel rail.

The fuel injector is electrically connected to the engine ECU.

Fuel injector operation

When the solenoid coil is de-energized, the valve needle and valve ball are pressed against the cone-shaped valve seat by the spring and the force exerted by the fuel pressure. The fuel-supply system is thus sealed off from the intake manifold. When the solenoid coil is energized, this generates a magnetic field which attracts the valve-needle solenoid armature. The valve ball lifts up from the valve seat and the fuel is injected. When the excitation current is switched off, the valve needle closes again due to spring force.

Fuel outlet

The fuel is atomized by means of an injection-orifice plate in which there are a number of holes. These holes (injection orifices) are stamped out of the plate and ensure that the injected fuel quantity remains highly constant. The injection-orifice plate is insensitive to fuel deposits. The spray pattern of the fuel leaving the injector is produced by the number of injection orifices and their configuration.

The injector is efficiently sealed at the valve seat by the cone/ball sealing principle. The fuel injector is inserted into the opening provided for it in the intake manifold. The lower sealing ring provides the seal between the fuel injector and the intake manifold.

Essentially, the injected fuel quantity per unit of time is determined by

- ▶ The primary pressure in the fuel-supply system
- ▶ The back pressure in the intake manifold and
- ▶ The geometry of the fuel-exit area

Electrical activation

An output module in the Motronic ECU actuates the fuel injector with a switching signal (Fig. 4a). The current in the solenoid coil rises (b) and causes the valve needle (c) to lift. The maximum valve lift is achieved after the time t_{pk} (pickup time) has elapsed. Fuel is sprayed as soon as the valve ball lifts off its seat. The total quantity of fuel injected during an injection pulse is shown in Figure 4d.

Current flow ceases when activation is switched off. Mass inertia causes the valve to close, but only slowly. The valve is fully closed again after the time t_{dr} (dropout time) has elapsed.

When the valve is fully open, the injected fuel quantity is proportional to the time. The non-linearity during the valve pickup and dropout phases must be compensated for throughout the period that the injector is activated (injection dura-

tion). The speed at which the valve needle lifts off its seat is also dependent on the battery voltage. Battery-voltage-dependent injection-duration extension (Fig. 5) corrects these influences.

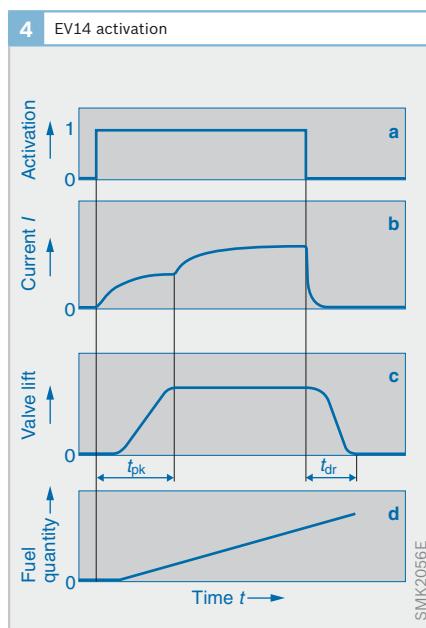
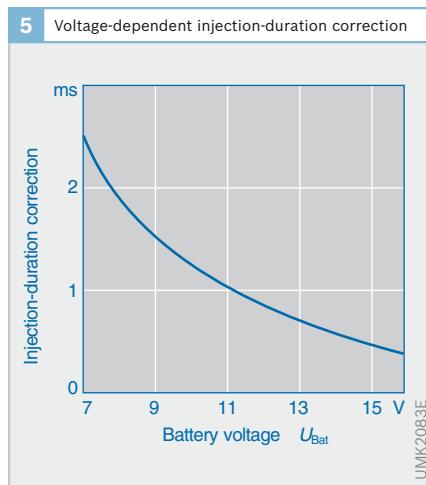


Fig. 4

- a Activation signal
- b Current curve
- c Valve lift
- d Injected fuel quantity



High-pressure fuel injector for gasoline direct injection

Function

It is the function of the high-pressure fuel injector (HDEV) on the one hand to meter the fuel and on the other hand by means of its atomization to achieve controlled mixing of the fuel and air in a specific area of the combustion chamber. Depending on the desired operating status, the fuel is either concentrated in the vicinity of the spark plug (stratified charge) or evenly distributed throughout the combustion chamber (homogenous distribution).

Design and operating principle

The high-pressure fuel injector (Fig. 6) comprises the following components:

- Inlet with filter (1)
- Electrical connection (2)
- Spring (3)
- Coil (4)
- Valve sleeve (5)
- Nozzle needle with solenoid armature (6) and
- Valve seat (7)

A magnetic field is generated when current passes through the coil. This lifts the valve needle off the valve seat against the force of the spring and opens the injector outlet bores (8). The primary pressure now forces the fuel into the combustion chamber. The injected fuel quantity is essentially dependent on the opening duration of the fuel injector and the fuel pressure.

When the energizing current is switched off, the valve needle is pressed by spring force back down against its valve seat and interrupts the flow of fuel.

Excellent fuel atomization is achieved thanks to the suitable nozzle geometry at the injector tip.

Requirements

Compared with manifold injection, gasoline direct injection differs mainly in its higher fuel pressure and the far shorter time which is available for directly injecting the fuel into the combustion chamber.

6 Design of HDEV5 high-pressure fuel injector

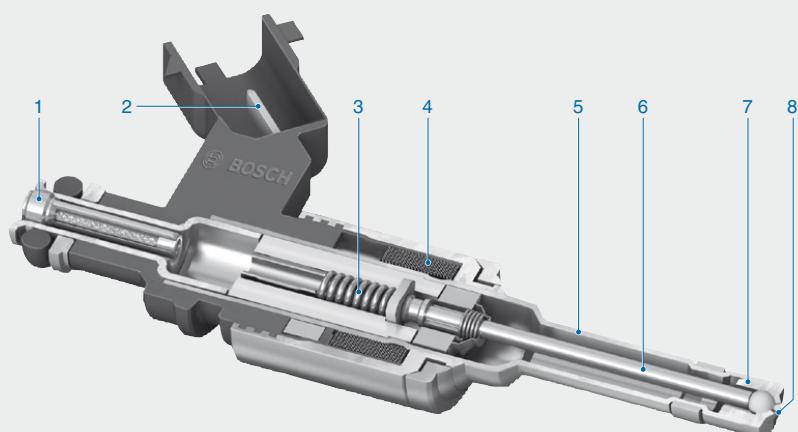


Fig. 6

- 1 Fuel inlet with filter
- 2 Electrical connection
- 3 Spring
- 4 Coil
- 5 Valve sleeve
- 6 Nozzle needle with solenoid armature
- 7 Valve seat
- 8 Injector outlet bores

Figure 7 underlines the technical demands made on the fuel injector. In the case of manifold injection, two revolutions of the crankshaft are available for injecting the fuel into the intake manifold. This corresponds to an injection duration of 20 ms at an engine speed of 6,000 rpm.

In the case of gasoline direct injection, however, considerably less time is available. In homogeneous operation, the fuel must be injected during the induction stroke. In other words, only a half crankshaft rotation is available for the injection process. At 6,000 rpm, this corresponds to an injection duration of 5 ms.

With gasoline direct injection, the fuel requirement at idle in relation to that at full load is far lower than with manifold injection (factor 1:12). At idle, this results in an injection duration of approx. 0.4 ms.

Actuation of HDEV high-pressure fuel injector

The high-pressure fuel injector must be actuated with a highly complex current

curve in order to comply with the requirements for defined, reproducible fuel-injection processes (Fig. 8). The microcontroller in the engine ECU only delivers a digital triggering signal (a). An output module (ASIC) uses this signal to generate the triggering signal (b) for the fuel injector.

A DC/DC converter in the engine ECU generates the booster voltage of 65 V. This voltage is required in order to bring the current up to a high value as quickly as possible in the booster phase. This is necessary in order to accelerate the injector needle as quickly as possible. In the pickup phase (t_{pk}), the valve needle then achieves the maximum opening lift (c). When the fuel injector is open, a small control current (holding current) is sufficient to keep the fuel injector open.

With a constant valve-needle displacement, the injected fuel quantity is proportional to the injection duration (d).

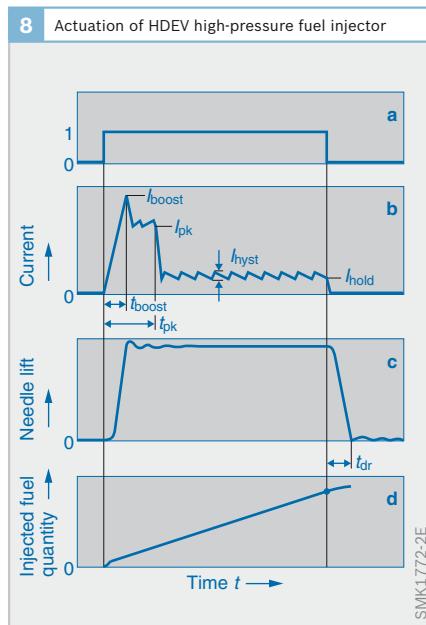
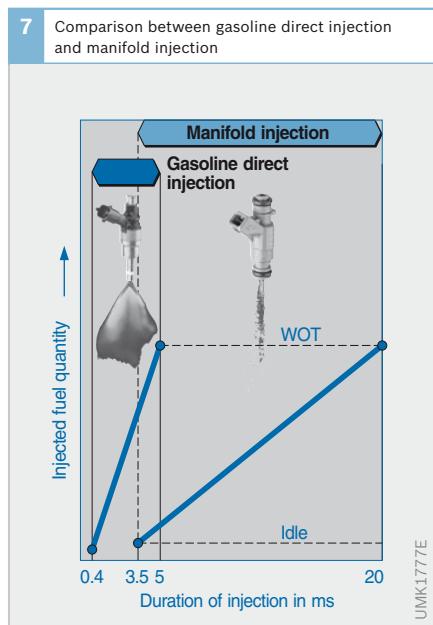


Fig. 7
Injected fuel quantity as a function of injection duration

- Fig. 8**
- a Triggering signal
 - b Current curve in injector
 - c Needle lift
 - d Injected fuel quantity

Inductive ignition System

Ignition of the air-fuel mixture in the gasoline engine is electric; it is produced by generating a flashover between the electrodes on a spark plug. The ignition-coil energy converted in the spark ignites the compressed air-fuel mixture immediately adjacent to the spark plug, creating a flame front which then spreads to ignite the air-fuel mixture in the entire combustion chamber. The inductive ignition system generates in each power stroke the high voltage required for flashover and the spark duration required for ignition. The electrical energy drawn from the vehicle electrical system is temporarily stored in the ignition coil.

Design

Figure 9 shows the principle layout of the ignition circuit of an inductive ignition system. It comprises the following components:

- ▶ Ignition driver stage (4), which is integrated in the Motronic ECU or in the ignition coil
- ▶ Ignition coils (3)
- ▶ Spark plugs (5) and
- ▶ Connecting devices and interference suppressors

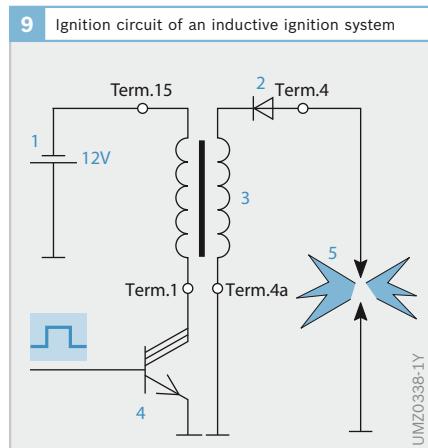
Fig. 9

- 1 Battery
- 2 AAS diode (integrated in ignition coil)
- 3 Ignition coil with iron core and primary and secondary windings
- 4 Ignition driver stage (integrated either in Motronic ECU or in ignition coil)
- 5 Spark plug

Term.1, Term.4,
Term.4a, Term.15
Terminal
designations

Fig. 10

- K Spark head
S Spark tail
 t_f Spark duration



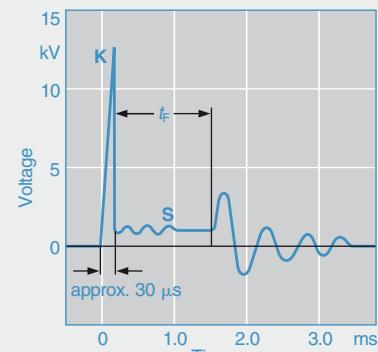
Generating the ignition spark

A magnetic field is built up in the ignition coil when a current flows in the primary circuit. The ignition energy required for ignition is stored in this magnetic field.

The current in the primary winding only gradually attains its setpoint value because of the induced countervoltage. Because the energy stored in the ignition coil is dependent on the current ($E = \frac{1}{2}LI^2$), a certain amount of time (dwell period) is required in order to store the energy necessary for ignition. This dwell period is dependent on, among others, the vehicle system voltage. The ECU program calculates from the dwell period and the moment of ignition the cut-in point, and cuts the ignition coil in via the ignition driver stage and out again at the moment of ignition.

Interrupting the coil current at the moment of ignition causes the magnetic field to collapse. This rapid magnetic-field change induces a high voltage (Fig. 10) on the secondary side of the ignition coil as a result of the large number of turns (turns ratio approx. 1:100). When the ignition voltage is reached, flashover occurs at the spark plug and the compressed air-fuel mixture is ignited.

10 Voltage curve at the electrodes



Flame-front propagation

After the flashover, the voltage at the spark plug drops to the spark voltage (Fig. 10). The spark voltage is dependent on the length of the spark plasma (electrode gap and deflection due to flow) and ranges between a few hundred volts and well over 1 kV. The ignition-coil energy is converted in the ignition spark during the combustion time; this ignition spark duration lasts from as little as 100 µs to over 2 ms. Following the breakaway of the spark, the damped voltage decays.

The electrical spark between the spark-plug electrodes generates a high-temperature plasma. When the air-fuel mixture at the spark plug is ignitable and sufficient energy input is supplied by the ignition system, the arc that is created develops into a self-propagating flame front.

Moment of ignition

The instant at which the ignition spark ignites the air-fuel mixture within the combustion chamber must be selected with extreme precision. This variable has a decisive influence on engine operation and determines the output torque, exhaust-gas emissions and fuel consumption.

The influencing variables that determine the moment of ignition are engine speed and engine load, or torque. Additional

variables, such as, for example, engine temperature, are also used to determine the optimal moment of ignition. These variables are recorded by sensors and then relayed to the engine ECU (Motronic). The moment of ignition is calculated and the triggering signal for the ignition driver stage is generated from program maps and characteristic curves.

Combustion knocks occur if the moment of ignition is too advanced. Permanent knocking may result in engine damage. For this reason, knock sensors are used to monitor combustion noise. After a combustion knock, the moment of ignition is delayed to too late and then slowly moved back to the pilot control value. This helps to counteract permanent knocking.

Voltage distribution

Voltage distribution takes place on the primary side of the ignition coils, which are directly connected to the spark plugs (static voltage distribution).

System with single-spark ignition coils

Each cylinder is allocated an ignition driver stage and an ignition coil (Figs. 11a and 11b). The engine ECU actuates the ignition driver stages in specified firing order. However, the system does also have to be synchronized with the camshaft by means of a camshaft sensor.

System with dual-spark ignition coils

One ignition driver stage and one ignition coil are allocated to every two cylinders (Fig. 11c). The ends of the secondary winding are each connected to a spark plug in different cylinders. The cylinders have been chosen so that when one cylinder is in the compression cycle, the other is in the exhaust cycle (only possible with engines with an even number of cylinders). It does not therefore need to be synchronized with the camshaft. Flashover occurs at both spark plugs at the moment of ignition.

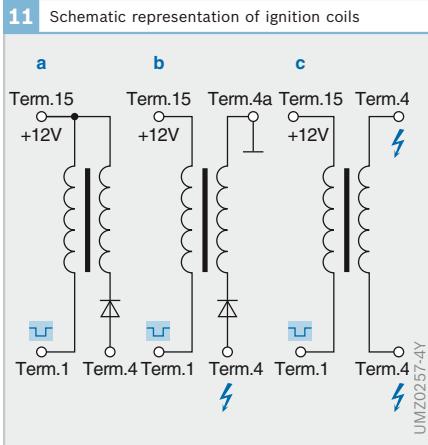


Fig. 11

- a Single-spark ignition coil in economy circuit
- b Single-spark ignition coil
- c Dual-spark ignition coil

Ignition coils

Compact ignition coil

Design

The compact coil's magnetic circuit consists of the O core and the I core (Fig. 12), onto which the primary and secondary windings are plugged. This arrangement is installed in the coil housing. The primary winding (I core wound in wire) is electrically and mechanically connected to the primary plug connection. Also connected is the start of the secondary winding (coil body wound in wire). The connection on the spark-plug side of the secondary winding is also located in the housing, and electrical contacting is established when the windings are fitted.

Integrated within the housing is the high-voltage contact dome. This contains the contact section for spark-plug contacting, and also a silicone jacket for insulating the high voltage from external components and the spark-plug well.

Following component assembly resin is vacuum-injected into the inside of the housing, where it is allowed to harden. This produces high mechanical strength, good protection from environmental influences and outstanding insulation of the high voltage. The silicone jacket is then pushed onto the high-voltage contact dome for permanent attachment.

Remote and COP versions

The ignition coil's compact dimensions make it possible to implement the design shown in Figure 12. This version is called COP (Coil On Plug). The ignition coil is mounted directly on the spark plug, thereby rendering additional high-voltage connecting cables superfluous. This reduces the capacitive load on the ignition coil's secondary circuit. The reduction in the number of components also increases operational reliability (no rodent bites in ignition cables, etc.).

In the less common remote version, the compact coils are mounted within the engine compartment using screws. Attachment lugs or an additional bracket are provided for this purpose. The high-voltage connection is effected by means of a high-voltage ignition cable from the ignition coil to the spark plug.

The COP and remote versions are virtually identical in design. However, the remote version (mounted on the vehicle body) is subject to fewer demands with regard to temperature and vibration conditions due to the fact that it is exposed to fewer loads and strains.

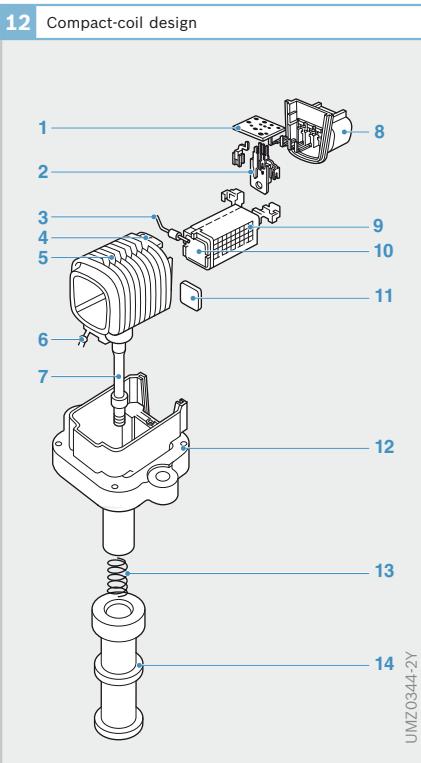


Fig. 12

- 1 Printed-circuit board
- 2 Ignition driver stage
- 3 AAS diode (activation arc suppression)
- 4 Secondary winding body
- 5 Secondary wire
- 6 Contact plate
- 7 High-voltage pin
- 8 Primary plug
- 9 Primary wire
- 10 I core
- 11 Permanent magnet
- 12 O core
- 13 Spring
- 14 Silicone jacket

Pencil coil

The pencil coil makes optimal use of the space available within the engine compartment. Its cylindrical shape makes it possible to use the spark plug well as a supplementary installation area for ideal space utilization on the cylinder head.

Because pencil coils are always mounted directly on the spark plug, no additional high-voltage connecting cables are required.

Design and magnetic circuit

Pencil coils operate like compact coils in accordance with the inductive principle. However, the rotational symmetry results in a design structure that differs considerably from that of compact coils.

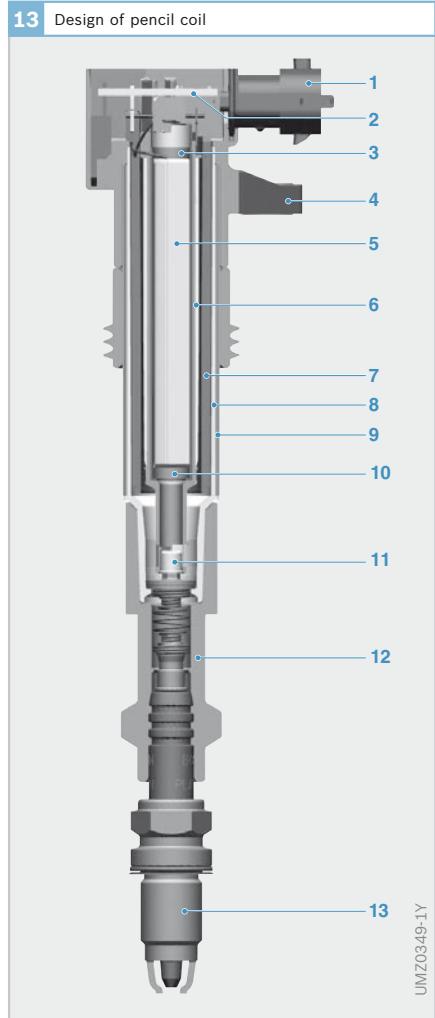
Although the magnetic circuit consists of the same materials, the central rod core (Fig. 13, Item 5) consists of laminations in various widths stacked in packs that are virtually circular. The yoke plate (9) that provides the magnetic circuit is a rolled and slotted sleeve - also in electrical sheet steel, sometimes in multiple layers.

Another difference relative to compact coils is the primary winding (7), which has a larger diameter and is above the secondary winding (6), while the body of the winding also supports the rod core. This arrangement brings significant benefits in the areas of design and operation. Owing to restrictions imposed by their geometrical configuration and compact dimensions, pencil coils allow only limited scope for varying the magnetic circuit (rod core, yoke plate) and windings.

In most pencil-coil applications, the limited space available dictates that permanent magnets be used to increase the spark energy.

The arrangements for electrical contact with the spark plug and for connection to the engine wiring harness are comparable with those used for compact pencil coils.

13 Design of pencil coil



UMZ0349-1Y

Fig. 13

- 1 Plug connection
- 2 Printed-circuit board with ignition driver stage
- 3 Permanent magnet
- 4 Attachment arm
- 5 Laminated electrical-sheet-steel core (rod core)
- 6 Secondary winding
- 7 Primary winding
- 8 Housing
- 9 Yoke plate
- 10 Permanent magnet
- 11 High-voltage dome
- 12 Silicone jacket
- 13 Attached spark plug

Electronic diesel control (EDC)

System overview

Electronic control of a diesel engine enables precise and differentiated modulation of fuel-injection parameters. This is the only means by which a modern diesel engine is able to satisfy the many demands placed upon it. Electronic diesel control (EDC) is subdivided into three system blocks: sensors/setpoint generators, ECU, and actuators.

Requirements

The lowering of fuel consumption and exhaust emissions (NO_x , CO, HC, particulates) combined with simultaneous improvement of engine power output and torque are the guiding principles of current development work on diesel-engine design. Conventional indirect-injection engines (IDI) were no longer able to satisfy these requirements.

State-of-the-art technology is represented today by direct-injection diesel engines (DI) with high injection pressures for efficient mixture formation. The fuel-injection systems support several injection processes: pre-injection, main injection, and secondary injection. These injection pro-

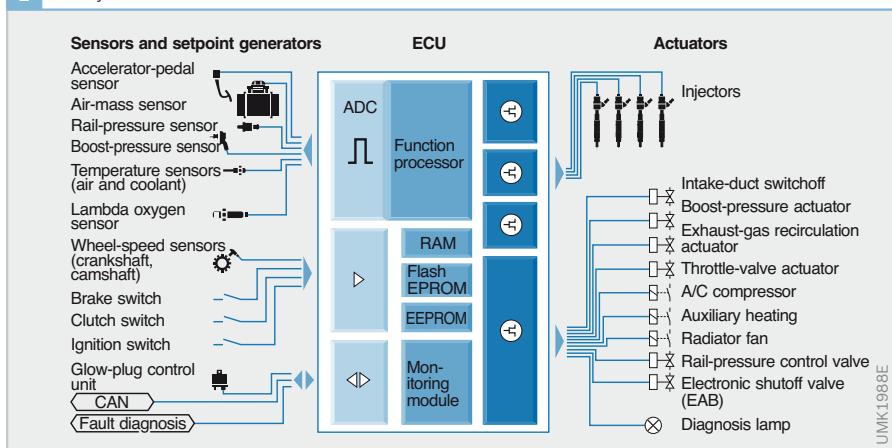
cesses are for the most part controlled electronically (pre-injection, however, is controlled mechanically on UIS for cars).

In addition, diesel-engine development has been influenced by the high levels of driving comfort and convenience demanded in modern cars. Exhaust and noise emissions are also subject to ever more stringent demands.

As a result, the performance demanded of the fuel-injection and management systems has also increased, specifically with regard to:

- ▶ High injection pressures
- ▶ Rate shaping
- ▶ Pre-injection and, if necessary, secondary injection
- ▶ Adaptation of injected fuel quantity, boost pressure and start of injection at the respective operating status
- ▶ Temperature-dependent excess-fuel quantity
- ▶ Load-independent idle speed control
- ▶ Controlled exhaust-gas recirculation
- ▶ Cruise control
- ▶ Tight tolerances for start of injection and injected-fuel quantity and maintenance of high precision over the service life of the system (long-term performance)
- ▶ Support of exhaust-gas treatment systems

1 EDC system blocks



Conventional mechanical RPM control uses a number of adjusting mechanisms to adapt to different engine operating statuses and ensures high-quality mixture formation. Nevertheless, it is restricted to a simple engine-based control loop and there are a number of important influencing variables that it cannot take account of or cannot respond quickly enough to.

As demands have increased, EDC has developed into a complex electronic engine-management system capable of processing large amounts of data in real time. In addition to its pure engine-management function, EDC supports a series of comfort and convenience functions (e.g. cruise control). It can form part of an overall electronic vehicle-speed control system ("drive-by-wire"). And as a result of the increasing integration of electronic components, complex electronics can be accommodated in a very small space.

Operating principle

Electronic diesel control (EDC) is capable of meeting the requirements listed above as a result of microcontroller performance that has improved considerably in the last few years.

In contrast to diesel-engine vehicles with conventional in-line or distributor injection pumps, the driver of an EDC-controlled vehicle has no direct influence, for instance through the accelerator pedal and Bowden cable, upon the injected fuel quantity. Instead, the injected fuel quantity is determined by a number of influencing variables. These include:

- ▶ Driver command (accelerator-pedal position)
- ▶ Operating status
- ▶ Engine temperature
- ▶ Interventions by other systems (e.g. TCS)
- ▶ Effects on exhaust emissions, etc.

The ECU calculates the injected fuel quantity on the basis of all these influencing variables. Start of injection can also be var-

ied. This requires a comprehensive monitoring concept that detects inconsistencies and initiates appropriate actions in accordance with the effects (e.g. torque limitation or limp-home mode in the idle-speed range). EDC therefore incorporates a number of control loops.

Electronic diesel control allows data communication with other electronic systems, such as the traction-control system (TCS), electronic transmission control (ETC), or electronic stability program (ESP). As a result, the engine-management system can be integrated in the vehicle's overall control system, thereby enabling functions such as reduction of engine torque when the automatic transmission changes gear, regulation of engine torque to compensate for wheel slip, etc.

The EDC system is fully integrated in the vehicle's diagnosis system. It meets all OBD (On-Board Diagnosis) and EOBD (European OBD) requirements.

System blocks

Electronic diesel control (EDC) is divided into three system blocks (Fig. 1):

1. *Sensors and setpoint generators* detect operating conditions (e.g. engine speed) and setpoint values (e.g. switch position). They convert physical variables into electrical signals.

2. The *ECU* processes the information from the sensors and setpoint generators in mathematical computing processes (open- and closed-loop control algorithms). It controls the actuators by means of electrical output signals. In addition, the ECU acts as an interface to other systems and to the vehicle diagnosis system.

3. *Actuators* convert the electrical output signals from the ECU into mechanical variables (e.g. solenoid-valve needle lift).

Data processing

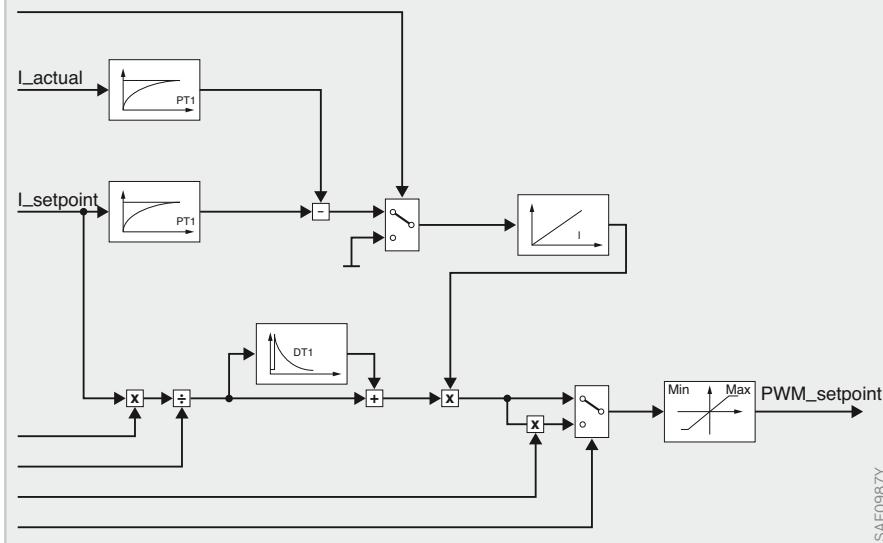
The main function of the electronic diesel control (EDC) is to control the injected fuel quantity and the injection timing. The common-rail accumulator injection system also controls injection pressure. Furthermore, on all systems, the engine ECU controls a number of actuators. The EDC functions must be matched to every vehicle and every engine. This is the only way to optimize component interaction (Fig. 3).

The control unit evaluates the signals sent by the sensors and limits them to the permitted voltage level. Some input signals are also checked for plausibility. Using this input data together with stored program maps, the microprocessor calculates the position and duration for injection timing. This information is then converted to a signal characteristic which is aligned to the engine's piston strokes. This calculation program is termed the "ECU software".

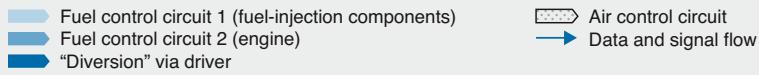
The required degree of accuracy together with the diesel engine's outstanding dynamic response requires high-level computing power. The output signals are applied to driver stages which provide adequate power for the actuators (for instance, the high-pressure solenoid valves for fuel injection, exhaust-gas recirculation positioner, or boost-pressure actuator). Apart from this, a number of other auxiliary-function components (e.g. glow relay and air-conditioning system) are triggered.

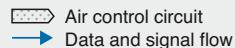
The driver-stage diagnosis functions for the solenoid valves also detect faulty signal characteristics. Furthermore, signals are exchanged with other systems in the vehicle via the interfaces. The engine ECU monitors the complete fuel-injection system as part of a safety strategy.

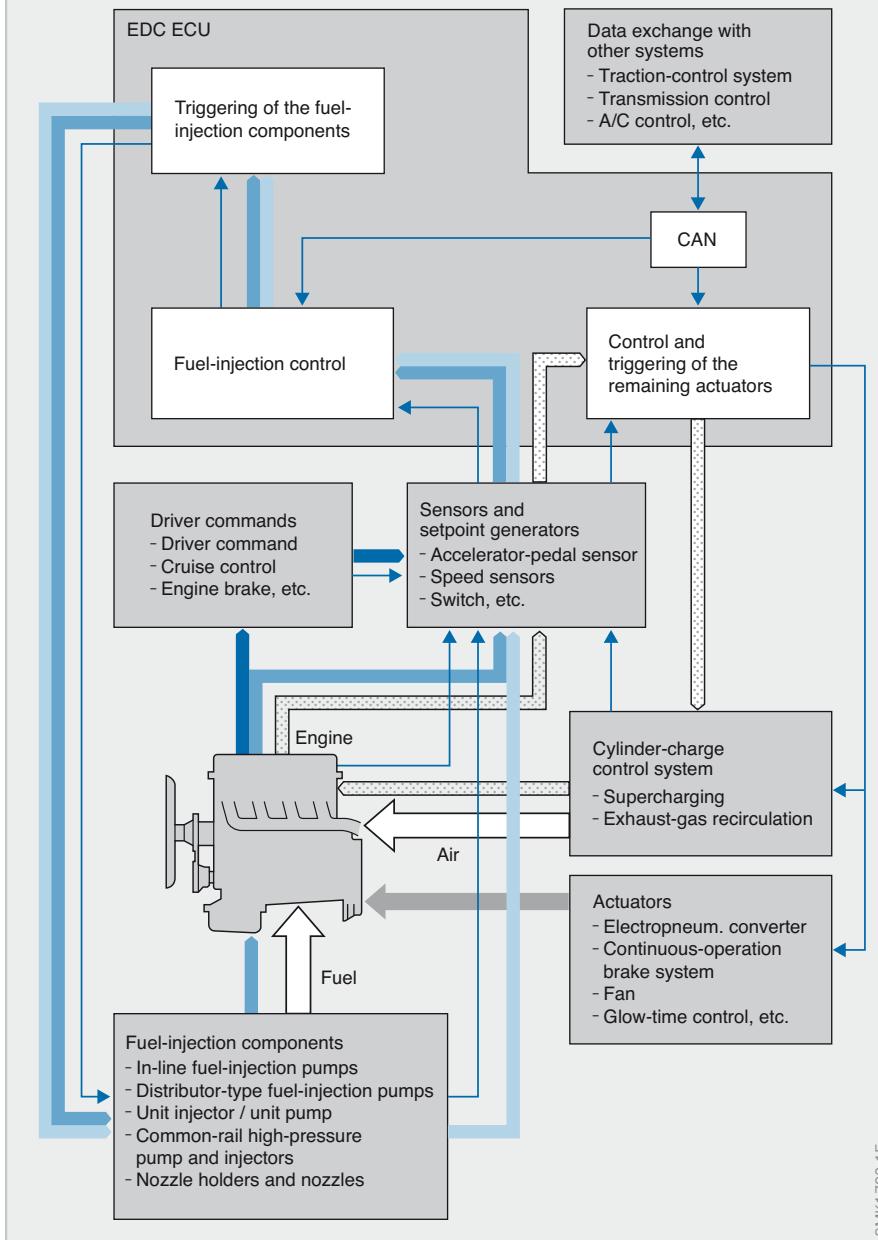
2 Schematic using the example of a current regulator



3 Basic sequence of electronic diesel control


 Fuel control circuit 1 (fuel-injection components)
 Fuel control circuit 2 (engine)
 "Diversion" via driver


 Air control circuit
 Data and signal flow



Fuel-injection control

Table 1 provides an overview of the EDC functions which are implemented in the various fuel-injection systems. Figure 4 shows the sequence of fuel-injection calculations with all functions, a number of which are optional extras. These can be activated in the ECU by the after-sales service when retrofit equipment is installed.

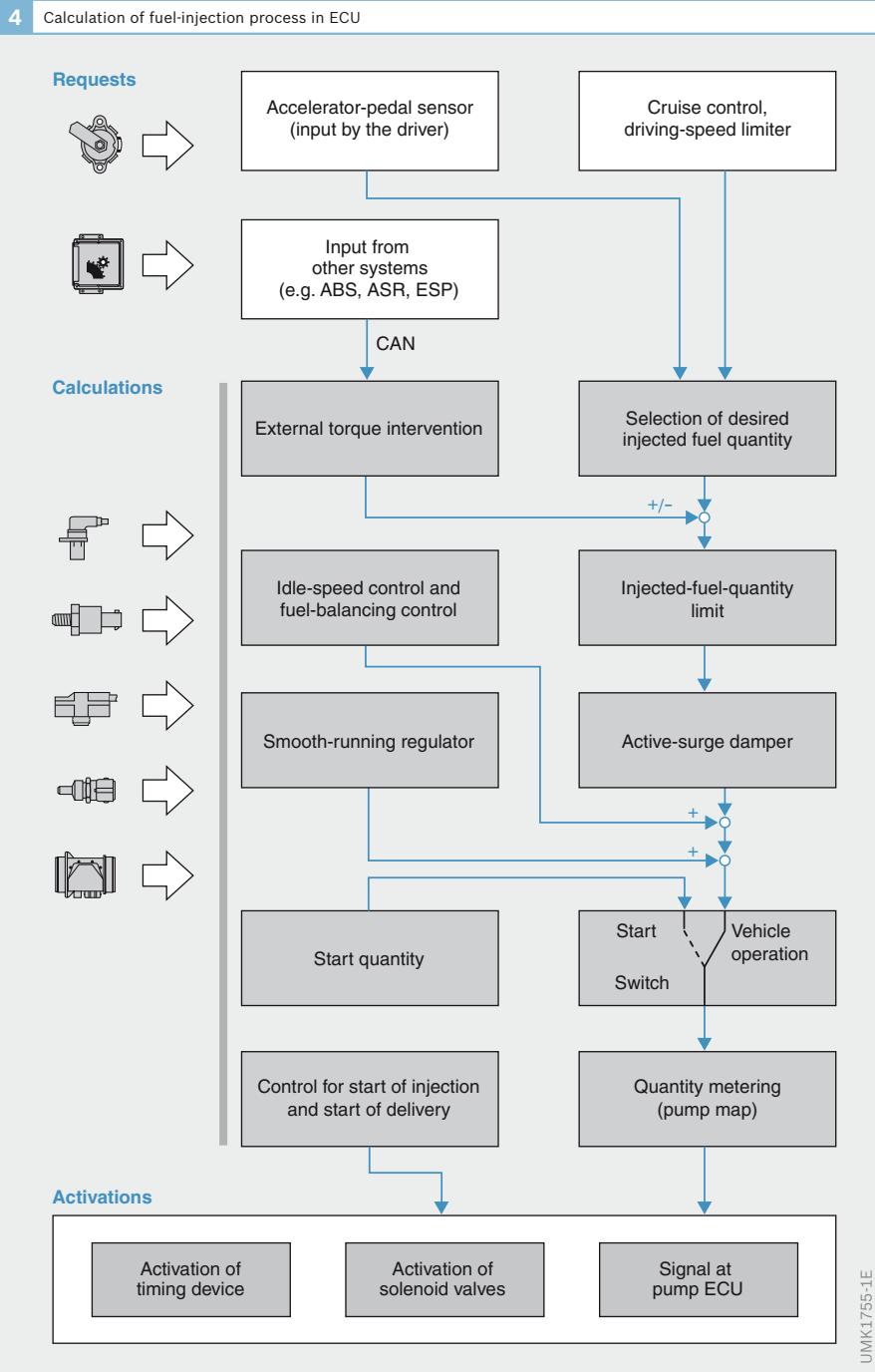
In order that the engine can run with optimal combustion under all operating conditions, the ECU calculates exactly the right injected fuel quantity for all conditions. Here, a number of parameters must be taken into account. On a number of solenoid-valve-controlled distributor-type injection pumps, the solenoid valves for injected fuel quantity and start of injection are triggered by a separate pump ECU.

1 Overview of functions of EDC variants for motor vehicles

Fuel-injection system	In-line fuel-injection pumps PE	Helix-controlled distributor-type injection pumps VE-EDC	Solenoid-valve-controlled distributor injection pumps VE-M, VR-M	Unit injector system and unit pump system UIS, UPS	Common-rail system CR
Function					
Injected-fuel-quantity limit	•	•	•	•	•
External torque intervention	• ³⁾	•	•	•	•
Driving-speed limitation	• ³⁾	•	•	•	•
Cruise control	•	•	•	•	•
Altitude correction	•	•	•	•	•
Boost-pressure control	•	•	•	•	•
Idle-speed regulation	•	•	•	•	•
Intermediate-speed regulation	• ³⁾	•	•	•	•
Active surge damping	• ²⁾	•	•	•	•
BIP control	—	—	•	•	—
Intake-port shutoff	—	—	•	• ²⁾	•
Electronic immobilizer	• ²⁾	•	•	•	•
Controlled pre-injection	—	—	•	• ²⁾	•
Glow control unit	• ²⁾	•	•	• ²⁾	•
A/C switch-off	• ²⁾	•	•	•	•
Auxiliary coolant heating	• ²⁾	•	•	• ²⁾	•
Smooth-running control	• ²⁾	•	•	•	•
Fuel-balancing control	• ²⁾	—	•	•	•
Fan activation	—	•	•	•	•
EGR control	• ²⁾	•	•	•	•
Start-of-injection control with sensor	• ¹⁾ ³⁾	•	•	•	•
Cylinder shutoff	—	—	• ³⁾	• ³⁾	• ³⁾
Increment-angle learning	—	—	—	•	•
Increment-angle rounding	—	—	—	• ²⁾	—

Table 1

- ¹⁾ Control-sleeve in-line fuel-injection pumps
- ²⁾ Cars only
- ³⁾ Commercial vehicles only



Torque-controlled EDC systems

The engine-management system is continually being integrated more closely into the overall vehicle system. Vehicle-dynamics systems (e.g. TCS), comfort and convenience systems (e.g. cruise control/Temposomat), and transmission control influence electronic diesel control (EDC) via the CAN bus. Apart from this, much of the information registered or calculated in the engine-management system must be passed on to other ECUs via the CAN bus.

In order to be able to incorporate EDC even more efficiently in a functional alliance with other ECUs, and implement other changes rapidly and effectively, it was necessary to make radical changes to the newest-generation controls. These changes resulted in torque-controlled EDC, which was introduced with the EDC16. The main feature is the changeover of the module interfaces to the parameters as commonly encountered in practice in the vehicle.

Engine characteristics

Essentially, an engine's output can be defined using the three characteristics: power P , engine n , and torque M .

Figure 5 compares typical curves of torque and power as a function of the engine speed of two diesel engines. Basically speaking, the following formula applies:

$$P = 2 \cdot \pi \cdot n \cdot M$$

It is sufficient therefore, for example, to specify the torque as the reference variable while taking into account the engine speed. Engine power then results from the above formula. Since power output cannot be measured directly, torque has turned out to be a suitable reference variable for engine management.

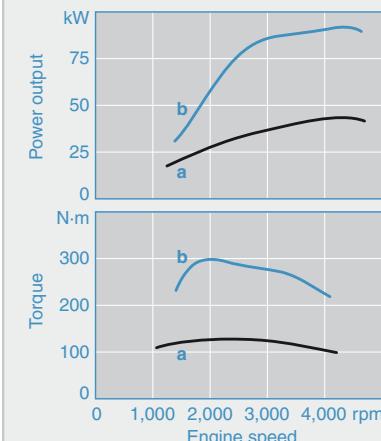
Torque control

When accelerating, the driver uses the accelerator-pedal (sensor) to directly

demand a given torque from the engine. Independently of the driver's requirements, other external vehicle systems submit torque demands via the interfaces resulting from the power requirements of the particular component (e.g. air-conditioning system, alternator). Using these torque-requirement inputs, the engine-management system calculates the output engine torque to be generated and controls the fuel-injection and air-system actuators accordingly. This has the following advantages:

- ▶ No system has a direct influence on engine management (boost pressure, fuel injection, preglow). The engine management system can thus also take into account other higher-level optimization criteria for the external requirements (e.g. exhaust-gas emissions, fuel consumption) and then control the engine in the best way possible.
- ▶ Many of the functions which do not directly concern the engine management system can be designed to function identically for diesel and gasoline engines.
- ▶ Expansions to the system can be implemented quickly.

5 Example of the torque and power-output curves as a function of engine speed for two car diesel engines with approx. 2.2 l engine displacement



Sequence of engine management

The setpoint values are processed further in the engine ECU. In order to fulfill their assignments efficiently, the engine management system's control functions all require a wide range of sensor signals and information from other ECUs in the vehicle.

Propulsion torque

The driver's input (i.e. the signal from the accelerator-pedal sensor) is interpreted by the engine management system as the request for a propulsion torque. The inputs from the cruise control and the vehicle-speed limiter are processed in exactly the same manner.

Following this selection of the desired propulsive torque, should the situation arise, the vehicle-dynamics system (TCS, ESP) increases the desired torque value when there is the danger of wheel lockup and decreases it when the wheels show a tendency to spin.

Further external torque demands

The drivetrain's torque adaptation must be taken into account (drivetrain transmission ratio). This is defined for the most part by the ratio of the particular gear, or by the torque-converter efficiency in the case of automatic transmissions. On vehicles with an automatic transmission, the transmission control stipulates the torque demand during the gearshift. This is reduced in order to produce a comfortable, smooth gearshift, thus protecting the engine. In addition, the torque required by other engine-powered auxiliary systems (e.g. air-conditioning compressor, alternator, servo pump) is determined. This torque demand is calculated either by the auxiliary systems themselves or by the engine management system. Calculation is based on the required power and engine speed, and the engine management system adds up the various torque requirements.

The vehicle's driveability remains unchanged notwithstanding varying require-

ments from the auxiliary systems and changes in the engine's operating states.

Internal torque demands

At this stage, the idle-speed control and the active surge damper intervene.

For instance, if demanded by the situation, in order to prevent mechanical damage, or excessive smoke due to the injection of too much fuel, the torque limitation reduces the internal torque demand. In contrast to previous engine-management systems, limitations are no longer only applied to the injected fuel quantity, but instead, depending on the required effects, also to the particular physical quantity involved.

The engine's losses are also taken into account (e.g. friction, drive for the high-pressure pump). The torque represents the engine's measurable effects to the outside. However, the engine management system can only generate these effects in conjunction with the correct fuel injection together with the correct injection point, and the necessary marginal conditions as apply to the air system (e.g. boost pressure and exhaust-gas recirculation rate). The required injected fuel quantity is determined using the current combustion efficiency. The calculated fuel quantity is limited by a protective function (e.g. protection against overheating), and if necessary can be varied by smooth-running control. During engine start, the injected fuel quantity is not determined by external inputs such as those from the driver, but rather by the separate "start quantity" control function.

Actuator triggering

The resulting setpoint value for the injected fuel quantity is used to generate the triggering data for the injection pumps and/or the fuel injectors, and for defining the optimum operating point for the intake-air system.

Lighting technology

Automotive light sources

The most important light sources for the lighting systems on the vehicle front and rear are halogen lamps, bulbs, gas-discharge lamps and LEDs.

Thermal radiators

Thermal radiators generate light from heat energy. The major liability of the thermal radiator is its low working efficiency (below 10 %) which, relative to the gas-discharge lamp, leads to very low potential for luminous efficiency.

Incandescent (vacuum) bulb

Among the thermal radiators is the bulb (Fig. 1) whose tungsten filament (2) is enclosed by glass (1). A vacuum is created inside the glass, which is why the incandescent bulb is also known as a vacuum bulb.

At 10 to 18 lm/W (lumen/Watt), the luminous efficiency of an incandescent bulb is comparatively low. During bulb operation, the tungsten particles of the filament vaporize. The glass consequently darkens over the course of the bulb's service life. The vaporization of the particles ultimately leads to the filament breaking and thus failure of the lamp. For this reason,

incandescent bulbs as light sources for the headlamps have been replaced by halogen lamps. For cost reasons, however, incandescent bulbs continue to be used for other lights and as light sources in the passenger compartment. Even the lighting of passive display elements (e.g. fan, heating and air-conditioning controllers, LCD displays) is generally performed by incandescent bulbs, the color of which is changed by means of color filters for the application and design concerned.

Halogen lamp

There are two types of halogen lamp: with one or two tungsten filaments. The halogen lamps H1, H3, H7, HB3 and HB4 (see table at the end of the chapter) only have one filament. They are used as light sources for the low-beam, high-beam and fog lights.

The bulb is made of quartz glass. The quartz glass filters out the low UV content of the beam that halogen lamps emit. Unlike an incandescent bulb, the glass of a halogen lamp contains a halogen charge (iodine or bromine). This makes it possible for the filament to heat up to temperatures approaching tungsten's melting point (around 3,400 °C), thereby achieving commensurately high levels of luminous power.

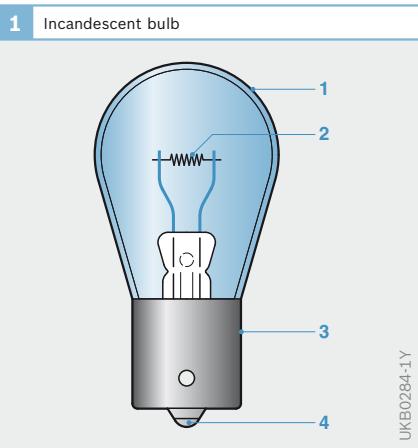
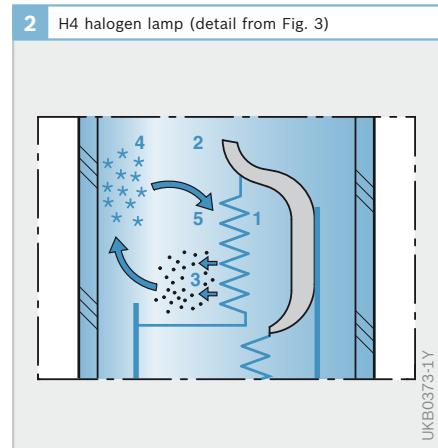


Fig. 1

- 1 Glass bulb
- 2 Filament
- 3 Lamp socket base
- 4 Electrical connection

Fig. 2

- 1 Tungsten filament
- 2 Halogen charge (iodine or bromine)
- 3 Evaporated tungsten
- 4 Halogenated tungsten
- 5 Tungsten deposits



Close to the hot bulb wall, vaporized tungsten particles combine with the filler gas to form a transparent gas (tungsten halide). This is stable within a temperature range of approximately 200 to 1,400 °C. Tungsten particles re-approaching the filament respond to the high temperatures at the filament by dispersing to form a consistent tungsten layer. This cycle (Fig. 2) limits the wear rate of the filament. In order to maintain this cycle, an external bulb temperature of approx. 300 °C is necessary. The glass therefore encloses the filament tightly. It remains clear throughout the entire service life of the lamp.

The rate of filament wear is also limited by the high pressure that is generated in the bulb, limiting the vaporization rate of the tungsten.

The H4 halogen lamp generates the light beam in the same way but has two filaments (Fig. 3, Items 2 and 3). This means

that only one lamp is required for each low-beam and high-beam headlamp.

The lower part of the low-beam filament is masked by a screen integrated in the headlamp. As a result, the light is only emitted into the upper part of the reflector (Fig. 8) and thereby prevents dazzling other road users.

Switching from low beam to high beam activates the second filament. Halogen lamps with an output of 60/55 W¹⁾ emit around twice as much light as incandescent bulbs with an output of 45/40 W. The high luminous efficiency of around 22 to 26 lm/W is primarily the result of the high filament temperature.

¹⁾ High beam/low beam

Gas-discharge lamps

Gas discharge describes the electrical discharge that occurs when an electrical current flows through a gas and causes it to emit radiation (examples: sodium-vapor lamps for street lighting and fluorescent lamps for interior lighting).

The discharge chamber of the gas-discharge lamp (Fig. 4, Item 3) is filled with the inert gas xenon and a mixture of metal halides. The electrical voltage is applied between two electrodes (4) protruding into the bulb. An electronic ballast unit is required for switching on and operation. Application of an ignition voltage in the 10 to 20 kV range ionizes the gas between the electrodes, producing an electrically conductive path in the form of a luminous arc. With the alternating current (400 Hz) applied, the metallic charge is vaporized due to the temperature increase inside the bulb and light is radiated.

Under normal circumstances the lamp requires several seconds to ionize all of the particles and generate full illumination. To accelerate this process, an increased starting current flows until this point.

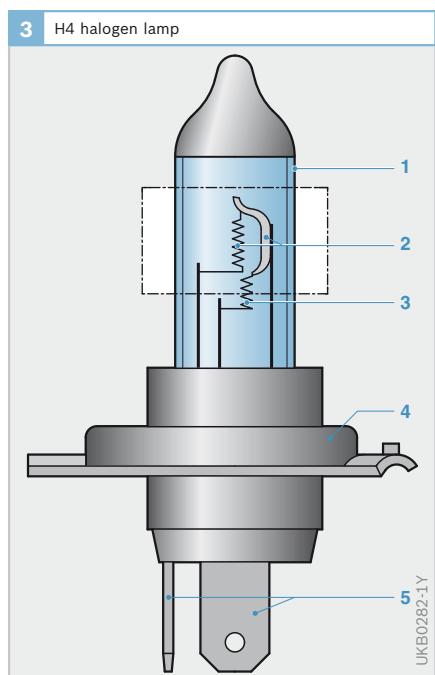


Fig. 3

- 1 Glass bulb
- 2 Low-beam filament with cap
- 3 High-beam filament
- 4 Lamp base
- 5 Electrical connection

When maximum luminous power is achieved, limitation of lamp current commences. A sustained operating voltage of only 85 V is sufficient to maintain the arc.

Light sources relying on the gas-discharge concept acquired new significance for automotive applications with the advent of the "Litronic" electronic lighting system. This concept features several crucial benefits compared with conventional bulbs:

- ▶ Greater range of the headlamp beam
- ▶ Brighter and more even carriageway illumination
- ▶ Longer service life, as there is no mechanical wear
- ▶ High luminous efficiency (approximately 85 lm/W) due to the emission spectrum being predominately in the visible spectral range

- ▶ Improved efficiency thanks to lower thermal losses
- ▶ Compact headlamp designs for smooth front-end styling

The D2/D4-series automotive gas-discharge lamps feature high-voltage-proof sockets and UV glass shielding elements. On the D1/D3-series models, the high-voltage electronics necessary for operation are also integrated in the lamp socket. All series can be broken down into two subcategories:

- ▶ Standard lamp (S lamp) for projection headlamps (Fig. 4) and
- ▶ Reflection lamp (R lamp) for reflection headlamps (Fig. 5). They have an integrated shutter (3) to create the light-dark cutoff, comparable with the shutter in the H4 lamp.

Until now, gas-discharge lamps with the type designations D1x and D2x were used. From 2007, the D3/D4-series will also be fitted as standard. These have a lower operating voltage, a different charge gas composition, and different arc geometries.

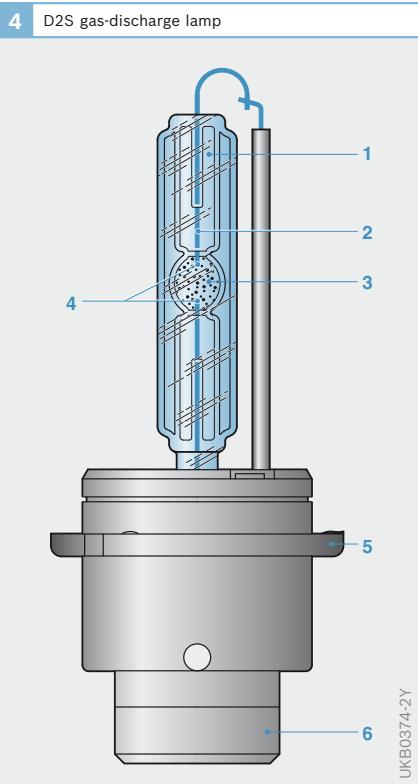


Fig. 4

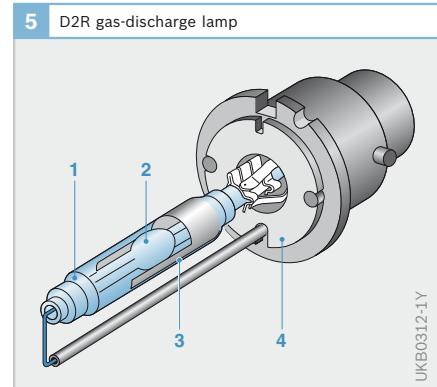
Gas-discharge lamp for projection headlamps

- 1 Glass capsule with UV shield
- 2 Electrical lead
- 3 Discharge chamber
- 4 Electrodes
- 5 Lamp base
- 6 Electrical connection

Fig. 5

Gas-discharge lamp for reflection headlamps

- 1 UV inert-gas bulb
- 2 Discharge chamber
- 3 Shutter
- 4 Lamp base



Light emitting diodes

The light emitting diode (LED) is an active light element. If an electrical voltage is applied, current flows through the chip. The electrons of the atoms of the LED chip are highly energized by the voltage. As light is emitted, they return to their initial state of low energy charge.

The 0.1 to 1 mm small semiconductor crystal is seated on a reflector that directs the light with pin-point precision.

LEDs are commonly used as light sources for lights on the rear of the vehicle, especially the additional stop lamps located in the center. They make it possible for a narrow, linear beam to be emitted.

By comparison with incandescent bulbs, LEDs are beneficial in that they emit maximum output in less than a millisecond. An incandescent bulb takes approximately 200 ms. LEDs, for example, are therefore able to emit the brake signal sooner and thus shorten the response time to the brake signal (brake pedal depressed) for drivers behind.

In the motor vehicle, LEDs are used as illuminators or in displays, in the interior they are used for lighting, in displays or display backlighting. In the lighting system, they find use as auxiliary stop lamps and tail lamps, and, increasingly in future, as daytime running lamps and in headlamps.

► Technical lighting variables

Luminous intensity

The brightness of light sources can vary. Luminous intensity serves as an index for comparing them. It is the visible light radiation that a light source projects in a specific direction.

The unit for defining levels of luminous intensity is the candela (cd), roughly equivalent to the illumination emitted by one candle. The brightness of an illuminated surface varies according to its reflective properties, the luminous intensity and the distance separating it from the light source.

Examples of permissible values

Stop lamp (individual): 60 to 185 cd
Tail lamp (individual): 4 to 12 cd
Rear fog lamp (individual): 150 to 300 cd
High beam (total, maximum): 225,000 cd

Luminous flux

Luminous flux is that light emitted by a light source that falls within the visible wavelength range.

Values are expressed in lumen (lm).

Illuminance

The illuminance is the luminous flux arriving at a given surface. It increases proportionally along with the light intensity, and decreases with the square of the distance.

Illuminance is expressed in lux (lx):
 $1 \text{ lx} = 1 \text{ lm/m}^2$

Range

The range is defined as the distance at which the illuminance in the light beam still has a given value (e.g. 1 lx). The geometric range is the distance at which the horizontal part of the light-dark cut-off is shown on the road surface with the headlamps on low beam.

¹⁾ The PES (Poly-Ellipsoid System) headlamp system works with an imaging optical lens. Unlike with conventional headlamps, the light pattern generated by the reflector is reproduced on the roadway by the lens together with a screen for creating the light-dark cutoff.

²⁾ Reflectors with small short focal length whose shape is calculated using special programs (CAL: Computer Aided Lighting). In this way, three separate reflectors for low beam, high beam and fog lamp can be accommodated within the same space needed by a conventional parabolic reflector, while luminous efficiency is increased at the same time.

³⁾ With faceted reflectors, the surface is divided into individually optimized segments. This results in reflector surfaces with high levels of homogeneity and sideways beam spread.

Main headlights (Europe)

Function

On the one hand, the main headlights must provide maximum visual range while at the same time ensuring that the glare effect for oncoming traffic is kept to a minimum and that light distribution immediately in front of the vehicle remains in line with the requirements of safe operation. It is vital to provide the lateral illumination needed to safely negotiate bends, i.e. the light must extend outward to embrace the verge of the road. Although it is impossible to achieve absolutely consistent luminance across the entire road surface, it is possible to avoid sharp contrasts in light density.

High beam

The high beam is usually generated by a light source located at the reflector's focal point, causing the light to be reflected outward along a plane extending along the reflector's axis (Fig. 7). The maximum luminous intensity which is available during high-beam operation is largely a function of the reflector's mirrored surface area.

In four and six-headlamp systems, in particular, purely parabolic high-beam reflectors can be replaced by units with complex geometrical configurations for simultaneous use of high and low beams.

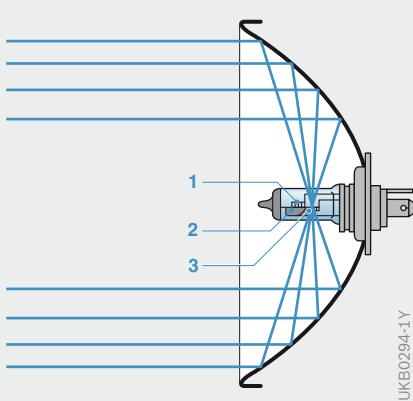
In these systems the high-beam component is designed to join with the low beam (simultaneous operation) to produce a harmonious overall high-beam distribution pattern. This strategy abolishes the annoying overlapping sector that would otherwise be present at the front of the light pattern.

Low beam (dipped beam)

The high traffic density on modern roads severely restricts the use of high-beam headlamps. The low beams serve as the primary source of light under normal conditions. Basic design modifications implemented within recent years are behind the substantial improvements in low-beam performance. Developments have included:

- ▶ Introduction of the asymmetrical low-beam pattern, characterized (RHD traffic) by an extended visual range along the right side of the road.
- ▶ Introduction of new headlamp systems featuring complex geometrical configurations (PES¹⁾, free-form surfaces²⁾, faceted reflectors³⁾) offering efficiency-level improvements of up to 50 %.
- ▶ Headlamp leveling control (also known as vertical aim control) devices adapt the attitude of the headlamps to avoid dazzling oncoming traffic when the rear

7 High beam (beam projection)



8 Low beam (beam projection)

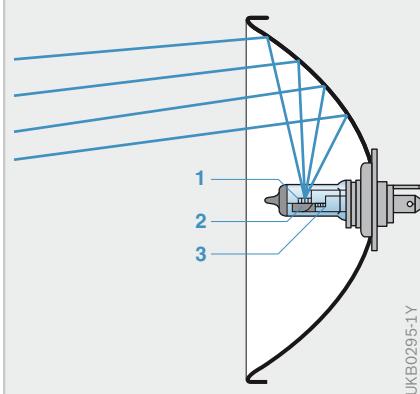


Fig. 7

- 1 Low-beam filament
- 2 Cap
- 3 High-beam filament at focal point

Fig. 8

- 1 Low-beam filament
- 2 Cap
- 3 High-beam filament

of the vehicle is heavily laden. Vehicles must also be equipped with headlamp washer systems.

- “Litronic” gas-discharge lamps supply more than twice as much light as conventional halogen lamps.

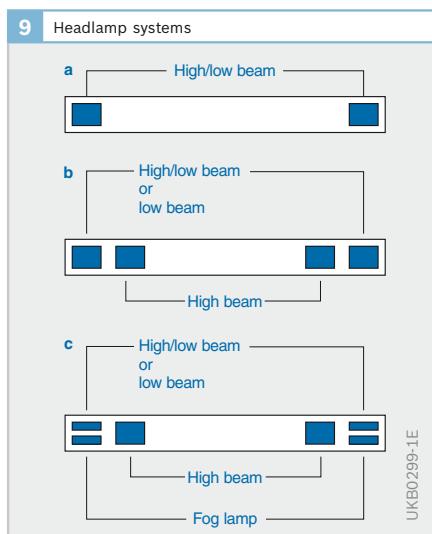
Operating concept

Low-beam headlamps need a light-dark cutoff in the light pattern. In the case of H4 halogen headlamps and Litronic headlamps with D2R bulbs, this is achieved by the image from the shield (H4) or the shutter (D2R). On headlamps for all-round use (H1-, H7-, HB11 bulbs), the light-dark cutoff is achieved by the special imaging of the filament.

Headlamp systems

Dual-headlamp systems rely on a single shared reflector for low- and high-beam operation, e.g. in combination with a dual-filament H4 bulb (Fig. 9 a).

In quad headlamp systems one pair of headlamps may be switched on in both modes or during low-beam operation only, while the other pair is operated exclusively for high-beam use (Fig. 9 b).



Six-headlamp systems differ from the quad configuration by incorporating a supplementary fog lamp within the main headlamp assembly (Fig. 9 c).

Main headlamps (North America)

High beam

The designs for high-beam headlamps are the same as in Europe. Facetted reflectors with, for example, HB5 or H7 lamps are used.

Low beam (dipped beam)

Headlamps with a light-dark cutoff that rely on visual/optical adjustment procedures have been approved in the USA since 1 May, 1997. This has made it possible to equip vehicles for Europe and the USA with headlamps of the same type and, in some cases, even the same reflectors.

Regulations

The regulations for the attachment and wiring of main headlamps are comparable with the European regulations (Federal Motor Vehicle Safety Standard [FMVSS] No. 108 and SAE Ground Vehicle Lighting Standards Manual).

An amendment to FMVSS 108 that entered effect in 1983 made it possible to start using headlamp units of various shapes and sizes with replaceable bulbs. These were known as the RBH, or Replaceable Bulb Headlamps.

Headlamp systems

North America mirrors European practice in employing dual, quad and six-headlamp systems.

Fig. 9

- a Dual-headlamp system
- b Quad-headlamp system
- c Six-headlamp system

Litronic

Overview

The "Litronic" (Light-Electronics) headlamp system uses xenon gas-discharge lamps that produce a powerful lighting effect despite the low front-end surface area requirement. The illumination of the carriageway represents a substantial improvement over that provided by conventional halogen units (Fig. 10).

The light generated contains a higher proportion of green and blue and is thus more similar to the spectral distribution of sunlight. Night-time driving is therefore less exacting for the driver.

Design

The components of the Litronic headlamp system are:

- ▶ Optical unit with xenon gas-discharge lamp (S lamp, R lamp; see "Gas-discharge lamps" section)
- ▶ Electronic ballast unit with igniter and ECU

10 Light pattern on the road (comparison)

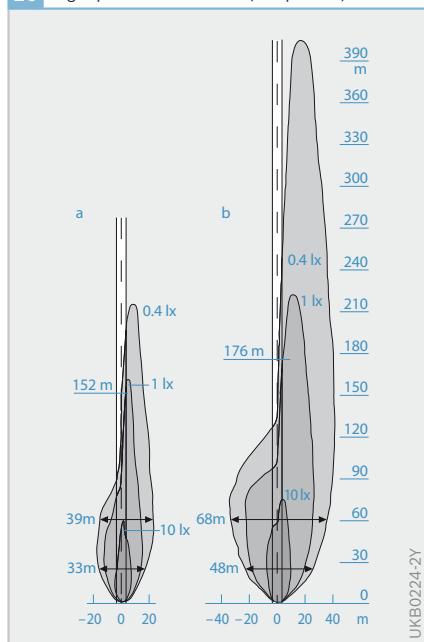


Fig. 10

- a H4 lamp
- b Litronic PES D2S lamp

Fig. 11

Electronic ballast unit
for 400 Hz alternating
current supply and
pulse ignition of the
lamp

- 1 Control unit
- 1a DC/DC converter
- 1b Shunt
- 1c DC/AC converter
- 1d Microprocessor
- 2 Igniter
- 3 Lamp socket
- 4 D2S lamp
- U_B Battery voltage

For low beam, the headlamps with xenon gas-discharge lamps are installed in a quad system that is combined with the high-beam headlamps of the conventional design.

With the Bi-Litronic system, however, the low and high beams are generated by only one gas-discharge lamp from a dual-headlamp system.

An integral part of the headlamp is the electronic ballast unit responsible for activating and monitoring the lamp.

Its functions include:

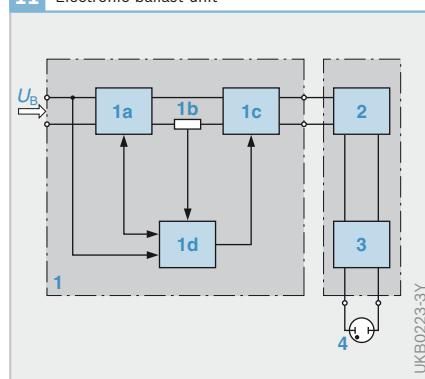
- ▶ Ignition of the gas discharge (voltage 10 to 20 kV)
- ▶ Regulated power supply during the warm-up phase when the lamp is cold
- ▶ Demand-oriented supply in continuous operation

The control units for the individual lamp types are generally developed for a specific design type and are not universally interchangeable.

Operating principle

In the gas-discharge lamp, the arc is ignited when the light is switched on. A high voltage of 18 to 20 kV is required for this to be possible. 85 V are required to maintain the arc after ignition. The voltage is generated and regulated by an electronic

11 Electronic ballast unit



ballast unit (igniter, Fig. 11). After ignition, the gas-discharge lamp is operated for approximately 3 secs with an elevated starting current (approximately 2.6 A) so that it achieves maximum luminosity with minimal delay. The bulb's output in this period is anywhere up to 75 W. During continuous-running operation, it is 35 W.

The maximum luminous efficiency of approximately 90 lm/W is achieved once the plasma has heated the quartz glass to approximately 900 °C. Once the gas-discharge lamp has achieved maximum luminosity, the ballast unit reduces the current output to the bulb to approximately 0.4 A for continuous-running operation.

Fluctuations in the vehicle system voltage are for the most part compensated for by the ballast unit to prevent luminous flux variations. If the bulb goes out, e.g. due to an extreme voltage drop (below 9 V) or increase (above 16.5 V) in the vehicle electrical system, it is automatically reignited without delay. The reignition is limited to five attempts for safety reasons. The power supply is then interrupted by the ballast unit.

Bi-Litronic "Reflection"

The "Reflection" Bi-Litronic system makes it possible to generate the low and high beams using only one gas-discharge lamp (DR2 lamp) from a dual-headlamp system. The concept relies on an electromechanical positioner that responds to the high/low-beam switch by varying the attitude of the gas-discharge lamp within the reflector. It alternates between two different positions to generate separate projection patterns for low and high beam (Fig. 12).

This layout gives Bi-Litronic the following major advantages:

- ▶ Xenon light for high-beam operation
- ▶ Visual guidance provided by the continuous shift in light distribution from close to extended range

12 Bi-Litronic "Reflection"

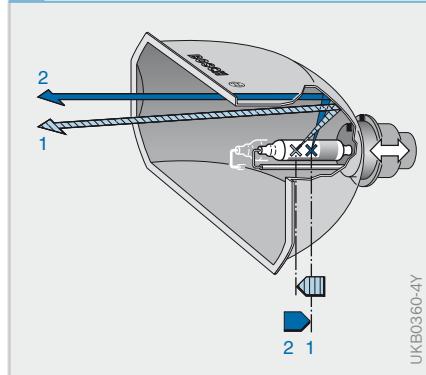


Fig. 12

- 1 Low beam
- 2 High beam

13 Litronic 4 system in reflection headlamp with integral dynamic headlamp leveling control (example)

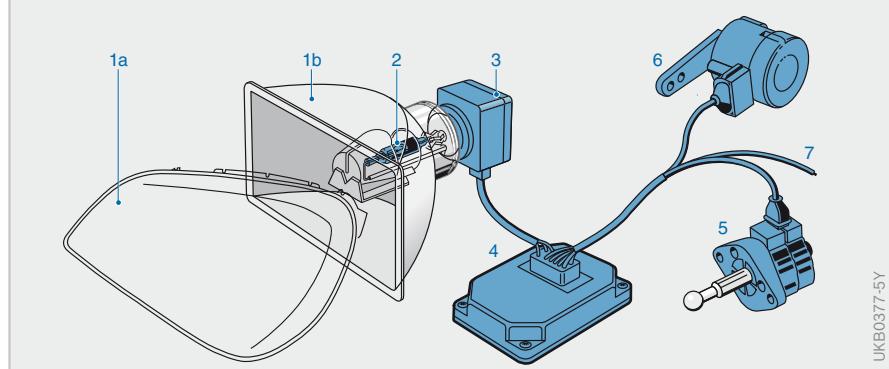


Fig. 13

- 1a Lens with or without scatter optics
- 1b Reflector
- 2 Gas-discharge lamp
- 3 Igniter
- 4 Control unit
- 5 Stepper motor
- 6 Axle sensor
- 7 To the vehicle electrical system

- ▶ Substantial reduction in space requirements as compared to a conventional quad headlamp system
- ▶ Lower costs through the use of just one gas-discharge bulb and one ballast unit per headlamp
- ▶ Greater freedom in headlamp design due to the individual reflector shape.

Special design variants of the Bi-Litronic "Reflection" lamp involve solutions in which the entire reflector is moved or individual components of the bulb cover are opened.

Bi-Litronic "Projection"

The Bi-Litronic "Projection" system is based on a PES Litronic headlamp. It shifts the position of the shutter for the light-dark cutoff to provide xenon light for high-beam operation.

With lens diameters of 60 and 70 mm, the Bi-Litronic "Projection" is the most compact combined low- and high-beam headlamp on the market, yet it still provides superb illumination.

The essential advantages of the Bi-Litronic "Projection" are:

- ▶ Xenon light for high-beam operation
- ▶ Most compact solution for high and low beams
- ▶ Modular system

Headlamp leveling control

Function

Without headlamp leveling control, the range of the headlamps would alter with a change in load or operating condition of the vehicle (constant-speed travel, stationary, acceleration, braking). The headlamp leveling control adjusts the tilt angle of the low beam to the tilt angle of the vehicle body. This results in a permanently good visual range with no dazzling of oncoming traffic under all load conditions.

Fig. 14

- 1 Low beam
- 2 High beam

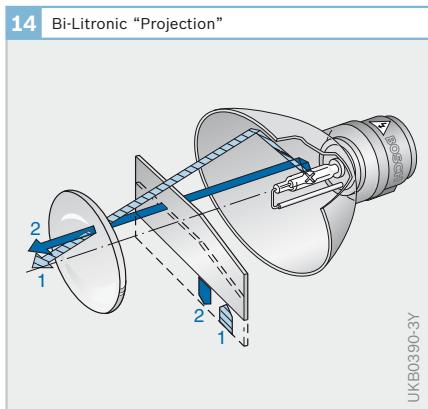
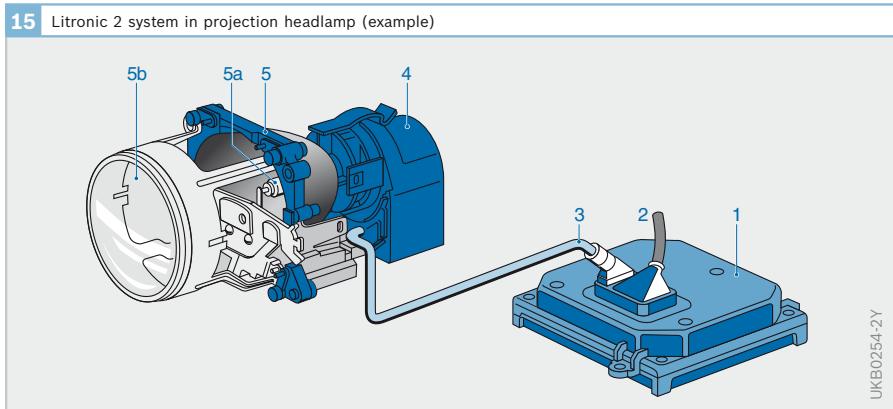


Fig. 15

- 1 Control unit
- 2 To the vehicle electrical system
- 3 Shielded cable
- 4 Igniter
- 5 Projection module
- 5a D2S lamp
- 5b Lens



Designs

All headlamp leveling controls feature actuators that move the headlamp reflector (housing-type design) or headlamp unit up and down. Automatic systems rely on sensors that monitor suspension travel as the basis for generating proportional signals for transmission to the aiming actuators. Manually operated units employ a switch near the driver's seat to control the setting.

Automatic headlamp leveling control

Automatic headlamp-leveling control systems fall into two categories: static and dynamic. While static systems compensate for load variations in the luggage and passenger compartments, dynamic systems also correct headlamp aim during acceleration – both from standing starts and when underway – and when braking.

The components of a typical headlamp-leveling control system include (Fig. 16):

- Sensors on the vehicle axles (Items 3 and 6) to measure the body's inclination or tilt angle.

- An ECU (5) that uses the sensor signals as the basis for calculating the vehicle's pitch angle. The ECU compares this data with the specified values and responds to deviations by transmitting appropriate triggering signals to the headlamps' servomotors.

- Servomotors (2) to adjust the headlamps to the correct angle.

Static system

In addition to the signals from the suspension sensors, the static system's control unit also receives a speed signal from the electronic speedometer. The controller relies on this signal to decide whether the vehicle is stationary, undergoing a dynamic change in speed, or proceeding at a constant speed. Automatic systems based on the static concept always feature substantial response inertia, so the system corrects only those vehicle inclinations that are consistently registered over relatively long periods.

Each time the vehicle has pulled away, it corrects the headlamp adjustment as a function of the vehicle's load. This adjustment is checked again when the vehicle steadies into constant-speed travel and is then corrected if necessary. Deviations between the target and actual position are evened out by the system.

The static system only generally requires a sensor on the rear axle of the vehicle. A DC motor is used as each headlamp's actuator.

16 Principle of an automatic headlamp-leveleing control system (dynamic system)

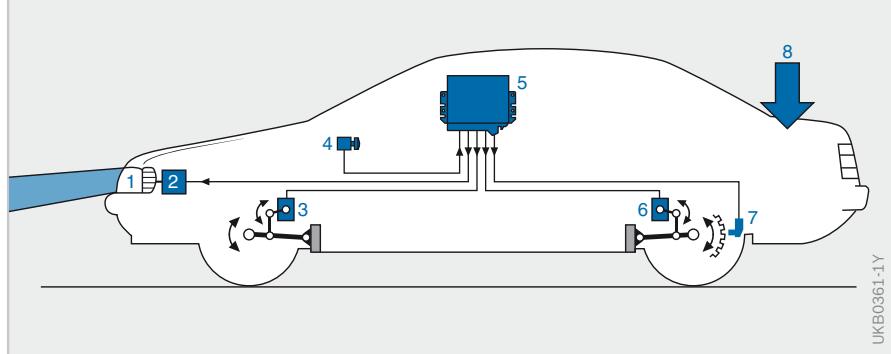


Fig. 16

- | | |
|---|--------------------------------|
| 1 | Headlamp |
| 2 | Actuator |
| 3 | Front-suspension travel sensor |
| 4 | Light switch |
| 5 | Electronic control unit |
| 6 | Rear-suspension travel sensor |
| 7 | Speed sensor |
| 8 | Load |

Dynamic system

The dynamic automatic system relies on two distinct operating modes to ensure optimal headlamp orientation under all driving conditions. Supplementary capabilities in speed-signal analysis over the static headlamp leveling control endow the system with the ability to differentiate between acceleration and braking.

With the vehicle driving at constant speed, the dynamic system, like the static system, remains in the range that features a high level of damping but as soon as the controller registers acceleration or braking, the system immediately switches to its dynamic mode. Faster signal processing and the higher servomotor adjustment speed allow the headlamp range to be re-adjusted within fractions of a second. Following acceleration or braking, the system automatically reverts to operation in its delayed-response mode.

Due to the greater dynamics requirements, the dynamic system needs one sensor per vehicle axle and rapid stepping motors to adjust the headlamps.

Adaptive lighting systems

Adaptive frontlighting system (AFS)

From 2007, function enhancements for headlamp systems based on a new EC control are permitted. The vehicle may

then also have motorway lights, adverse weather lights and city lights. The optimum light pattern for each of the functions is identified and automatically selected by the vehicle electronics in response to evaluation of various vehicle sensors.

The first vehicles with AFS systems were registered back in mid-2006 thanks to an EU waiver for road traffic.

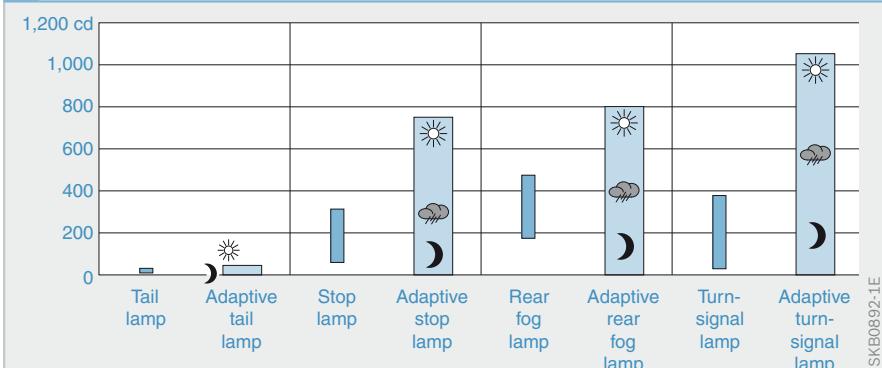
Adaptive rearlighting system (ARS)

Until now, the rear lights for vehicle perimeter lighting were equipped with single level switching. Depending on the type and design, these produced an invariable luminous intensity within the legal limit values.

Today, a multitude of sensors are used to determine environmental parameters and light conditions (brightness, dirt, visual range, wet conditions, etc.). To achieve optimum visibility (sufficient luminous intensity without excessive glare), the rear lights may in future vary luminous intensity to suit the vehicle surrounding (Fig. 17).

A stop lamp, for example, would be lit with high luminous intensity in sunlight and with low luminous intensities at night to ensure that other road users are able to recognize and draw the correct conclusion from the action of the vehicle.

17 Permitted luminous intensity of adaptive rear lights by comparison with conventional systems



Cornering lights (Europe)

The cornering lights function that has been approved for use in Germany since 2003 improves visual range on corners and in turning situations. This is made possible by a variation of the horizontal illumination of the area in front of the vehicle. With static cornering lights, this is achieved by supplementary reflectors being switched; with dynamic cornering lights, the headlamp module is pivoted laterally (Fig. 18).

During the control process, the light module or the reflector elements are pivoted by a stepping motor. The pivot angle and pivot speed are calculated by the cornering lights ECU as a function of vehicle speed and the steering angle. Sensors detect the adjustment angle of the headlamps and use failsafe algorithms to prevent dazzling of oncoming traffic in the event of a system malfunction.

18 Aiming strategy of the basic and turning modules of a static/dynamic cornering headlamp (left side)

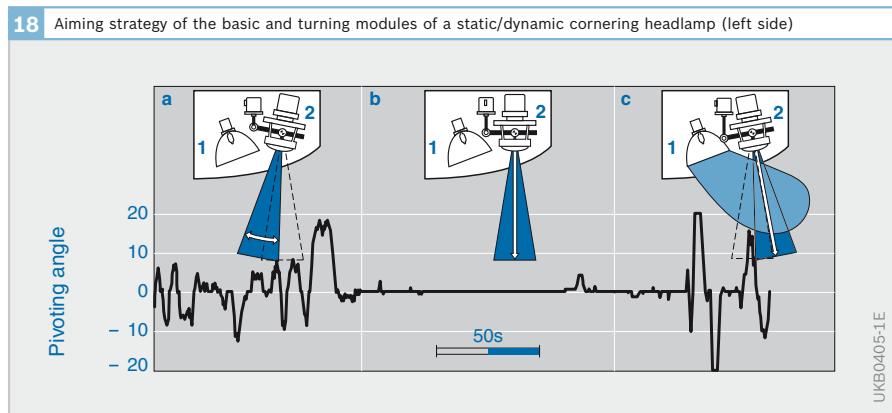


Fig. 18

- a "Country road/bends" position
- b "Motorway" position
- c "Town/turning" position
- 1 Turning module
- 2 Basic module

19 Improvement in the adaptive light pattern of the cornering headlamps

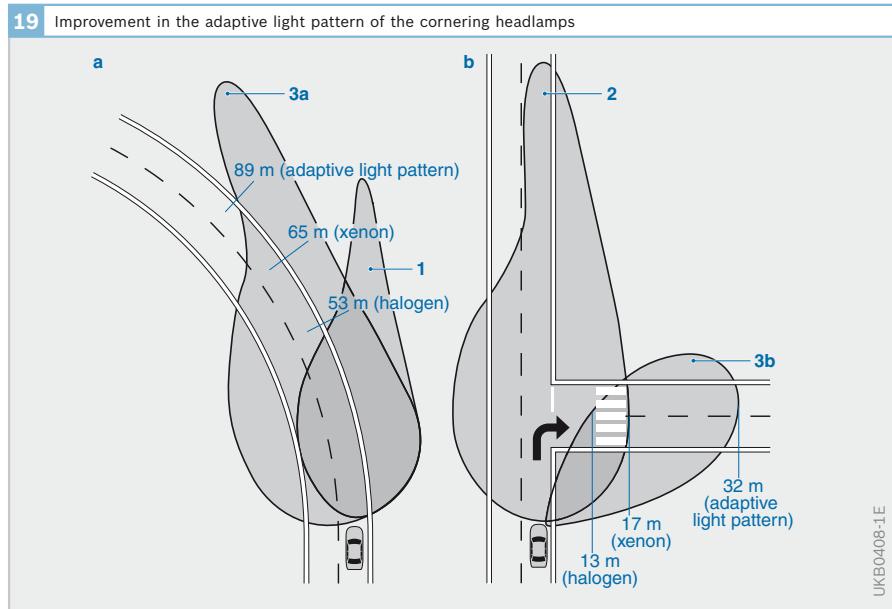


Fig. 19

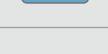
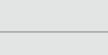
- a Dynamic cornering lights, cornering to the left
- b Static cornering light, turning to the right
- 1 Light pattern of halogen headlamps
- 2 Light pattern of xenon headlamps
- 3a Adaptive light pattern: dynamic cornering lights
- 3b Adaptive light pattern: static cornering lights

1 Specifications for motor-vehicle bulbs (2-wheeled vehicles not included)

Application	Category	Voltage rated value V	Power rated value W	Luminous flux setpoint value Lumen	IEC base type	Illustration
High beam, low beam	R2	6 12 24	45/40 ¹⁾ 45/40 55/50	600 min/ 400–550 ¹⁾	P 45 t-41	
Fog lamp, high beam, low beam in 4-headlamps	H1	6 12 24	55 55 70	1,350 ²⁾ 1,550, 1,900	P14.5 e	
Fog lamp, high beam	H3	6 12 24	55 55 70	1,050 ²⁾ 1,450 1,750	PK 22 s	
High beam, low beam	H4	12 24	60/55 75/70	1,650/ 1,000 ^{1), 2)} , 1,900/1,200	P 43 t-38	
High beam, low beam in 4-headlamps, fog lamp	H7	12 24	55 70	1,500 ²⁾ 1,750	PX 26 d	
Fog lamp, static cornering light	H8	12	35	800	PGJ 19-1	
High beam	H9	12	65	2,100	PGJ 19-5	
Fog lamp	H10	12	42	850	PY 20 d	
Low beam, fog lamp	H11	12 24	55 70	1,350 1,600	PGJ 19-2	
Low beam in 4-headlamps	HB4	12	55	1,100	P 22 d	
High beam in 4-headlamp	HB3	12	60	1,900	P 20 d	
Low beam, high beam	D1S	85 12 ⁵⁾	35 approx. 40 ⁵⁾	3,200	PK 32 d-2	
Low beam, high beam	D2S	85 12 ⁵⁾	35 approx. 40 ⁵⁾	3,200	P 32 d-2	
Low beam, high beam	D2R	85 12 ⁵⁾	35 approx. 40 ⁵⁾	2,800	P 32 d-3	
Stop, turn-signal, rear fog, reversing lamp	P 21 W PY 21 W ⁶⁾	6 12 24	21	460 ³⁾	BA 15 s	

Table 1

1 Specifications for motor-vehicle bulbs (continued)

Application	Category	Voltage rated value V	Power rated value W	Luminous flux setpoint value Lumen	IEC base type	Illustration
Stop lamp/ tail lamp	P 21/5 W	6	21/5 ⁴⁾	440/35 ³⁾ , ⁴⁾	BAY 15 d	
		12	21/5	440/35 ³⁾ , ⁴⁾		
		24	21/5	440/40 ³⁾		
Side-marker lamp, tail lamp	R 5 W	6	5	50 ³⁾	BA 15 s	
		12				
		24				
Tail lamp	R 10 W	6	10	125 ³⁾	BA 15 s	
		12				
		24				
Daytime running light	P 13 W	12	13	250 ³⁾	PG 18.5 d	
Stop lamp, turn signal	P 19 W	12	19	350 ³⁾	PGU 20/1 PGU 20/2	
	PY 19 W	12	19	215 ³⁾		
Rear fog lamp, reversing lamp, front turn signal	P 24 W	12	24	500 ³⁾	PGU 20/3 PGU 20/4	
	PY 24 W	12	24	300 ³⁾		
Stop, turn-signal, rear fog, revers- ing lamp	P 27 W	12	27	475 ³⁾	W 2.5 x 16 d	
	P 27/7 W	12	27/7	475/36 ³⁾		
Stop lamp/ tail lamp	C 5 W	6	5	45 ³⁾	SV 8.5	
		12				
		24				
Reversing lamp	C 21 W	12	21	460 ³⁾	SV 8.5	
Side-marker lamp	T 4 W	6	4	35 ³⁾	BA 9 s	
		12				
		24				
Side-marker lamp, license-plate lamp	W 5 W	6	5	50 ³⁾	W 2.1 x 9.5 d	
		12				
		24				
Side-marker lamp, license-plate lamp	W 3 W	6	3	22 ³⁾	W 2.1 x 9.5 d	
		12				
		24				

¹⁾ High/low beam. ²⁾ Setpoint values at test voltage of 6.3; 13.2 or 28.0 V.³⁾ Setpoint values at test voltage of 6.75; 13.5 or 28.0 V. ⁴⁾ Main/secondary filament.⁵⁾ With ballast unit. ⁶⁾ Yellow-light version.

Electronic stability program (ESP)

The electronic stability program (ESP) is a closed-loop system designed to improve driveability through programmed intervention in the brake system and/or drivetrain. The integrated functionality of ABS¹⁾ prevents the wheels from locking when the brakes are applied, while TCS²⁾ inhibits wheel spin during acceleration. The overall role of ESP is to prevent the vehicle's tendency to "plow" or become unstable and break away to the side, provided the vehicle remains within its physical limits.

Braking is activated on individual wheels in a targeted manner, such as the inner rear wheel to counter understeer, or the outer front wheel during oversteer, and helps to keep the vehicle's course stable under all driving conditions. ESP can also accelerate the driven wheels by specific engine-control interventions to ensure the stability of the vehicle.

Using this *individual control* concept, the system has two options for steering the vehicle: it can brake selected wheels (selective braking) or accelerate the driven wheels.

Figure 1 shows ESP control in a schematic diagram with

- ▶ The sensors that determine the controller input parameters
- ▶ The ESP control unit with its hierarchically-structured controller, featuring a higher-level vehicle dynamics controller and the subordinate slip controllers
- ▶ The actuators used for control of braking, drive and side forces

Controller hierarchy of ESP

Higher-level vehicle dynamics controller

Function

The vehicle dynamics controller is responsible for

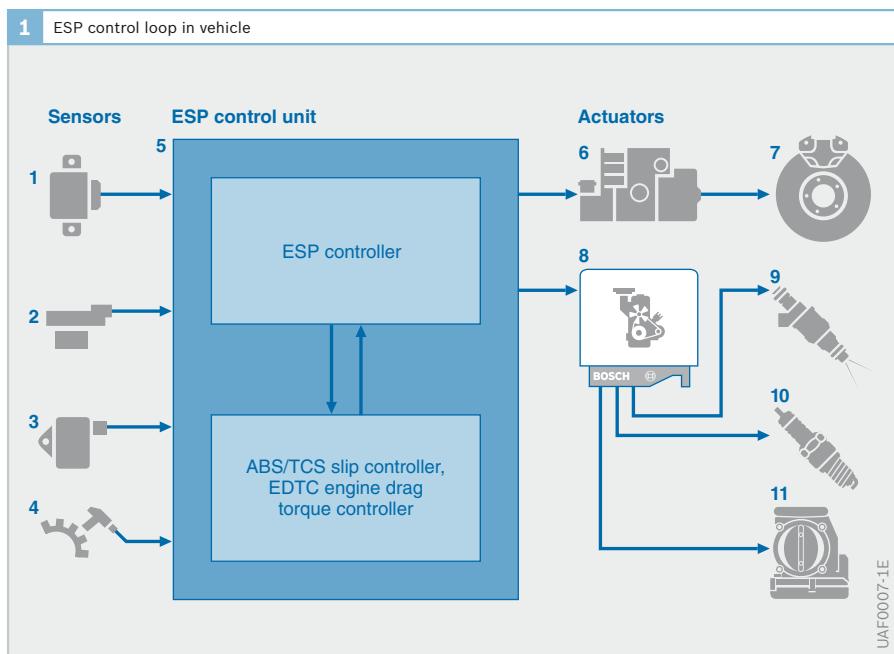
- ▶ Determining current vehicle status based on the yaw-rate signal and the sideslip angle estimated by the "monitor", and then

- 1) Antilock brake system
- 2) Traction-control system

Fig. 1

- 1 Yaw-rate sensor with lateral-acceleration sensor
- 2 Steering-wheel-angle sensor
- 3 Brake-pressure sensor
- 4 Wheel-speed sensors
- 5 ESP control unit
- 6 Hydraulic modulator
- 7 Wheel brakes
- 8 Engine-management system ECU
- 9 Fuel injection

- Only for gasoline engines:
- 10 Ignition-timing intervention
- 11 Throttle-valve intervention (ETC)



UAF0007-1E

- ▶ achieving maximum possible convergence between vehicle response in the limit range and its characteristics in the normal operating range (nominal behavior).

The following components register driver commands and the system evaluates their signals as the basis for defining nominal behavior:

- ▶ Engine-management system (e.g. apply accelerator pedal)
- ▶ Brake-pressure sensor (e.g. apply brakes) or
- ▶ Steering-wheel-angle sensor (turning the steering wheel)

At this point the driver command is defined as the specified response. The coefficient of friction and the vehicle speed are also included in the calculation. The "monitor" estimates these factors based on signals transmitted by the sensors for

- ▶ Wheel speed
- ▶ Lateral acceleration
- ▶ Braking pressures and
- ▶ Yaw velocity

The desired vehicle response is brought about by generating a yaw moment acting on the vehicle. In order to generate the desired yaw moment, the system influences the tire-slip rate, and thus indirectly the longitudinal and side forces. The system influences the tire slip by varying the desired specifications for slip rate, which must then be executed by the subordinated ABS and TCS controllers.

The intervention process is designed to maintain the handling characteristics that the vehicle manufacturer intended the vehicle to have and to serve as the basis for ensuring consistently reliable control.

The vehicle dynamics controller generates the specified yaw moment by relaying corresponding slip-modulation commands to the selected wheels.

The subordinate-level ABS and TCS controllers trigger the actuators governing the brake hydraulics and the engine-management system using the data generated in the ESP controller.

Antilock brake system (ABS)

The antilock brake system (ABS) detects incipient lock on one or more wheels and makes sure that the brake pressure remains constant or is reduced. By so doing, it prevents the wheels from locking up and the vehicle remains steerable.

Wheel-speed sensors

The speed of rotation of the wheels is an important input variable for the ABS control system. Wheel-speed sensors detect the speed of rotation of the wheels and pass the electrical signals to the control unit.

A car may have three or four wheel-speed sensors depending on which version of the ABS system is fitted (ABS system versions). The speed signals are used to calculate the degree of slip between the wheels and the road surface and therefore detect incipient lock on individual wheels.

Electronic control unit

The ECU processes the information received from the sensors according to defined mathematical procedures (control algorithms). The results of these calculations form the basis for the triggering signals sent to the hydraulic modulator.

Hydraulic modulator

The hydraulic modulator incorporates a series of solenoid valves that can open or close the hydraulic circuits between the master cylinder (Fig. 2, Item 1) and the wheel-brake cylinders (4). In addition, it can connect the wheel-brake cylinders to the return pump (6). Solenoid valves with two hydraulic connections and two valve positions are used (2/2 solenoid valves).

The intake valve (7) between the master cylinder and the wheel-brake cylinder controls pressure build-up, while the exhaust valve (8) between the wheel-brake cylinder and the return pump controls pressure release. There is one such pair of solenoid valves for each wheel-brake cylinder.

Under normal conditions, the solenoid valves in the hydraulic modulator are at the “pressure build-up” setting. That means the intake valve is open. The hydraulic modulator then forms a straight-through connection between the master cylinder and the wheel-brake cylinders. Consequently, the brake pressure generated in the master cylinder when the brakes are applied is transmitted directly to the wheel-brake cylinders at each wheel.

As the degree of brake slip increases due to braking on a slippery surface or panic braking, the risk of the wheels locking up also increases. The solenoid valves are then switched to the “maintain pres-

sure” setting. The connection between the master cylinder and the wheel-brake cylinder is shut off (intake valve is closed) so that any increase of pressure in the master cylinder does not lead to an increase in brake pressure.

If the degree of slip of any of the wheels increases further despite this action, the pressure in the wheel-brake cylinder(s) concerned must be reduced. To achieve this, the solenoid valves are switched to the “pressure release” setting. The intake valve is still closed, and in addition, the exhaust valve opens to allow the return pump integrated in the hydraulic modulator to draw brake fluid from the brake(s) concerned in a controlled manner. The brake pressure in the wheel-brake cylinder is thus reduced so that wheel lock-up does not occur.

Fig. 2

- 1 Master cylinder with expansion tank
- 2 Brake booster
- 3 Brake pedal
- 4 Wheel brake with wheel-brake cylinder

Hydraulic modulator with

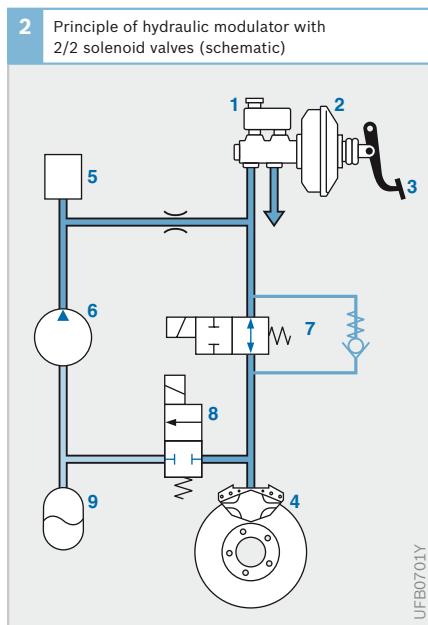
- 5 Damping chamber
- 6 Return pump
- 7 Intake valve
- 8 Exhaust valve
- 9 Brake-fluid accumulator

Intake valve:
shown in open
setting

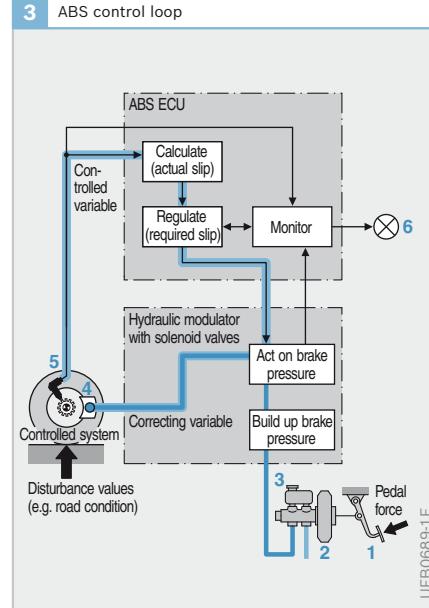
Exhaust valve:
shown in closed
setting

Fig. 3

- 1 Brake pedal
- 2 Brake booster
- 3 Master cylinder with expansion tank
- 4 Wheel-brake cylinder
- 5 Wheel-speed sensor
- 6 Warning lamp

**ABS control loop****Overview**

The ABS control loop (Fig. 3) consists of the following:



Controlled system

- ▶ The vehicle and its wheel brakes
- ▶ The wheels and the friction pairing of tire and road surface

The disturbance values affecting the control loop

- ▶ Changes in the frictional connection between the tires and the road surface caused by different types of road surface and changes in wheel load, e.g. when cornering
- ▶ Irregularities in the road surface causing the wheels and axles to vibrate
- ▶ Lack of circularity of the tires, low tire pressure, worn tire tread, differences in circumference between wheels, (e.g. spare wheel)
- ▶ Brake hysteresis and fade
- ▶ Differences in master-cylinder pressure between the two brake circuits

Controller

- ▶ The wheel-speed sensors
- ▶ The ABS control unit

Controlled variables

- ▶ Wheel speed and, derived from it, wheel deceleration
- ▶ Wheel acceleration and brake slip

The reference variable

- ▶ The foot pressure applied to the brake pedal by the driver - amplified by the brake booster - generates the brake pressure in the brake system

The correcting variable

- ▶ Brake pressure in the wheel-brake cylinder

Controlled system

The data-processing operations performed by the ABS control unit are based on the following simplified controlled system:

- ▶ A non-driven wheel
- ▶ A quarter of the vehicle's mass apportioned to that wheel
- ▶ Wheel brake

- ▶ An idealized coefficient of friction slip curve (substitute for the friction pairing of tire and road)

That curve is divided into a stable zone with a linear gradient and an unstable zone with a constant progression (μ_{HFmax}).

As an additional simplification, there is also an assumed initial straight-line braking response that is equivalent to a panic-braking reaction.

Figure 4 shows the relationships between brake torque M_B (the torque that can be generated by the brake through the tire), or road frictional torque M_R (torque that acts against the wheel through the friction pairing of tire and road surface), and time t , as well as the relationships between the wheel deceleration ($-a$) and time t , whereby the brake torque increases in linear fashion over time. The road frictional torque lags slightly behind the brake torque by the time delay T , as long as the braking sequence is within the stable zone

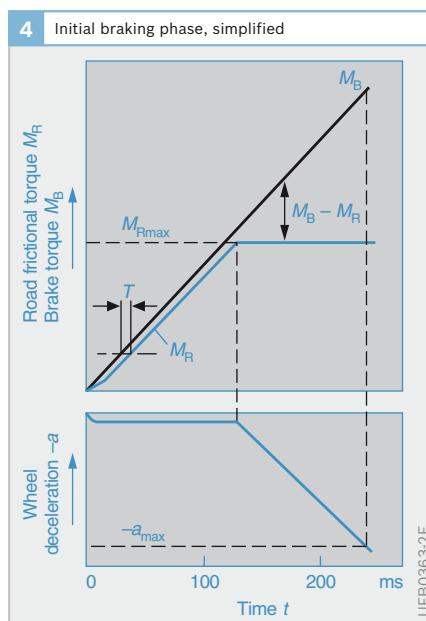


Fig. 4

- (-a) Wheel deceleration
- (-a_{max}) Maximum wheel deceleration
- M_B Brake torque
- M_R Road frictional torque
- M_{Rmax} Maximum road frictional torque
- T Time delay

of the curve for coefficient of friction versus brake slip. After about 130 ms, the maximum level (μ_{HFmax}) – and therefore the unstable zone – of the curve for coefficient of friction versus brake slip is reached. From that point on, the curve for coefficient of friction versus brake slip states that while the brake torque M_B , continues to rise at an undiminished rate, the road frictional torque M_R , cannot increase any further and remains constant. In the period between 130 and 240 ms (this is when the wheel locks up), the minimal torque difference $M_B - M_R$, that was present in the stable zone rises rapidly to a high figure. That torque difference is a precise measure of the wheel deceleration ($-a$) of the braked wheel (Fig. 4, bottom). In the stable zone, the wheel deceleration is limited to a small rate, whereas in the unstable zone it increases rapidly, according to the amount. As a consequence, the curve for coefficient of friction versus wheel slip reveals opposite characteristics in the stable and unstable zones. The ABS exploits these opposing characteristics.

Controlled variables

An essential factor in determining the effectiveness of an ABS control system is the choice of controlled variables. The basis for that choice is the wheel-speed sensor signals from which the ECU calculates the deceleration/acceleration of the wheel, brake slip, the reference speed and the vehicle deceleration. On their own, neither the wheel deceleration/acceleration nor the brake slip are suitable as controlled variables because, under braking, a driven wheel behaves entirely differently to a non-driven wheel. However, by combining those variables on the basis of appropriate logical relationships, good results can be obtained.

As brake slip is not directly measurable, the ECU calculates a quantity that approximates to it. The basis for the calculation is the reference speed, which represents the

speed under ideal braking conditions (optimum degree of brake slip). So that speed can be determined, the wheel-speed sensors continuously transmit signals to the ECU for calculating the speed of the wheels. The ECU takes the signals from a pair of diagonally opposed wheels (e.g. right front and left rear) and calculates the reference speed from them. Under partial braking, the faster of the two diagonally opposite wheels generally determines the reference speed. If the ABS cuts in under emergency braking, the wheel speeds will be different from the vehicle speed and can thus not be used for calculating the reference speed without adjustment. During the ABS control sequence, the ECU provides the reference speed based on the speed at the start of the control sequence and reduces it at a linear rate. The gradient of the reference-speed graph is determined by analyzing logical signals and relationships.

If, in addition to the wheel acceleration/deceleration and the brake slip, the vehicle deceleration is brought into the equation as an additional quantity, and if the logical circuit in the ECU is modulated by computation results, then ideal brake control can be achieved. This concept has been realized in the Bosch antilock brake system (ABS).

Traction-control system (TCS)

The antilock brake system (ABS) prevents the wheel lock when the brakes are applied by lowering the wheel brake pressures. The traction-control system (TCS) prevents wheel spin by reducing the drive torque at each driven wheel.

In addition to this safety-relevant task of ensuring the stability and steerability of the vehicle when accelerating, TCS also improves the traction of the vehicle by regulating the optimum slip. The upper limit here is, of course, set by the traction requirement stipulated by the driver.

The TCS regulates the slip of the driven wheels as quickly as possible to the optimum level. To do this the system first de-

termines a setpoint value for the slip. This value depends on a number of factors that represent the current driving situation. These factors include:

- ▶ The basic characteristic for TCS reference slip (based on the slip requirement of a tire during acceleration)
- ▶ Effective coefficient of friction
- ▶ External tractive resistance (deep snow, rough road, etc.)
- ▶ Yaw velocity, lateral acceleration, and steering angle of the vehicle

TCS interventions

The measured wheel speeds and the respective drive slip can be influenced by changing the torque balance M_{tot} at each drive wheel. The torque balance M_{tot} at each driven wheel results from the drive torque $M_{\text{Kar}}/2$ at this wheel, the respective brake torque M_{Br} and the road torque M_{Str} (Fig. 5).

$$M_{\text{tot}} = M_{\text{Kar}}/2 + M_{\text{Br}} + M_{\text{Str}}$$

(M_{Br} and M_{Str} are negative here.)

This balance can obviously be influenced by the drive torque M_{Kar} provided by the engine as well as by the brake torque M_{Br} .

Both these parameters are therefore correcting variables of the TCS which can be used to regulate the slip at each wheel to the reference slip level.

In gasoline-engine vehicles, the drive torque M_{Kar} can be controlled using the following engine-control interventions:

- ▶ Throttle valve (throttle valve adjustment)
- ▶ Ignition system (ignition-timing advance)
- ▶ Fuel-injection system (phasing out individual injection pulses)

In diesel-engine vehicles, the drive torque M_{Kar} is influenced by the electronic diesel control system (EDC) (reduction in the injected fuel quantity).

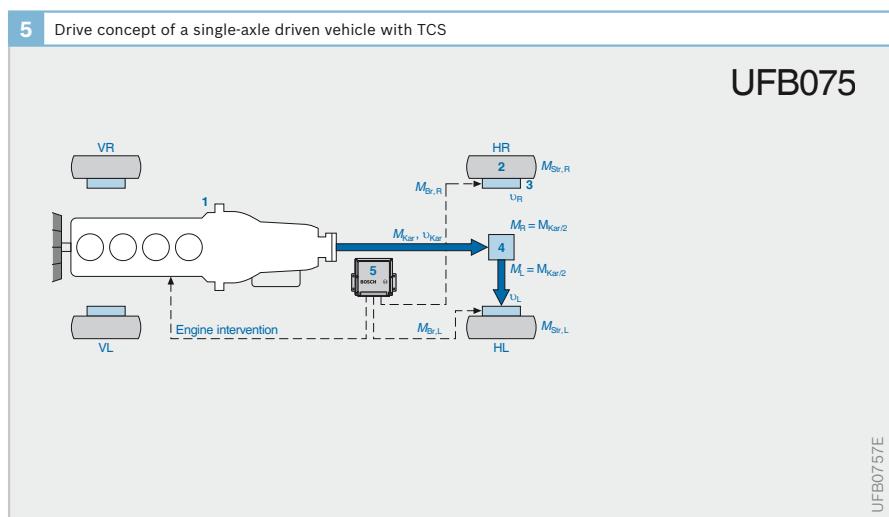
The brake torque M_{Br} can be regulated for each wheel via the brake system. The TCS function requires the original ABS hydraulic system to be expanded because of the need for active pressure build-up.

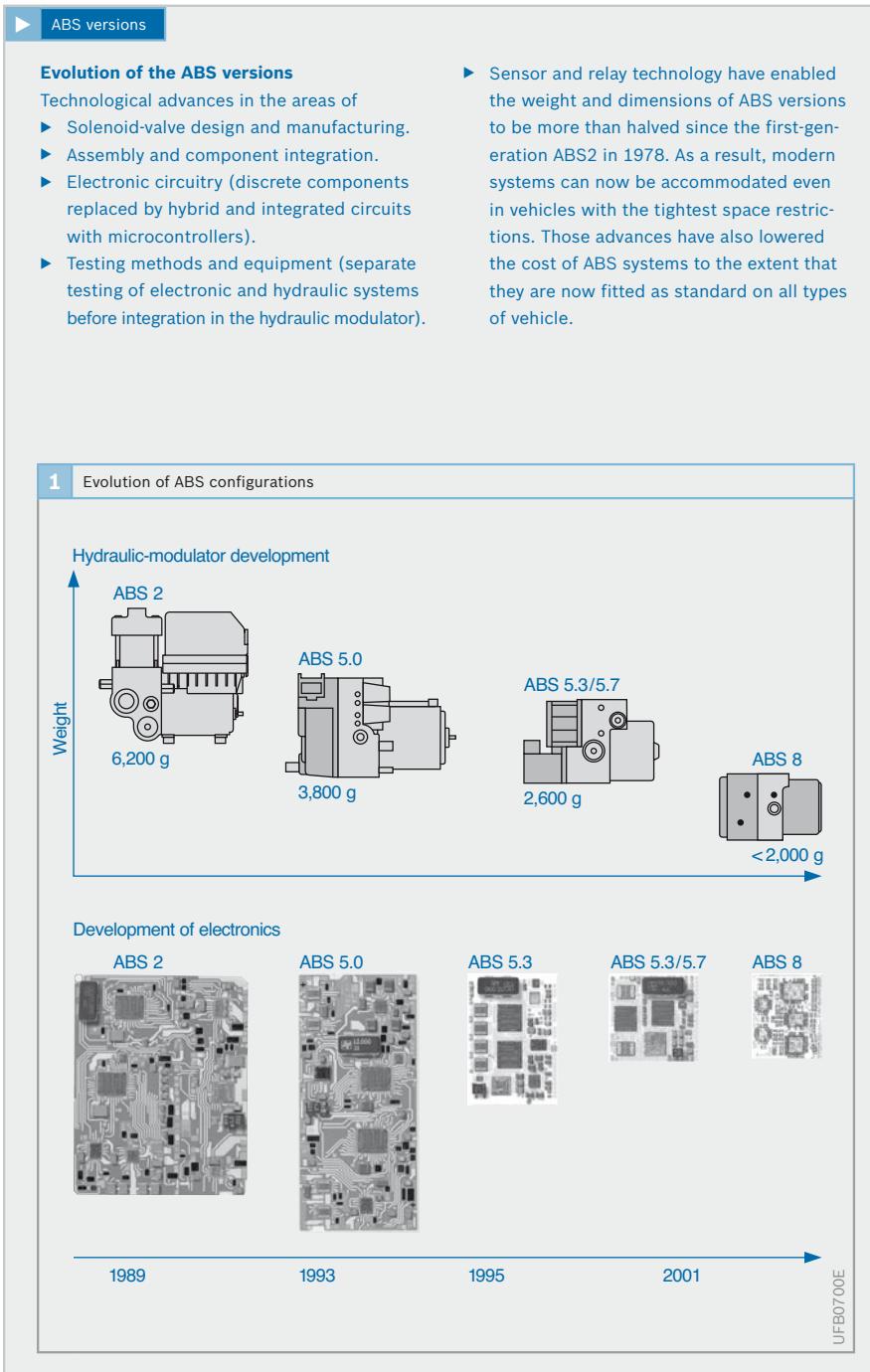
Fig. 5

- 1 Engine with transmission
- 2 Wheel
- 3 Wheel brake
- 4 Transversal differential
- 5 Electronic control unit with TCS functionality

Engine, transmission, gear ratio of differential and losses are combined in one unit.

UFB075







Technology to mimic the animals

Radar (RAdiation Detecting And Ranging) is a wireless method for locating objects that has traditionally been used mainly in air travel and shipping. Since radar-supported air defence was introduced in World War 2, radar has also been a feature of weapons technology. More recently, radar has found application in space travel, weather forecasting and ultimately in road traffic as part of measuring distances between vehicles using ACC (Adaptive Cruise Control).

The basis for the development of radar was the sonar system (Sound Navigation and Ranging) possessed by animals for navigation and determining distances. Echolocating bats, for example, emit orientation sounds in the form of shrill whistles in the 30 to 120 kHz *ultrasound range*. They then pick up the echo from obstacles or their prey with their ears. They use the information contained in the echo to decide what to do next.



UFS0038Y

Radar functions in a similar way but uses *radio signals* not sound waves. The distance measurement method of radar is based on the measurement of the time that elapses between the emission of the *electromagnetic waves* and the reception of a signal echo reflected by an object.

While air travel and shipping radar systems, for example, operate in the 500 MHz to 40 GHz frequency range, ACC is authorized to use the 76 to 77 GHz frequency band.

Stages in the development of radar

The development of electromagnetic, long-range search facilities was a great challenge for design engineers. Only a small part of the energy originally emitted was reflected back. It is therefore necessary to emit a large amount of energy and to concentrate it in as narrow a beam bundle as possible. The only devices suitable for this are very sensitive transmitters and receivers for waves that are shorter than the dimensions of the target.

The development path that led to radar technology is punctuated by the following historical milestones and characters:

1837 Morse: Message transmission over long distances by means of electrical pulses sent by the telegraph was the first method to find widespread use

1861/1876 Reis and Bell: Replacement of telegraphs by the telephone makes message transfer much more direct and user-friendly

1864 Maxwell, Hertz and Marconi: Existence of "radio waves" proven theoretically and experimentally. Radio waves reflect off metallic objects in the same way as light waves reflect off a mirror

1922 Marconi: The pioneer of the radio proposes further investigations into earlier research approaches to radio measuring technology

1925 Appleton and Barnett: The principle of radio measuring technology is used to detect conducting strata in the atmosphere

Breit and Tuve: Development of pulse modulation that permits exact development measurements

1935 Watson Watt: Invention of the radar

1938 Ponte: Invention of the magnetron (velocity-modulated tube for generating high-frequency vibrations)

Adaptive cruise control (ACC)

System description

Like the basic cruise-control system that has been available as a standard feature for many years, ACC (Adaptive Cruise Control) can be categorized as a driver-assistance system. Cruise control regulates driving speed to maintain the desired speed selected by the driver using the cruise-control unit. In addition to the basic cruise-control function, ACC measures the distance to the vehicle in front and its relative speed, and uses this information together with other collected data (position of other vehicles in the same or different lane; in future, even stationary objects) to regulate the time gap between the vehicles. ACC is thus able to adapt the vehicle's speed to match the speed of the vehicle traveling in front and maintain a safe distance from it. The driver is able to override or switch off the ACC function at any time (e.g. by depressing the gas or brake pedal).

Fig. 1

- 1 ACC sensor and control unit
- 2 Engine-management system ECU (ME or DI Motronic) for gasoline engines or electronic diesel control (EDC) for diesel engines
- 3 Active brake intervention via ESP
- 4 Control and display unit
- 5 Engine-control intervention by means of electrically adjustable throttle valve (ME or DI Motronic)
- 6 Sensors
- 7 Transmission-shift control by means of electronic transmission control (optional)

Distance sensor

In the main, ACC systems currently have a radar sensor (Fig. 1, Item 1) that operates in a frequency range of between 76 and 77 GHz and emits four radar lobes. Once

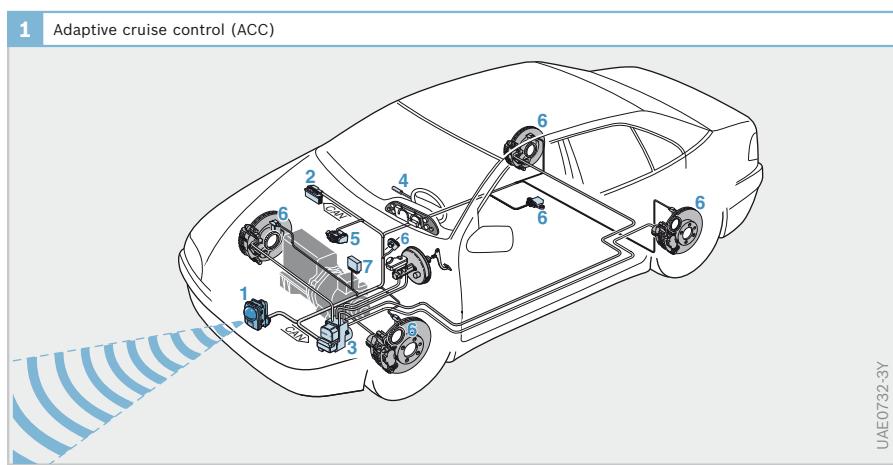
activated, ACC detects a range of up to approximately 200 m in front of the vehicle. The radar beams reflected by vehicles in front are analyzed for timing, Doppler shift and amplitude ratio. These factors are used to calculate distance, relative speed and angle position relative to vehicles in front.

Network architecture

The ACC function cannot be represented independently as a stand-alone system; various subsystems (engine-management system, electronic stability program, transmission control, instrument cluster) must be networked with each other. The evaluation and control electronics (control unit) of the ACC are integrated in the sensor housing. They receive and send data on a CAN data bus from and to other electronic control units.

Course setting

To ensure reliable ACC operation no matter what the situation - e.g. also on bends - it is essential that the preceding vehicles can be allocated to the correct lane(s). For this purpose, the information from the ESP sensor system (yaw rate, steering angle, wheel speeds and lateral acceleration) is evaluated with regard to the ACC-equipped vehicle's own curve status.



Setting options

The driver inputs the desired speed and the desired time gap; the time gap available to the driver usually ranges from 1 to 2 s. The time gap to the vehicle in front is calculated from the radar signals and compared with the desired time gap specified by the driver. If this value is shorter than the desired value, the ACC system responds in a manner appropriate to the traffic situation by initially reducing engine torque, and only if necessary by automatically braking the vehicle. If the desired time gap is exceeded, the vehicle accelerates until either the speed of the vehicle in front or the desired speed set by the driver is reached.

Engine-control intervention

Speed control requires an electronic engine-performance control system. The ME or DI Motronic engine-management system and the electronic diesel control (EDC) are integrated with this function. This system allows the vehicle to be accelerated to the desired speed or, if an obstacle appears, to be decelerated by means of automatic throttle closing.

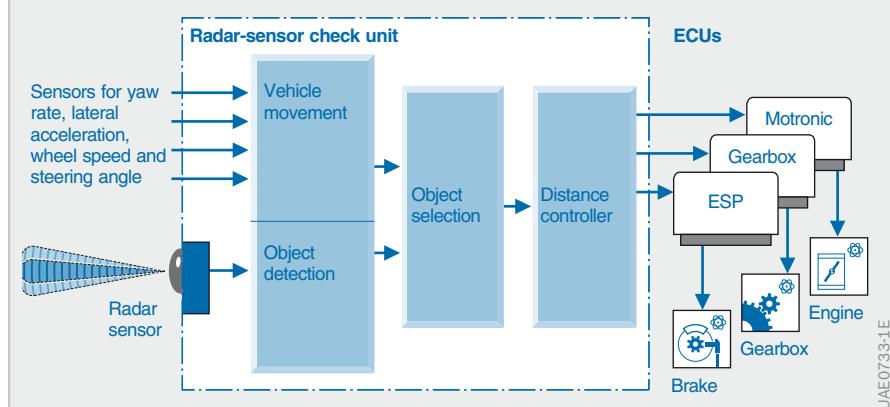
Brake intervention

If the rate of deceleration achieved by easing off the gas is not sufficient, the vehicle will have to be braked. The electronic stability program (ESP) is required here as it is able to initiate the brake intervention.

Due to the design of ACC as a comfort system, the braking deceleration calculated by the ACC controller is limited to approximately 2 to 3 m/s² with current ACC systems. If this is not sufficient for the current traffic situation (e.g. if the vehicle in front brakes sharply), the vehicle audibly requests the driver to take over responsibility for braking. The necessary braking deceleration is then to be achieved using the service brake. Safety functions, such as panic braking, are not part of the ACC system.

The other stabilizing systems of ABS, TCS or ESP may be active as normal during an active ACC control intervention as necessary. Depending on the parameter settings of ACC, stabilization interventions may result in ACC being deactivated.

2 Basic structure and components of the ACC control system



Control and display

Controls include switches, push-buttons or thumbwheels for

- ▶ Activating the function and
- ▶ Setting the desired speed and
- ▶ Desired time gap

The following information may be displayed to the driver in the instrument cluster:

- ▶ Desired speed
- ▶ Information on the activation status
- ▶ The time gap selected by the driver
- ▶ Indication of the follow-up mode, which informs the driver as to whether the system is controlling the distance to a detected target object or not

System limits

ACC does not yet permit control operations in city environments. This system can only be activated at speeds in excess of 30 km/h.

Control algorithms

The control system basically consists of three control modules:

▶ *Control module 1: cruise control*

If the radar sensor has not detected any vehicles in front, the system maintains the desired speed set by the driver.

▶ *Control module 2: follow-up control*

The radar sensor has detected vehicles in front. Control essentially maintains the time gap to the nearest vehicle at a constant setting.

▶ *Control module 3: control when cornering*

When negotiating tight bends, the radar sensor can “lose sight” of the vehicle in front because of the limited width of its “field of vision”. Until the vehicle comes in sight of the radar again, or until the system is switched to normal cruise control, special measures come into effect. Depending on the manufacturer, the speed would then be maintained, the current rate of lateral acceleration adapted or the ACC function deactivated, for example.

Object detection and lane allocation

The central task of the radar sensor and its integrated electronics is to detect objects and allocate them either to the same lane as the one on which the vehicle is traveling, or to a different lane. Firstly, lane allocation demands the precise detection of vehicles in front (high angle resolution and accuracy), and secondly, a precise knowledge of the motion of the system's own vehicle. Vehicle motion is calculated from the signals sent by sensors also used for the electronic stability program (ESP) (course prediction). These include the wheel-speed sensors and driving-dynamics sensors for the yaw rate and lateral acceleration. Optionally, information supplied by a steering-angle sensor may also be processed. The decision as to which of the detected objects is used as the reference for adaptive cruise control is essentially based on a comparison between the positions and motion of the detected objects and the motion of the system's own vehicle.

Adjustment

The radar sensor is fitted at the front end of the vehicle. Its radar lobes are aligned relative to the vehicle longitudinal axis. This is done using adjusting screws at the fastening part of the sensor. If it is moved out of alignment by physical force, e.g. deformation of the mounting due to accident damage or any other effect, realignment must be carried out. Small degrees of misalignment are automatically corrected by the permanently active alignment routines implemented in the software. If manual realignment is required, this is indicated to the driver.

Ranging radar

The radar (RAdiation Detecting And Ranging) transceiver unit transmits packets of electromagnetic waves using an antenna. These reflect off an object made of electrically conductive materials (e.g. vehicle body) and are then received. The signals received in this manner are “compared” with the transmitted signals with respect to their propagation time and/or frequency.

Measuring principles

Propagation time measurement

For all radar methods, the distance measurement is based on the direct or indirect propagation time measurement for the time between when the radar signal is transmitted and when the signal echo is received. The direct propagation time measurement is used to measure period τ . With direct reflection, this is equal to twice the distance d to the reflector divided by the speed of light c :

$$\tau = 2d/c$$

For a distance of $d = 150$ m and $c \approx 300,000$ km/s, the propagation time is
 $\tau \approx 1 \mu\text{s}$.

Frequency modulation

Direct propagation time measurement requires much effort; an indirect propagation time measurement is simpler. The method is known as FMCW (Frequency Modulated Continuous Wave). Rather than comparing the times between the transmitted signal and received echo, the FMCW radar compares the frequencies of the transmitted signal and received echo. The prerequisite for a meaningful measurement is a modulated transmit frequency.

With the FMCW method, radar waves linearly modulated in their frequency are transmitted for a duration of typically a few milliseconds and in a cycle of a few hundred MHz (f_s , continuous curve in Fig. 3). The signal reflected off a vehicle

in front is delayed in accordance with the signal propagation time (f_e , dashed line in Fig. 3). In the rising ramp, the frequency is lower; in the falling ramp, the frequency is higher by the same amount. The difference in frequency Δf is a direct measure for the distance.

If there is additionally a relative speed between the vehicles, the receive frequency f_e is increased (f_e' , dotted line in Fig. 3) by a specific amount Δf_d in both the rising and falling ramp due to the Doppler effect. This produces two different frequency differences Δf_1 and Δf_2 . Their addition produces the distance between the vehicles, and their subtraction the relative speed of the vehicles.

The signal processing in the frequency range therefore delivers a frequency for each object that as a linear combination produces a term for distance and relative speed. From the measured frequencies of two ramps with a different gradient, it is possible to determine the distance to and the relative speed of an object. For situations involving several targets, several ramps with a different gradient are required.

3 FMCW method

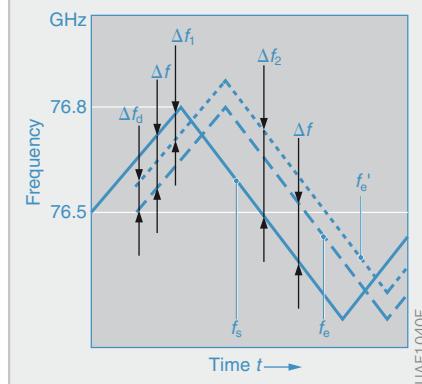


Fig. 3
 f_s Transmission signal
 f_e Return signal at same road speed
 f_e' Return signal with relative road speed

Doppler effect

Although the relative speed of the other vehicle can be measured using a number of subsequent distance measurements, it is calculated more quickly, more reliably and more accurately when the Doppler effect is utilized in the measurement.

For an object moving relative to the radar sensor with a relative speed (relative speed v_{rel}), the signal echo undergoes a frequency shift f_D compared to the emitted signal. At the relevant differential speeds, this is represented as:

$$f_D = -2f_c \cdot v_{\text{rel}}/c$$

f_c is the carrier frequency of the signal.

At the radar frequencies commonly used for ACC, $f_c = 76.5$ GHz, there is a frequency shift of $f_D \approx -510 \cdot v_{\text{rel}}/\text{m}$, and thus 510 Hz at a relative speed of -1 m/s (approaching).

Angle measurement

The third basic dimension which is needed is the side offset (angle) of the preceding vehicle. The only way this can be measured is by radiating the radar beam in a number of different directions. The (reflected) signals are then applied to determine from which direction the strongest reflection came. This method needs either high-speed back-and-forth movement of

the beam (scanning), or the installation of a multi-beam antenna array.

High-frequency part of the ACC sensor

The high-frequency part can be broken down into four functional groups.

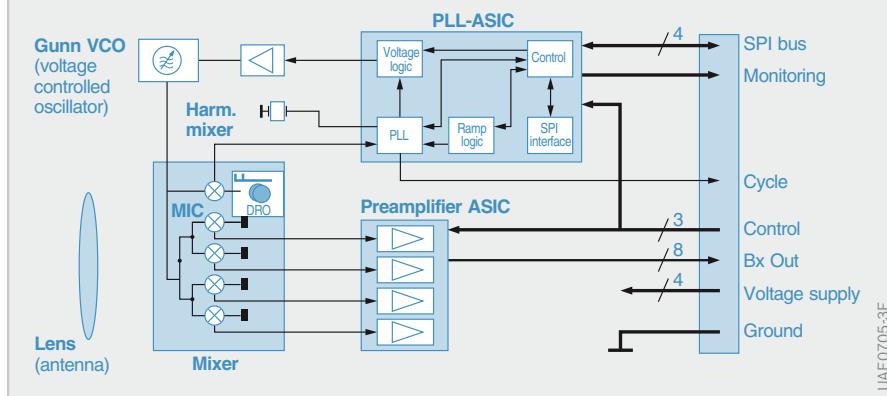
HF generation

The HF generation and control section makes the high frequency available for transmission (Fig. 4). The HF output of between 76 and 77 GHz is generated using a voltage-controlled oscillator (VCO) comprising a Gunn diode in a mechanical resonator. A small part of the output generated is downmixed into an intermediate frequency band using a dielectrical resonance oscillator (DRO) with harmonic mixer and supplied to the control electronics (PLL-ASIC, PLL = Phase Locked Loop). The latter controls the VCO by means of an output driver and provides frequency stabilization and modulation.

Transmission and reception circuit

In the transmission and reception circuit, the HF output is divided between four transmission/reception channels by three Wilkinson splitters. Using bandpass mixers, this output is supplied to the antenna while the return signal is downmixed to the basic band.

4 Block diagram of a 4-channel FMCW radar



Amplification

Signals in the basic band are amplified in an ASIC. It has four channels, switching amplification, and a special characteristic curve. To some extent, this compensates for the large signal dynamics in that high frequencies (correspondingly high distances) are more strongly amplified. In addition, the characteristic is integrated with a low-pass anti-aliasing filter for subsequent scanning.

Antenna system

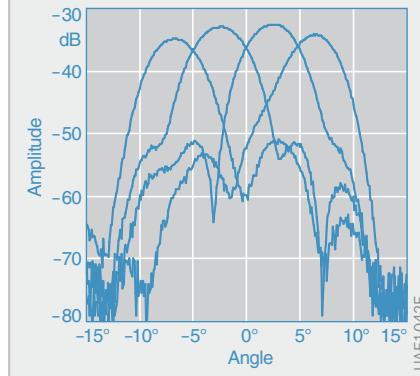
The antenna system is a monostatic system. It comprises four combined transmitter and receiver patches on the HF substrate, four polyrods (plastic cones) for prefocusing and a plastic lens for beam concentration. As part of the housing, the lens also acts as a radar-optical window and shield. The radar waves are emitted simultaneously and coherently by the four antenna patches to produce a single transmission wave. The beams are only actually split into four separate beams on the receiver side. Four separate receiver channels are used here.

The transmit frequency is modulated by the voltage-controlled oscillator (VCO) in a linearly ramped fashion (Fig. 3) with a gradient of $m = df/dt$. While the received signal returns after the propagation time $\tau = 2d/c$ the transmit frequency has changed in the meantime by the differential frequency $f_D = \tau \cdot m$. Therefore, the propagation time, and thus the distance, can be measured indirectly by ascertaining the differential frequency between the transmitted and received signals. The differential frequency, in turn, can be ascertained using a mixer, followed by low-pass filtering. To determine the frequency, the signal is digitized and converted into a frequency spectrum using an FFT (Fast Fourier Transformation).

However, the information about the difference frequency does not only contain information for the propagation time but also for the Doppler shift. This situation means that there is at first a certain ambiguity in the evaluation. It can be resolved by applying multiple FMCW modulation cycles with different gradients.

To determine the angle at which the radar locates an object, multiple radar lobes are transmitted and evaluated. At least two overlapping radar beams are required to measure the angle. No conclusions on the angle of sight can be drawn from the relationships between the amplitudes that are measured for an object in adjacent beams. If, for example, four radar beams are used, the horizontal angular dependence of which is shown in Figure 5 in the form of a two-way antenna diagram as an example, it is possible to determine the horizontal angle of sight by comparing amplitude and phase of the measured radar signals using the antenna diagram.

5 Two-way antenna diagram



Radar signal processing

The low-frequency part of the FMCW radar comprises several components (Fig. 6).

A dual-core processor is used for digital data processing. The digital signal processor (DSP) contained in this module is used for data acquisition, calculating the fast fourier transform (FFT) and for other basic signal processing. The processor also contains a microcontroller in which additional signal processing, the application software and control unit functions are executed. Furthermore, various peripherals are integrated in the dual processor: serial interfaces, two CAN controllers (Controller Area Network), an analog/digital converter and various digital ports.

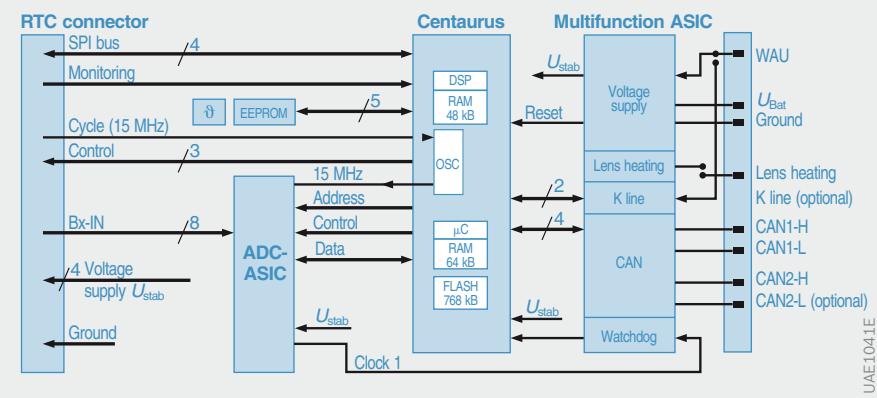
The processor is surrounded by various peripheral modules. The analog radar signals from the HF PCB are converted into digital sample values in an analog/digital converter (ADC). This takes place simultaneously for four channels. A digital low-pass filter is also integrated in this module to ensure limitation to the Nyquist bandwidth. An EEPROM is used as an external, non-volatile memory. Application parameters and, if applicable, fault codes are stored here. A multifunction ASIC is used to generate the supply voltages (different DC voltages) and as an output driver (K line,

CAN, lens heating to prevent icing). In addition, a watchdog is also integrated. Using a temperature sensor, it is possible to measure the internal temperature of the system.

The unit is connected to the vehicle by an eight-pin connector. The connector supplies battery voltage (approximately 12 V), ground (GND), two CAN buses, or alternatively a wakeup or K line, or a radome heater, or a time gap signal.

The low-frequency circuit can be designed using standard printed-circuit board technology. Figure 7 provides a look inside the unit.

6 Low-frequency block diagram



7 Exploded view of an FMCW radar with integrated signal processing

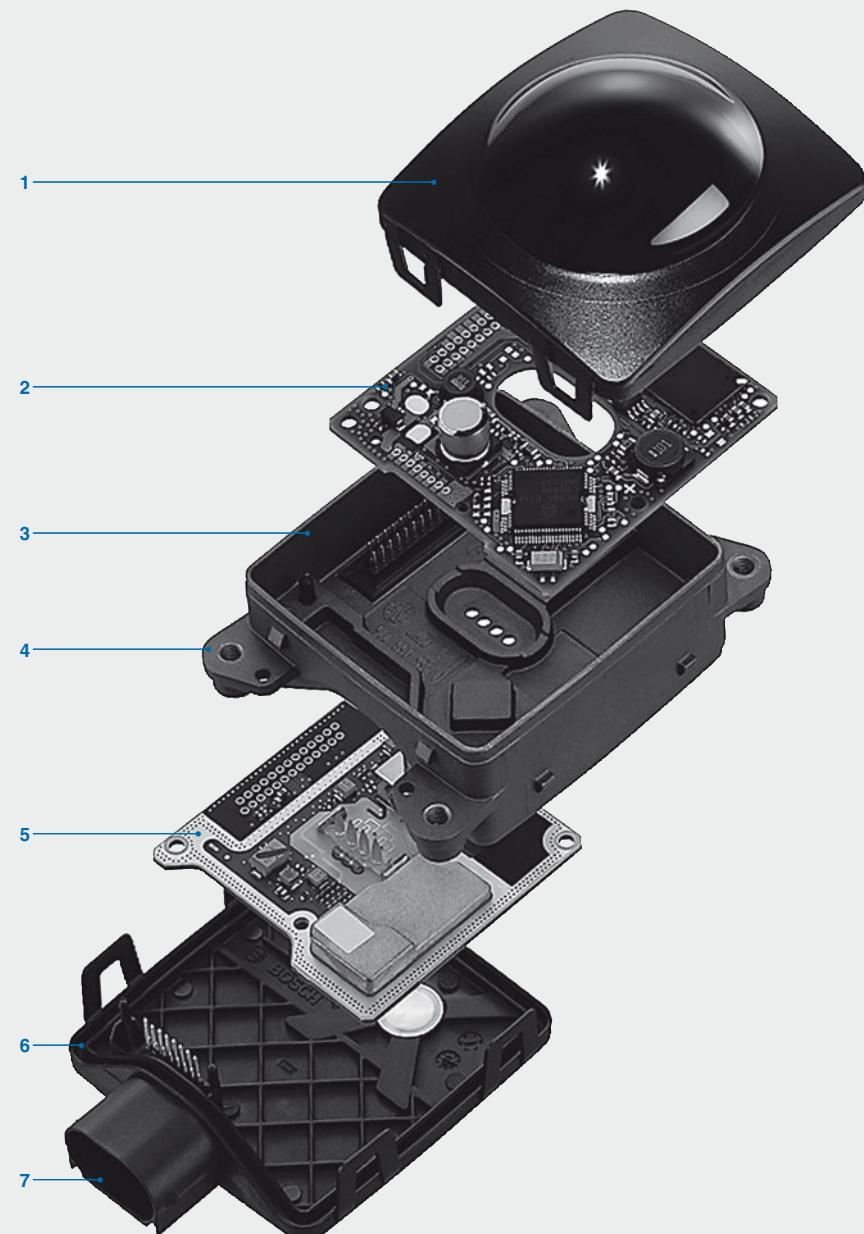


Fig. 7

- 1 Upper housing section with lens
- 2 HF PCB
- 3 Bearing points for alignment
- 4 Intermediate carrier
- 5 LF PCB
- 6 Housing base
- 7 Plug

Occupant-protection systems

In the event of an accident, occupant-protection systems are intended to keep the accelerations and forces that act on the passengers low and lessen the consequences of the accident. These passive vehicle safety systems include:

- ▶ Seat belts with seat-belt pretensioners
- ▶ Airbags and
- ▶ Rollover protection systems (on cabriolets)

Seat belts and seat-belt pretensioners provide the greater part of the protective effect since they absorb 50 to 60 % of impact energy. With front airbags, the energy absorption is about 70 % if deployment timing is properly synchronized.

In order to achieve optimum protection, the response of all components of the complete occupant-protection system must be adapted to one another.

Seat belts and seat-belt pretensioners

Function

The function of seat belts is to restrain the occupants of a vehicle in their seats when the vehicle impacts against an obstacle. In the event of a frontal impact, seat-belt pretensioners pull the seat belts tighter against the body and hold the upper body as closely as possible against the seat backrest. This prevents unimpeded forward displacement of the occupants caused by inertia. Seat-belt pretensioners thus improve the restraining characteristics of a three-point inertia-reel belt and increase protection against injury.

Operating principle

In an impact, the shoulder-belt tightener eliminates the seat belt slack and the "film-reel effect" by rolling up and tightening the belt webbing. On activation, the system electrically fires a pyrotechnic propellant charge. The rising pressure acts on a piston,

1 Electronic impact protection system

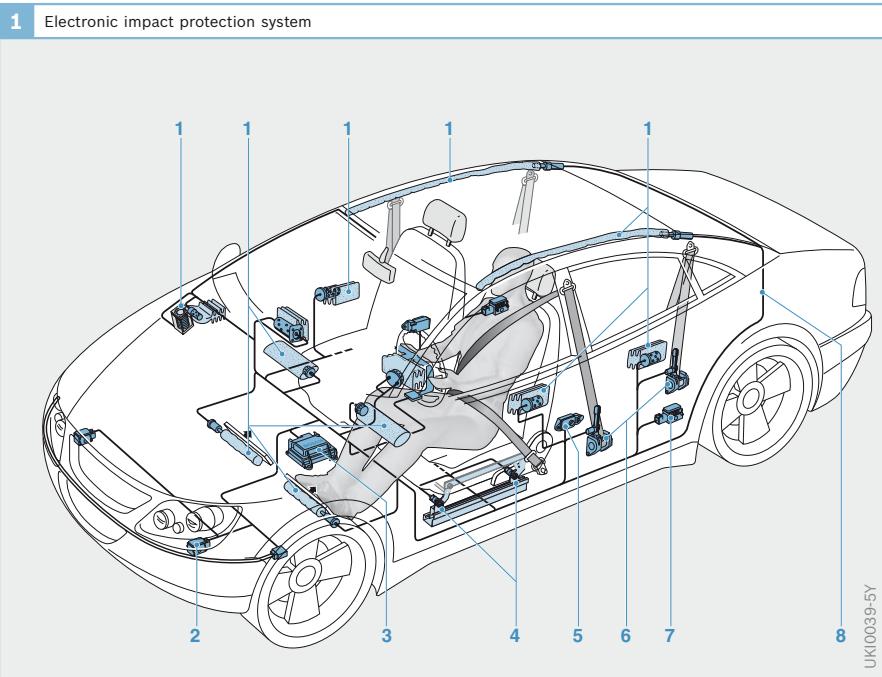


Fig. 1

- 1 Airbag with gas inflator
- 2 Upfront sensor
- 3 Central electronic control unit with integrated rollover sensor
- 4 iBolt™
- 5 Peripheral pressure sensor (PPS)
- 6 Seat-belt pretensioner with propellant charge
- 7 Peripheral acceleration sensor (PAS)
- 8 Bus architecture (CAN)

which turns the belt reel via a steel cable in such a way that the belt is held tightly against the body (Fig. 2).

Variants

In addition to the shoulder-belt tighteners, there are variants which pull the seat-belt buckle back (buckle tighteners) and simultaneously tighten the shoulder and lap belts. Buckle tighteners further improve the restraining effect and the protection to prevent occupants from sliding forward under the lap belt ("submarining effect").

A further improvement is achieved by the use of belt-force limiters. In this case, the seat belt tighteners initially tighten fully (using the maximum force of approx. 4 kN, for example) and restrain the occupants. If a certain belt tension is exceeded, the belt gives thereby extending the degree of forward movement. The kinetic energy is converted into deformation energy, which prevents the occurrence of acceler-

ation peaks. Deformation elements include a torsion rod in the inertia reel shaft. However, there is also an electronically controlled single-stage belt-force limiter, which reduces the belt tension to 1 to 2 kN by firing a detonator a specific period after deployment of the second front airbag stage and after a specific extent of forward movement is reached.

Front airbag

Function

The function of front airbags is to protect the driver and front passenger against head and chest injuries in a vehicle impact with an obstacle. In a serious accident, a seat-belt pretensioner cannot keep the head from striking the steering wheel.

Operating principle

To protect driver and front passenger, pyrotechnical gas inflators inflate the driver and passenger airbags highly dynamically after a vehicle impact detected by sensors. In order to provide maximum protection, the airbag must be fully inflated before the occupant plunges into it. Once the occupant makes contact with it, the airbag partly deflates in order to "gently" absorb impact energy acting on the occupant with noncritical (in terms of injury) surface pressures and declaration forces.

The maximum permissible forward displacement of the driver before the airbag on the driver's side has inflated is approximately 12.5 cm. This equates to a time of approximately 40 ms from the start of impact (in the case of an impact with a hard obstacle at 50 km/h). 10 ms elapse before the electronics detect the impact and trigger the electronic ignition system. A further 30 ms is required for the airbag to inflate. The airbag is deflated through the outlet openings after another 80 to 100 ms. The entire process takes little more than a tenth of a second.

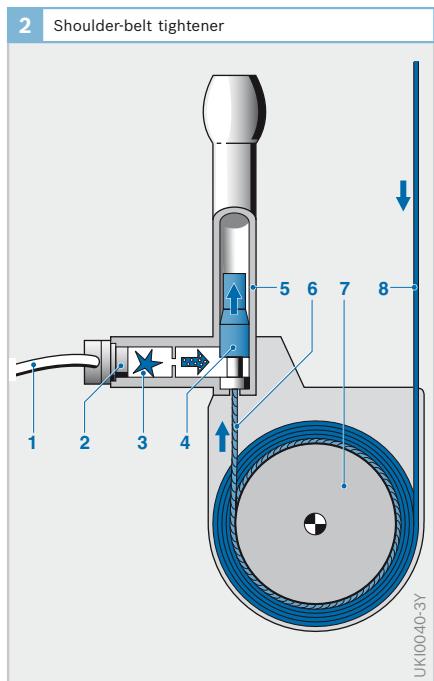


Fig. 2

- 1 Firing cable
- 2 Firing element
- 3 Propellant charge
- 4 Piston
- 5 Cylinder
- 6 Wire rope
- 7 Belt reel
- 8 Belt webbing

Impact detection

The deceleration arising from the impact is detected by one (or two) longitudinal acceleration sensor(s) and the change in speed is calculated from it. In order to be able to better detect oblique and offset impacts, the deployment algorithm can also evaluate the signal from the lateral acceleration sensor.

The impact must also be analyzed in addition to the crash sensing. The airbag should not trigger from a hammer blow in the workshop, gentle impacts, bottoming out, driving over curbstones or potholes. With this goal in mind, the sensor signals are processed in digital analysis algorithms whose sensitivity parameters have been optimized with the aid of crash data simulations. Depending on the vehicle manufacturer's production concept, the deployment parameters and the vehicle's equipment level can also be programmed into the ECU at the end of the assembly line ("end-of-line programming").

In order to prevent airbag-related injuries to "out-of-position" occupants (e.g. leaning too far forward) or to small children in reboard (rearward-facing) child seats, it is essential that the front airbags be triggered and inflated in accordance with the particular situations. The following measures are implemented for this purpose:

Deactivation switch

This can be used to disable the passenger airbag. The status of the airbag function is indicated by lamps.

Intelligent airbag systems

The introduction of improved sensing functions and control options for the airbag inflation process, with the accompanying improvement in protective effect, is intended to result in a gradual reduction in the risk of injury. Such functional improvements are:

- ▶ Impact severity detection through further optimization of the deployment algorithm or the use of one or two upfront sensors (Fig. 4). The latter are acceleration sensors installed in the vehicle's crumple zone (e.g. on the radiator cross-member) which facilitate early detection of and distinction between different types of impact, such as ODB (Offset Deformable Barrier) crashes, pole or underride impacts. They also allow an assessment of the impact energy.
- ▶ Seatbelt usage detection.
- ▶ Occupant presence, position and weight detection.
- ▶ Seat position and backrest inclination detection.
- ▶ Use of front airbags with two-stage gas inflators or with single-stage gas inflators and pyrotechnically triggered gas-discharge valves.
- ▶ Use of seat-belt pretensioners with occupant-weight-dependent belt-force limiters.
- ▶ Through data exchange with other systems, e.g. ESP (Electronic Stability Program), and environment sensors, it is possible to use information from the phase shortly before the impact to further optimize the deployment of the restraints.

Side airbag

Function

Side airbags, which inflate along the length of the roof lining for head protection (inflatable tubular systems, window bags, inflatable curtains) or from the door or seat backrest (thorax bags, upper body protection) are designed to cushion the occupants and protect them from injury in the event of a side impact.

Operating principle

Due to the lack of a crumple zone, and the minimum distance between the occupants and the vehicle's side structural components, it is particularly difficult for side airbags to inflate in time. For this reason, the time required for crash sensing and activating of the side airbags is approximately 5 to 10 ms for hard side impacts. The inflation time of the approximately 12 l capacity thorax bag is not permitted to be more than 10 ms.

These requirements can be fulfilled through evaluation of peripheral (at suitable points on the body, e.g. b-pillar or door), lateral (sideways) acceleration and pressure sensors.

Rollover protection systems

Function

In the event of an accident where the vehicle rolls over, open-top vehicles such as convertibles lack the protecting and supporting roof structure of closed-top vehicles. Extendable rollover bars or the extendable head restraints provide protection against injury for occupants.

Operating principle

Current sensing concepts no longer trigger the system at a fixed threshold but rather at a threshold that conforms to a situation and only for the most common rollover situation, i.e. about the longitudinal axis. The Bosch sensing concept involves a surface micromechanical yaw-rate sensor and high-resolution acceleration sensors in the vehicle's transverse and vertical axes (y and z axes).

The yaw-rate sensor is the main sensor, while the y and z-axis acceleration sensors are used both to check plausibility and to detect the type of rollover (slope, gradient, curb impact or "soil-trip" rollover). On Bosch systems, these sensors are incorporated in the airbag triggering unit.

Deployment of occupant-protection systems is adapted to the situation according to the type of rollover, the yaw rate and the lateral acceleration, i.e. systems are triggered after between 30 and 3,000 ms by automatic selection and use of the algorithm module appropriate to the type of rollover.

Combined ECUs for seat-belt pretensioners, front and side airbags and rollover protection equipment

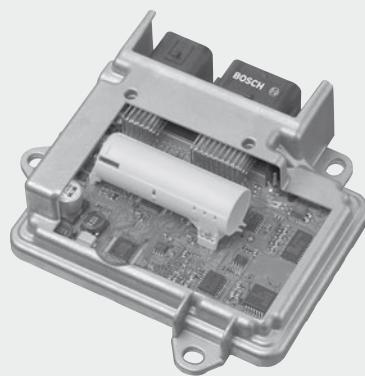
Optimum occupant protection against the effects of frontal, offset, oblique or pole impact is obtained through the precisely coordinated interaction of electronically detonated pyrotechnical front airbags and seat-belt pretensioners. To maximize the effect of both protective devices, they are activated with optimized time response by a common ECU installed in the passenger cell.

The following functions are currently incorporated in the central ECU, also referred to as the trigger unit:

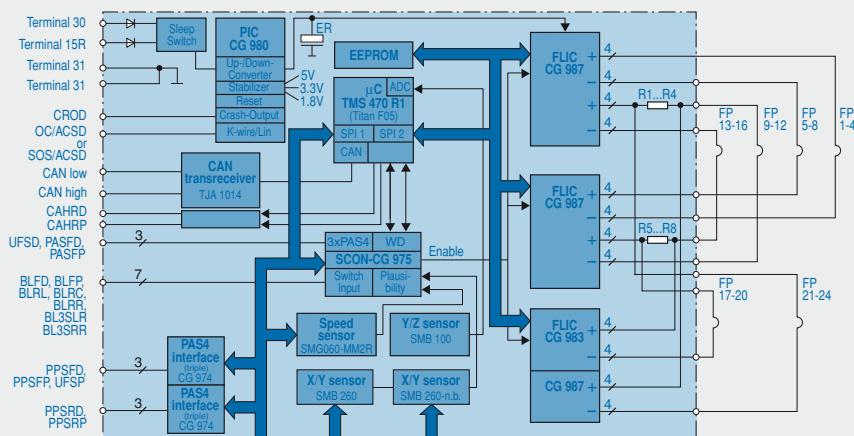
- ▶ Crash sensing by acceleration sensor and safety switch or by two acceleration sensors without safety switch (redundant, fully electronic sensing).
- ▶ Rollover detection by yaw rate and acceleration sensors that record y and z axis acceleration in the low-g range (up to approximately 5 g).
- ▶ Prompt activation of front airbags and seat-belt pretensioners in response to different types of impact in the vehicle longitudinal direction (e.g. frontal, oblique, offset, pole, rear-end).
- ▶ Control of rollover protection equipment.
- ▶ For the side airbags, the ECU operates in conjunction with a central lateral sensor and two or four peripheral acceleration sensors. The peripheral acceleration sensors (PAS) transmit the triggering command to the central ECU via a digital interface. The central ECU triggers the

4 Central combined airbag 9 ECU (block diagram)

Terminal designations:	
Terminal 30 Direct battery positive, not fed through ignition lock	BL3SRL Belt Lock (switch) 3rd Seat Row Left
Terminal 15R Switched battery positive when ignition lock in "radio", "ignition on" or "starter" position	BL3SRR Belt Lock (switch) 3rd Seat Row Right
Terminal 31 Body ground (at one of the device mounting points)	PPSFD Peripheral Pressure Sensor Front Driver
	PPSFP Peripheral Pressure Sensor Front Passenger
	UFSP UpFront Sensor Passenger
	PPSRD Peripheral Pressure Sensor Rear Driver
	PPSRP Peripheral Pressure Sensor Rear Passenger
Abbreviations:	ZP Firing pellets 1...4 or 21...24
CROD CRash Output Digital	FLIC Firing Loop Integrated Circuit
OC/ACSD Occupant Classification/ Automatic Child Seat Detection	PIC Periphery Integrated Circuit
SOS/ACSD Seat-Occupancy Sensing/ Automatic Child Seat Detection	SCON Safety CONtroller
CAN low Controller Area Network, low level	μ C Microcontroller
CAN high Controller Area Network, high level	
CAHRD Crash Active Head Restraint Driver	
CAHPR Crash Active Head Restraint Passenger	
UFSD UpFront Sensor Driver	
PASFD Peripheral Acceleration Sensor Front Driver	
PASFP Peripheral Acceleration Sensor Front Passenger	
BLFD Belt Lock (switch) Front Driver	
BLFP Belt Lock (switch) Front Passenger	
BLRL Belt Lock (switch) Rear Left	
BLRC Belt Lock (switch) Rear Center	
BLRR Belt Lock (switch) Rear Right	



UK100350Y



UK100364E

side airbags provided the internal lateral sensor has confirmed a side impact by means of a plausibility check. Since the central plausibility confirmation arrives too late in the case of impacts into the door or above the sill, peripheral pressure sensors (PPS) inside the door cavity are used to measure the adiabatic pressure changes caused by deformation of the door. This will result in rapid detection of door impacts. Confirmation of "plausibility" is now provided by PAS mounted on supporting peripheral structural components. This is now unquestionably faster than the central lateral-acceleration sensors.

- ▶ Voltage transformer and energy accumulator in case the supply of power from the vehicle battery is interrupted.
- ▶ Selective triggering of the seat-belt pretensioners, depending on monitored seat-belt buckle status: firing of the airbag only takes place if the belt is engaged in the belt buckle. At present, proximity-type seat-belt buckle switches are used, i.e. Hall IC switches which detect the change in the magnetic field when the buckle is fastened.
- ▶ Setting of multiple trigger thresholds for two-stage seat-belt pretensioners and two-stage front airbags depending on the status of belt use and seat occupation.
- ▶ Reading of signals from the interior sensors and appropriate triggering of restraints.
- ▶ Watchdog (WD): airbag triggering units must meet high safety standards with regard to false activation in non-crash situations and correct activation when needed (crashes). For this reason, ninth-generation (AB 9) airbags launched in 2003 incorporate three independent hardware watchdogs (WDs): WD1 monitors the 2 MHz system eClock using a dedicated, independent oscillator. WD2 monitors the realtime processes

(time base 500 µs) for correct and complete sequence. For this reason, the safety controller (SCON, see Fig. 4) sends the microcomputer eight digital messages to which it must respond by sending eight correct replies to the SCON within a time window of 1 ± 0.3 ms. WD3 monitors the background processes such as the built-in self-test routines of the ARM core for correct operation. The microcomputer's response to the SCON in this case must be provided within a period of 100 ms.

- ▶ With AB 9, sensors, analyzer modules and driver stages are linked by two SPIs (Serial Peripheral Interfaces). The sensors have digital outputs and their signals can be transmitted directly via SPIs. Shunts therefore remain on the printed-circuit board without effect, unlike with analog sinusoidal transfers, and a high level of functional reliability is achieved. Deployment is only permitted if an independent hardware plausibility channel has also detected the impact and enables the driver stages for a limited period.
- ▶ Diagnosis of internal and external functions and of system components.
- ▶ Storage of failure modes and durations with crash recorder; readout via the diagnosis or CAN-bus interface.
- ▶ Warning-lamp activation.

Acceleration sensors

Acceleration sensors for crash sensing can be located at the following points in the vehicle:

- ▶ Directly integrated in the ECU (seat-belt pretensioners, front airbag)
- ▶ At selected points on the right and left-hand side of the vehicle on supporting structural parts such as seat crossmembers, door sills, b and c-pillar (side airbag) or
- ▶ In the deformation zone at the front end of the vehicle (upfront sensors for "intelligent airbag systems")

These sensors are surface micromechanical sensors consisting of fixed and moving finger structures and spring pins. Since the sensors only have low working capacitances (approximately 1 pF), it is necessary to accommodate the evaluation electronics in the same housing in the immediate proximity of the sensor element so as to avoid stray capacitance and other forms of interference.

Gas inflators

The pyrotechnical propellant charges of the gas inflators for generating the airbag inflation gas and for actuating seat-belt pretensioners are activated by an electrical firing element. The gas inflator in question inflates the airbag with charge gas. The driver airbag built in the steering-wheel hub (volume approx. 60 l) or, as the case may be, the passenger airbag fitted in the glove compartment space (approx. 120 l) is inflated in approx. 30 ms from firing.

AC firing

In order to prevent inadvertent triggering through contact between the firing element and the vehicle system voltage (e.g. faulty insulation in the wiring harness),

the firing element is fired by alternating-current pulses with a frequency of approx. 80 kHz (AC firing). A small ignition capacitor with a capacitance of 470 nF incorporated in the firing circuit in the firing element plug galvanically isolates the firing element from the DC current. This isolation from the vehicle system voltage prevents inadvertent triggering, even after an accident when the airbag remains untriggered and the occupants have to be freed from the deformed passenger cell by emergency services. It may even be necessary to cut through the (permanent +) firing circuit wires in the steering column wiring harness and short-circuit them according to positive and ground.

Passenger-compartment sensing

For passenger classification, an absolute weight measuring method is available with the iBolt (intelligent bolt). These force-measuring iBolts (Fig. 1) secure the seat frame (seat link) to the sliding rail and replace the four mounting screws otherwise used. They measure the weight-dependent change in the gap between the bolt sleeve and the internal bolt with integral Hall-element IC connected to the sliding base.

Micromechanics

Micromechanics is defined as the application of semiconductor technology in the production of mechanical components from semiconductor materials (usually silicon). Not only silicon's semiconductor properties are used but also its mechanical characteristics. This enables sensor functions to be implemented in the smallest-possible space. The following techniques are used:

Bulk micromechanics

The silicon wafer material is processed using anisotropic (alkaline) etching and, where needed, an electrochemical etching stop. The material is etched away from the reverse side of the silicon layer (Fig. 1, Item 2) in those areas where it is not protected by the etching mask (1). This method can be used to create very small diaphragms (a) with typical thicknesses of between 5 and 50 µm, holes (b) and bars and ridges (c), e.g. for pressure or acceleration sensors.

Surface micromechanics

The substrate material here is a silicon wafer on whose surface very small mechanical structures are formed (Fig. 2). First of all, a "sacrificial layer" is applied and shaped (A) using semiconductor production processes (e.g. etching). An approx. 10 µm polysilicon layer is then deposited on top of this (b) and structured vertically using a mask and etching (c). In the final processing step, the "sacrificial" oxide layer underneath the polysilicon layer is removed by means of gaseous hydrogen fluoride (d). In this way, structures such as flexible electrodes (Fig. 3) for acceleration sensors can be created.

Wafer bonding

Anodic and seal glass bonding are methods used to join wafers together by the action of electricity and heat or heat and pressure in order, for instance, to hermetically seal a reference-vacuum chamber or to protect sensitive structures by placing a cap over them.

1 Structures produced by bulk micromechanics

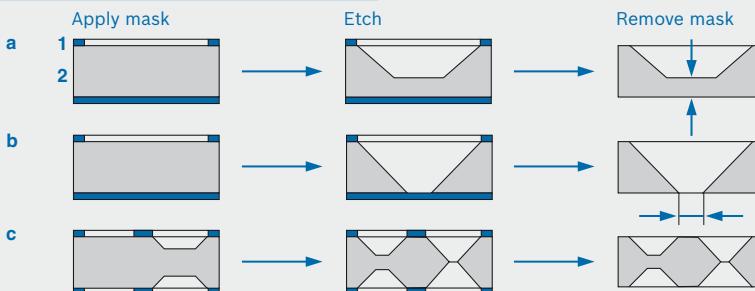
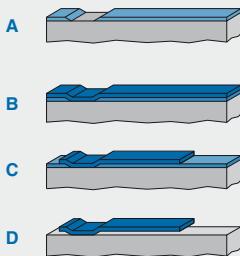


Fig. 1

- a Production of diaphragms
- b Production of holes
- c Production of bars and ridges

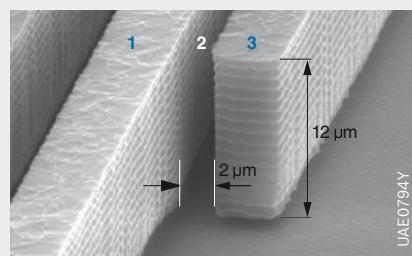
1 Etching mask
2 Silicon

2 Surface micromechanics (processing steps)



UAE0793Y

3 Surface micromechanics (structure details)



UAE0696-1Y
UAE0793Y

Fig. 2

- A Cutting and structuring the sacrificial layer
- B Cutting the polysilicon
- C Structuring the polysilicon
- D Removing the sacrificial layer

Fig. 3

- 1 Fixed electrode
- 2 Gap
- 3 Spring electrode

Basic principles of networking

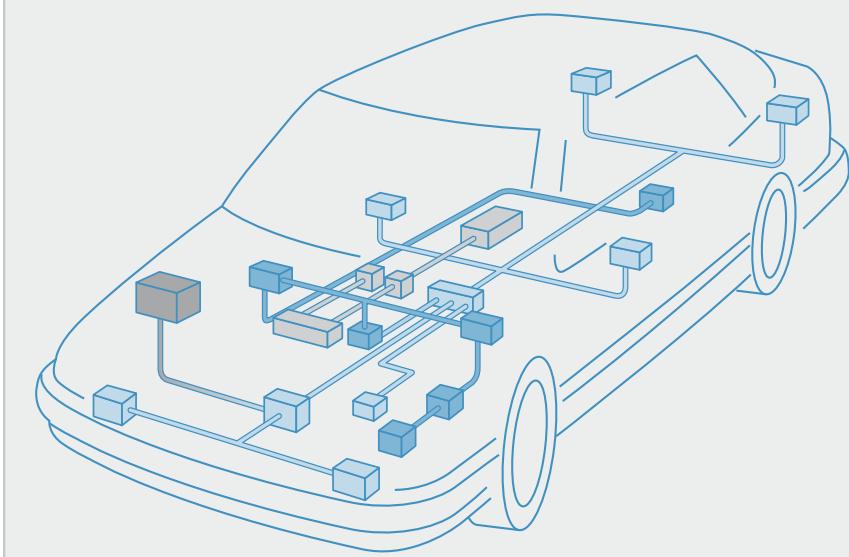
With the tremendous speed at which computer technology is advancing, the number of electronic systems in use is increasing more and more. This growth is also continuing in automotive engineering. However, this also means that the complexity of an overall system (the vehicle in this case) is on the increase. Individual systems such as engine management have been improved over the last few years. However, innovations are mainly achieved by means of interaction between several individual systems. The individual components need to be networked so that the multitude of information that is managed by the individual systems can also be used elsewhere throughout the system. Different communication systems are used depending on requirements (e.g. transmission reliability, fault tolerances, costs).

Network topology

A network is understood to be a system in which a group of elements can exchange information via a transportation medium. If the elements are visualized as nodes and the communication relationships as lines, a picture is created of a network where many nodes are related to several other nodes. The nodes in a communication network are also often referred to as network subscribers or stations.

In motor vehicles, complex control units such as those for the engine management system (Motronic or electronic diesel control, EDC), the electronic stability program (ESP), the transmission control system or the door modules can be network subscribers (Fig. 1). However, a sensor with a conditioning circuit that merely prepares and digitizes a measured value can also act as a network subscriber and make the measured signals available to other net-

1 Vehicle system networking



UVAG0141V

work subscribers. The transport medium via which the communication takes place is referred to as a bus or a data bus.

A network topology is understood to be the structure consisting of network nodes and connections. This merely shows which nodes are interconnected, but does not depict underlying details such as the length of the connection. Every network subscriber must have at least one connection to another network subscriber in order to participate in network communication. Different requirements are made of the network topology for a variety of communication network applications, while the topology determines some of the characteristics of the overall network. All network topologies are based on the following four basic topologies

- ▶ Bus topology
- ▶ Star topology
- ▶ Ring topology and
- ▶ Mesh topology

Other structures (hybrid topologies) can be created by combining these basic topologies.

Bus topology

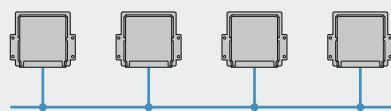
This network topology is also referred to as a linear bus. The core element of a bus topology is a single cable to which all nodes are connected via short connecting cables (Fig. 2). This topology makes it extremely easy to add other subscribers to the network. Information is transmitted by the individual bus subscribers in the form of so-called messages and distributed over the entire bus.

Nodes transmit and receive messages. If a node fails, the data that is expected from this node is no longer available to the other nodes on the network. However, the remaining nodes can continue to exchange information. However, a network with a bus topology fails completely if the main line is defective (due to a cable break, for example).

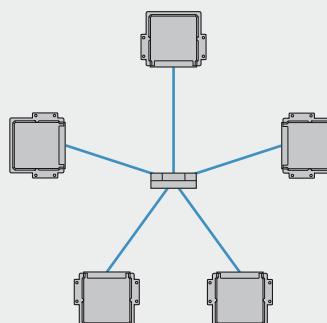
Star topology

The star topology consists of a main node (repeater, hub) to which all other nodes are coupled via a single connection (Fig. 3). A network with this topology is therefore easy to extend if free capacity is available (connections, cables).

2 Linear bus topology



3 Star topology



SVA0002Y

SVA0003Y

In star topologies, data is exchanged between the individual node connections and the main node, whereby a distinction is made between active and passive star topologies. In active star topologies, the main node contains a computer that processes and relays information. The performance capability of a network is essentially determined by the performance capability of this computer. However, the main node does not have to have special control intelligence. In passive star systems, it merely connects the bus lines of the network subscribers together.

The following applies to active and passive stars: if a network subscriber fails or a connecting line to the main node is defective, the rest of the network continues to operate. However, if the main node fails the entire network is disabled.

In the automotive area, star structures are under discussion for safety and security systems such as brakes and steering. In this case, the risk of a complete network failure is prevented by designing the main node to be physically redundant. This means that several main nodes are used to which the nodes whose information is needed for safe operation of the vehicle can be connected in parallel.

Ring topology

In the ring topology, each node is connected to its two neighbors. This creates a closed ring (Fig. 4). A distinction can be made between single rings and double rings.

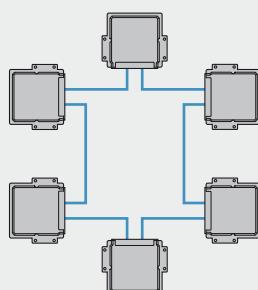
In a single ring, data transfers are unidirectional from one station to the next. The data is checked when it is received. If the data is not intended for this station it is repeated (repeater function), boosted and relayed to the next station. The data that is being transferred is therefore relayed from one station to the next in the ring until it has reached its destination or arrives back at its point of origin, where it will then be discarded.

As soon as a station in a single ring fails, the data transfer is interrupted and the network breaks down completely.

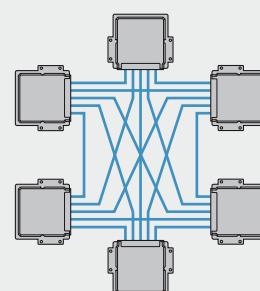
Rings can also be set up in the form of a double ring (e.g. FTTI), in which the transfer of data is bidirectional. In this topology, the failure of a station or a connection between two stations can be overcome, since all data is still transferred to all operational stations in the ring.

However, if several stations or connections fail, the possibility of a malfunction cannot be ruled out.

4 Ring topology



5 Mesh topology



Mesh topology

In a mesh topology, each node is connected to one or more other nodes (Fig. 5). In a fully meshed network, each node is connected to every other node.

If a mode or connection fails it is possible for the data to be rerouted. This type of network therefore has a high degree of system stability. However, the cost of networking and transporting the message is high.

Radio networks form a type of mesh topology, since the transmissions from each station are received by every other station that is within range.

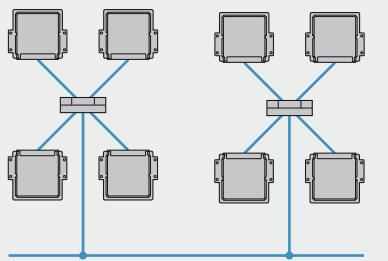
A mesh topology is bus-like as far as exchanging messages is concerned, and star-like regarding data transfers, since every station receives all transmissions from every other station, but connection failures can be overcome.

Hybrid topologies

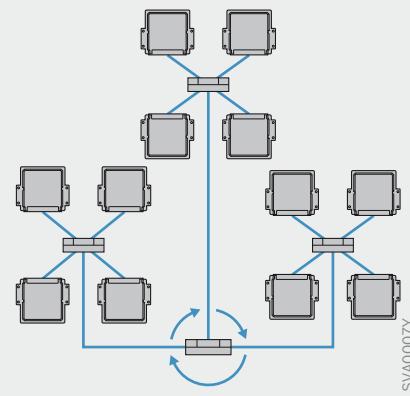
Hybrid topologies are a combination of different network topologies. Examples of such combination are:

- ▶ Star bus topology: the hubs of several star networks are interconnected as a linear bus (Fig. 6).
- ▶ Star ring topology: the hubs of several star networks are connected to the main hub (Fig. 7). The hubs of the star network are connected in the form of a ring in this main hub.

6 Star bus topology



7 Star ring topology



Network organization

Addressing

In order to make it possible to transmit messages via a network and evaluate the contents thereof, the useful data (payload) that is transmitted is also accompanied by data transfer information. This can be explicitly contained within the transmission or implicitly defined using preset values. Addressing represents important information for data transfer information. It is needed in order for a message to be sent to the correct recipient. There are different ways of doing this.

Subscriber-oriented method

The data is exchanged on the basis of node addresses. The message sent by the transmitter contains the data to be transmitted and also the destination node address (Fig. 8a). All receivers compare the transmitter receiver address to their own address, and only the receiver with the correct address evaluates the message.

The majority of conventional communication systems (such as Ethernet) operate using the subscriber addressing principle.

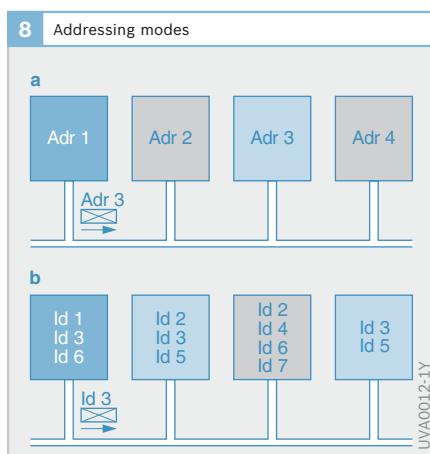


Fig. 8

- a Subscriber-oriented method
- b Message-oriented method

Message-oriented method

In this method it is not the receiver node that is addressed, but the message itself (Fig. 8b). Depending on the content of the message, it is identified by a message identifier that has been predefined for this message type. In this method, the transmitter does not need to know anything about the destination of the message, since each individual receiver node decides whether or not to process the message. Of course, the message can be received and evaluated by several nodes.

Transmission-oriented method

Transmission characteristics can also be used to identify a message. If a message is always transmitted within a defined time window, it can be identified on the basis of this position. By way of a safeguard, this addressing is often combined with message or subscriber-oriented addressing.

Bus access method

A node must access the bus in order to transmit a message. In the bus access method a distinction is made between

- ▶ Predictable methods in which the bus access is determined by certain time-dependent network characteristics, whereby only one node can transmit at a time and
- ▶ Random methods whereby any node can attempt to transmit data if the bus appears to be free

In the predictable method the bus access right is determined before bus access. It can thereby be ensured that only one subscriber is using the bus at a time.

Access collisions because of simultaneous bus usage will be prevented if all subscribers use this method.

In the random method, the nodes can simultaneously attempt to use the bus as soon as it appears to be free. The timing of the bus access is therefore random. There is a risk of transmission collisions using this method, which will require attention. This can be dealt with by repeating transmissions after a collision has been detected (e.g. Ethernet), by giving the transmissions different codings (CDMA), controlling communication via a master or prioritizing message types or transmitters.

Time division multiple access (TDMA)

TDMA is a deterministic (predictive) access method. In this case each node is assigned a time window in which it is allowed to transmit (a priori). A fixed schedule is therefore required for the network. There is not usually a main communication subscriber controlling the communication procedure. However, concepts exist in which it is possible to switch between different schedules if necessary. The internal clocks of the different stations must run extremely synchronously with TDMA, since the transmit windows have to be adhered to with extreme precision.

Master-slave

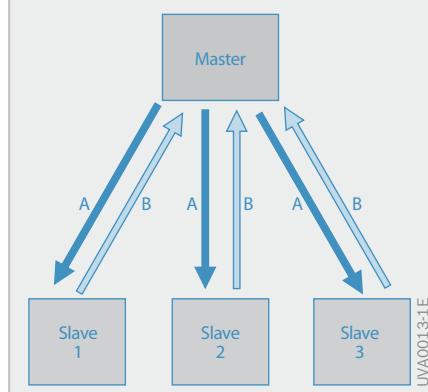
In the master-slave system, one node on the network operates as the master. This node determines the communication frequency by interrogating its subordinate nodes (slaves). A slave only replies if it is spoken to by the master (Fig. 9). However, some master-slave protocols allow a slave to contact a master in order to transmit a message (e.g. transmit information about the position of the power-window unit to the door module).

Multimaster

In a Multimaster network, several nodes can access the transport medium independently without the assistance of another node. Bus access is uncontrolled. Every node can access the bus and transmit a message if the bus appears to be free. This means that each node is its own master, and that any node can start a message transfer with equal status. However, this also means that collision detection and handling methods have to be in place. For example, this may be in the form of a decision-making phase with prioritization or delayed transmission repeats. The use of priority control prevents a bus conflict if several nodes wish to use the bus at the same time, since the network node that has high priority or wishes to transmit a message with high priority forces its way through in the event of a conflict and transmits its message first. Normal message transmission resumes when the line is free again.

The Multimaster architecture has a positive effect on the availability of the system, since no individual node is in control of communication whose failure would lead to total communication breakdown.

9 Master-slave method



OSI reference model

Network protocols are usually defined in layers, which combine properties and tasks. The properties of the deeper-lying layers are assumed in the next level up. This has the advantage that individual layers are exchangeable, provided that the interfaces that are provided between the layers remain unchanged.

The ISO OSI reference model (Open Systems Interconnection) provides a basis for describing and comparing many communication protocols. This was developed by the ISO (International Standardization Organization) and led to the adoption of international standards by ISO and IEEE (Institute of Electrical and Electronic Engineers).

In the OSI model, data communication systems are depicted in different layers (Fig. 10). The complex task of data communication is distributed among clearly ar-

ranged functional areas (layers). Not all of the layers in the OSI model are needed in a simple communication system. Layers can also be combined for many applications. Network protocols in the automotive area are often divided up into

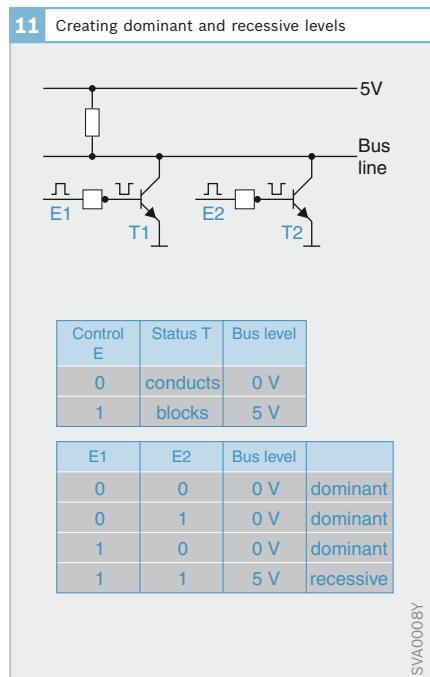
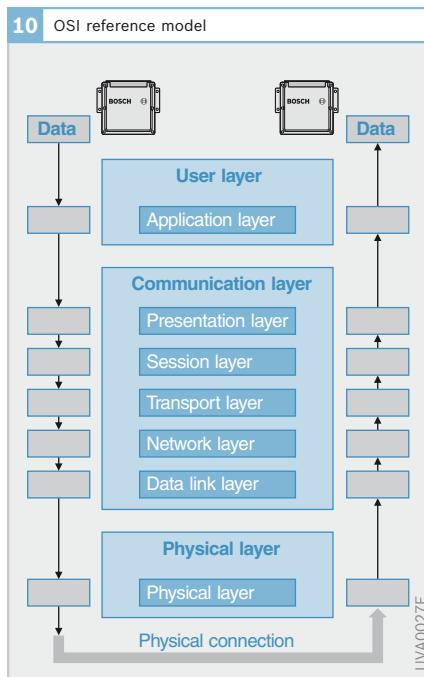
- ▶ Physical layer
- ▶ Communication layer and
- ▶ Application layer (user layer)

Physical layer

The electrical and procedural parameters of the physical connection between the network subscribers is defined in the physical layer.

Signal level

In digital technology, data is represented by sequences of the two binary statuses, 0 and 1. In order to transmit the data on a bus, these statuses must be represented on the transmission agent. It is particularly important to avoid short-circuits on the



bus when one node is transmitting a status of 1 and another is transmitting a status of 0.

The binary statuses can be depicted in many different ways. The serial interface of the PC, for example, uses +12 V and -12 V, and CAN-B uses voltages of 0 V and 5 V. The voltages of the serial interface are unsuitable for a bus, since short-circuits can occur if several subscribers wish to transmit conflicting binary statuses simultaneously.

If the coding allows one level to overwrite another, the overwriting level is referred to as dominant, and the subordinate level as recessive.

It is also possible to depict dominant and recessive levels using visual media. A status of 1 (recessive) then corresponds to e.g. dark, and a status of 0 (dominant) corresponds to light. In an optical fiber, an individual node can override all of the others by feeding light into the conductor.

Bit stream

The application information cannot usually be transmitted directly. In order to make transmission possible, the information is first incorporated as a payload in the frame of a message that contains information to be transmitted. Since all protocols have been developed in accordance with different requirements, the frame format differs from protocol to protocol.

The frame needs to be converted into a bit stream to actually transmit the information, i.e. a sequence of bits that can be transmitted via the transport medium as physical states.

Example of the UART interface

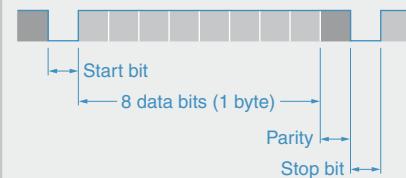
The microcontrollers that are used in control units have a simple interface (UART, Universal Asynchronous Receiver/Transmitter) on the chip, via which they can communicate with the outside world (e.g. with a PC). The essential features of a data transmission are read out via this interface.

If no data is being exchanged, the bus level is 5 V (microcontroller operating voltage, Fig. 11). When the start bit is transmitted (dominant level), the other station connected to the bus (receiver) is notified that a data transfer is starting (Fig. 12). The length of the start bit determines a bit time that represents the basis for the entire data transfer. Every subsequent data bit has the same length. The reciprocal of this time corresponds to the data transfer rate, i.e. the number of bits that can be transmitted in one second in a continuous data stream. All participating stations must be set to the same data transfer rate.

After the start bit has been received, the transmission of an 8-bit data word commences (1 byte) with the lowest significant bit (LSB, Low Significant Bit). The receiver that has synchronized itself to the start bit scans the data bus between each data bit and therefore assembles the transferred data byte.

The eighth data bit is followed by the parity bit. This bit indicates whether the number of transmitted ones is odd or even. It therefore allows the receiver to perform a simple check for possible transmission errors. The sequence is completed with the stop bit, which is placed onto the bus

12 UART interface transmission frame



SVA009Y

with a dominant level. The next data transfer can then take place.

Communication layer

Control units can only interconnect and exchange data if they speak the same “language”. This language determines the rules that are used to exchange data between the individual network subscribers.

The communication layer accepts data from the application layer, prepares it for transmission and forwards it to the physical layer.

The essential features of this protocol layer are:

- ▶ Message frame format
- ▶ Bus access control
- ▶ Message addressing
- ▶ Detection and handling of collisions
- ▶ Network node synchronization
- ▶ Checksum calculation

Application layer

The application layer consists of the application that processes and provides the information. The application layer is the only protocol layer to be affected by user or sensor input.

Control mechanisms

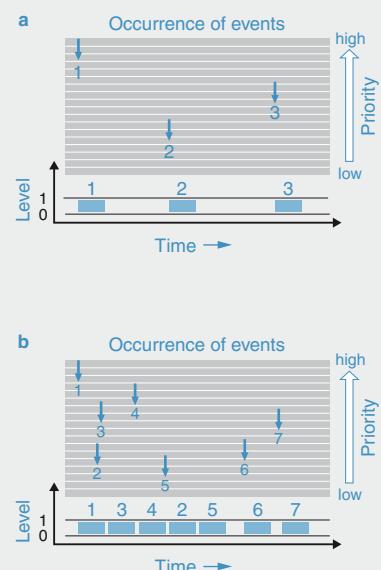
Event control

In an event-driven bus system, messages are transmitted as soon as an event that triggers the transmission of a message has occurred (Fig. 13a). Examples of such events are:

- ▶ Pressing a button on the air conditioning system control panel
- ▶ Operating the hazard warning flasher switch
- ▶ Incoming message that requires a reaction (e.g. information from rpm sensor to speedometer needle motor)
- ▶ Expiration of a fixed time period (time frame, e.g. 100 ms), after which messages are transmitted cyclically

Since the stations are not synchronized with each other, situations where several stations wish to access the bus simultaneously are unavoidable. In order to allow

13 Event control



a message to be transmitted without falsification, only one station at a time can transmit data on the bus. Collision avoidance mechanisms are available for preventing or solving bus conflicts.

If a node wishes to transmit a message whilst the bus is occupied, the transmission is delayed (Fig. 13b). A station that is ready to transmit must then wait until the transmission that is currently in progress has been completed.

Since bus access is subsequently renegotiated, the transmission may be delayed yet again. These delays become problematic if the bus becomes overloaded by a large number of network subscribers that wish to transmit messages. In this case messages may be lost if the transmitter abandons the transmission due to excessive delays.

Event-driven bus systems are suitable for reacting to asynchronous (unforeseen) events as quickly as possible. In an ideal case, they reduce the delay between the occurrence of the event and the message transmission (latency time) compared to time-driven systems. However, the latency time can vary considerably depending on the network loading.

Advantages

- ▶ High level of flexibility and capability of retrofitting new nodes in the network
- ▶ Good response time to asynchronous external events
- ▶ Bus usage depending on event frequency in line with requirements
- ▶ No network loading by unused events, since only events that have actually occurred trigger a transmission

Disadvantages

- ▶ Static bus occupancy, non-deterministic (i.e. not possible to prove that a message was transmitted at the right time)

Timer control

In the most recent developments in dynamic driving systems such as brakes and steering, an increasing number of mechanical and hydraulic components are being replaced with electronic systems (x-by-wire). Mechanical connections such as the steering column are becoming superfluous, and the functionality thereof is being taken over by sensors and actuators. The reliability, safety and failure tolerance requirements of these systems are extremely high. This means:

- ▶ Messages must be received on time
- ▶ The latency time of critical messages must be extremely small
- ▶ The system must have a redundant design
- ▶ The failure of a node must affect the rest of the system as little as possible and
- ▶ It must be possible to achieve a safe operating status from any fault situation

X-by-wire systems require close networking by the various components. The external increase in complexity places new demands on the safety, failure tolerance and availability of the communication system. The demands that are made of the electronic and network architecture therefore also increase. A reliable, fault-tolerant network architecture is required so that data is transmitted with guaranteed transmission characteristics, and electronic system malfunctions are handled in the most efficient way.

System architectures for real-time applications meet these requirements because their behavior is predictable and verifiable because of the way in which they are constructed. In these protocols, time windows within which a node is permitted to transmit are assigned to the control units in the communication network (nodes) during network planning (Fig. 14). In order to comply with the time window, the nodes must be synchronized as precisely as possible.

All transmissions are processed sequentially in accordance with the network planning (without collisions). Once each node has transmitted its message, the cycle restarts with the first transmitter. This makes it possible to determine how chronologically up-to-date the data is at any time. Since missing messages are detected immediately, time-triggered concepts are more reliable than event-driven systems.

If a fast data rate is required in a time-triggered system, the time delay between the occurrence of an event and the transmission of the data can be so small that the system complies with strict real-time requirements.

The bus can be protected from unauthorized access by a bus guardian. The bus guardian prevents a defective node from interfering with network communication by transmitting messages outside the relevant transmit window.

These characteristics make it possible to create redundant, fault tolerant systems in which transmission errors can be remedied and faults in the network can be picked up by network nodes that can provide the functionality without errors.

Advantages

- ▶ Deterministic system
- ▶ Punctual data transmission
- ▶ Reliable detection and isolation of defective network nodes

Disadvantages

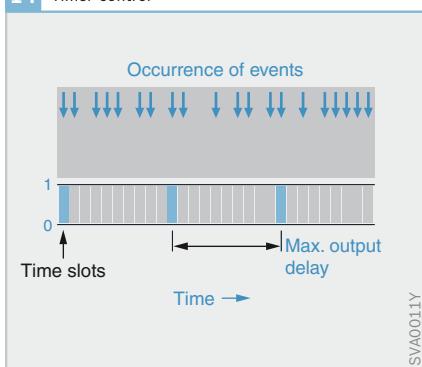
- ▶ Overall system must be planned for distributed developments
- ▶ Capacity for expanding the communication system must be planned in
- ▶ Good response time to asynchronous external events

Composability

If a communication system allows independently developed subsystems to be integrated in an overall system, it is said to support composability. An important criterion when doing this is that the properties that have been assured for the functionality of a subsystem are not adversely affected by adding other subsystems. If this has been ensured, the checking of system functionality is restricted to subsystem checking that can be carried out by the constructor of the subsystem.

If a communication system supports composability, changes can be made to a control unit without affecting the functionality of other control units. It is therefore not necessary to recheck the entire system after integrating a modified control unit - it is sufficient to check that the individual subsystems are operating reliably. Composability therefore reduces the time and cost of integrating new subsystems. This is the only way to increase the complexity of the electronics in the vehicle.

14 Timer control



► Overview of bus systems used in vehicles				
	CAN-C high-speed CAN	CAN-B low-speed CAN	LIN	TTP
Definition	Controller area network	Controller area network	Local interconnect network	Time-triggered protocol
Bus type	Conventional bus	Conventional bus	Conventional bus	Conventional and optical bus
Domains	Drivetrain	Comfort/convenience	Comfort/convenience	Safety-related networking
Applications	Engine management, transmission control and ABS/ESP networking	Body and comfort and convenience electronics networking	Low-cost expansion of CAN bus for simple applications in the comfort and convenience electronics area	Networking in safety-related environments such as brakes, steering, railway signal boxes or aircraft landing gear
Most frequently used topology	Linear bus	Linear bus	Linear bus	Star topology
Data transfer rate	10 kbit/s to 1Mbit/s	Max. 125 kbit/s	Max. 20 kbit/s	Unspecified, typ. 10 Mbit/s
Max. number of nodes	10	24	16	Unspecified
Control mechanism	Event-driven	Event-driven	Time-driven	Time-driven
Bus lines	Copper conductors (twisted pair)	Copper conductors (twisted pair)	Copper conductor (single wire)	Copper conductors (twisted pair)
Deployment	in all vehicles	in all vehicles	in all vehicles	Premium class vehicles, aircraft, rail control systems
Standard	ISO 1198	ISO 11519-2	LIN consortium	TTA group
SAE classification	Class C	Class B	Class A	Drive-by-wire

	MOST Bus	Bluetooth	Flexray
Definition	Media oriented systems transport	Proprietary name (Danish king)	Proprietary name
Bus type	Optical bus	Wireless	Conventional and optical bus
Domains	Multimedia and Infotainment	Multimedia and Infotainment	Deployment across all domains
Applications	Transmission of control, audio and video information	Data transfers over short distances, e.g. mobile phone integration in the infotainment system	A network system for use in safety-related and simple applications
Most frequently used topology	Ring topology	Network topology (radio)	Star topology
Data transfer rate	Max. 22.5 Mbit/s	Max. 3 Mbit/s (v2.0) Max. 723 kbit/s (v1.2)	Typ. 10 Mbit/s Max. 20 Mbit/s
Max. number of nodes	64	8 active (up to 256 passive)	Theoretically up to 2,048 Max. 22 per passive bus/star
Control mechanism	Time and event-driven	Event-driven	Time and event-driven
Bus lines	Plastic or glass optical waveguides	Electromagnetic radio waves	Copper conductors (twisted pair)
Deployment	Premium class vehicles made by European manufacturers	All vehicles, connection between multimedia equipment and infotainment system	Pilot application
Standard	MOST cooperation	Bluetooth SIG	Flexray consortium
SAE classification	Mobile Media	Wireless	Drive-by-wire

Table 1

Automotive networking

Electrical and electronic systems in motor vehicles are often not independent of each other but influence and complement each other. For this reason, signal lines were used in previous injection and ignition systems in order to simplify communication between these two systems. However, the increasing number of electronic systems rapidly increased the demand for and the scope of the information that was being exchanged. The number of signal lines and plug connections that is required increased accordingly, meaning that the technology that has so far been used was approaching the limit of its capability.

The solution was provided by the development of serial bus systems, with which large volumes of data from different sources can be transferred. A serial bus system was first used in a vehicle in 1991, when the CAN bus was used in the Mercedes-Benz 500E.

The demand for additional driving safety, convenience, economy and stricter legal requirements on the environmental compatibility of motor vehicles can only be achieved with the aid of additional electronics. The number of electronic systems in vehicles is therefore increasing all the time (Fig. 2).

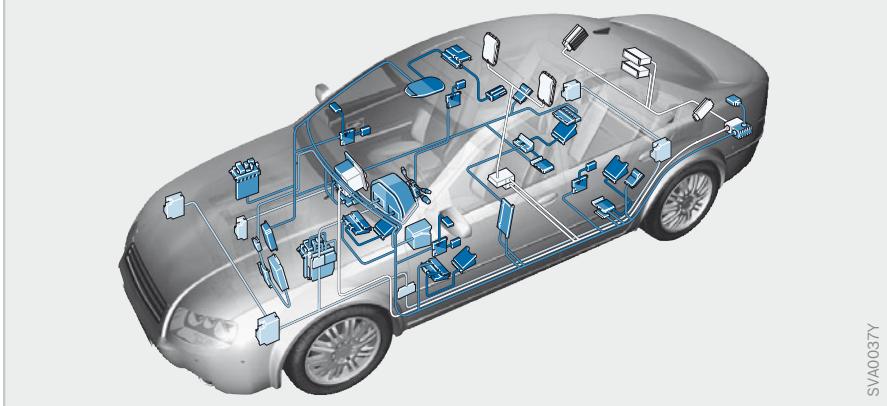
Cross-system functions

If you examine the signals that are processed in the individual systems, it becomes evident that many signals are needed in several control units. For example, the driving speed is evaluated in the electronic stability program (ESP) for the vehicle dynamics control, in the engine management for the automatic speed control (cruise control) and in the car sound system for the speed-dependent volume control.

The preparation of these variables from sensor signals requires computing power and therefore hardware and software resources. It is therefore advisable for these variables always to be calculated in a control unit and transmitted to other control units via a communication network.

However, in order to implement cross-system functions, electronic systems can also exchange information and therefore influence each other. Intelligent sensors are also considered to be electronic systems that prepare the sensor signal in an evaluation circuit and put the information on the data bus via a bus interface. Pre-crash sensors detect a pending collision, for example; the airbag control unit then sends the door modules and the overhead control panel a request to close the windows

1 Automotive networking



SYA037Y

and the sliding roof. This protects the occupants from penetrating objects.

Another example of a system-encompassing function is the adaptive cruise control (ACC), in which the radar sensor, the engine management, the electronic stability program (ESP) and the transmission control communicate with each other. Distance control from the vehicle in front that is adapted to the flow of traffic is made possible in this system by means of engine torque adjustments, automatic brake system intervention and gear selection.

Coordination between the individual systems is therefore required for cross-system functions. Large volumes of data must be exchanged to do this. As well as powerful components, a powerful communication system is also required, with a low-cost network that is suitable for automotive vehicles. Special serial databus systems have been developed for this purpose.

The use of bus systems has the following advantages in comparison to a solution that uses conventional wiring:

- ▶ Reduced costs with less weight and installation space because of fewer cables in the wiring harness
- ▶ Better reliability and functional reliability due to fewer plug-in connections
- ▶ Simplification of vehicle assembly during production
- ▶ Multiple use of sensor signals
- ▶ Simple connection of system components to a bus
- ▶ Easier handling of equipment and special equipment variants in a vehicle

Requirements for bus systems

Both financial (e.g. cable costs, component costs) and technical constraints must be taken into consideration when a bus system is being selected. The most important technical selection criteria are explained in the following.

Data transfer rate

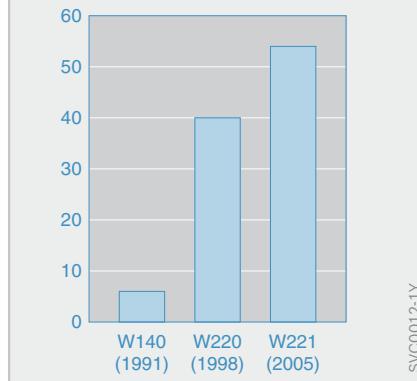
This variable specifies the volume of data that is transmitted during a time unit. The smallest unit of data is the bit, and the data transfer rate is usually specified in bits/seconds. Alternative names for this expression are transfer rate, data rate or bit rate.

The required data rate is dependent on the application. A slower transfer rate is required to switch the air-conditioning compressor on and off than to transfer audio signals, for example.

Interference immunity

Ideally, the data should be transferred without interference. However, this cannot be guaranteed in a motor vehicle because of electromagnetic effects. The interference immunity requirements that are made depend on the safety relevance of the electronic systems concerned. Lesser requirements are made of comfort and

2 Number of control units in the Mercedes S-class that are networked via CAN



convenience systems than the antilock brake system (ABS), for example.

In order to meet these requirements, mechanisms that detect transmission errors are incorporated in the network protocols. A simple check can be carried out using the parity bit, which is calculated in the transmitter and is transmitted together with the useful data. This specifies whether the number of 1's in the transferred byte is even or off. This information is checked by the receiver. Single errors can be detected using this method.

Another method is the checksum check. If several data bytes are being transmitted, the transmitter calculates a checksum from the individual data bytes using a pre-defined formula and transmits this value. The receiver also calculates the checksum of the data bytes that have been received and compares it with the checksum that has been received. If a data transmission error is detected, the received data is not used and a repeat transmission is requested.

Real-time capability

A real-time system guarantees that its results are calculated within a fixed time interval. The duration of the time interval depends on the application. The antilock brake system (ABS) must react to the incipient locking of a wheel within a few milliseconds (wheel speed reduction), whereas response times of 100 ms are adequate for actuating the power-window motor. Human beings cannot perceive delay periods of less than 100 ms.

Different demands are made of real-time behavior depending on the application:

- Soft real-time requirement: the system generally adheres to the specified response time, and if these times are occasionally exceeded, it does not produce any serious effects (e.g. image jerking during picture transmission).

- Hard real-time requirement: the time specification must be strictly adhered to. If the specified response time was exceeded, the calculated result would not be able to be used. This can lead to serious problems in safety-critical systems.

For example, if time allowances were exceeded in the ABS system, the incipient locking of the wheels would not be detected soon enough and the pressure in the master cylinder would not be reduced in time. This would result in locked wheels.

The time allowances must also be strictly adhered to for many engine-management system functions. Delays in transmitting injection and ignition signals could lead to engine judder and even misfiring. These reactions must be avoided, since they represent a potential danger. Hard real-time requirements must therefore be made of these systems.

However, this does not necessarily mean that the transmission of data via a bus system also has to be subject to these hard real-time requirements. Adherence to soft real-time requirements is usually sufficient. If signals from other control units are needed for functions (e.g. a torque reduction request during a shift operation), the bus system must transmit the data at a faster data transfer speed and with a smaller time delay so that the overall system complies with the specified real-time requirements.

Number of network nodes

The maximum number of nodes to be integrated varies for different areas of vehicle operation. The number of nodes for comfort and convenience systems may be high due to servomotor networking (e.g. seat adjustment) and intelligent sensors (e.g. rain sensors). Several identical busses can be used if necessary.

Classification of bus systems

Because of differing requirements, bus systems can be subdivided into the following classes.

Class A	
Transfer rates	Low data rates (up to 10 kBit/s.)
Applications	Actuator and sensor networking
Representative	LIN
Class B	
Transfer rates	Average data rates (up to 125 kBit/s.)
Applications	Complex mechanisms for error handling, control unit networking in the comfort functions
Representative	Low speed CAN
Class C	
Transfer rates	High data rates (up to 1 MBit/s.)
Applications	Real-time requirements, control unit networking in the drive and running gear functions
Representative	High speed CAN
Class C+	
Transfer rates	Extremely high data rates (up to 10 MBit/s.)
Applications	Real-time requirements, control unit networking in the drive and running gear functions
Representative	FlexRay
Class D	
Transfer rates	Extremely high data rates (> 10 MBit/s.)
Applications	Control unit networking in the telematics and multimedia functions
Representative	MOST

Applications in the vehicle

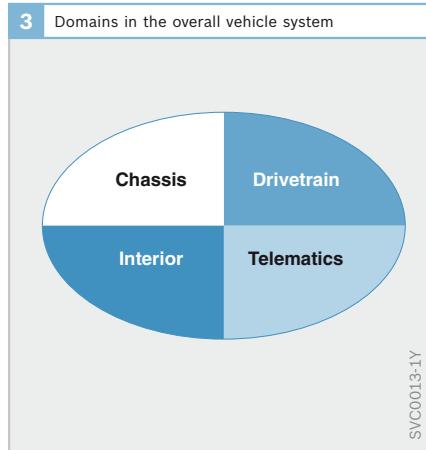
The overall vehicle system can be divided into four domains or functional areas from the point of view of electrics/electronics:

- Drivetrain
- Chassis
- Interior and
- Telematics

In the drivetrain and chassis domains, the emphasis is primarily on real-time applications. In the interior domain, the main focus is on multiplex aspects in networking. Mainly multimedia and infotainment applications are networked in the telematics domain.

Real-time applications

The networking of these systems makes considerable demands of the performance capability of the communication system. Crankshaft-synchronous processes or processes within a fixed time frame with cycle times of a few milliseconds are typical. If the system response times are adequate for the task in hand, it is described as having real-time capability (e.g. rapid ignition-timing advance in the Motronic after a request from the traction-control system for reducing torque and therefore preventing the wheel from spinning).



The drivetrain and chassis systems are assigned to class C. These require fast transfer rates in order to ensure the real-time behavior that is required for these applications. They also make considerable fault tolerance demands. These requirements are met by the event-driven CAN bus with a transfer rate of 500 kBaud (high-speed CAN).

Examples:

- ▶ Engine-management system (Motronic or electronic diesel control, EDC)
- ▶ Transmission control
- ▶ Antilock brake systems, ABS
- ▶ Vehicle dynamics control (e.g. electronic stability program, ESP)
- ▶ Chassis control systems (e.g. active body control, ABC)
- ▶ Support systems (e.g. adaptive cruise control, ACC)

Multiplex applications

The multiplex application is suitable for controlling and regulating components in the body and comfort and convenience electronics area (class B), such as

- ▶ Displays
- ▶ Lighting
- ▶ Access authorization with anti-theft warning device
- ▶ Air-conditioning
- ▶ Seat and mirror adjustment
- ▶ Door module (power-window unit, door-mirror adjustment)
- ▶ Windshield wipers
- ▶ Headlamp adjustment

The transfer rate requirements are not as high for class B systems as they are for class C systems. For this reason, low-speed CAN with a transfer rate of 125 kBit/s or single-wire CAN with 33 kBit/s. can be used.

If the transfer rate requirements drop to less than 20 kBit/s, the low-cost LIN is more frequently used. Applications are mainly in the mechatronics area; examples being the transfer of switch information or the activation of actuators.

Multimedia networking

Mobile communication applications combine components such as

- ▶ Car sound system
- ▶ CD changer
- ▶ Navigation system
- ▶ Driver-information systems
- ▶ Telephone
- ▶ Video system
- ▶ Voice input
- ▶ Internet, E-mail
- ▶ Back-up camera

The networking of these components makes it possible to have a centrally located display and control unit for several applications. Operating procedures can be standardized in this way, and status information can be summarized. Driver distraction is therefore minimized.

A distinction must be made between control data and audio/video data in multimedia networking. Transfer rates of up to 125 kBit/s are sufficient for control tasks (such as CD changer control), meaning that the low-speed CAN bus can be used, for example. The direct transmission of audio or video data requires extremely high transfer rates of more than 10 MBit/s. The MOST bus is used for this purpose, for example.

Coupling of networks

The network topologies and protocols that are most suitable for requirements are used for the different applications. However, the different network protocols are incompatible, meaning that data cannot be simply exchanged between networks.

In this case, help is provided by a gateway. A gateway can be compared to an interpreter that receives the “data” from a discussion partner, translates it and passes it to another discussion partner. Technically speaking a gateway is a computer that reads in the data that is transmitted by the networks and converts it into another format. The use of gateways therefore makes it possible to exchange information between different networks.

A central gateway (Fig. 4a) or several distributed gateways can be used (Fig. 4b) to interconnect the bus systems. All bus lines are routed to the central gateway. In the other case, one gateway connects two or more busses.

Examples of networked vehicles

Topology

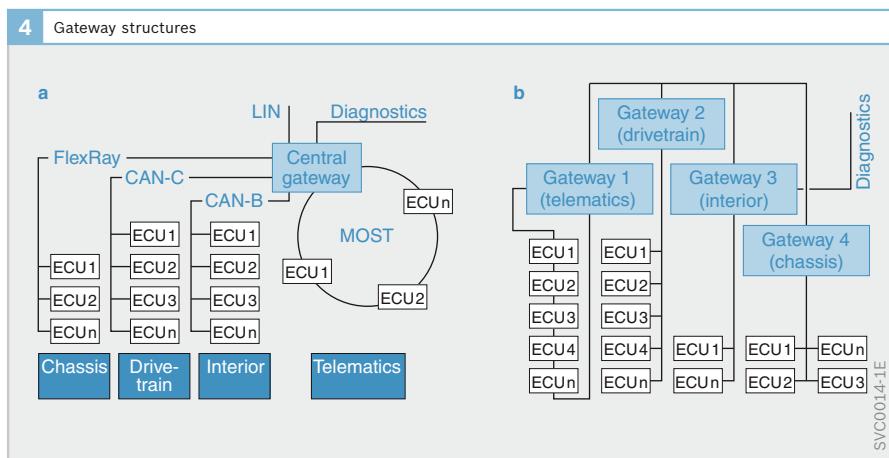
The topologies of the communication networks can differ considerably depending on the vehicle equipment. Figures 5 and 6 show examples of how the network can be designed for different vehicle classes. In some cases, different car manufacturers use different bus systems for communication.

Signal transmission

Signal types

A wide variety of information can be transmitted in a communication network in a vehicle. Some examples are:

- ▶ Engine operating conditions (e.g. engine temperature, engine speed, engine load)
- ▶ Physical measurements recorded by sensors (e.g. outside temperature)
- ▶ Control signals for activating servomotors (e.g. power-window units) or
- ▶ Control element switch positions (e.g. for the windshield wiper)



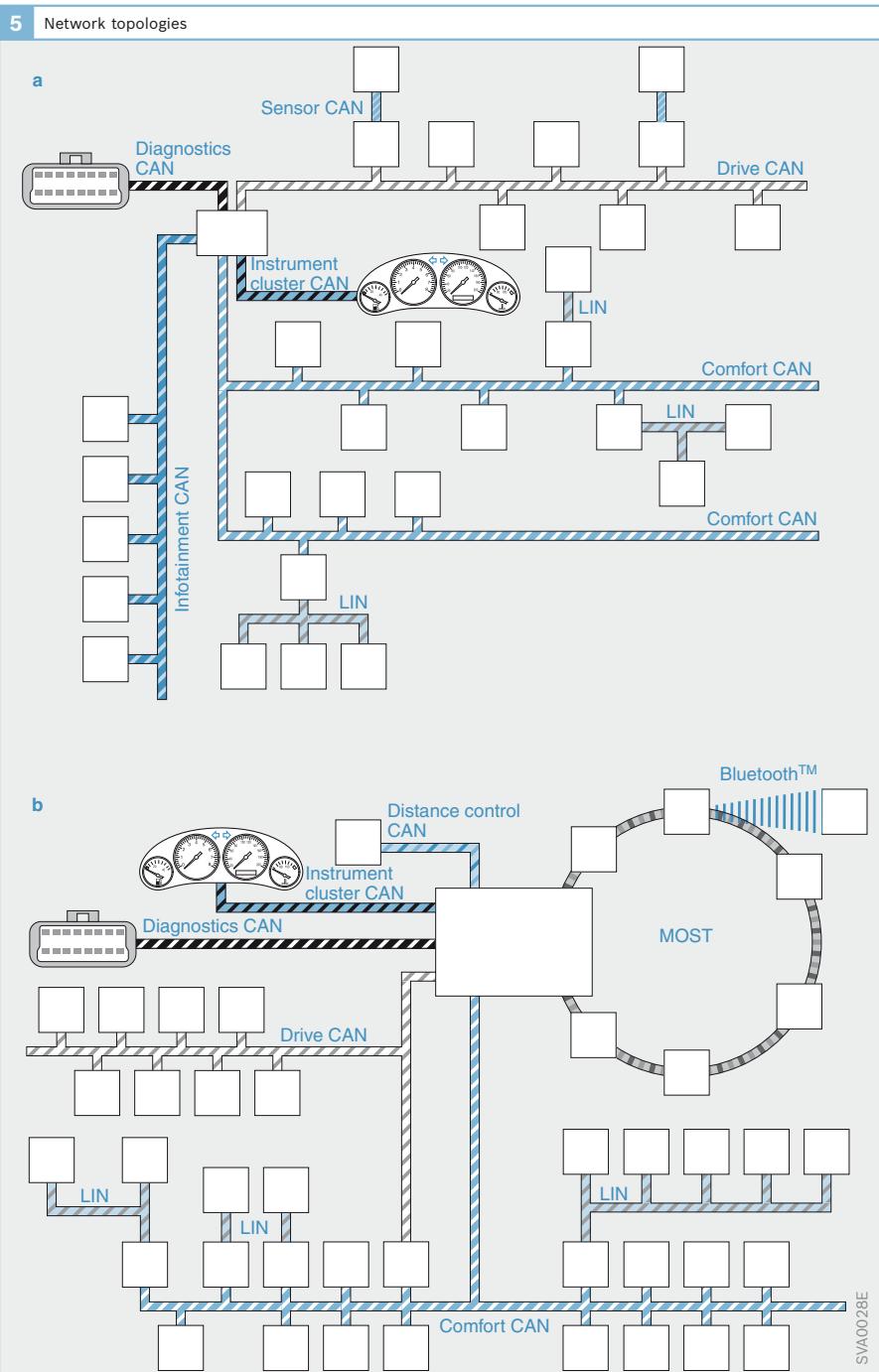


Fig. 5

Typical network topologies for new generation vehicles
 a Compact class
 b Luxury class

Resolution

The signals must be available with a suitable resolution. Switch positions can simply be shown as a 1-bit value (0 for switch open, 1 for switch closed). Other signals such as digitized analog voltages from the engine-temperature sensor or the calculated engine speed are shown as 1-byte or 2-byte values, for example, depending on resolution requirements. One byte can be used to represent 256 values, and two bytes can be used to represent 65,536 values (= 256·256). The resolution for sensor signals with a voltage range of 0 to 5 V is approximately 20 mV with a 1 byte representation (= 5 V/256). A resolution of 5 mV requires a 10-bit data representation.

The conversion between the binary value and the physical value must be uniform so that the transferred signals represent the same physical value in all systems. A resolution of 30 rpm is sufficient for accessing the ignition map as far as the engine speed n is concerned. A value range of 0 to 255·30 rpm (= 7,650 rpm) and therefore the entire speed range can consequently be represented by one byte (8 bits). An incrementation of 30 rpm is too little for idle-speed control, on the other hand. More bits are required to represent the signal with a higher resolution, provided that the same measuring range is being recorded.

Output

In event-driven systems such as the ones that are primarily used in the automotive area, the signals can be transferred on the data bus when an event occurs. Examples of such events are the operation of switches for switching on the air-conditioning system or the windshield wiper.

Signals that represent the operating state of the engine, for example, are not necessarily associated with an event. The engine temperature, which only changes slowly, is cyclically measured in a fixed time frame by the engine control unit (e.g. 1 second). The engine speed, on the

other hand, can change extremely rapidly. The times at which the engine management carries out measurements and calculations depends on the crankshaft position and is performed once per combustion cycle. At fast engine speeds, this corresponds to a time interval of a few milliseconds, i.e. approx. 3.3 ms at an engine speed of 6,000 rpm for a 6-cylinder engine. However, not every system that needs the speed information for its control and regulation functions depends on this availability. The engine control unit therefore does not need to output the speed information on the data bus as soon as it is calculated. In this case too, data transmission takes place according to a cyclic time frame. A time frame of 10 ms is normal in the engine control area. This means that the speed information is transmitted on the bus 100 times per second.

Data transfer: examples

The following examples show which signals are measured and evaluated in which systems.

Driving speed

The ESP control unit calculates the driving speed from the wheel-speed sensors. This variable is transmitted on the CAN-C bus (drive CAN). The engine-management system needs this value for the cruise control, among other things, and the transmission control unit determines gear changes from the driving speed. The adaptive cruise control (ACC) needs the current driving speed to calculate the necessary distance from the vehicle in front and use it as a setpoint value.

A gateway transmits the speed information via another CAN bus (instrument cluster CAN) to the instrument cluster, which displays the value via a needle instrument.

The CAN-B bus (comfort CAN) is also connected to the network via the gateway. Some luxury class vehicles are equipped with dynamic seats. The padding of the seats is inflated depending on the speed

and the acceleration, counteracting the centrifugal force of the driver. This increases comfort considerably when cornering.

The speed information is sent to the Infotainment CAN via the gateway and relayed to the car sound system. This allows the volume to be adapted to the driving speed. The navigation system needs the speed to calculate the position if the GPS signal is missing (e.g. in a tunnel).

The diagnosis interface is directly connected to the engine and transmission control unit via the serial K-line. All other control units are connected to the diagnosis interface via a virtual K-line that is simulated on the CAN bus. This allows the driving speed to be read out in the workshop via the connected diagnostic tester (example: the correct assignment of the wheel-speed sensors must be checked for the ABS functional test).

Engine speed

The injection and (with a gasoline engine) the ignition timing are output with a resolution of less than 1° of the crankshaft angle. In order to ensure real-time behavior, the crankshaft position must be recorded in the engine control unit. The enginespeed sensor scans the crankshaft trigger wheel and relays the signal to the control unit, which calculates both the crankshaft position and the engine speed. This variable is used to calculate the injection time and the ignition angle, for example.

The engine speed is a variable that is needed in many other systems. The engine control unit therefore outputs it on the data bus. The shifting points are defined in the transmission control unit depending on the speed. The engine speed is needed for the TCS function (Traction-Control System) in the electronic stability program (ESP) – TCS intervention (torque reduction) must not make the engine stall.

As in the previous example, the engine speed is transmitted to the diagnosis inter-

face and the instrument cluster (display on the rev counter).

Turn signaling

The driver operates the turn signal lever (Fig. 6, Item 1). A signal is relayed to the steering column control unit (2) via a discrete line depending on whether the driver is indicating a right or left turn. This may be a resistance-coded signal, for example. The control unit evaluates the signal and detects that the driver is indicating a left turn, for example.

The comfort CAN relays this information to the vehicle electrical system control unit (3). The indication direction is defined on the basis of the received information (normal flash frequency, increased flash frequency in the event of bulb failure). The front left and rear left turn-signal lamps (4, 5) are then actuated via discrete lines. The vehicle electrical system control unit also transmits the “left turn signaling” information on the comfort CAN. The gateway (6) relays the information to the instrument cluster CAN. The indicator lamp then flashes on the instrument cluster (9).

If the vehicle has a trailer hitch, the information goes to the trailer recognition control unit (7) via the comfort CAN. This actuates the turn-signal lamp on the trailer (8) via cables.

6 Data transfer during turn signaling

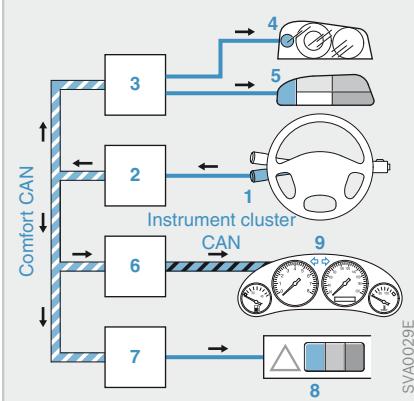


Fig. 6

- 1 Turn signal lever
- 2 Steering column control unit
- 3 Vehicle electrical system control unit
- 4, 5 Turn-signal lamp
- 6 Gateway
- 7 Trailer recognition control unit
- 8 Turn-signal lamp on trailer
- 9 Instrument cluster

Wiper stage 1

The wiper switch (Fig. 7, Item 1) transmits a signal via a discrete line to the steering column control unit (2), which evaluates the information (e.g. wiper stage 1). The control unit transmits this information on the comfort CAN bus. The vehicle electrical system control unit (3) picks up the information and relays it to the wiper motor (4) via the LIN bus. The vehicle electrical system control unit acts as a gateway between the comfort CAN and the wiper LIN.

Load management

At low revs (idle speed) and when a considerable amount of power is being used up by the electrical consumers that are switched on, the battery or alternator voltage can drop to a low value. The vehicle electrical system control unit calculates the current status of the vehicle electrical system from the current battery voltage, the DF signal from the alternator (alternator utilization) and the information about heavy current consumers that are

switched on with a short switch-on duration. The vehicle electrical system control unit requests an idle speed increase via the CAN bus if the vehicle electrical system is insufficient at idle speed. The engine control unit implements the request. If this action does not solve the problem, the vehicle electrical system control unit switches off specific consumers such as the heated rear windshield, the seat heating or the heated outside mirror. These consumers are connected to the vehicle electrical system control unit via discrete lines.

Multimedia application

Audio and video data is transmitted in digital form via the optical MOST bus (Media Orientated System Transport) (Fig. 5). The information is therefore transmitted with high quality. A fast transfer rate ensures that large volumes of data are transferred reliably. A wide variety of multimedia information can be transmitted to different terminals via a cable.

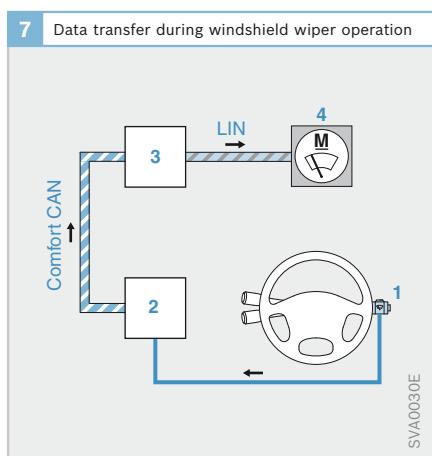


Fig. 7

- 1 Windshield wiper lever
- 2 Steering column control unit
- 3 Vehicle electrical system control unit
- 4 Wiper motor

Bus systems

CAN bus

In 1991 the CAN bus (Controller Area Network) was the first bus system to be introduced to a motor vehicle in mass production. It has since established itself as the standard system in the automotive sector, but the CAN bus is also commonly used as a field bus in automation engineering in general. In imitation of other network types, such as the local area network (LAN), wide area network (WAN) or personal area network (PAN), this bus system was given the name, CAN.

Applications

The CAN bus is used in various domains in the motor vehicle. These domains differ in the requirements they demand of the network. Due to the fast processes involved in the area of engine management, information is required much faster here than in the area of comfort/convenience where the controlled systems are located further apart and as such lines are more prone

to damage. As a result of these different requirements, buses with different data rates are used that offer an optimum cost-benefit ratio for the field of application concerned. A distinction is made between high-speed and low-speed CAN buses.

High-speed CAN (CAN-C)

CAN-C is defined in ISO Standard 11898-2 and operates at bit rates of 125 kBit/s to 1 MBit/s. The data transfer is therefore able to meet the real-time requirements of the drivetrain.

CAN-C buses are used for networking the following systems:

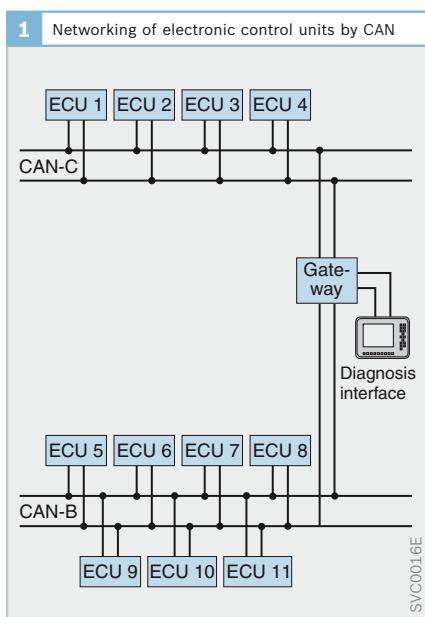
- ▶ Engine-management system (Motronic for gasoline engines or EDC for diesel engines)
- ▶ Electronic transmission control
- ▶ Vehicle stabilization systems (e.g. ESP)
- ▶ Instrument cluster

Low-speed CAN (CAN-B)

CAN-B is defined in ISO Standard 11898-3 and operates at a bit rate of 5 to 125 kBit/s. For many applications in the comfort/convenience and body area, this speed is sufficient to meet the real-time requirements demanded in this area. Examples of such applications are:

- ▶ Control of the air-conditioning system
- ▶ Seat adjustment
- ▶ Power-window unit
- ▶ Sliding-sunroof control
- ▶ Mirror adjuster
- ▶ Lighting system
- ▶ Control of the navigation system

The CAN bus is finding ever more use in vehicle diagnostics. Here, the electronic control unit is connected directly to the CAN bus and thus receives the information it needs for diagnostics immediately. Previous diagnosis interfaces (e.g. KWP2000) are becoming less important.



Topology

Bus topology

In the development of the CAN, particular focus was placed on eliminating the need for a central control element for communication. This approach is most effectively supported by a bus topology in which all network nodes are connected to a bus and each node is able to receive all information sent on the bus. The bus topology is most commonly selected during the conceptual design of the communications system.

In addition to offering favorable electrical properties, the linear bus topology has the advantage that the failure of one station would not affect the functionality of the data transmission system. Furthermore, additional stations can be connected to the system with little extra effort.

Star topology

The use of a central coupler makes it possible to build star topologies. Active as well as passive couplers can be used. The use of a star topology achieves a high level of flexibility in adapting to the networking task.

The coupler is used to build a star topology and simply forwards the messages to the individual segments. Since the signal transit times remain unchanged, the circuit length of the star topology is the same as that of a bus topology.

Data transmission system

Network nodes

A network node (Fig. 2) comprises the microcontroller for the application software, the CAN controller and the CAN transceiver (bus driver). The CAN controller is responsible for the transmit and receive modes. It generates the bit stream for data communication from the binary data to be transmitted and forwards it to the transceiver on the Tx D line. This amplifies the signals, generates the voltage level required for differential data transfer and transmits the processed bit stream serially on the bus line (CAN_H and CAN_L).

Incoming messages are processed by the transceiver and sent to the CAN controller on the Rx D line.

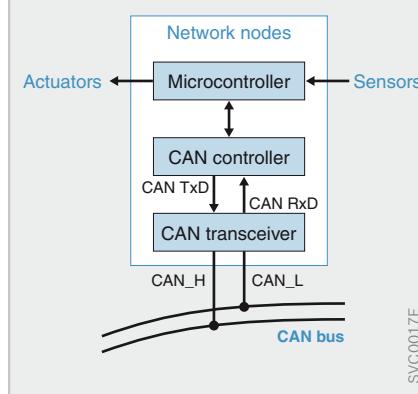
The microcontroller, which runs the application program (e.g. Motronic), controls the CAN controller, prepares the data to be sent and evaluates the data received.

Logic bus states and coding

CAN uses two states for communication, dominant and recessive, with which the information bits are transmitted. The dominant state represents a binary “0”, the recessive a binary “1”. NRZ (Non-Return to Zero) is used as the encoding method for the data transmission. With this method, there is no compulsory return to zero between two transmission states of the same value.

When it receives messages, the CAN transceiver converts the signal level back to logical states. In the process, a differential amplifier subtracts the CAN_L level from the CAN_H level (Fig. 3). If lines become twisted, disturbance pulses (e.g. from the ignition system) have the same effect on both lines. Differential data transfer therefore makes it possible to filter out interference on the line.

2 Network nodes in the CAN



Some transceivers also evaluate the voltage level on the CAN_H and CAN_L line separately. It would then be possible for operation to continue in single-line mode if one of the two bus lines were to fail as a consequence of a short-circuit or cable break. However, the bus subscribers would have to share a common ground that would assume the function of the failed line.

Transmission agent and bus coupling

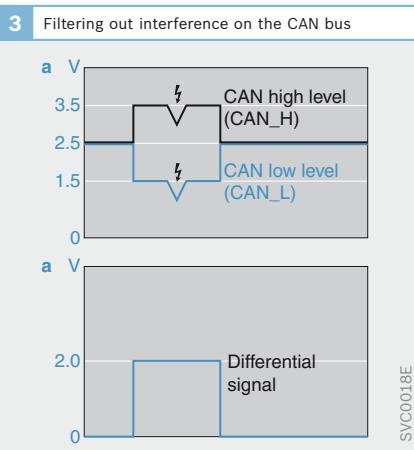
Two-wire line

For the CAN, any transmission agent on which dominant and recessive states can be transmitted is worthy of consideration. An untwisted or twisted pair of wires, depending on the ambient conditions, is usually used, and the wires are either galvanically coupled or decoupled. The two bus lines are designated CAN_H and CAN_L. The two-wire line supports symmetrical data transfer whereby the bits are sent on both bus lines and represented by different voltages. This reduces sensitivity to in-phase interference because the interference affects both lines equally and can be filtered out. Additional shielding of the lines reduces their own radiation emissions, especially at high baud rates.

Single-wire line

The single-wire line is a means of reducing manufacturing costs by dispensing with the second line. For this to be possible, however, all bus subscribers must share a common ground that would assume the function of the second line. The single-wire version of the CAN bus is therefore only an option for a communications system of limited spatial dimensions.

The data transfer on the single-wire line is more prone to interference radiation because it is not possible to filter out disturbance pulses as it is with the two-wire line. For this reason, a higher level increase is required on the bus line to improve the signal-to-noise ratio. This in turn has a negative effect on interference radiation. The flank steepness of the bus signals must therefore be reduced in comparison to the two-wire line. This is accompanied by a lower data transfer rate. Consequently, the single-wire line is only used for the low-speed CAN in the body electronics and comfort/convenience electronics. Thanks to this feature, a low-speed CAN with two-wire line would still remain functional in the event of a line failure (CAN_H or CAN_L).



Voltage level

The CAN transceiver converts the logical states 0 and 1 received by the CAN controller into voltage levels that are fed to the CAN_H and CAN_L bus lines.

The high-speed and low-speed CANs use different voltage levels for the transmission of dominant and recessive states. The voltage levels of the low-speed CAN are shown in Figure 4a, those of the high-speed CAN in Figure 4b.

In the recessive state, the high-speed CAN uses a voltage of 2.5 V on both lines. In the dominant state, a voltage of 3.5 V is present on CAN_H and a voltage of 1.5 V is present on CAN_L.

On the low-speed CAN, a voltage of 0 V is present on CAN_H in the recessive state, and 5 V on CAN_L. In the dominant state, voltages of 3.6 V and 1.4 V are present on the CAN_H and CAN_L respectively.

Reflection-free termination

Reflections of the electrical signals at open ends of lines would interfere with communication. To dampen these reflections, the bus lines are terminated at each end with a resistor of 120Ω .

Alternatively, the terminating resistors may be integrated into the electronic control units themselves.

Emission limits

To permit correct evaluation of bit transmission, the signal must arrive at each node within the bit time without interference, and yet still at the relevant sampling point. Delays arise from the signal transit time on the data bus. The maximum permissible bit rate thus depends on the total length of the bus. ISO 11898 specifies the bit rate for a defined circuit length. The following recommendations exist for longer lines.

- ▶ 1 MBit/s for 40 m (specified)
- ▶ 500 kBit/s up to 100 m (recommendation)
- ▶ 250 kBit/s up to 250 m
- ▶ 125 kBit/s up to 500 m
- ▶ 40 kBit/s up to 1,000 m

It is possible to connect at least 30 network nodes to the bus without the need for additional measures.

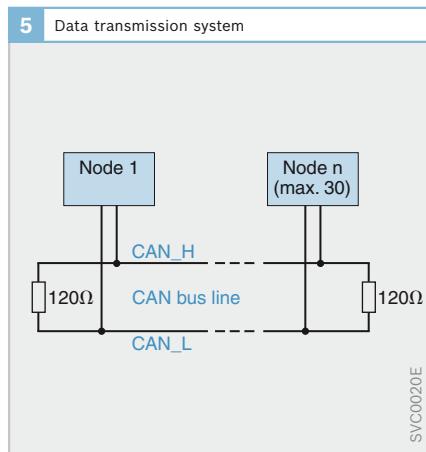
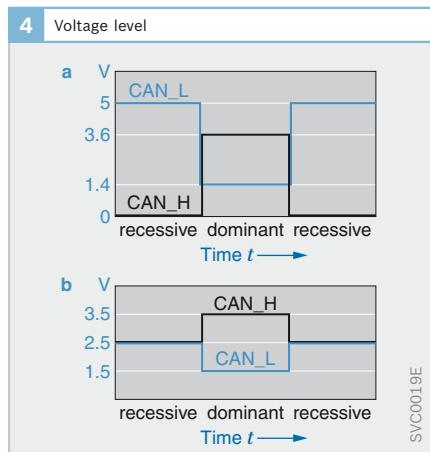


Fig. 4

- a Voltage level of the low-speed CAN (CAN-B)
- b Voltage level of the high-speed CAN (CAN-C)

CAN protocol

Protocol layers

For communications protocols, it is standard practice to arrange thematically coherent tasks as a set of layers, which affords a high level of flexibility in the implementation of a bus system. With the CAN, both the CAN hardware and the CAN software are subdivided into several layers (Fig. 6).

Application layer

The application layer represents the information in the form of data structures used by the application. These sets of data which are to be transmitted are forwarded to the object layer for this purpose.

Object layer

The task of the object layer is to manage the messages. The functions of this layer are used to decide which message should be sent at which time. For incoming messages, this layer is responsible for message filtering.

Transport layer

The transport layer furnishes the object layer with received messages and processes the messages prepared by the object layer for sending in such a way that the physical layer is able to transmit this

information. To fulfill this purpose, the transport layer is responsible for such functions as arbitration or fault detection and signaling.

Physical layer

The physical layer is the lowest level in the transport stack. It consists of the physical components of the network, such as the wiring and the voltages used to send the information.

Multimaster principle

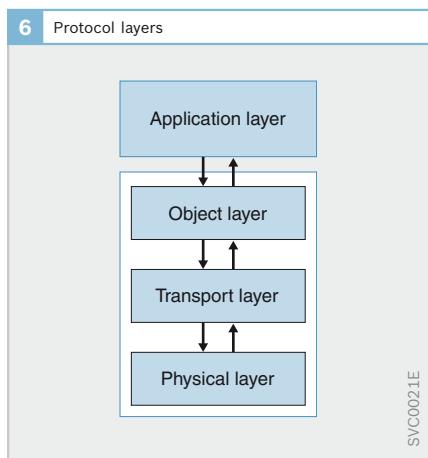
The CAN protocol supports communication between network nodes without the need for a central control unit. Each node may attempt to send messages at any time. Whether this attempt is successful or not essentially depends on two factors:

- ▶ Is the bus free before the start of transmission?
- ▶ Has the arbitration phase been passed successfully?

This design ensures that, even if any number of nodes were to fail, it can still be determined whether a node is authorized to send.

Content-based addressing

Unlike other networks, the CAN does not address the individual network nodes but the messages that have been sent. Each message has a unique marker, or identifier. The identifier classifies the content of the message (e.g. engine speed or power-window position). A station is therefore able to broadcast a message to all other stations (multicast or broadcast method). These stations read only those messages whose identifiers are stored in their acceptance list (message filtering, Fig. 7). In this way, each station decides for itself whether or not it needs a message sent on the bus. The identifier has 11 bits (standard format, CAN 2.0 A) or 29 bits (extended format, CAN 2.0 B). With 11 bits in the standard format, it is possible to distinguish between 2,048 different CAN messages;



in the extended format, this number rises to over 536 million.

The advantage of this addressing method is that the network nodes do not require any information about system configuration and are thus free to operate fully independently of each other. This results in a highly flexible complete system, which makes it easier to manage equipment variants. If one of the ECUs requires new information which is already on the bus, all it needs to do is call it up from the bus. It is possible to integrate additional stations into the system (provided they are receivers) without having to modify the existing stations.

Controlling bus access

Arbitration phase

If the bus is unoccupied (recessive state) and messages are available for sending, each station is free to initiate the sending of its message. The message begins with a dominant bit (start-of-frame bit), followed by the identifier. When several stations start to transmit simultaneously, the system responds by employing “wired-and” arbitration (arbiter = logical AND operator) to resolve the resulting conflicts over bus access. The message with the highest pri-

ority (lowest binary value of the identifier) is assigned first access, without any data loss or delay (non-destructive protocol).

The arbitration principle permits the dominant bits transmitted by a given station to overwrite the recessive bits of the other stations (Fig. 8). Each station outputs the identifier of its message onto the bus bit by bit, with the most significant bit first. During this arbitration phase, each station wishing to send data compares the level present on the bus with the level it actually possesses. Each station that attempts to send a recessive bit but encounters a dominant bit loses the arbitration process. The station with the lowest identifier, i.e. the highest priority, makes its way onto the bus without having to repeat the message (non-destructive access control). The transmitters of lower-priority messages automatically become recipients of the message just sent by another station. They repeat their attempt to send as soon as the bus is free again.

Without this access control, bus collisions would result in faults. To guarantee unequivocal bus arbitration, therefore, it is not permissible for more than one node to send a message with the same identifier.

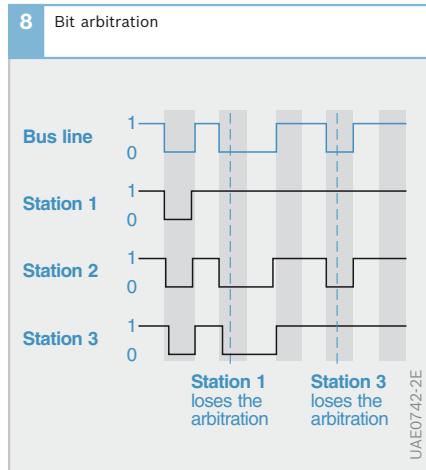
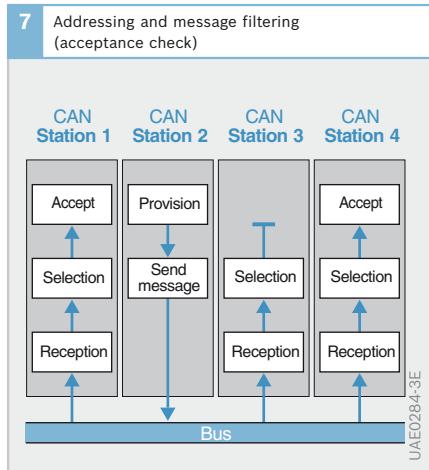


Fig. 7
Station 2 transmits
Stations 1 and 4 accept
the data

Fig. 8
Station 2 gains first
access (signal on the
bus = signal from
station 2)

0 Dominant level
1 Recessive level

With this access method, maximum-priority messages only have to wait for the transfer of the message currently being sent and, with bit times of 130 (CAN 2.0 A) or 150 (CAN 2.0 B), they have the lowest latency. At a data transfer rate of 500 kBit/s, this equates to 260 µs or 300 µs. The higher the load on the bus, the greater the data transfer's temporal offset becomes for messages of lower priority, and thus the greater the uncertainty over when a message to be sent will arrive at the recipient.

In order that all messages have a chance of accessing the bus, the data transfer rate must be matched to the number of bus subscribers.

Priority assignments

The direct consequence of the arbitration process is that the identifier also has the role of prioritizing the frame during transmission in addition to identifying the frame content. An identifier corresponding to a low binary number has high priority and vice versa. Message priorities are derived from the speed at which the message content changes or from the importance of the message to safety considerations, for example. It must not be possible for messages to have the same priority.

Message format

The message transfer on the CAN bus is based on four different frame formats:

► *Data frame:*

The transmitted message contains data (e.g. current engine speed) that is provided by the transmitting station (data source).

► *Remote frame:*

Stations can request the data they need from the data source (example: the windshield wiper requests how wet the windshield is from the rain sensor). The data source responds by sending the relevant data frame.

► *Error frame:*

If a station detects a fault or error, it communicates this to the other stations using an error frame.

► *Overload frame:*

This can be used to create a delay between a preceding and subsequent data frame or remote frame. The transmitting node reports that it cannot currently process another frame.

For data transfer on the data bus, a message frame is created. It contains the information to be transmitted arranged in a defined sequence. CAN supports two different formats of frame, which are specified in CAN 2.0 A and CAN 2.0 B. The most important difference between these two frame formats is the length of the identifier. A CAN 2.0 A frame has an 11-bit identifier, while a CAN 2.0 B frame has a 29-bit identifier divided into two parts (11 bit and 18 bit).

Both formats are compatible with each other and can be used together in a network. The frames compliant with CAN 2.0 A and B are shown in Figure 9. They have a maximum length of 130 bits (standard format) or 150 bits (extended format).

9 CAN message format

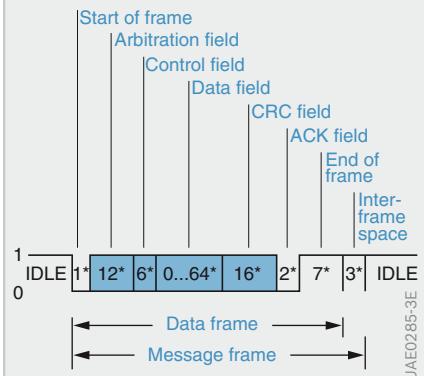


Fig. 9

- 0 Dominant level
- 1 Recessive level

In CAN 2.0 A and CAN 2.0 B data frames, the sequence of transferred information is arranged in the same way. The start-of-frame bit is followed by the arbitration, control, data and CRC fields. A frame is completed by the ACK field and the end-of-frame.

Start of frame

The bus is recessive in idle state. The start of frame, represented by a dominant bit, indicates the start of a transmission and, together with the process of bit stuffing, serves to synchronize all stations.

Arbitration field

With a CAN 2.0 A frame, the arbitration field comprises the 11-bit identifier and a control bit, the RTR bit (Remote Transmission Request).

With a CAN 2.0 B frame, the arbitration field comprises the 11-bit identifier, followed by the SRR bit (Substitute Remote Request) and the IDE (Identifier Extension Bit). Both bits are sent recessively and thereby ensure that a CAN 2.0 A frame always takes priority over a CAN 2.0 B frame if it has the same 11-bit identifier. The second 18-bit identifier follows. The RTR bit completes the arbitration field.

The RTR bit indicates whether the transmitted frame is a data or remote frame. The RTR bit is dominant in the data frame and recessive in the remote frame. If station A, for example, happens to send a message by data frame and station B requests this message by remote frame at the same time, the arbitration conflict cannot be resolved in this situation by means of the message identifier: the RTR bit is the decisive factor for access authorization. First of all, station A wins arbitration with the sending of the dominant RTR bit and continues to transmit the message. Station B, which has requested precisely this message, prepares to receive and is able to read the other data sent by station A.

Control field

In a CAN 2.0 A frame, the control field comprises the IDE bit (Identifier Extension Bit), which is always sent as dominant here, followed by a reserved bit for future extensions, which is sent recessively. The remaining four bits in this field define the number of data bytes in the next data field. This enables the receiver to determine whether all data has been received.

The structure of a CAN 2.0 B frame is practically identical. However, since the IDE bit already belongs to the arbitration field, a further reserved bit for future extensions takes its place and this reserved bit is sent recessively.

Data field

The data field contains the actual message information comprised of between 0 and 8 bytes. A data field in which the length of the data is expressed in 0 results in the shortest possible data frame with a length of 44 or 64 bits. A frame like this can be used to synchronize distributed processes. A number of signals can be transmitted in a single message (e.g. engine temperature and engine speed).

CRC field

The CRC field (Cyclic Redundancy Checksum) contains a 15-bit checksum (frame check word) across the preceding frame from the start bit to the final bit of the data field. The 16th bit (CRC delimiter) in this field is recessive and closes the checksum. The checksum is a means of detecting possible transmission interference.

ACK field

Unlike all preceding fields, the ACK field (acknowledgment) is not set by the sender of the frame but by a different node that is able to acknowledge receipt of the frame directly after the data field. This field comprises the ACK slot and the recessive ACK delimiter. The ACK slot is also transmitted recessively by the sender and over-

written as dominant by a receiver upon the message being correctly received. Here, it is irrelevant whether the message is of any significance for the particular receiver in the sense of an acceptance check (message filtering). Only correct receipt is confirmed. This signals to the sender that no malfunction occurred during data transfer.

End of frame

The end of frame marks the end of the message and comprises seven recessive bits. With seven bits of the same value in succession, the stuffing rule is deliberately broken (see Error detection).

Interframe space

The interframe space comprises a succession of three recessive bits which serve to separate successive messages. After a total of 10 of these recessive bits, the stations are permitted to begin transmitting within the network again. Until then, the bus remains in idle state.

Only data and remote frames need to include the interframe space. Error and overload frames can be sent immediately after the last frame. This enables immediate signaling of errors and problems.

Transmitter initiative

The transmitter will usually initiate a data transmission by sending a data frame. It is also possible for a receiving station to call in data from a sending station by depositing a remote frame. The network node that is able to deliver the information requested makes the information available in response to this query.

Error detection

The bus line may suffer interference, e.g. electromagnetic interference. To eliminate the risk of faulty behavior, the transmitted data must be checked for correctness. A number of control mechanisms for detecting errors are integrated in the CAN protocol.

Cyclic redundancy check

For every transmitted message, the transmitter calculates a check sequence from the start-of-frame, arbitration, control and data field. A 15-bit checksum is derived from the bit sequence by means of a generator polynomial. The checksum is then used to detect errors in the data transfer.

With CRC generation, a defined generator polynomial is used to carry out polynomial division across a given frame range. The remainder forms the checksum.

Once the CRC field has been received, the receiver is able to check that the frame has been transferred correctly by carrying out its own polynomial division of the received frame range using the generator polynomial and checking whether the received checksum matches the calculated remainder.

Frame check

The frame check involves all bus subscribers, senders as well as receivers, checking the sent/received data frame for compliance with the predefined frame structure. The CAN protocol contains a number of fixed-format fields (start and end-of-frame, delimiter) which are checked by all stations.

ACK check

With the ACK check, a receiver acknowledges that the frame has been received correctly by sending a dominant bit in the ACK slot. The sender of the frame can then verify whether a message was transferred correctly. An absence of this bit indicates that a transmission error has been detected.

Monitoring

The sender of a frame continuously monitors the bus level. It is able to detect a transmission error by comparing the sent bit and the sampled bit.

Bit stuffing

Compliance with bit stuffing is checked by means of the code check. The stuffing convention stipulates that in every data frame or remote frame a maximum of five successive bits of equal state may be sent between the start of frame and the end of the CRC field. As soon as five identical bit states have been transmitted in succession, the sender includes a bit with the opposite state. The receiving station clears all of these inserted bits after the message has been received (destuffing).

This measure permits detection of line faults, e.g. short-circuit or burst interference. Rare signal changes would affect the possibility for synchronization in the nodes.

Error handling

If a CAN controller detects a fault or format error, it interrupts the current transmission by sending an error frame comprising six successive dominant bits. This breaks the stuffing rule that prohibits this type of bit sequence. If the sender detects that its message has been interrupted by an error frame, it stops transmitting and makes another attempt at a later time. This effect prevents other stations from accepting the erroneous message and thereby ensures consistency of data across the entire system.

Fault confinement in the event of failures

Defective stations could weigh heavily on bus traffic if they were to frequently send erroneous messages or interrupt the transmission of correct messages by repeatedly sending an error frame. The CAN protocol localizes station failures by means of statistical error analysis. A station recognizes the probability of its own malfunction by how often it aborts messages before other stations send an error frame. The protocol's first measure in this case is to prevent a station such as this from continuing to abort transmissions. In an emergency, the station shuts down automatically.

Hardware

CAN controller of a subscriber

The CAN controller of a subscriber generates a frame from the data to be sent containing all the fields required for compliance with the CAN protocol. It then converts the frame into a bit stream.

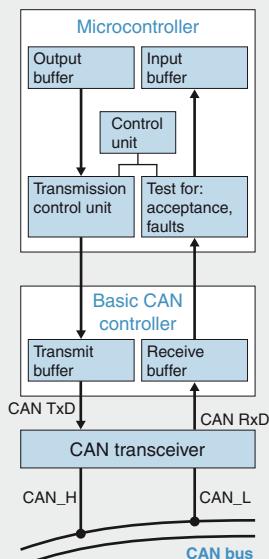
CAN only specifies the physical, transportation and object layers. The interfaces that are provided in the application layer can be arranged differently to suit the field of application. In order to provide proper CPU support (electronic control unit microcontroller for application software) for a wide range of different requirements, semiconductor manufacturers have brought onto the market products that provide a variety of capabilities. They differ neither in the frame format they produce, nor in their error-handling methods, but solely in the type of CPU support required for message administration.

If messages on the bus are managed in extended format (29-bit identifier), the CAN controllers must be compatible with the CAN 2.0B specification. There are controllers that only support CAN 2.0A and send and receive messages in the standard format, but that generate errors with messages in the extended format. Other controllers are able to tolerate the extended format without generating an error. These modules can be used in a CAN together with controllers that administer the sending and receiving of messages in the extended format.

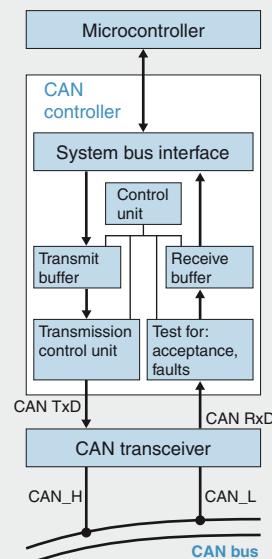
Basic CAN

In modules with basic-CAN implementation, only the basic function of the CAN protocol for the generation of the bit stream is implemented in the hardware. For the management of the messages to be sent and received, there is an intermediate

10 Basic-CAN module



11 Full-CAN module



buffer to which the local computer (application software microcontroller) has access (Fig. 10). Since the buffer capacity is limited, the computer must read the received data before new messages are received. Message filtering also takes place in this computer. A part of the computer's capacity is therefore used for CAN management. Since the computers do not usually have sufficient processing capacity, modules with basic CAN are primarily suitable for low bit rates, or for the transmission of fewer messages but at higher bit rates.

The advantage of these modules, in comparison to modules with full CAN, is the smaller chip surface and the lower manufacturing costs.

Full CAN

Full CAN implementation is the protocol of preference in cases where a station has to manage several messages at high bit rates and the local computer has no free capacity for communication tasks. They contain several "communication objects", each of which contains the identifier and the data of a particular message. During the initialization of the CAN module by the local computer, it is decided which messages the CAN controller should send and which received messages it should process further. Received messages are only accepted (message filtering) if the identifier matches one of the communication objects.

CAN controllers with full-CAN implementation relieve the burden on the local computer by performing all of the communication including message filtering in the controller (Fig. 11).

The CAN controller can be coupled to the microcontroller in the electronic control unit as a stand-alone module by the address/data bus. Powerful microcontrollers have the CAN controller integrated on-chip. This type of bus coupling is the more cost-effective and thus the more common solution.

Modules without local computer

A further category of CAN module is one that is supplied without a local computer. These SLIOs (Serial Linked Input/Output) are able to input and output data via ports. They are therefore suitable for making sensors and actuators bus-compatible at low cost, but they do need a master that controls them.

Transceiver

The bit stream generated by the CAN controller is made up of binary signals. They do not yet correspond to the required voltage levels of the CAN bus. The CAN-bus interface module, or transceiver, generates the differential signals CAN_H and CAN_L and the reference voltage U_{ref} from the binary data stream.

Sleep mode

The CAN comfort bus must remain ready for operation even with the ignition switched off so that functions such as the radio, power windows or parking lamp may continue to operate. The bus subscribers must therefore be supplied by terminal 30 (permanent positive). After terminal 15 has been switched off (ignition off), a CAN node may enter sleep mode (standby) to relieve the vehicle electrical system of as much load as possible. The transmitter part of the transceiver module is switched off in this condition to minimize the power consumption in this mode of operation. However, the receiver part remains active and checks whether messages are being sent on the bus. In this way, the CAN controller, which also enters standby mode, is able to react to a wake-up message and fully activate the CAN node.

Data transfer sequence

Using the transfer of engine speed as an example, it shall be demonstrated how data is transferred on a CAN with full-CAN module.

Transmission

The engine management's application software calculates the engine speed from the signal from the engine-speed sensor. This value is calculated once for every combustion cycle. The measured value enters both the receive buffer and the transmit buffer of the microcontroller (Fig. 12a).

The engine-management microcontroller is coupled to the CAN controller by a parallel interface. The contents of the transmit buffer are cyclically passed (e.g. every 10 ms) to the transmit buffer of the CAN controller. A flag notifies the CAN controller that a message is ready to be sent. With this transmit instruction, the engine management has completed its part of the task.

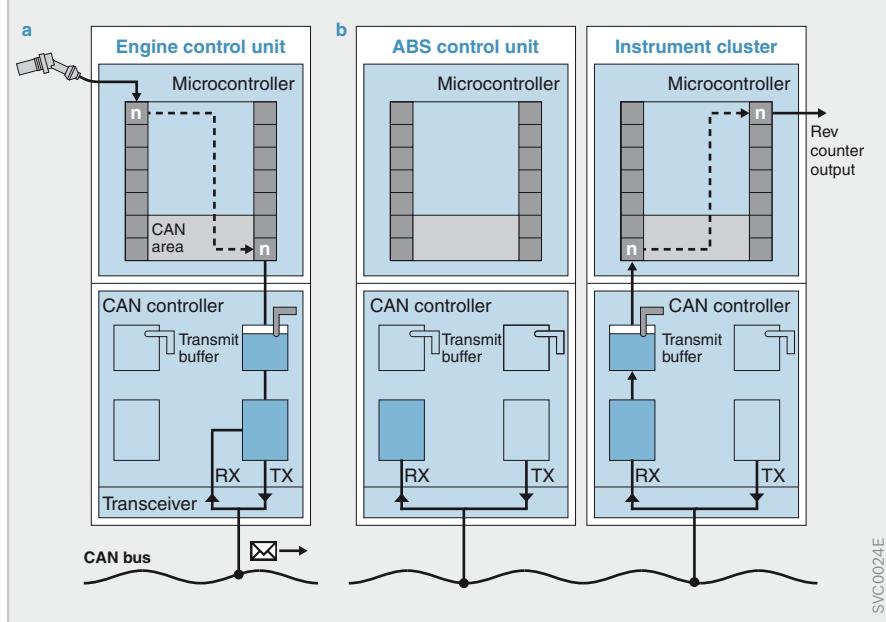
Once the engine-speed information has been stored in the transmit buffer, the CAN controller begins to compile the CAN frame. Over the RxD line, the CAN controller now monitors the bus. If the bus is free, it immediately starts to generate the bit stream and directs it to the transceiver on the TxD line; if the bus is occupied, it waits. From the bit stream, the transceiver creates a signal with the requisite voltage level.

Reception

The message originally sent by the engine management and forwarded on by the transceiver is received by all stations connected to the bus (Fig. 12b). The message reaches the CAN controller on the RxD lines in the form of a bit stream.

At the first stage (monitoring layer), the CAN controller checks the incoming message for errors using the CRC checksum. If the message is free of errors, each

12 Data transfer sequence



station responds with an acknowledge (ACK check).

At the second stage, or acceptance layer, the message undergoes message filtering. Each station checks whether the received identifier is addressed to the particular station and whether the message is required in the application software. If not, the message is rejected. Otherwise, it makes its way to the receive buffer. A flag notifies the application software that a new message is ready for processing.

The instrument cluster, for example, calls up the available message, processes the engine-speed information and calculates triggering signals for the actuator of the rev counter.

Standardization

The International Organization for Standardization (ISO) and SAE (Society of Automotive Engineers) have issued CAN standards for data exchange in automotive applications:

- ▶ For low-speed applications up to 125 kBit/s: ISO 11 519-2 and 11 898-3
- ▶ For high-speed applications faster than 125 kBit/s: ISO 11 898-2 and SAEJ 22 584 (passenger cars) or SAEJ 19 39 (commercial vehicles)
- ▶ An ISO Standard for diagnosis via CAN has also been published as ISO 15 765

Standardization makes it possible for components of different manufacturers to function together. No adaptations are required.

CAN is also widely used in industrial automation. These applications are supported by an alliance of companies in the “CAN in Automation” users group (CiA).

Bosch has concluded contracts with its licensees that guarantee that any CAN implementations will be able to communicate with each other. Users will be able to rely on the interaction of any CAN modules.

Characteristics

- ▶ Standardized in accordance with ISO 11898
- ▶ Prioritized communication
- ▶ Data transfer rates: up to 1 MBit/s
- ▶ Data capacity: up to 8 bytes per message
- ▶ Real-time response: the data protocol is sufficient for the real-time requirements in the motor vehicle
- ▶ Non-destructive bus-access method
- ▶ Low power consumption
- ▶ Flexibility of configuration
- ▶ Simple and economical design with twisted line pairs
- ▶ Very high reliability of data transfer
- ▶ Fault detection and signaling
- ▶ Localizing of failed stations
- ▶ Handling of intermittent and permanent faults
- ▶ Short-circuit resistance
- ▶ The number of nodes is theoretically unlimited. However, a limit arises in practice from the capacitive load of the bus and the increasing latencies of messages when a high number need to be sent.

LIN bus

Overview

The increasing use of mechatronic systems in the motor vehicle gave rise to the idea of designing a cost-effective bus system as an alternative to the low-speed CAN. In 1998, several automotive manufacturers founded a consortium with the aim of developing a specification for a serial bus for the networking of sensors and actuators in the body electronics area.

It was believed that a bus system with simple bus protocol and a simple sequence control would make it possible to use even low-capability microcontrollers without additional hardware for the communication interface.

The workgroup's resulting LIN bus specification was introduced into mass production with the Mercedes-Benz SL as early as 2001.

The name, LIN (Local Interconnect Network), is derived from the fact that all electronic control units are located within a demarcated installation space (e.g. in the door). The LIN, therefore, is a local subsystem for supporting the vehicle network by means of superordinate CAN networks.

The LIN bus is suitable for low data rates of up to 20 kBit/s and is typically limited to a maximum of 16 bus subscribers.

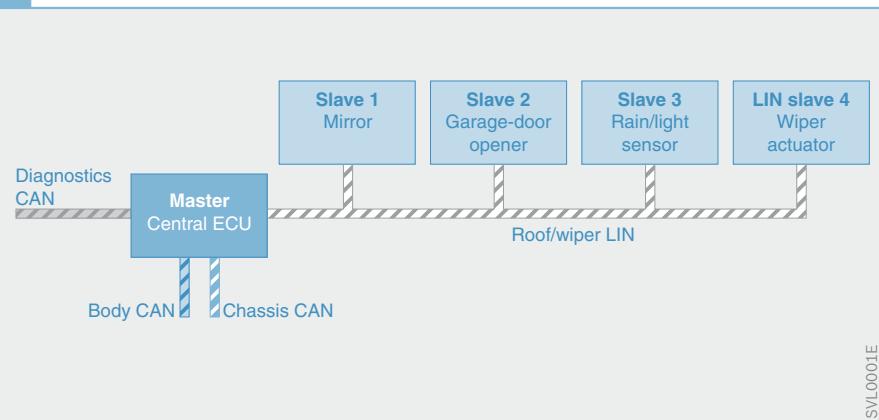
The electrical interface can be created easily and cost-effectively in the network nodes. As far as the nodes are concerned, a distinction is made between the master, which is generally an electronic control unit connected to a superordinate bus system, and the slaves. These are intelligent actuators, intelligent sensors or, quite simply, just switches with additional hardware for the LIN-bus interface.

The bus subscribers are usually arranged in a linear bus topology and connected to each other by a single-wire line. This topology, however, is not explicitly specified.

Communication on the LIN bus takes place in a time-synchronous manner, whereby the master defines the time grid. Consequently, there arises a strictly deterministic LIN bus response.

Figure 1 shows an example of a LIN network as a subbus in the roof/wiper area of the motor vehicle. Here, the bus comprises a central electronic control unit, as the master, and the four slaves: mirror, garage-door opener, rain/light sensor and wiper actuator. The master also functions as a gateway to the Chassis CAN, the Body CAN and the Diagnostics CAN.

1 LIN bus with master and slave nodes



Applications

The LIN bus as a means of networking mechatronic systems can be used for many applications in the motor vehicle for which the bit rates and variability of the CAN bus are not essential. Examples of LIN applications:

- ▶ Door module with door lock, power-window drive and door-mirror adjustment
- ▶ Control of the power-sunroof drive unit
- ▶ Control of the wiper motor for the windshield wiper
- ▶ Sensor for rain and light detection
- ▶ Air-conditioning system (transmission of signals from the control element, activation of the fresh-air blower)
- ▶ Headlight electronics
- ▶ Control of motors for seat adjustment
- ▶ Theft deterrence
- ▶ Garage-door opener

Data transmission system

The LIN bus is designed as an unshielded single-wire line. The bus may adopt two logical states:

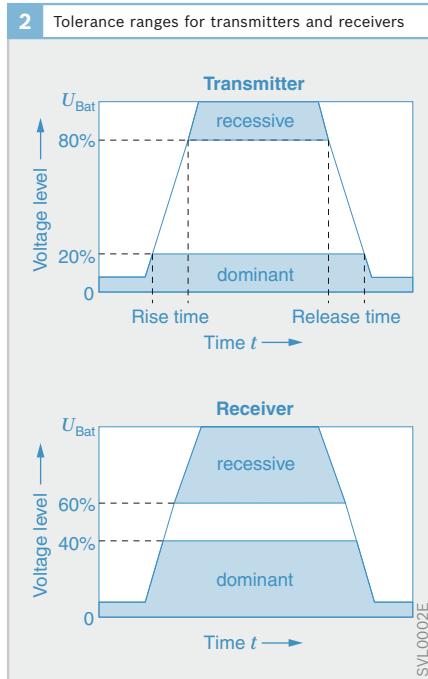
- ▶ The dominant level corresponds to the electrical voltage of approx. 0 V and represents logical 0
- ▶ The recessive level corresponds to battery voltage U_{Bat} and represents the logical 1 state

The recessive level is characterized by a resistance to battery voltage of $1k\Omega$ in the master and $30k\Omega$ in the slaves.

Due to circuitry variations, there may be differences in the voltage levels. A stable data transfer is ensured by the defining of tolerances for sending and receiving in the areas of the recessive and dominant levels. The tolerance zones on the reception side are wider (Fig. 2) to make it possible for valid signals to be received despite interference radiation.

The data rate of the LIN bus is limited to a maximum of 20 kBit/s. This is the outcome of a compromise between the demand for high edge steepness for easy synchronization of the slaves on the one hand, and the demand for lower edge steepness for improved EMC characteristics on the other. The standard bit rates of 2,400 Bit/s, 9,600 Bit/s and 19,200 Bit/s are recommended. The minimum permissible value for the bit rate is 1 kBit/s to prevent timeout conflicts. The edge steepness itself is defined in the LIN specification as 1 to 3 V/ μ s.

The maximum number of nodes is not specified in the LIN specification. It is theoretically limited by the number of available identifiers. In practice, the number of subscribers is restricted to 16 due to the maximum permissible total capacity of the bus system.



Bus access

Access to the LIN bus is determined by the master-slave principle. Each message is initiated by the master. The slave has the possibility to respond. The messages are exchanged between the master and one, several or all of the slaves (point-to-point, multicast, broadcast).

The following relationships are possible in the communication between master and slave:

- ▶ Message with slave response: the master sends a message to one or more slaves and asks for data (e.g. switch states or measured values)
- ▶ Message with master instruction: the master gives a slave a control instruction (e.g. switch on a servomotor)
- ▶ The master initiates communication between two slaves

No arbitration or conflict management is required because, with the master-slave access control, the master alone controls access to the data line.

LIN protocol

Frame

The information transferred on the LIN bus is embedded in a defined frame (Fig. 3). A message initiated by the master always begins with a header. The response (message field) contains different information depending on the type of message. If the master wishes to transmit control instructions for a slave, it will populate the response with data to be used by the slave. If the master is transmitting a data request, the slave that it is addressing will populate the response with the data requested by the master.

Header

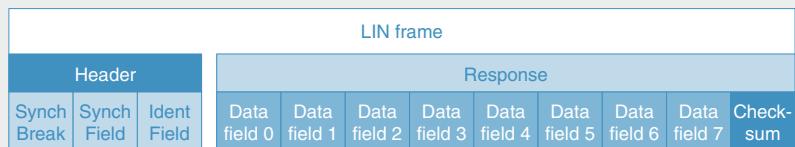
The head is composed of the following parts:

- ▶ The synchronization break
- ▶ The synchronization field
- ▶ The identifier field

Synchronization

To guarantee consistent data transfer between master and slaves, a synchronization takes place at the beginning of each frame. First of all, the beginning of a frame is unambiguously marked by the synchronization break (SynchBreak).

3 LIN message format



The SynchBreak comprises at least 13 consecutive dominant levels and one recessive level.

At the end of the synchronization break, the master sends the synchronization field (SynchField) consisting of the bit sequence 01010101. The slaves are then able to adjust themselves to the time basis of the master and thus synchronize.

The synchronization method described permits a loose specification of the timing of the bus subscribers. The clocking of the master should not deviate more than $\pm 0.5\%$ from the nominal value. The clocking of the slaves is permitted to deviate by up to 15 % before synchronization as long as the synchronization reduces the deviation to a maximum of 2 % before the end of the message.

In this way, the slaves can be built without an expensive quartz oscillator, e.g. using a simple RC circuit.

Identifier

The third byte in the header is used as the LIN identifier. As with the CAN bus, a content-based addressing method is used – the identifier therefore provides information about the content of a message (e.g. engine speed). Based on this information, all nodes connected to the bus decide whether they would like to receive and process the message further or simply ignore it. This process is known as acceptance filtering.

Six of the eight bits of the identifier field determine the identifier itself. Their permutations give rise to a possible 64 different identifiers (ID). They have the following meanings:

- ▶ ID = 0 to 59: transmission of signals.
- ID = 60: master request for the commands and diagnosis
- ▶ ID = 61: slave response to ID 60
- ▶ ID = 62: reserved for manufacturer-specific communication
- ▶ ID = 63: reserved for future extensions to the protocol

Of the 64 possible messages, 32 may only contain two data bytes, 16 may only contain four data bytes, and the remaining 16 eight data bytes each.

The last bits in the identifier field contain two checksums, which are used to check the identifier for transmission errors and any resulting incorrect message assignments.

Data field

Once the header sent by the master node has been transmitted, it is time for the transfer of the actual data to begin. The slaves know from the transmitted identifier whether or not they are being addressed and, if they are, they reply with their response in the data field.

Several signals can be packed into one frame. In this case, each signal has precisely one generator, i.e. it is always written by the same node of the network.

During the data transfer of the bytes, it is always the least significant bit (LSB) that is output first. Each byte (8 bits) is preceded by a start bit and followed by a stop bit, which means that each byte involves the transmission of ten bits. The purpose of the start and stop bits is to resynchronize the nodes and thereby prevent transmission errors.

The data response of the slaves is verified by means of a checksum.

LIN description file

The configuration of the LIN bus, in other words the specification of network subscribers, signals and frames is managed in the LIN description file, or ldf. For this purpose, the LIN specification provides for an appropriate configuration language.

From the ldf, a set of C codes or header files is automatically generated using suitable tools. These are used as the basis for implementing the master and slave functions in the electronic control units connected to the bus.

The ldf is therefore a means of configuring the entire LIN network. It represents a common interface between the vehicle manufacturer and the supplier of the master or slave modules.

Message scheduling

The scheduling table in the ldf defines the sequence and time grid in which messages are sent. Frequently needed information is sent more often. Once the table has been worked through, the master begins with the first message again. The sequence in which the table is worked through may change depending on the operating status (e.g. diagnostics active/inactive, ignition on/off).

The transfer grid of each message is therefore known. This deterministic behavior is guaranteed by the fact that, with the master-slave access control principle, all messages are initiated by the master.

Network management

The nodes of an LIN network can be forced into sleep mode to minimize the no-load current of the entire electronics and electrical system in the vehicle. Sleep mode can be achieved in two ways:

- ▶ The master sends the go-to-sleep command with the reserved identifier 60
- ▶ The slaves enter sleep mode of their own accord if no data transfer has taken place on the bus for a relatively long period (4 seconds)

Both the master and the slaves are able to wake up the network. To do so, they must send the wake-up signal. This comprises data byte 128. After a break of 4 to 64 bit times (wake-up delimiter), all nodes must have been initialized and be ready to respond to the master.

Example: air-conditioning control

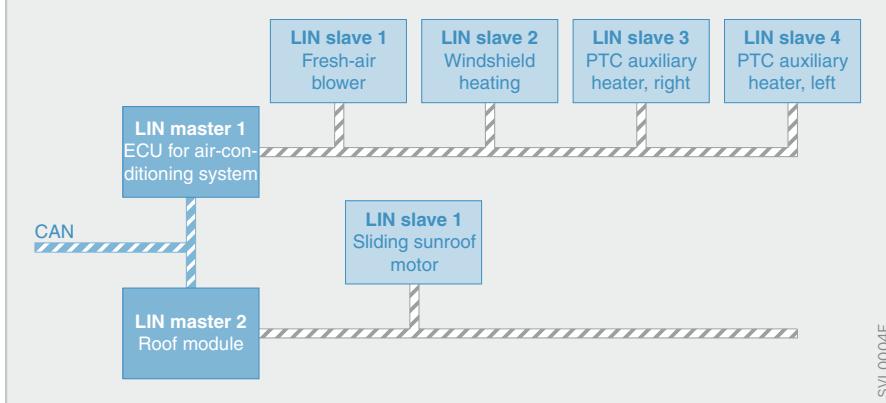
The LIN bus is typically used in the control of the air-conditioning system. The operating and display unit functions as the bus master. This is where the software for the open and control algorithms are stored.

One of its tasks is to adjust the speed of the fresh-air blower. The controlled variable for this is the actual temperature in the passenger cell and the desired temperature set by the driver. The electronic control unit receives the interior temperature from a temperature sensor located in an appropriate place in the interior. The desired temperature is set at the control unit, e.g. using a sensor ring.

From the input variables, the electronic control unit calculates that the blower speed needs to be increased to a value of 200 rpm, for example. The electronic control unit transmits a message containing the master instruction to the LIN bus at a defined time interval. The identifier in this example would be "set the blower speed". In this subbus, this instruction corresponds to identifier 25, for example. Following the header containing this identifier, the master transmits a numerical value in the data field that equates to the physical value of 200 rpm. Each slave possesses a list of identifiers that are relevant for this node. The fresh-air blower is the one and only slave that responds to the "set blower speed" identifier and executes the master's request.

At speeds below idle, the fresh-air blower must be switched off. At this low speed, the heavy load could cause the engine to stall. The LIN bus is coupled to the CAN bus via the gateway and thus receives the current engine speed on a continuous basis. If the speed falls below the specified engine-speed threshold, the LIN master sends the message containing the "set blower speed" identifier where the data field contains the value 0. The fresh-air blower switches off in response.

4 LIN networking (example)



SVL0004E

Intelligent actuators can send up-to-date operating-status information to the actuating unit. The fresh-air blower records the speed by means of a sensor and sends it back on the LIN bus as a numerical value. Due to the master-slave access method, the value is only contained in a message with slave response initiated by the master. The feedback signal with the current engine speed makes it possible to control, and therefore accurately maintain, the desired value.

Summary

The essential characteristics of the LIN bus are:

- ▶ Single-master/multiple-slave concept
- ▶ Master/slave access control
- ▶ Independent synchronization of the slave possible even without quartz
- ▶ Deterministic signal transmission
- ▶ Communication in the form of very short messages
- ▶ Character-based transmission (UART)
- ▶ Bit rate max. 20 kBit/s
- ▶ Data transfer over an unshielded single-wire line
- ▶ The reference potentials of the data line are battery voltage and ground
- ▶ Maximum bus length 40 m
- ▶ Maximum number of nodes 16, typically fewer than 12

MOST bus

Introduction

The MOST bus (Media Oriented Systems Transport) was specifically developed for the networking of infotainment systems in motor vehicles. In addition to traditional entertainment functions, such as radio tuners or CD players, these infotainment systems (which are becoming an increasingly common feature of modern vehicles, and premium class vehicles in particular) also offer video functions (DVD and TV), route guidance capabilities and access to mobile communications and information.

Requirements

The functions of an infotainment system place high demands on a bus system. The transmission of multimedia data, both audio and video, requires a high data rate and a synchronization of the data transfer between source and sink as well as between sinks themselves.

For the transmission of informational data, e.g. detailed information about the music tracks on an mp3 player, and the sending of software updates, it is necessary for the data transfer to be flexible with varying and, at certain times, equally very fast data rates.

Likewise, the requirements for use in a motor vehicle, and the associated requirements of electromagnetic compatibility (EMC), must also be fulfilled.

MOST Cooperation

It was with these requirements in mind that the MOST bus was developed by the MOST Cooperation. This was founded in 1998 by BMW, DaimlerChrysler, Harman/Becker and Oasis SiliconSystems (now part of SMSC). By 2006, the MOST Cooperation had 16 automotive manufacturers and 67 suppliers and tool makers among its membership.

The MOST Cooperation creates and administers the specifications on which the MOST bus is based. Furthermore, it defines the requirements for implementation of MOST devices and provides appropriate compliance tests through accredited test houses.

Type of use

At present, the MOST bus is almost exclusively used for the networking of infotainment systems in motor vehicles. It can currently be found in over 35 vehicle models of various manufacturers, especially in the premium and mid-range classes (as at 2006).

Features of the MOST bus

The MOST bus supports the logical networking of up to 64 devices, at which point constraints associated with the chipsets may be encountered.

In its current version, the MOST bus offers a data rate of 24.8 MBit/s (MOST 25). Versions with higher data rates of 50 MBit/s (MOST 50) and 150 MBit/s (MOST 150) are already available as a development model.

Data transmission channels

For the transmission of data, the MOST bus supports the following channels and their different attributes:

- ▶ The control channel is used for the simple transmission of control commands, for the signaling of device statuses and for the sending of messages necessary to system management. With MOST 25, the control channel has a gross bandwidth of 705.6 kBit/s.
- ▶ For the transmission of multimedia data, the MOST bus has a flexible number of synchronous channels that are able to carry both audio and video data. In MOST 25, there are a maximum of 15 audio channels available in stereo quality.

- On the asynchronous channel, data is sent in packets. It is therefore suitable for transferring information that does not require a fixed data rate but does need a high data rate at certain times. Examples could be the transfer of track information of an mp3 player or a software update. With MOST 25, the asynchronous channel has a gross bandwidth of up to 12.7 MBit/s.

The available bandwidth can be flexibly distributed between the asynchronous channel and the synchronous channels by means of the “boundary descriptor”, which can also be shifted to the transit period if relevant preconditions are met.

Topology

A MOST system is arranged in a ring structure whereby a device is connected to its predecessor or successor in the ring by an input or output respectively (Fig. 1). One of the devices acts as the “timing master” and generates the data frames for data transfer with which the other devices synchronize.

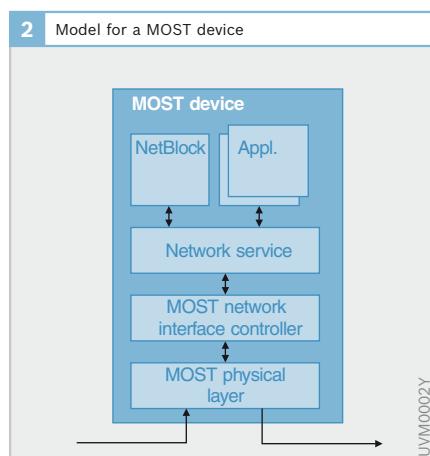
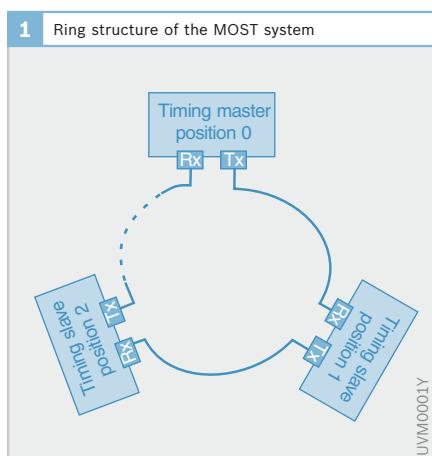
Each device has a bypass. When this is closed, the device forwards the signal directly and is therefore invisible to the MOST system. This service is useful, for example, when the system is starting and a device requires more time for initialization, or in the event of a temperature-dependent shutoff.

There are also other possible structures for a MOST system, e.g. a star topology. Internally, however, these must always be arranged to form a (logical) ring. With a star structure, this can be achieved through use of a central hub, for example.

Device model

The MOST standard defines the model shown in Figure 2 for a MOST device for which the following elements are required:

At the lowermost level, the physical layer provides access to the transmission agent. Departing from the original definition that provided for optical transmission by POF cable (Plastic Optical Fiber), there are today various physical layers available with optical and electrical transmission as well as various speeds of transfer.



The MOST network interface controller (NIC) is a hardware controller that is responsible for controlling the physical layer and implements the basic transfer services. While in the original version of the NIC many control tasks had to be carried out by the main processor of the device, the current version (Intelligent Network Interface Controller, INIC) is, for the most part, able to handle tasks autonomously and already implements substantial parts of the network service. Typical controllers for the MOST are the OS8104 (NIC) and the OS81050 (INIC), manufactured by SMSC.

The MOST network service is the driver layer through which applications and system services have access to the NIC. The lower layer of the network service (layer 1) offers basic communication and management functions, while the layer above (layer 2) offers functions for supporting the development of applications, such as a translator for the application protocol or support for the notification service.

The applications of the device are implemented on top of the network service. In a MOST system, the interface of an application is realized as a function block (FBlock). Each device must at least implement a special FBlock, the NetBlock, which is required for management functions within the MOST system. In addition, a device will usually implement one or more applications that can be used as FBlocks of applications on other devices. A device of a vehicle infotainment system is often integrated with several functions, e.g. those of a radio tuner and those of an amplifier, which can be represented in the MOST system as independent FBlocks.

Transmission agent

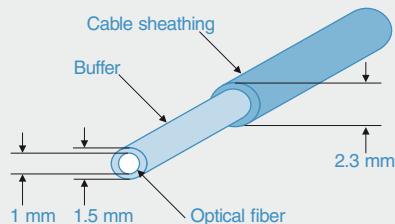
The MOST bus was originally defined as an optical transmission agent that used a plastic optical fiber (POF). The optical signal used here with a wavelength of 650 nm (in the red wavelength range) is generated by an LED on the transmit side, which is designated Tx-FOT (FOT = Fiber Optic Transceiver). On the receive side, Rx-FOT, the optical signal is converted back into an electrical signal by a PIN photodiode.

The POF cables used for automotive applications (Fig. 3) consist of a 980 µm thick optical core insulated by 20 µm thick optical cladding with a low refractive index. In total, the optical conductor therefore has a diameter of 1 mm. The optical fiber is insulated with a black buffer, which is in turn surrounded by a protective cable sheathing. This gives the cable a total diameter of 2.3 mm.

POF lines offer the following benefits when used in the motor vehicle:

- ▶ No electromagnetic interference radiation
- ▶ Insensitivity to interference irradiation
- ▶ Lower weight than equivalent shielded electrical lines
- ▶ More flexible routing in comparison to equivalent shielded electrical lines

3 POF cable



To date, modern vehicles are equipped exclusively with MOST 25, with POF data transmission via POF connections. In addition to this, however, further transmission techniques have been defined for the MOST:

- ▶ The optical transmission by glass fiber cable (PCS) with laser diodes (VCSEL), which offers a greater damping reserve, supports higher speeds and is less sensitive to high temperatures.
- ▶ An electrical transmission by copper cable, which is also less sensitive to temperature and is more economical by comparison, requires additional shielding measures at higher bandwidths. This has consequences for costs and cabling.

Data transfer

Data transfer on the MOST bus is organized into data frames, which are generated by the timing master with a fixed data rate and passed on by subsequent devices in the ring.

Data frames

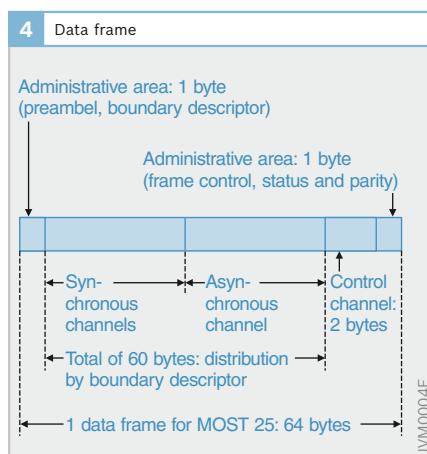
The timing master normally generates data frames with a cycle of 44.1 kHz, and in rarer cases 48 kHz. The cycle is defined by the system manufacturer (i.e. the vehicle manufacturer in conjunction with its device suppliers) to suit the predominant media formats in the system. The size of the data frames, by extension, determines the bus speed of a MOST system. With MOST 25, the size of a data frame is 512 bits.

A data frame is made up of the following ranges (Fig. 4):

- ▶ Areas for administrative information at the start and end of the data frame, the tasks of which include the synchronization of the data transfer, determination of the devices' ring positions and the transmission of the boundary descriptor

- ▶ An area for the control channel (with MOST 25, two bytes per data frame)
- ▶ An area for the synchronous channels, whereby one or more bytes are always assigned to a specific synchronous channel
- ▶ An area for the asynchronous channel

In the case of MOST 25, the synchronous and asynchronous area demand a 60-byte share of the data frame. The distribution between the synchronous channels and the asynchronous channel is determined by the value of the boundary descriptor with a resolution of 4 bytes. The synchronous area must have at least 24 bytes (6 stereo channels). This means that 24 to 60 bytes are permitted for the synchronous area and 0 to 36 bytes are permitted for the asynchronous area.



System events

The fundamental system events described below indicate the current status of the data transfer in the MOST system. These are monitored by the NIC and forwarded by the network service to the applications, which must then respond accordingly.

- ▶ Lock, stable lock: a lock occurs as soon as communication is established in the ring, i.e. the correct device receives the correct data frames. A stable lock is set when the system has been in lock status for a specific period.
- ▶ Unlock, critical unlock: an unlock denotes the loss of the lock, i.e. that no correct data frames are being received. If the system is in the unlock state for a certain period, the status switches to critical unlock.
- ▶ Network change event: if the number of devices in a MOST system changes because one device has opened or closed its bypass, this is signaled by a network change event.

Addressing

Devices on the MOST bus are addressed by means of a 16-bit address. The following different addressing types are supported:

- ▶ Logical addresses: each device has a logical address that is unique in the system and is set and distributed when the system starts. The logical addresses can be set dynamically based on the ring position, for which, beginning with the timing master, the hexadecimal range 0x100 to 0x13F is reserved. Or they can be defined statically by the system manufacturer in a range reserved specifically for this purpose.
- ▶ Physical addresses: for system administration services, e.g. determination of logical addresses, each device also has a physical address that is derived from its current ring position. The physical addresses are assigned the range 0x400 to 0x43F.

- ▶ Group addresses: each device may also be assigned a group address in the 0x300 to 0x3FF range, making it possible to address all devices in a given group at the same time.
- ▶ Broadcast address: the MOST additionally supports a special broadcast address, hexadecimal 0x3C8, which makes it possible to address all devices in the MOST system.

Transmission of control commands

The control channel permits the transmission of control messages with a maximum length of 32 bytes. These have to be shared among consecutive data frames. With MOST 25, the 16 data frames required here are also known as a block. The format of the control messages is defined by the protocol described in the next section.

The sending of messages on the control channel is supported by the NIC. It provides the necessary services for data protection and channel arbitration, i.e. for determining the next opportunity for a message to be sent. The reliable transmission of control messages is safeguarded by the use of a cyclic redundancy code (CRC) and an automatic repeat request.

The control channel's gross bandwidth of 705.6 kBit/s is reduced to a net data rate of 406 kBit/s following the deduction of approximately 3 % for system administration overheads (two of 64 control messages) and less the extra overhead to account for arbitration and data protection. Since arbitration only allows a particular device to populate every third available control message, an individual device has a maximum data rate of 135 kBit/s at its disposal for sending.

To enable the sending of messages of more than 32 bytes, the network service supports the segmentation and desegmentation of application messages up to 65,556 bytes in size on the control channel. This service, also known as the appli-

cation message service (AMS), differs from the control message service (CMS), which only allows 32-byte control messages to be sent.

Transmission of multimedia data

In the MOST system, the transmission of multimedia data takes place on the synchronous channels. The data transfer is controlled by relevant control commands on the control channel. A synchronous channel can be assigned a specific bandwidth, which is achieved with a resolution of one byte of a data frame. A stereo audio channel with a resolution of 16 bits requires four bytes, for example. In the case of MOST25, a maximum of 60 bytes (depending on the value of the boundary descriptor) are available for the synchronous channels, which equates to 15 stereo audio channels.

The sinks and sources for multimedia data are each assigned to an FBlock, which provides the necessary functions for their management. An FBlock may therefore contain several sources and sinks, which are numbered consecutively with a source and sink number respectively.

An FBlock has functions that supply information about the quantity and type of sources and sinks that it makes available (*SyncDataInfo*, *SourceInfo* and *SinkInfo*). Furthermore, each FBlock with a source has an “Allocate” function with which it requests a synchronous channel and associates its source to it. Its “Deallocate” function, on the other hand, is used to free up the channel again.

Similarly, an FBlock with a sink has a “Connect” function for connecting the sink to a particular synchronous channel, and a “DisConnect” function for its disconnection. Several sinks may also be connected to a particular channel at the same time.

With MOST 25, the channel assignment for synchronous channels is managed by the timing master. A source wishing

to occupy a synchronous channel sends an appropriate system message request (*ResourceAllocate*) to the timing master. A corresponding message (*ResourceDeallocate*) is used to free the channel.

Transmission of packet data

On the asynchronous channel, data is sent in the form of packets. The asynchronous channel currently supports two modes: a slower 48-byte mode, where 48 bytes are available in each packet for the net data transfer but which places less heavy demands on device implementation, and a 1,014-byte mode that is more complex to implement.

With MOST 25, between 0 and a maximum of 36 bytes of a data frame can be assigned to the asynchronous channel, which corresponds to a maximum gross bandwidth of 12.7 MBit/s. For the 1,014-byte mode, this produces a maximum net bandwidth of almost 11 MBit/s, and a data rate of almost 3 MBit/s for the 48-byte mode. In practice, however, substantially lower data rates are achieved mainly due to restrictions in device implementations.

The transmission of data packets on the asynchronous channel is also directly supported by the NIC. Likewise, a packet on the asynchronous channel contains a CRC to permit detection of simple transmission errors. Unlike the control channel, however, there is no issuing of automatic repeat requests on the data link layer.

To ensure reliable transfer and flow control in the large data volumes typical of the asynchronous channel, it is conventional to use yet another transport protocol that is implemented in a higher driver layer. This may either be the MOST high protocol (MHP), developed specifically for the MOST, or the popular TCP/IP protocol, which is set up on an appropriate adaptation layer known as the MOST asynchronous medium access control (MAMAC).

Administrative functions

The MOST standard defines the following administrative services that are required for the operation of a MOST system.

Configuration status

A valid configuration status in the MOST system is the prerequisite for communication in the application layer. The configuration of the MOST system is managed by a single device, the network master. An applicative communication is only permitted once this device has signaled a conflict-free configuration status.

- ▶ OK configuration status: the network master transmits the OK configuration status as soon as a conflict-free configuration is achieved.
- ▶ NotOK configuration status: if a conflict arises, e.g. because two devices have the same address, the network master transmits its NotOK configuration status. In response to this, all devices initialize their addresses and communication settings.
- ▶ New and Invalid configuration statuses: these configuration statuses indicate that a new application (FBlock) has been registered in the system or that a previously registered application has dropped out.

NetBlock

It is mandatory for each device to implement an FBlock called the NetBlock, which covers a range of different administrative functions (e.g. for address initialization) and supplies information about the device and the FBlocks that it implements.

Network master

The network master is implemented by a special device in a MOST system and is responsible for system configuration. In present-day systems, the network master is usually realized by the headunit (i.e. the control element) of the infotainment system. This device often assumes the role of timing master too, but this does

not necessarily have to be the case. The other devices in the MOST system in this relationship are known as network slaves.

The network master manages an image of the current configuration of a MOST system in the central registry. In addition to the addresses of all devices, this also contains all FBlocks implemented by them (down to the system FBlocks such as the NetBlock). A network slave searching for the device address for a particular FBlock may request it from the network master.

If a network slave frequently accesses other devices, it may, for faster access, also store a local copy of the central registry – a decentral registry – which it must update accordingly in the event of any change in the configuration.

To build the central registry, the network master carries out a network scan when the system starts and as soon as a network change event has occurred. During this scan, it queries the NetBlocks of all devices for the FBlocks implemented by this device. If it detects a conflict while doing so, e.g. duplicate device address or duplicate instance of an FBlock, it takes measures to resolve the conflict. To do this, the network master may initiate a recalculation of addresses, convert the InstId of an FBlock or, if necessary, ignore a device.

Connection master

The connection master manages the synchronous connections existent in the MOST system at a given time. It makes available a table containing information about all the connections currently present in the MOST system (SyncConnectionTable). It also makes it possible to query the bandwidth remaining for further synchronous connections (AvailableChannels).

On request, the connection master builds connections between a specific source and sink (BuildSyncConnection) or removes them (RemoveSyncConnection) by sending the necessary commands to the associated FBlocks.

It is even possible to dispense with the connection master completely if a central entity in a MOST system, e.g. the headunit, is responsible for the administration of the synchronous channels.

MOST application layer

The MOST standard defines a suitable protocol in the application layer for the transmission of control commands, status information and events. This protocol makes it possible to address a specific function of an application interface (i.e. of an FBlock) that is provided by any device within the MOST system. For example, it is possible to start playback of a CD on a separate CD player or to query the number of the track currently playing in it. While the MOST protocol is mainly applied to the control channel, it can be transferred across the asynchronous channel if necessary.

Overview

The protocol for MOST control messages provides for the following elements of a control message (Table 1):

- ▶ The address of a device in the MOST system (DeviceID)
- ▶ An identifier for an FBlock implemented by this device (FBlockID) and its instance in the MOST system (InstID)
- ▶ The identifier for the requested function within the FBlock (FktID)
- ▶ The type of operation (OpType) that should be applied to this function, e.g. the setting or querying of a property of the FBlock
- ▶ A data range containing the parameters of the function call (Data) and a corresponding length value (Length)

1 Elements of a control message		
Field	Size	Description
DeviceID	16 bits	Device address
FBlockID	8 bits	FBlock identifier
InstID	8 bits	Instance of the FBlock
FktID	12 bits	Function identifier
OpType	4 bits	Type of operation
Length	16 bits	Length of the data field
Data	0 to 65,535 bytes	Data field

Table 1

Function block (FBlock)

A function block (FBlock) defines the interface of a specific application or system service. An FBlock is assigned an address comprising an 8-bit FBlockID, which specifies the type of FBlock, and an additional 8-bit InstID. The latter is a means of distinguishing between several instances of FBlocks of the same type in a MOST system.

An FBlock that is controlled by an associated application, and the device that implements the FBlock, are known as the slave in this relationship (e.g. an application that addresses an external CD player through the associated FBlock). The application that controls the FBlock is called the controller.

The functions that an FBlock possesses are defined by the function catalog. In the case of system services, the function catalogs are defined by the MOST standard. The MOST standard even defines the interfaces for common applications from the area of vehicle infotainment systems (e.g. for an amplifier or a CD player/changer). However, these interfaces tend to be extended by the system manufacturer. Other proprietary FBlocks are defined in their entirety by the system or device manufacturer.

Functions and operations

A function block is made up of several functions that are addressed with a 12-bit FktID. Depending on the type of function, the function may be assigned differently predefined operations, which are described by a 4-bit OpType identifier. In the case of functions, a further distinction is made between methods and properties.

A method describes an action that an FBlock is able to execute, e.g. starting the station scan of a radio tuner. A method can be initiated by the controller using a Start or StartResult operation. In response, the slave returns a corresponding result message if applicable (for StartResult). If this result is not made available within a specific period, the controller is notified that the method is still being processed by means of a processing message. In the end, a method may even be terminated by the sending of an abort message.

A property describes a particular attribute of the FBlock, such as the number of the track currently playing. A property such as this can be requested (Get operations) or set (Set or SetGet operations) by the controller. As the result of such a query, the slave responds with a status message, if applicable (for Get and SetGet), containing the current contents of the property.

Furthermore, the MOST standard provides for a notification service with which a controller may register with a slave for certain properties. If the value of one of these properties changes, all controllers that have registered for notification are informed of the change by a status message.

Data field

The data field of the control message contains the parameter values for a function call or its results. The data field is interpreted as specified in the definition of the function concerned and may contain one or more parameter values.

The MOST standard defines the following parameter types:

- ▶ Boolean and BitField for individual boolean values (1 byte) or sequence of individual bits (1, 2 and 4 bytes).
- ▶ Enum for enumerations.
- ▶ Unsigned and signed byte, word and long for integer values with or without sign and a size of 1, 2 or 4 bytes.
- ▶ String for character strings. These are zero terminated and contain a description of the character coding in the first byte.
- ▶ Stream, ClassifiedStream and ShortStream for various types of byte sequences of any length.

A parameter type for real numbers does not exist. The MOST standard provides for real numbers represented as fixed-point values by the integer parameter types. More complex data types are not defined in the standard either. These have to be realized by function classes instead.

Function classes

To standardize the way in which functions are defined, the MOST standard specifies a series of function classes for properties. These determine which properties the function has and which operations are permissible.

In addition to simple function classes (with a single parameter for numerical values or texts), there also exist complex, composite function classes. These are used to represent composite data structures such as records and arrays. They can also be nested at a specific nesting depth, which makes it possible to create two-dimensional arrays or arrays of records, for example. These nested data structures are, for example, used to represent the phone book of a mobile phone.

With multidimensional function classes, the two parameters PosX and PosY are used to address a particular position in the first or second dimension, where 0 always stands for all elements of this dimension. This makes it possible to access not only a specific element but also an entire line of a twodimensional array.

Applications

As well as defining the lower layers necessary for data transfer, the MOST standard defines the interfaces for typical applications from the area of vehicle infotainment systems, e.g. a CD changer, amplifier or radio tuner.

The FBlocks defined by the MOST Cooperation are compiled in a function catalog. The function catalog defines in detail all functions for the FBlocks contained in them as well as the permissible operations and their parameters.

For the function catalog, there is a machine-readable description available in XML format that can be used to import the catalog into various MOST tools.

In addition to the function catalog's description of all functions of an FBlock, developments are underway to specify the dynamic behavior for the use of an FBlock. This involves the use of message sequence charts (MSCs) compliant with the MSC 2000 standard but with minor, MOST-specific enhancements.

Functions required by various applications, e.g. the aforementioned functions for managing sources for multimedia data, are defined in the GeneralFBlock. This is used as a function compilation, i.e. when a new FBlock is defined, all necessary functions are copied from the GeneralFBlock to the new FBlock.

Bluetooth

Overview

Mobile communication is gaining ever more importance in all domains. For convenient communication between various devices, a wireless link is indispensable. Infrared connections, as often used in the past, tended to be manufacturer-specific, they required a direct line of sight and imposed constraints on the area of movement. Only with a standardized, wireless link could mobile devices of different manufacturers communicate with each other without problems.

The development of the Bluetooth standard began in 1994 at the telecommunications company, Ericsson. As part of an initial study, Ericsson investigated the possibilities for completely replacing the cable connections between the cell phone and additional devices. In cooperation with other industry partners, the Bluetooth SIG (Special Interest Group) was formed in 1998 to create a uniform, internationally-accepted standard. The SIG set itself the aim of specifying a wireless technology with low manufacturing costs, low energy consumption and robust resistance to interference. The wireless interface had to be suitable

for transmitting data for multimedia applications.

In the meantime, the Bluetooth SIG has expanded to involve around 2,000 companies from the telecommunications, data processing and automotive engineering fields.

Applications

Bluetooth is an industry standard for the networking of mobile multimedia devices, such as a car sound system, cell phone, headset, PDA (Personal Digital Assistant), PC and peripheral equipment (Fig. 1). Bluetooth stands for the simple exchange of data between portable terminals and the wireless transmission of audio and video signals for entertainment and information.

The short-range wireless connectivity of Bluetooth eliminates the need for connection cables between the cell phone and the hands-free system: no more tangling of cables when making a call. It is in this area that Bluetooth has been the most well received.

Meanwhile, a large number of consumer electronics devices, be they cell phones, PDAs, notebooks or car phones, now offer an address book function.

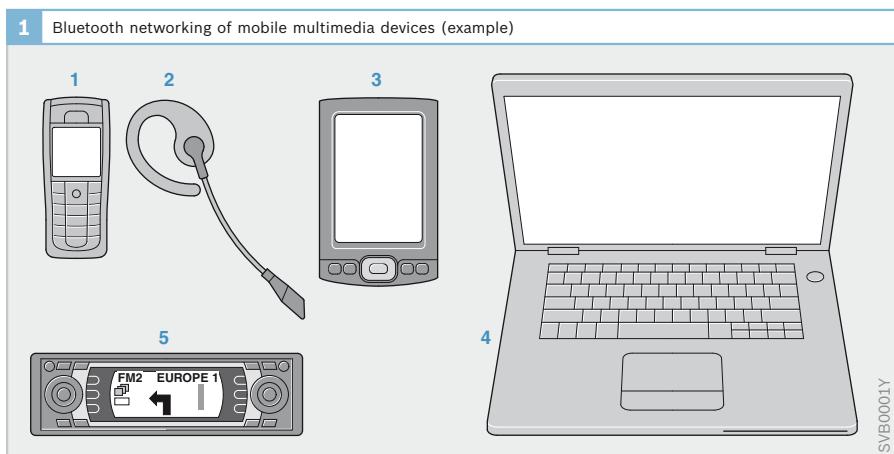


Fig. 1

- 1 Cellular phone
- 2 Headset
- 3 Personal digital assistant (PDA)
- 4 Laptop computer
- 5 Car sound system with navigation unit

SYB0001Y

With Bluetooth, it is possible to conveniently synchronize the entries stored in the various address books. The car driver is spared the laborious task of typing the phone number stored in the cell phone or PDA or inputting an address into the navigation system.

Bluetooth also makes it possible to listen to the radio through a headset in the rear of the vehicle and to control the car sound system using a PDA.

Bluetooth is predominantly used for the transfer of data in the multimedia domain. However, this technology can quite easily be used for diagnostic and service purposes. Workshops or breakdown services would be able to conveniently read vehicle information during fault diagnosis and record the condition of the vehicle.

Bluetooth versions

The first specification (Bluetooth 1.0) of the Bluetooth standard was adopted in July 1999. In December 1999 this specification was replaced by Bluetooth 1.0b, which contained improvements and clarifications for interoperability. February 2001 saw the release of the Bluetooth 1.1 specification, which fixed fundamental interoperability problems between the first devices and chips and is used in most Bluetooth devices today. Three years later, in November 2003, version 1.2 of the specification was published. This used the IEEE (Institute of Electrical and Electronics Engineers) language definition for technical documentation and introduced adaptive frequency hopping (AFH) as a technical advancement that improves resistance to interference in the frequency band. The most significant modification was introduced with version 2.0 of the specification in November 2004. With the enhanced data rate (EDR), it was possible to triple the bit rate from 1 MBit/s to 3 MBit/s and reduce power consumption. With each new version of

the specification, great emphasis was placed on backward compatibility with the previous versions.

Transmission technology

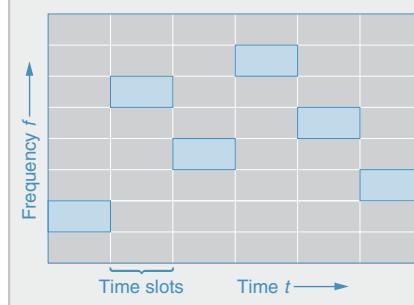
Frequency band

Bluetooth operates in the worldwide-license-free 2.4 GHz ISM band (Industrial Scientific Medicine). This is available practically worldwide without approval - apart from certain national limitations - and is reserved for applications in industry, science and medicine. For this reason, the possibility of interference from garage-door openers, microwave ovens or other appliances which operate using the same frequency band has to be accepted. To minimize interference, the best possible use should be made of the frequency spectrum. Devices that use the ISM band therefore implement spread-spectrum techniques.

Frequency-hopping method

Bluetooth divides the 2.4 GHz band into 79 channels at intervals of 1 MHz ($f = 2,402 + k$, $k = 0 \dots 78$). The band is terminated at both the upper and lower end by two guard bands with a width of 3.5 MHz (upper guard band) and 2 MHz (lower guard band). The transmission

2 Frequency-hopping method



uses a combined frequency-hopping/time division duplex (FH/TDD) method: the frequency-hopping spread spectrum (FHSS, Fig. 2). The channels are switched 1,600 times per second (i.e. in 625 µs slots). The slots are assigned in accordance with the TDD method, i.e. the transmitter and receiver are authorized to send alternately. The frequency-hopping method guarantees optimal and uniform use of the ISM band and makes Bluetooth resistant to interference from other transmitters in the same frequency band (e.g. WLAN). In addition, it also offers some security because a hacker would not know the destination of a frequency switch or indeed the time frame for which a particular frequency would be valid. This makes interception of a connection much more difficult.

The introduction of adaptive frequency hopping (AFH) in the Bluetooth 1.2 specification further increased Bluetooth's resistance to interference and reduced the effect of Bluetooth itself as a source of interference on other devices in the 2.4 GHz ISM band. The number of effective channels is dynamically adaptable to the conditions in the frequency band. Frequencies used by other transmitters are then removed from the list of 79 possible channels.

Modulation method

Bluetooth uses the Gaussian frequency-shift keying (GFSK) method of modulation. This is a special variant of frequency-shift keying (FSK) supplemented with a Gaussian low-pass filter. GFSK varies the frequency and delivers 1 bit per keying interval. Bluetooth has a symbol rate of 1 MBit/s. With the enhanced data rate (EDR) introduced with the Bluetooth 2.0 specification, it has been possible to increase this data rate to 3 MBit/s. This is made possible by the use of two variants of phase-shift keying (PSK): the $\pi/4$ differential quaternary phase-shift keying ($\pi/4$ DQPSK) and the

8 differential phase-shift keying (8DPSK), which modulate the phase of the carrier.

$\pi/4$ DQPSK represents one of four combinations of a bit pair (00, 01, 10 and 11) and so transfers a maximum of two bits; 8DPSK can represent any combination of three bits. This is used to increase the bit rate from 1 MBit/s (net 727 kBit/s) to 2 MBit/s (net 1,446 kBit/s) with GFSK, or 3 MBit/s (net 2,169 kBit/s) with EDR.

Power classes

Bluetooth devices are categorized into different power classes. Manufacturers are free to decide which power class they implement. The present specifications define three power classes:

- ▶ Class 1 with 100 mW (20 dBm) transmission output for a range of 100 to 150 m
- ▶ Class 2 with 2.5 mW (4 dBm) transmission output for a range of 10 to 25 m
- ▶ Class 3 with 1 mW (0 dBm) transmission output for a range of 10 m

Topology

Bluetooth supports the automatic configuration of ad-hoc networks. This means that two or more devices are able to spontaneously form a network with no prior knowledge of each other. Bluetooth is based on the master-slave principle. In principle, any Bluetooth device may be a master or a slave. The master is a special device that coordinates communication between the devices.

Piconet

A network of Bluetooth devices is known as a piconet (Fig. 3). For the devices within a piconet, there are two roles: master or slave. Which device assumes which role is not decided until communication is established. A device sets up a piconet and transmits its device ID and the value of its internal clock. This device becomes the master.

Any device that is not participating in a piconet but is not switched off remains in standby, which is marked by a small current draw.

The first stage in the creation of a piconet involves the potential master searching for other devices in its range (inquiry scan). Special frequency-hopping sequences (inquiry sequences) are available for this purpose. Devices that wish to enter into communication with other devices respond with an inquiry response message. By the end of the inquiry sequence, the information about all devices willing to communicate is available. The potential master is now able to address a specific device by means of a special paging sequence. Paging is a process whereby a fixed connection is established between a master and a slave. The slave contains all the information it needs to synchronize with the master.

The piconet is defined by the frequency-hopping sequence, i.e. the sequence of switching between the 79 possible channels. The hopping sequence is specified by the master and is calculated from the master's device ID (a worldwide unique 48-bit identifier) and its internal clock. The slaves synchronize with the master's clock and its hopping sequence.

A master is able to manage up to seven active slaves. Each active slave receives a 3-bit active member address (AMA).

Active devices can send data or even simply remain connected. By deregistering, a device can switch back to standby. A Bluetooth device may also switch to any of three energy saving modes: hold mode, sniff mode and park mode. The devices continue to be synchronized with the hopping sequence of the piconet. The lowest energy saving mode is the park mode, where the device relinquishes its AMA and receives in its place an 8-bit parked member address (PMA). With the PMA, the device is still a subscriber in the piconet but makes way for another, active device. Parked devices remain synchronized with the hopping sequence of the piconet and can be addressed and reactivated by the master using the PMA.

Scatternet

In a piconet, only one slave is ever able to communicate with the master simultaneously. When more slaves enter the network, the data rate per device falls very quickly. This is where scatternets provide a solution. A scatternet consists of up to ten piconets in which the individual piconets partly overlap. In this way, it is possible to set up even larger networks in which all network subscribers may connect to one another. The scatternet in Figure 4 comprises two piconets. One of the Bluetooth devices is a member of both piconets. Through this device, the two networks are able to exchange data. However, this device, which is assigned

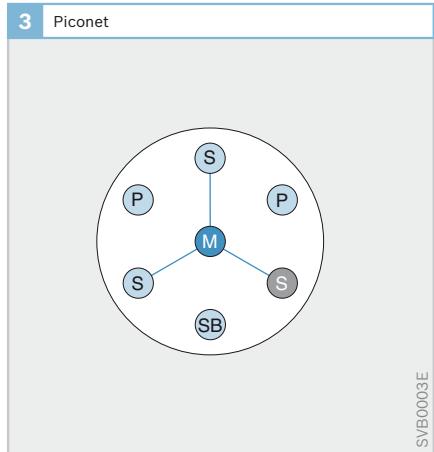


Fig. 3

- M Master
- S Slave
- P Park mode
- SB Standby

to more than one piconet, can only function as master in one of these networks.

A device that is assigned to two piconets in a scatternet must synchronize with the network with which it wishes to communicate. It cannot remain connected to both networks at the same time. Before the device leaves the old piconet, it must inform its master that it will be unavailable for a certain period of time. The remaining network subscribers can continue to communicate.

A master, too, may leave its own piconet and become a slave in another. All communication in its old network, however, is broken until it returns and assumes its master function again.

A master cannot become master of a second piconet. If this were to happen, both piconets would behave in an identical manner and form a single network.

Physical data channel

A data channel is represented by a hopping sequence between the 79 possible frequencies in the ISM band. Any device that is actively participating in a piconet must also hop at the same carrier frequency (frequency f_i). As soon as a master has sent data at frequency f_k , a slave can respond at frequency f_{k+1} . This procedure is illustrated in Figure 5.

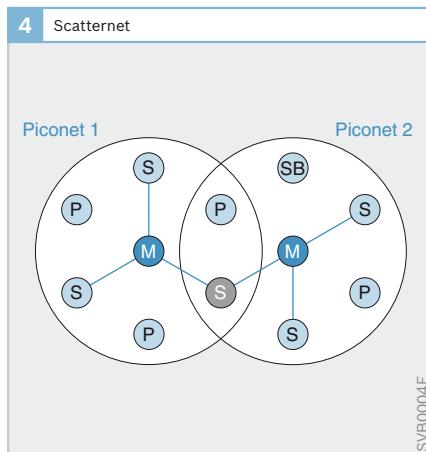


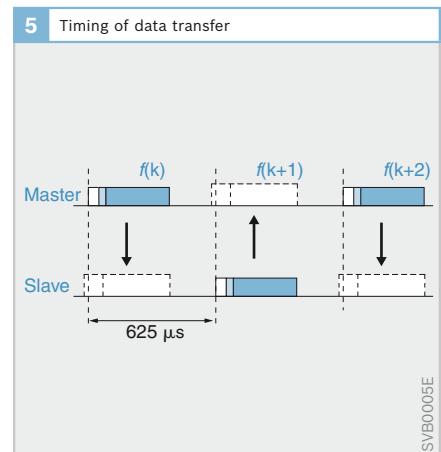
Fig. 4

- M Master
- S Slave
- P Park mode
- SB Standby

It is conventional for the transmissions of the master to begin in the even time slots and those of the slaves to begin in the odd time slots. This characteristic conforms to the regular time-division multiplexing method (TDD here) whereby the master uses one half of the time slots and the slaves the other half.

In addition to packets that may occupy one time slot, Bluetooth provides for packets that may occupy three or five time slots (multi-slot packets). As soon as a master or slave sends a packet three or five time slots long, this transmitter will remain on the same frequency. There is no frequency change within a packet (Fig. 6).

Once the packet has been transferred, the frequency hops as determined by the hopping sequence (independently of the transmission process). In the illustration, for example, the frequency hops from f_k to f_{k+3} after the transmission of a 3-slot packet. The reason for this behavior is that not every station may have received the transmission and cannot therefore respond specifically to the transfer of data in several time slots. For this reason, all stations that are not involved in the transmission always continue to hop in the sequence specified by the master.



Physical connections

Bluetooth supports both circuit-switched and packet-switched data channels. For circuit-switched, synchronous communication, Bluetooth offers a synchronous connection-oriented link (SCO), and an ACL (Asynchronous Connectionless Link) for packet-switched, asynchronous communication. Between two Bluetooth devices, there may only ever be one ACL link but there can be up to three SCO links. The ACL links are the foundation for the data transfer. The SCO links are purely voice channels with a data rate of 64 kBit/s. The voice is transmitted by continuously variable slope delta modulation (CVSD) or pulse-code modulation (PCM). A Bluetooth master can maintain several different data channels between its slaves even at the same time.

Logical data channels

Logical channels refer to different types of channels streamed over a physical connection. The data transmitted over a physical channel has different logical meanings. Bluetooth distinguishes between two categories of channel: link channels, which are used for the exchange of control information between master and slave; and user channels,

which transfer data from the user level – the applications.

Data packets

A Bluetooth data packet generally comprises three fields:

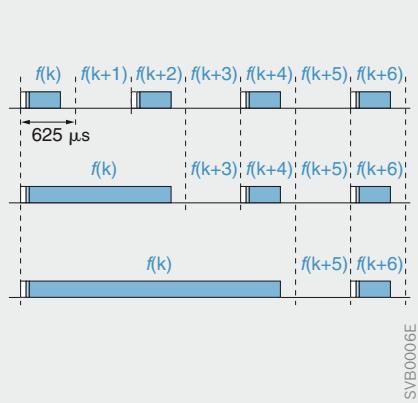
- ▶ Access code
- ▶ Header and
- ▶ Payload

A packet may simply comprise the access code, or the access code and a header, or the access code, header and payload (Fig. 7).

A packet begins with an access code. If a header follows, the access code has a length of 72 bits, otherwise it is 68 bits long. The access code is used for the synchronization and identification of a piconet, for device polling and for device calls. There are three different types of access code:

- ▶ Channel access code (CAC), the purpose of which is to synchronize and identify a piconet. The CAC is sent with each packet transmitted over the piconet.
- ▶ Device access code (DAC), which is used to transfer specific identifiers during the device call (e.g. when paging).
- ▶ Inquiry access code (IAC), which is used for device polling.

6 Frequency hopping with multi-slot packets



7 Structure of a packet

Access code 72 (68) bits	Header 54 bits	Payload 0 to 2,745 bits
4 preamble	3 AMA	
64 synchronization	4 type	
(4) appendix	1 flow	
	1 ARQN	
	1 SEQN	
	8 checksum	

The header of a Bluetooth packet is 54 bits long and contains information about the link (link control information). It is made up of the address, packet type, flow control, error monitoring and checksum fields. The actual length of the header (i.e. the sum of the bits in the header) is 18 bits. The header is protected by forward error correction (FEC). The bits of the header are sent three times (1/3 FEC), giving a header length of 54 bits. A receiver can then simply take the majority decision; each bit triplet is resolved to the value with majority in the triplet.

The payload of a Bluetooth packet can be up to 341 bytes in size (1,023 bytes with EDR). The data field of the payload consists of up to three segments: the payload header, the payload itself and, in certain circumstances, a checksum.

Bluetooth uses 16 different packet types, the common feature of all being the access code and the header. The structure of the payload depends on the packet type concerned. Along with packets for ACL and SCO links, there are also packet types for polling the slaves, synchronizing the hopping sequence and acknowledging data transfers.

Bluetooth device addresses

A Bluetooth device is assigned a worldwide unique Bluetooth device address. The address is derived from the IEEE-802 standard and implemented in the device by the manufacturer. The 48-bit addresses are subdivided into three parts

- ▶ LAP (Lower Address Part, containing 24 bits)
- ▶ UAP (Upper Address Part, containing 8 bits)
- ▶ NAP (Non-significant Address Part, containing 16 bits)

The LAP and UAP fields form the significant part of a Bluetooth device address.

Bluetooth architecture

The Bluetooth architecture is complex. This is inherently associated with the idea behind general cable emulation. The architecture is designed to enable new protocols and applications to be adapted for the use of Bluetooth. Accordingly, the Bluetooth protocol stack is very extensive (Fig. 8 shows a simplified version of the protocol stack).

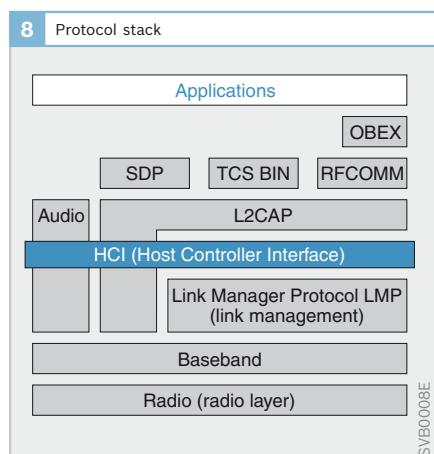
The Bluetooth technology can be subdivided into three logical sections:

- ▶ Hardware
- ▶ Host stack
- ▶ Applications

The hardware layer and the host stack are defined in the Bluetooth Core Specification; the applications are part of the Bluetooth Profile Specification.

Hardware-side protocols

The hardware is represented by the Bluetooth chip. This is where the radio layer, baseband and link management are contained. The radio layer deals with the radio transmission, used frequencies, modulations and transmission output. The baseband represents the link-establishment mechanisms, the packet structure and the timing. The link management establishes and manages the link between



two devices and also implements security and authentication functions.

The host controller interface (HCI) is the interface between the hardware of a Bluetooth module and the host-side protocols. The HCI is a hardware abstraction and realizes various interfaces for controlling the Bluetooth hardware and transferring data (e.g. via USB or UART).

Host stack

The host stack uses the HCI to control the Bluetooth hardware and transfer data. The format of this data is determined by the upper protocol layers. The logical link control and adaptation protocol (L2CAP) implements the abstraction of the hardware's properties, adapts the upper layers of the protocol stack to the capabilities of the baseband and hides transmission details such as the connectionless or connection-oriented transmission type. The L2CAP layer essentially has three main functions:

- ▶ It can receive packets with a length of up to 64 kB from the upper layers and decomposes them into smaller data packets (segments) for processing in the lower layers if necessary. At the other end, the segments are recomposed back into packets.
- ▶ It manages the multiplexing and demultiplexing of several packet sources. If a packet is being recomposed, the L2CAP layer determines to which protocol of the upper layers the packet is forwarded.
- ▶ L2CAP offers functions for negotiating quality of service and configuration parameters. This means, for example, that the maximum size of the payload can be negotiated so that a device with limited resources would not be overwhelmed by overly large packets. Using the configuration parameter for the quality of service, it is possible to define the properties of the data transfer: best effort (best attempt but with no guarantee of data transfer) or guaranteed.

In addition to L2CAP, there are also audio and control protocols that govern the handling of audio data and control data. Audio applications, for example, can make direct use of the baseband layer once the audio signals have been coded accordingly.

The service discovery protocol (SDP) is another important host-side protocol. It serves to identify and scan for services with specific properties and to describe services within range of a Bluetooth device. Bluetooth devices are meant to be able to interact on an ad hoc basis with other devices in different environments. It is therefore necessary to know which services are made available by which devices within range. All devices wishing to provide services must use an SDP server; for all other devices, an SDP client is sufficient. The SDP prepares the services available on a device in a service database. The service information in the possession of the SDP server is stored in a service record. The service record consists of a list with service attributes that describe the properties of the service more precisely and is identified by a 32-bit service record handle.

Two further protocols in the protocol stack form the fundamental basis for the interoperability of Bluetooth devices.

The RFCOMM (Radio Frequency Communication) cable emulation protocol above the L2CAP layer simulates up to 60 virtual serial interfaces derived in accordance with the ETSI 07.10 standard. As a result, almost any software that previously expected a serial interface can work with Bluetooth.

TCS BIN (Telephony Control Protocol Specification - Binary) is employed as the means of controlling telephone and telephony functions. This is a bit-oriented protocol for establishing voice and data connections between Bluetooth devices. The use of appropriate, indus-

try-wide standards has made it possible, here too, to ensure wide-ranging compatibility with legacy applications.

Applications

Many more protocols have been adapted for the Bluetooth standard (adapted protocols) and can be found in the protocol stack. Internet applications can, for example, continue to use TCP/IP through the point-to-point protocol (PPP) or the Bluetooth network encapsulation protocol (BNEP). For the exchange of vCalendars and vCards, it is possible to use the object exchange protocol (OBEX) covered by the IrDA standard.

Profiles

The uppermost layer contains the applications and profiles. Profiles represent “standard solutions” for a particular usage scenario. The Bluetooth specification currently brings together 13 different applications described as profiles. Each profile uses a certain choice of protocol; in principle, application profiles make a different protocol stack available for each application. Profiles describe the vertical slice through the Bluetooth protocol stack; the protocols represent the horizontal layers (Fig. 9). Within the profiles, the required and optional functions of the layers are defined. These standardized profiles make it possible to ensure interoperability between different devices.

The four fundamental profiles of the Bluetooth specification are:

- ▶ Generic access profile (GAP)
- ▶ Service discovery application profile (SDAP)
- ▶ Serial port profile (SPP) and
- ▶ Generic object exchange profile (GOEP)

The Bluetooth profiles form the hierarchy (Fig. 10). The GAP forms the basis and describes all the essential functions that a Bluetooth device must fulfill at the lowermost level. These include, for example, the functions for establishing and managing the link, the supported operating modes and the security of a link.

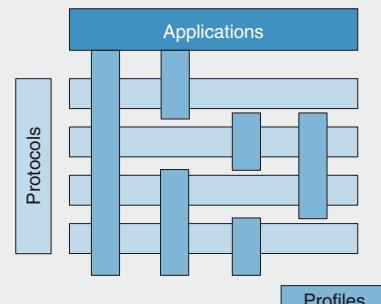
The SDAP defines the access interface for the service discovery protocol (SDP) with which devices can discover or poll the services offered by other devices.

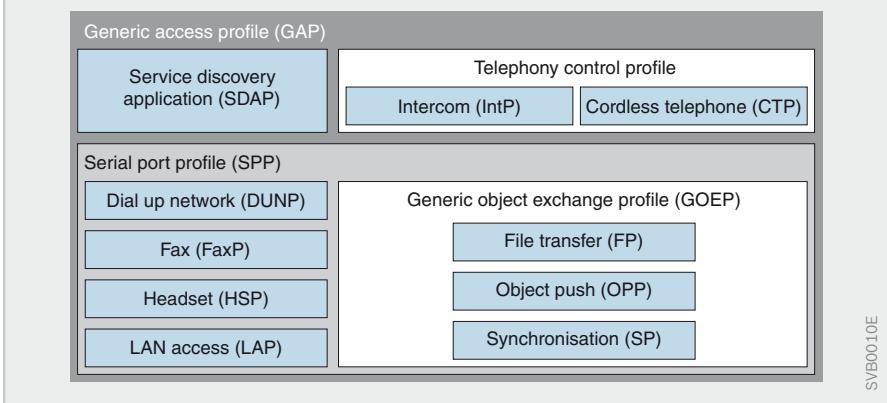
SDP builds on the GAP. All Bluetooth devices must implement these two profiles.

The SPP is used by most other Bluetooth profiles. One exception to this is the telephony control profile. This is always used when Bluetooth is being used for cable emulation or if a serial data connection is to be used. The SPP builds on the GAP and uses the RFCOMM protocol.

The GOEP defines the fundamental functions necessary for the exchange of complex objects. It defines a client-server relationship for the exchange of data. Like the SPP, the GOEP provides the foundation for further profiles.

9 Profiles and protocols



10 Profile hierarchy

SVB0010E

**Origin of the name, Bluetooth**

The name, Bluetooth, originates from King Harald Gormsen, a Danish Viking. Harald had the nickname, Blatand. In the Middle Ages, Bla actually meant dark, while Blatand made reference to the dark figure of Harald Gormsen. Bluetooth, therefore, is not a direct translation of Blatand.

In the 10th Century, Harald Gormsen united large parts of Denmark and Norway. The Bluetooth wireless system developed 1,000 years later unified a wide range of different information, data processing and wireless mobile devices. This is why it was named after King Harald.

FlexRay

Overview

FlexRay is a field that was designed to support open and closed-loop control technologies in the automotive sector. Development focused on the suitability for use in active safety systems in particular. FlexRay therefore offers transfers with guaranteed compliance with transfer properties, high bit rates and a fault-tolerant design.

The development of FlexRay, managed by the FlexRay Consortium, goes back to cooperation between German automotive manufacturers, BMW and DaimlerChrysler. In 1999 these two companies began to compile requirements for a new communication system. In the process, development would incorporate experience gained from the byteflight system of BMW and the prototype development of DaimlerChrysler.

The aim of FlexRay is to provide a system with high transfer rates that will work in a deterministic and fault-tolerant manner while being as flexible as possible to use and expand. The main fields of application of the FlexRay are drivetrain systems (drive) and active safety systems with no mechanical fall-back level (x-by-wire). However, the areas of passive safety systems and comfort/convenience and body electronics are also supported.

To support these very different domains, FlexRay uses two different types of bus access. For deterministic transfer properties, it offers time-controlled bus access. In the case of applications whose transfers place less demanding requirements on transfer properties, it is desirable to make as effective use of the available transfer capacity as possible. To combine these two approaches, communication takes place in cycles. In each cycle, there is firstly a static transmission component in which bus access is controlled by TDMA (Time Division Multiple Access). There then follows a dynamic transmission component in which

bus access is controlled by the use of mini-slots. Minislots are small time windows in which a defined message can be transferred. This access method is also known as FTDMA (Flexible Time Division Multiple Access).

With messages transmitted in the static component of the cycle, the transfer properties can be assured in accordance with known methods. In the dynamic component, messages are sent as required. In this case, messages are prioritized based on the message ID. The division between the static and dynamic component is freely configurable but it cannot be modified later during system operation. The same applies to the slot lengths in the static range, which, while configurable, must remain constant once operation is underway.

FlexRay systems can be equipped with two transmission channels, with each channel having separate lines. This makes it possible to create a redundant data transfer or increase the available bandwidth for specific applications. The latter can be used both for the parallel transmission of information of two network nodes and for the faster transfer of information of one node.

FlexRay operates at a maximum bit rate of 10 MBit/s, which is achievable in optimal environmental conditions. A rate of up to 20 MBit/s is achievable if two channels are used without redundant access. In addition to the bit rate of 10 MBit/s, which is the only bit rate defined in Specification 2.1 dated 15th December 2005, there are endeavors to support lower bit rates of 2.5 or 5 MBit/s.

Areas of application

The areas of application of the FlexRay are very diverse. By offering redundant, time-controlled and fault-tolerant transfers, not only is FlexRay suitable for use in active safety systems with no mechanical fall-back level and in the area of the drive-train, but its optional, dynamic transmission range means that it is able to support application in the areas of passive safety systems and networking in the body and comfort/convenience domains.

With the high bandwidth of up to 20 MBit/s for non-redundant transfers, it is even conceivable, in theory, that FlexRay could be used for the transfer of audio or highly compressed video.

Topology

Point-to-point

The simplest configuration of a FlexRay system is the direct link between two network nodes. The maximum distance between the two nodes is 24 m in this configuration.

Bus topology

Between 4 and 22 nodes can be connected in a bus topology. The maximum distance between any two nodes in a bus topology may be 24 m.

Star topology

A star topology is another architectural variant of the FlexRay network. Both active and passive couplers are supported.

Passive star topology

3 to 22 nodes can be connected in a passive star. Here, too, the distance between any two nodes cannot be greater than 24 m.

Active star topology

An active star topology can be viewed as a point-to-point link between a number of nodes and an active coupler. In this topology, therefore, only the distance from a node to the coupler is of importance. This cannot be more than 24 m.

Apart from the connection lengths and the number of connectable nodes, the most important difference between an active and a passive star topology is in the achievable bit rates. With the use of passive couplers, the achievable data rate (approx. 1 MBit/s) is below the rate that can be achieved with active couplers (up to 10 MBit/s).

Cascaded star topologies

Cascaded star topologies are the product of several active star topologies linked together. In a FlexRay system, a topology like this is limited to three cascaded star topologies.

Hybrid topologies

Hybrid topologies are a blend of topologies from bus and star topologies, e.g. the connection of several bus topologies to a star topology instead of individual nodes.

Two-channel topologies

Since it is possible to implement both channels of a FlexRay system independently, different topologies can be used for each channel. For example, one channel could be realized as an active star topology, the other as a bus topology.

Hardware

FlexRay controller of a subscriber

A FlexRay node (Fig. 1) comprises the host processor, a communication controller (CC) and one bus driver (BD) for each channel to which the node is connected. Optionally, each node may contain a bus guardian (BG) for monitoring each bus driver and the bus guardian may receive information through an additional host processor.

It is optional whether a network node is connected to one or both channels; here, at the very most, the requirements from the field of application of the network node take effect.

Transmission agents

The transmission agent for a FlexRay system is of a twisted-pair cabling design, where shielded (STP) or unshielded (UTP) cabling may be used.

Each of the two FlexRay channels consists of two wires designated bus plus (BP) and bus minus (BM).

Host processor

The host processor gathers information from the sensors, which is forwarded to the communication controller for trans-

mission. The information received by the communication controller is forwarded to the actuators responsible for processing.

Communication controller

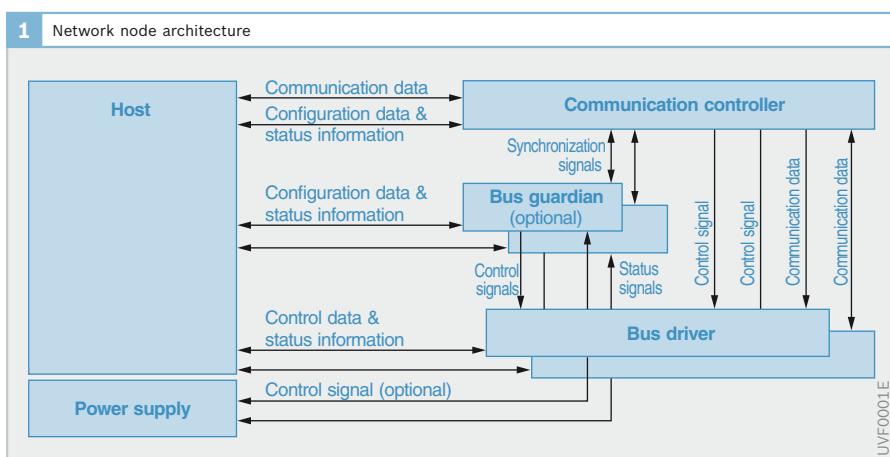
The communication controller (CC) realizes all aspects of the FlexRay system relating to the protocol. Its tasks include scheduling, synchronizing with other network nodes, creating a macrotick signal, creating a bit stream from the information of the host or controlling bus access.

Bus driver

The bus driver (BD) assumes the role of a transceiver. It converts logical information into physical voltages that are carried on copper wires, and vice versa. The role of the BD also includes a protective function against electrostatic discharge (ESD).

Furthermore, the BD monitors the states of the BP and BM and is thereby able to detect whether the bus is affected by physical faults.

The BD also processes wakeup signals and, optionally, is able to control the power management of the network node because functions of the node can be switched off.



Bus guardian

The bus guardian (BG) is a device that will only permit transmissions if the connected network node is authorized to send a message. With this system, it is theoretically possible for faulty network nodes to interfere at the very most with their own transmission but not with the transmissions of other network nodes.

The FlexRay protocol only provides the framework for use of a BG: its use is not a mandatory specification. Specification 2.0 only described a local bus guardian that would operate inside the network node and monitor access to one of the channels. A critical aspect of this concept was that correct operation of the bus guardian in a faulty network node could not be guaranteed as a consequence of the integration of the bus guardian in the node and the permission for the bus guardian to access the oscillator of the network node.

Specification 2.1 contains a refined concept for the local bus guardian and introduces the concept of a central bus guardian.

Local bus guardian

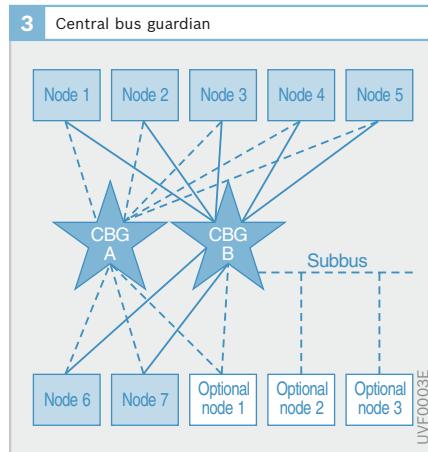
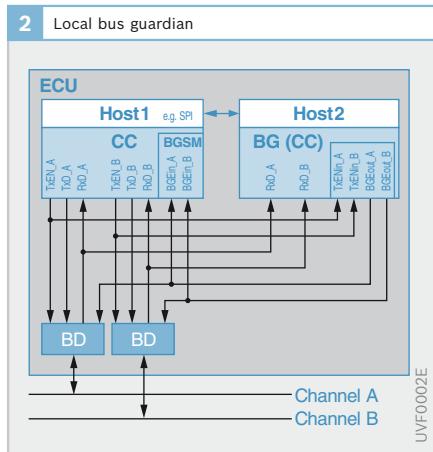
The local bus guardian (node-local bus guardian) is implemented as a stand-alone CC inside the network node that monitors operation of a CC (Fig. 2). Its synchronization process is independent of that of the CC's; in particular, the CC and BG have separate oscillators.

In addition, it is also possible for the BG and CC to be controlled by dedicated host processors that communicate with each other. This is, however, not mandatory.

As a consequence of direct control lines (bus-guardian enable, BGE) between the BG, CC and BD, information can only be transferred from the CC to the BD if the BG permits this transfer to take place.

Central bus guardian

The central bus guardian (CBG, Fig. 3) was a new feature introduced with Specification 2.1. It operates in the coupler of a star topology and monitors the operation of all connected nodes or networks. Faults, such as short-circuits on buses or network nodes, that transmit information like a “babbling idiot” with no consideration of the transmission window, and thus interfere with orderly communication, can be detected by a CBG and suitable counter-



measures may then be taken. As a result, the effects of these faults can be confined to small areas of the network.

Operating modes

In addition to normal operation, a FlexRay system also supports the complete initialization of the network from sleep to standby (Fig. 4). In order for this system start to function, the BDs of all network nodes have the capability to detect wakeup signals on the bus and to start the remaining other components of the dedicated network node.

Special network nodes that may be connected to the starter motor of a vehicle, for example, generate the wakeup signal.

The system start is split into the wakeup and startup phases. The wakeup phase involves the activation of the network nodes by the wakeup signal. In the startup phase, the network nodes are initialized and synchronized.

If the function of a node is not required for the time being, the node can be set to a standby mode in which all operations of the coding and decoding process are stopped. This reduces the energy consumption of the node. In sleep mode, the power consumption of a network node is at its minimum. In this mode, all functions of the network node are deactivated: only

the bus driver looks out for the transmission of a wakeup signal, in response to which the node is set to wakeup mode.

Sending and receiving

A bus driver connects a FlexRay network node to the channel that contains a receiver and transmitter. If a controller is connected to both channels, it follows that there must be two receivers and transmitters available in the network node.

Protocol

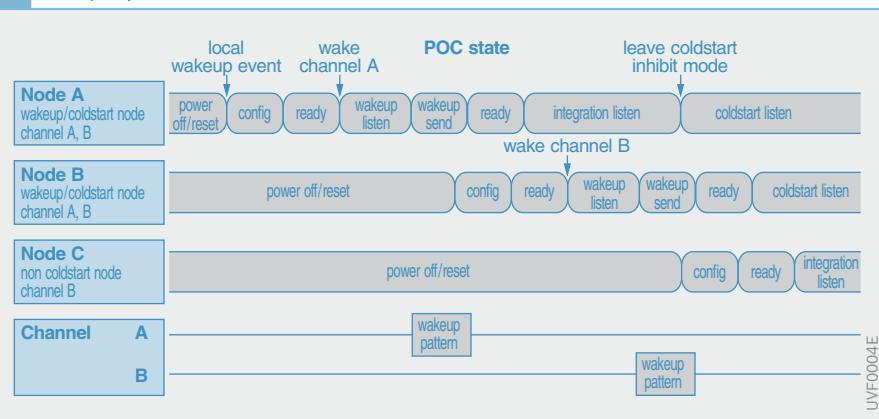
Protocol layers

The FlexRay protocol is built on five core mechanisms:

- ▶ Coding and decoding
- ▶ Control of bus access (media access control, MAC)
- ▶ Processing of frames and symbols
- ▶ Clock synchronization and
- ▶ Schedule monitoring by the bus guardian

Unlike in other network protocols, with FlexRay there are interfaces between all core mechanisms. This requires a process that coordinates and synchronizes the changes in the core mechanisms. This is the task of the protocol operation control (POC).

4 Wakeup sequence



FlexRay gives the host system the capability of influencing all five core mechanisms directly. To arrange this, communication takes place through an interface - the controller host interface (CHI) - between the network node and host.

Concentrating on the main interfaces, the task areas can be arranged as a protocol stack (Fig. 5). At the top level is the application, which forwards its commands to the CHI. Under the CHI is the POC, which in turn has access to the MAC, clock synchronization and frame & symbol processing. In the next lower layer are the processes for coding and decoding. The lowermost level is represented by the transfer characteristics of the physical layer.

The bottom three layers are in nodes that are connected to both channels: double presence because these functions have to be fulfilled for each channel separately. More precisely speaking, this means that for one node connected to both FlexRay channels there is double availability of synchronization, MAC, frame & symbol processing, coding and decoding, and physical layer.

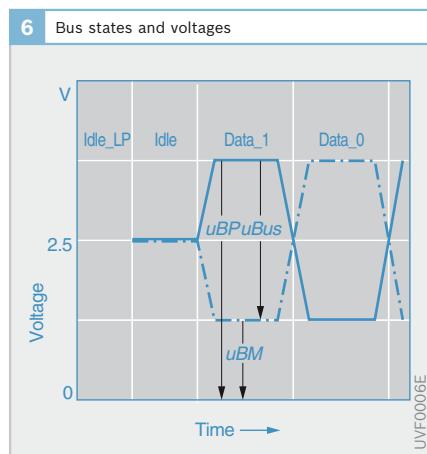
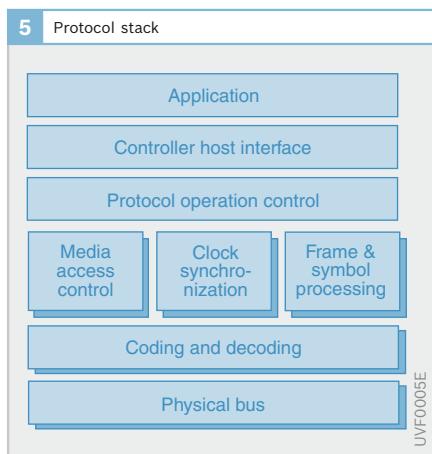
Coding on the physical layer

For coding, FlexRay uses a NRZ method (Non-Return to Zero) in which the two identical transmission states are not divided by a range in which the voltage falls back to a zero value. This type of coding makes it necessary to have mechanisms with which a network node can divide the transmitted states. To this end, FlexRay adds a sequence behind each transmitted byte: the byte start sequence (BSS). From this sequence, each node can detect when a byte has been transmitted and, with this information, encode the individual bits of the transmission.

By applying different voltages to the two wires of one channel, it is possible to create four bus states, which are designated Idle_LP, Idle, Data_0 and Data_1 (Fig. 6).

A bus state is identified by measuring the differential voltage. Here, the bus voltage (u_{Bus}) is made equal to the difference between u_{BM} and u_{BP} (voltages at BP and BM).

$$u_{Bus} = u_{BP} - u_{BM}$$



With this method, the data transfer is protected against external electromagnetic interference because these act equally on both wires and are canceled out in the difference.

Idle_LP (LP = low power) is the state in which a very low voltage of -200 mV and 200 mV is present at BP and BM. This state, for example, is used to identify the start of a transmission.

In Idle state, a voltage of 2.5 V with a tolerance of 500 mV is present at BP and BM.

To set the channel to Data_0 state, at least one transmitting node must apply a negative differential voltage of -600 mV to the channel.

For the channel to be set to Data_1 state, at least one transmitting node must apply a positive differential voltage of 600 mV to the channel.

If the transmission of an information signal is neither blocked by the bus guardian nor the communication controller, a HIGH bit is signaled by the Data_1 channel state and a LOW bit by Data_0.

Generation of a frame bit stream

Before a node can transmit a frame containing the data of the host, the frame is converted into a bit stream. To this end, the frame is first decomposed into individual bytes. The start of a frame is populated with a transmission start sequence (TSS) followed by a frame start sequence (FSS) (Fig. 7). From the bytes of the frames, an extended byte sequence (EBS) is then generated whereby each frame byte is preceded with a byte start sequence (BSS).

The 24-bit checksum (CRC) for this bit sequence (TSS+FSS+EBS) is now calculated and appended to the bit sequence. To finish the bit stream, a frame end sequence (FES) is appended.

If the frame belongs to the dynamic segment, a dynamic trailing sequence (DTS) can be additionally appended to prevent another node from beginning its transmission on the channel prematurely.

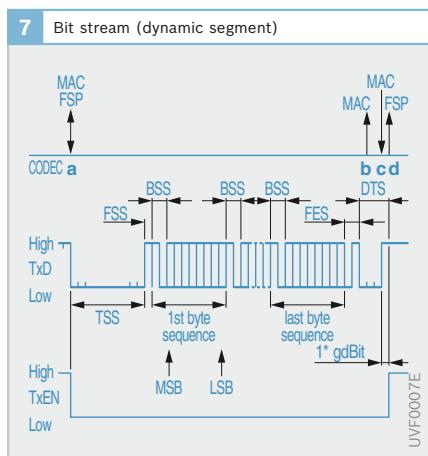
Communication cycle

Each FlexRay cycle (Fig. 8) contains a static segment that is transmitted as the first part of the cycle. The static segment contains a fixed number of transmission ranges, the “static slots”.

Optionally, there may be a dynamic segment in the FlexRay cycle and this is transferred in second place. Each dynamic segment contains a fixed, freely configurable number of “minislots”.

Also optional is the symbol window, which is transmitted as the third element of the cycle. It can be used for the transmission of an individual symbol and has the same size as a static slot.

To terminate the cycle, there is a phase - network idle time (NIT) – in which the bus is set to idle state. In general, the length of the NIT corresponds to the remaining macroticks not used by the static and dynamic segment or symbol window. This is not the case if it was detected during synchronization that an offset correction is required through which the length of the NIT can be increased or reduced.



Static segment

The static segment comprises a fixed number of equally sized transmission windows (static slots), which are sent simultaneously on both channels. Exactly one frame can be sent in each static slot. Both the size and quantity of the static slots in a static segment are configured during integration.

Bus access in the static segment is controlled by a TDMA method whereby, in each slot, the frame with the corresponding frame ID is sent. Frame ID 1 is therefore sent in slot 1, frame ID 2 in slot 2 and thus frame ID n in slot n.

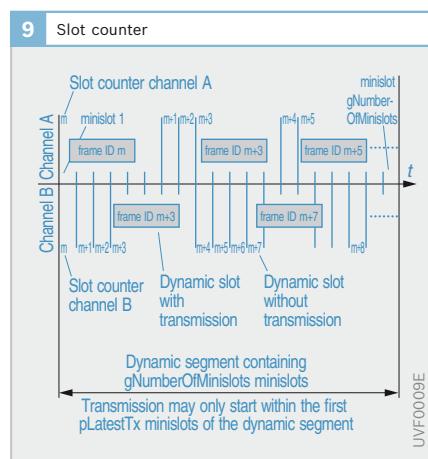
Dynamic segment

Unlike in the static segment, the transmissions in the dynamic segment are not of fixed length. To control bus access, “minislots” are used that hold the same fixed number of macroticks for all nodes connected to the network. Macroticks are ranges in which exactly one frame can be transmitted.

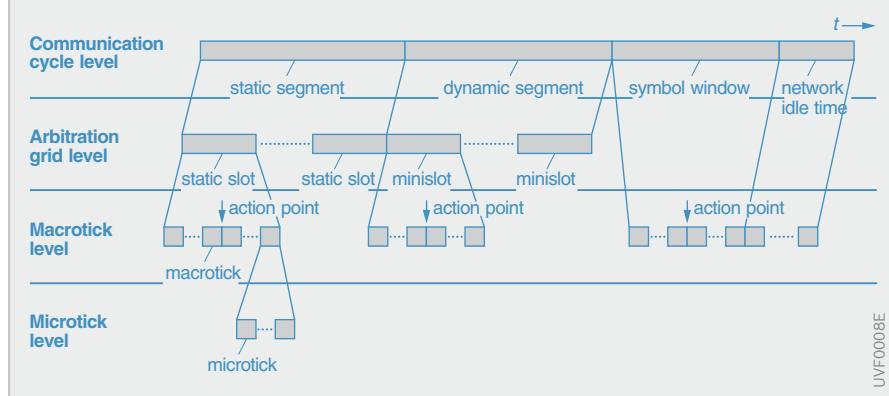
“Dynamic slots” (transmission windows of variable size) are now created based on the minislots. The maximum size of a dynamic slot is limited by a fixed number of minislots specified during configuration.

To identify the end of a transmission, a phase in which the channel is idle (dynamic slot idle phase) is added to the end of a dynamic slot. The phase in which the transmission takes place is called the dynamic slot transmission phase.

Another difference between the dynamic and static segment is that the counters of slots transferred are incremented synchronously on both channels in the dynamic segment. In the dynamic segment, the counter of both channels is incremented independently in line with the current transmission status (Fig. 9).



8 Communication cycle



Symbol window

The symbol window makes it possible to send a collision avoidance symbol (CAS) or a media access test symbol (MTS). These two symbols are identical and help to prevent collisions during the system-startup phase.

The third symbol defined in the FlexRay protocol – the wakeup symbol (WUS) – is not permitted to be sent in the symbol window. It serves only to generate a wakeup pattern (WUP), which is used during the system startup.

Network idle time (NIT)

During the NIT, a network node corrects the ascertained time deviations of its internal clock. It also gives the network node the opportunity in this phase to make implementation-dependent adjustments and changes to settings relating to the communication cycle.

Clock control

In both the static and the dynamic part of the communication cycle, FlexRay reverts to the identifiers of the transmission window in which a node may transmit. If this method is to be successful, it is necessary for all nodes connected to the network to have synchronous time information.

For this purpose, selected messages from the static part of the cycle are used (sync frames) and are transmitted by the network nodes that are connected to both channels.

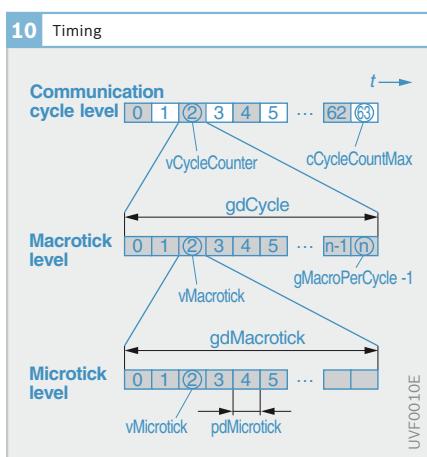
In a distributed communication system, each network node generally needs its own internal clock for setting transmission timing correctly. As a consequence of temperature and voltage fluctuations and production-related precision tolerances, these internal clocks can soon deviate from an abstract, system-wide global time. In the case of a system such as FlexRay, which controls bus access by means of time slots, the synchronization of all network nodes plays an especially important role because, without a standardized, global time preset for all nodes, the specified transmission windows cannot be maintained and may break down.

Clock control hierarchy

In a FlexRay system, clocking is controlled on three levels (Fig. 10). The bottom level is represented by the microticks, which are derived directly from the oscillator clock of the network node. A microtick is therefore not a system-wide variable, but only relevant to the node concerned.

At the next level up, the clock is controlled by means of macroticks – ranges in which exactly one frame can be sent. Within fixed tolerances, macroticks form the lowermost system-wide time unit. Each network node determines the duration of a macrotick in microticks. The number of microticks per macrotick does not have to be the same for all macroticks.

The top level is characterized by a clock control based on the communication cycle, which always contains the same, fixed number of macroticks.



Time synchronization

A synchronization method is required in order to guarantee a standardized time across the entire system. The basic parameters for timing are the number of microticks per communication cycle (*pMicroPerCycle*), the duration of a communication cycle (*gdCycle*) and the duration of a microtick (*pdMicrotick*), where the relationship between each is determined by the following equation:

$$pMicroPerCycle = \text{round}\left(\frac{gdCycle}{pdMicrotick}\right)$$

During operation of a FlexRay network, the clocks of the oscillators of various network nodes will not be equal or even simply remain constant. They will fluctuate in response to outside influences. This deviation is balanced by two corrective measures: the gradient correction and the inclusion of an offset in the communication cycle.

Fault-tolerant average algorithm

Since an individual node can only observe the differences between the expected time at which a transmission should have been actioned and the actual time at which the transmission took place, a fault-tolerant average algorithm (FTM), as described by J. L. Welch and N. A. Lynch, is used in both corrective measures. The measured time differences of all transmitted sync frames are sorted by their value. If more than seven sync frames were received, the two highest and two lowest values are removed from the list. If between three and seven sync frames were received, the highest and the lowest value are removed. For fewer than three sync frames, the average is derived from all measured differences.

Gradient correction

With gradient correction, deviations in the transmission frequency of a node are continuously corrected during the communication cycle. The node here checks the transmission frequencies of all other nodes. By means of the FTAA, these values are used to adjust the node's own transmission frequency.

Offset correction

Offset correction is a way to correct deviations in the communication-cycle phase by including an offset in the NIT. With this method, the node similarly employs the FTAA to determine its phase difference in relation to the other nodes of the network and uses it to determine the offset required to shorten or extend the NIT phase. Since all nodes use this approach, all nodes can initiate the transmission of a cycle in a synchronous manner.

Timing

Transmissions in the FlexRay system are timed by action points, which are specially-marked macrotick boundaries. A transmission in a static slot begins once a static-slot action point has been reached. The position of this action point in a macrotick is defined system-wide.

In a dynamic slot, a transmission begins once a minislot action point has been reached and ends at a subsequent minislot action point. This method ensures that there is always an idle period between individual transmissions.

Frame format

FlexRay uses the same format in both the static and dynamic part of the frame. The format can be broken down into three parts: header, payload segment and trailer segment (Fig. 11).

Header segment

The header segment contains a total of five bytes, at the start of which is a series of indicators.

The reserved bit is intended for future modifications to the protocol and is sent as a logical “0”.

The payload preamble indicator indicates whether or not the payload segment contains a network-management vector (NMVector). An NMVector enables the host processor to transmit data directly without first being processed and prepared by the CC.

The null frame indicator indicates a null frame that contains no usable information.

The sync frame indicator indicates that this frame is intended to be used for system synchronization. This indicator may only be set by network nodes that operate as sync nodes.

The startup frame indicator reveals that this frame is a startup frame. Startup frames are used in the network's starting phase and may only be sent by special coldstart nodes.

The indicators are followed by the frame ID, payload length, header CRC and cycle count.

The frame ID has a length of 11 bits and corresponds to the number of the slot in which the frame is sent. In one cycle, each frame ID exists only once on each channel.

The payload length expresses the size of the information sent in the payload segment. Since the payload in a FlexRay frame is always sent in 2-byte words, a 62 in the payload length field indicates, for example, that 124 bytes of payload are being transmitted in the payload segment.

In the static segment, the payload length field always contains the same value because the size of information here is constant for all frames. Of course, this property does not apply to frames in the dynamic segment.

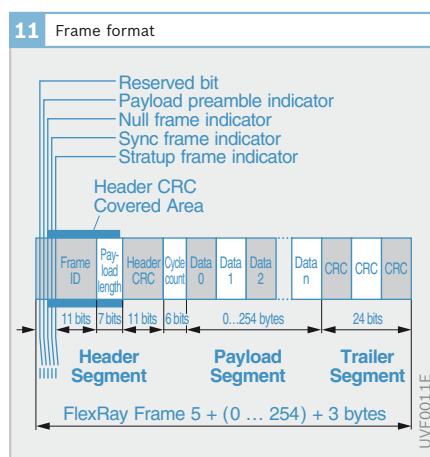
In the header CRC, an 11-bit CRC value is transmitted. This is formed from the polynomial

$$x^{11} + x^9 + x^8 + x^7 + x^2 + 1.$$

Since the information in the header segment is constant for all frames in the static segment, the header CRC value can be calculated offline and assigned to the network node through configuration settings.

If a frame in the dynamic segment is always of the same size, it is possible here also for the CRC value to be calculated off-line.

The final field in the header segment is the cycle count. This field contains the count number of the cycle in which the sending network node is currently active.



Payload segment

The payload segment has a maximum length of 254 bytes, which are transmitted in 2-byte words.

The payload segment customarily transfers the payload that is further processed by the host processors. Optionally, however, an NMVector may also be sent in the payload segment, or a 16-bit message ID.

With the assignment of message IDs, it is possible to send several data blocks in a frame. For this purpose, the message ID is placed at the front of the other data by the host processor as application data.

Trailer segment

The trailer segment contains a single field in which a 24-bit CRC checksum (frame CRC) is sent. The polynomial used to determine the frame CRC is:

$$x^{24} + x^{22} + x^{19} + x^{18} + x^{16} + x^{15} + x^{14} + \\ x^{13} + x^{11} + x^{10} + x^8 + x^7 + x^6 + x^3 + x + 1$$

Standardization

FlexRay Group

FlexRay is a protocol specified by the FlexRay Group. The FlexRay Group was founded in 2000 by the companies, BMW, DaimlerChrysler, Motorola and Philips Semiconductors (www.flexray-group.com).

The members of the FlexRay Group are organized in a hierarchy. They also have different levels of influence on the further

development of FlexRay, depending on the obligations that they have accepted. The uppermost group is represented by the core partners, which first and foremost include the founding members. Alongside the core partners are the premium associates and associate memberships.

By 2004 General Motors, Bosch and Volkswagen had joined the FlexRay Group as core partners. In 2004 Freescale took over Motorola's position in the group of core partners.

Characteristics

- ▶ Deterministic transfers possible
- ▶ Optional bus guardian monitors network access and protects against faults
- ▶ Differential signal transmission
- ▶ Collision-free transmission while operation is in progress
- ▶ Redundant transmission of information on two channels
- ▶ High transmission speed of up to 10 MBit/s; up to 20 MBit/s with parallel transmission on two channels
- ▶ Startup and initialization from an inactive sleep state are directly supported by the protocol
- ▶ Support for diverse fields of application
- ▶ Event and time-controlled transmission of information possible
- ▶ Published standardization by the FlexRay Group

Diagnosis interfaces

Diagnosis legislation (e.g. CARB, California Air Resources Board) demands constant monitoring of emission-related components in various electronic systems (OBD, On-Board Diagnosis). Faults (e.g. electrical short-circuits of sensors, implausible operating states) are stored in a fault memory in the control unit. These faults can be read using a scan tool (official testing station) or workshop tester. For this purpose, there is a socket in the footwell, dashboard or center console of the vehicle to which the tester is connected by means of a standard connector (ISO 15031-3) (Fig. 1).

Using a workshop tester, the after-sales service is able to read the entire fault memory: emission-related faults and vehicle-specific faults (e.g. from Motronic, ABS). With the information stored in the fault memory, it is possible to diagnose a fault and repair it efficiently. The tester can also be used to clear the fault memory. Furthermore, measured values and control-unit identification values can be evaluated. Using the workshop tester, it is also possible to control special diagnostic functions. With actuator diagnosis, for exam-

ple, individual actuators can be targeted for activation to check them for correct operation. At the end of the line (EOL), the diagnosis interface is used to test the electronic control unit and make changes to configurations (e.g. immobilizer, transmission type). It is even possible to program the entire flash EPROM using end-of-line programming.

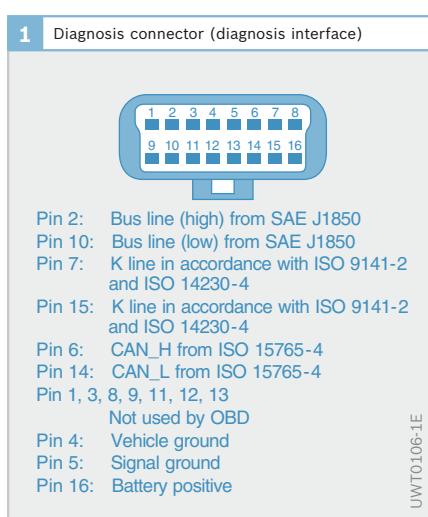
Diagnostics requires an interface for connecting the tester to the electronic control units. For communication, there are basically two options available at the present time:

- ▶ K line and
- ▶ CAN bus

The K-line network is not a bus in the true sense, which means that collisions may occur. Since most electronic control units have a CAN interface for the exchange of data between electronic systems, this option for communication with the tester is gaining ever more in importance.

Various protocols are implemented on the interfaces. These tend to be the CARB or official protocol and the manufacturer-specific communication, which can be closely related in their properties and functions. It is even possible to use both interfaces: CARB-relevant functions run on the CAN, manufacturer-specific functions on the K line.

Figure 2 provides an overview of the standards and the various layers of the OSI reference model.



Diagnostic protocol

The communication method in diagnostics is characterized by the fact that the tester (client) addresses one or several electronic control units (servers) and requests data output (e.g. from the fault memory) or an action (e.g. actuator diagnosis). These services and the transmission agent are defined in the various protocols.

KWP 71

This protocol was developed in collaboration with Bosch as the first standardized serial interface and is based on a 5-baud initialization and byte handshake. For addressing, the protocol provides for the use of a unidirectional line (L line). The line for data exchange (K line) is bidirectional. Alternatively, if the L line is not used at all in the system architecture, the K line can also be used for addressing because addressing and data exchange are two chronologically sequential states. The way in which communication is established is predefined, the services are standardized.

KWP 2000 (ISO 14230: 1-3)

This diagnostic protocol connects a tester to the electronic control units using a K line. In the KWP 2000 standard, there are various possibilities for the way in which communication is established (fast initialization and 5-baud initialization). The communication services are specified but there is some freedom for customer-specific adaptations.

ISO 15765: 1-3 (CAN)

This standard reproduces the existing protocols, ISO 14230: 1-4, on the CAN. It defines a way to leave untouched the fundamental format of messages exchanged between tester and electronic control unit and to transmit them on the CAN bus.

CARB

The Californian environmental authority (CARB) introduced guidelines to promote support for a standardized interface. The CARB protocol is built on existing standards and specifies the data flow and timing (transport layer). This is established in ISO 14230-4/ISO 9141-2 for the K line and in ISO 15765-4 for the CAN. The actual communication messages (services) in the application layer are precisely defined in ISO 15031-5/SAE J1979 and permit no manufacturer-specific deviations.

In California, every newly registered vehicle must support at least one of these protocols with all emission-related electronic control units. The European counterpart is the EOBD standard.

2 Standards for diagnosis communication				
Layer	CARB			
	K line		CAN	
7	ISO 15031-5	ISO 15031-5	ISO 15031-5	ISO 15031-5
6				
5				ISO 15765-4
4				
3				ISO 15765-2 ISO 15765-4
2	ISO 9141-2	ISO 14230-2 ISO 14230-4	SAE J1850	ISO 11898 ISO 15765-4
1	ISO 9141-2	ISO 14230-1 ISO 14230-4	SAE J1850	ISO 11898 ISO 15765-4

Layer	Manufacturer-specific	
	K line: KWP 2000	CAN / UDS
7	ISO 14230-3	ISO 15765-3 ISO 14229-1
6		ISO 15765-3
5		
4		
3		ISO 15765-2
2	ISO 14230-2	ISO 11898-1
1	ISO 14230-1	ISO 11898

- Fig. 2**
- a CARB communication
 - b Customer-specific communication

- Layers of the OSI reference model
- 7 Application
 - 6 Presentation
 - 5 Session
 - 4 Transport
 - 3 Network
 - 2 Data link
 - 1 Physical

Application protocols

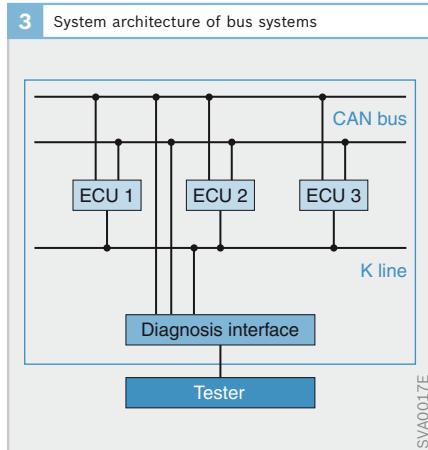
The increasing complexity of control-unit functions places high demands on the tuning and optimization of these systems to suit the types of engine and vehicle concerned. This process is known as application engineering. For this purpose, an application system is connected to the electronic control unit. The application protocol for the K line is the McMess protocol, and the CCP protocol for the CAN.

McMess

With the McMess protocol, parameters in the electronic control unit can be modified and operands recorded. In a special application mode, the protocol describes the method and contents of communication between application system and electronic control unit for the K line. With McMess, for example, the measuring device is able to read the contents of the variable memory (e.g. sensor values, measured values) quickly.

CCP (CAN Calibration Protocol)

The CCP protocol describes the contents of communication between an application system and an electronic control unit if these communicate on the CAN bus. CCP allows electronic control units in the network to function at higher payload rates than on the K line.



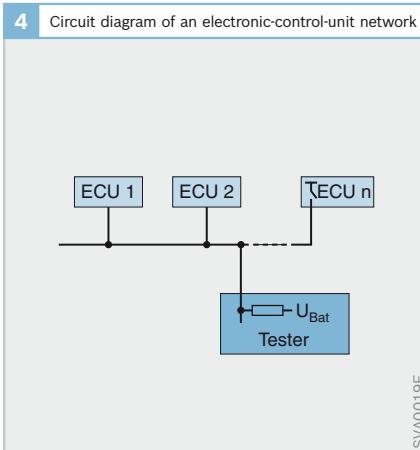
Communication on the K line

Physical layout

The diagnostic tester connects to one or several electronic control units using the K-line system (Fig. 3). All electronic control units have equal rights. Communication flows bidirectionally between the tester and the electronic control units. This means that data can be received and sent by all subscribers. However, only one subscriber is ever able to send at a given time.

The communication is an asynchronous process. There is no additional line for the transmission of a clock signal. The baud rate on the K line is between 1,200 and 10,400 baud in accordance with ISO standards. In special applications (e.g. manufacturer-specific flash programming), it could even be as high as 250 kBd.

Data transfer on the K line is realized as follows: the resistor in the tester increases the potential of the line – provided no electronic control unit is sending on it – to battery-voltage level (Fig. 4). If an electronic control unit is activated, it connects its K-line connection to the shared ground and therefore switches the potential of the K line to ground. Each of the connected electronic control units is able to detect that one of the bus subscribers is actively accessing the K line.



Message format

The message format of the KWP 2000 protocol comprises three parts (Fig. 5):

- ▶ Header
- ▶ Data bytes and
- ▶ Checksum

Header

The header is made up of no more than four bytes.

- ▶ The format byte (Fmt) contains information about the form and composition of the message. Two bits (A0 and A1) provide address information, while six bits (L0 to L5) provide the quantity of data bytes. With this coding, there can be up to 63 data bytes. If L0 to L5 are set to zero, the number of payload bytes must be coded using the length byte (Len).
- ▶ The target address (Tgt) indicates the communication partner for which the information is destined.
- ▶ The source address (Src) indicates the communication partner that sent the information.
- ▶ The length byte (Len) specifies the quantity of payload bytes (max. 255 bytes). Up to a value of 63, the coding can take place in Fmt. For more than 63, the length must be specified using this byte.

The existence of a target and source address as well as the length byte is dependent on the parameters from the format byte.

Data bytes

Up to 255 bytes of payload can be sent per message. The first byte of the payload is always a service-identification byte (SID). The subsequent data bytes are data that differ depending on the service concerned.

Checksum

The 1-byte checksum is always positioned at the end of a message. It contains the value of the modulo 256 sum across all bytes in the message - apart from the checksum itself.

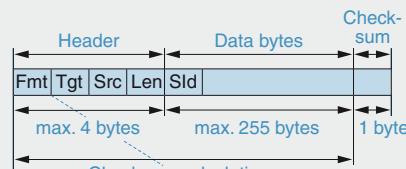
Initialization

To be able to send messages, each electronic control unit must support a subset of the possible message formats. At the beginning of communication, the format supported by the electronic control unit is communicated to the tester by means of key bytes as part of initialization.

The ISO standards specify which of the possible initialization procedures are customary for ECU diagnostics. The following initialization procedures are supported:

- ▶ 5-baud initialization and
- ▶ Fast initialization

5 Message format of the KWP 2000 protocol



0 0	No address information
0 1	Exception mode (CARB)
1 0	With address information, physical addressing
1 1	With address information, functional addressing

These different initialization methods are used either for CARB diagnostics or for after-sales diagnostics. Furthermore, a distinction can be made between physical initialization and functional initialization, which introduces the topic of “point-to-point” or “point-to-multi” communication. With functional initialization, a group of electronic control units is addressed and initialized. A function initialization can only be successful if all the electronic control units of this group support the same baud rate, the same transmission timing and the same protocol. By contrast, only one electronic control unit is initialized in the case of physical initialization.

5-baud initialization

Communication is initiated by means of a 5-baud address sent by the tester. It is thus possible for each electronic control unit to be addressed individually and addressing is clearly separate from the data transfer.

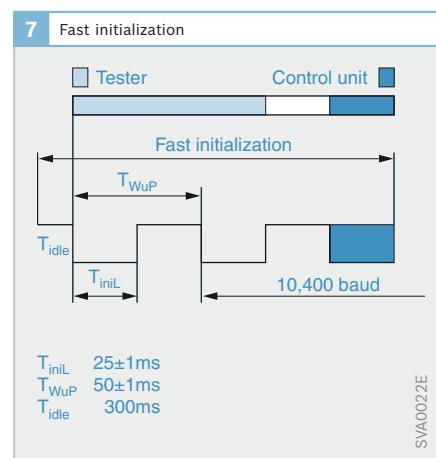
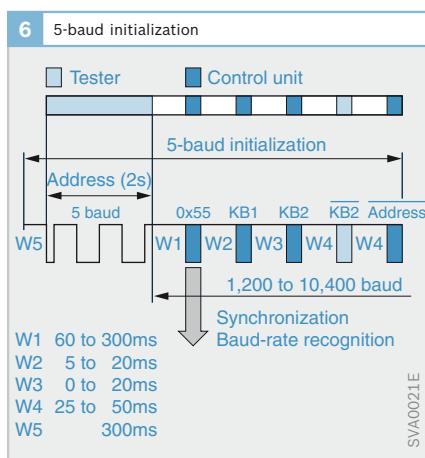
The tester outputs the address onto the K line at a transfer rate of 5 baud (Fig. 6). 10 bits (eight data bits, one start and one stop bit) take approx. 2 seconds.

On receiving the address, the electronic control unit responds with the baud-rate

recognition pattern. From the bit sequence 01010101 (binary format), the tester determines the baud rate of the electronic control unit, which may be between 1,200 and 10,400 baud. The electronic control unit then transmits the two key bytes, which communicate the header format and timing method that the electronic control unit supports. As acknowledgment that communication has been successfully established, the tester sends the complement of the second key byte back to the electronic control unit and, in return, receives from the electronic control unit the complement of the address. Initialization is then complete and the regular data transfer may begin.

Fast initialization

Fast initialization establishes communication in a shorter time than 5-baud initialization. With fast initialization, the tester sends a wakeup pattern (WuP) comprising a low and high phase, each lasting 25 ms (Fig. 7). At the end of the WuP, the tester sends the StartCommunication Service at a rate of 10,400 baud. The electronic control unit returns a positive reply containing the key bytes. Initialization is then complete and regular communication may begin.



CARB initialization

CARB initialization is a special case of 5-baud initialization, although CARB may also be activated by fast initialization. A functional initialization of all connected electronic control units is achieved with the fixed address 0x33. The baud rate is permanently set at 10,400 Bd. The requirements are described in more detail in ISO 14230-4 and ISO 9141-2.

One of the outcomes of these standards is that all emission-related electronic control units fitted in a vehicle may only support one protocol: ISO 14230-4, ISO 9141-2 or SAE J1850. A mixture of protocols in the same vehicle is not permitted.

Key bytes

The electronic control unit uses key bytes to inform the tester of which header bytes, length bytes and timing method are supported. The key-byte decoding process is defined in ISO 9141 and ISO 14230-2.

Arbitration

Arbitration is a means of preventing or detecting collisions of data sent by different electronic control units at the same time. Arbitration is necessary if a correct response to the tester request is to be received in the case of functional communication (e.g. reading the fault memory of all emission-related electronic control units using a scan tool). Collisions can be expected because more than one electronic control unit is permitted to send following initialization by the tester.

Arbitration is only relevant to data transfer on the K line.

Collision prevention

To prevent collisions following a tester request, each electronic control unit checks the K line for a falling edge after a time of P2min. A falling edge may be interference, or another electronic control unit has already been sending data.

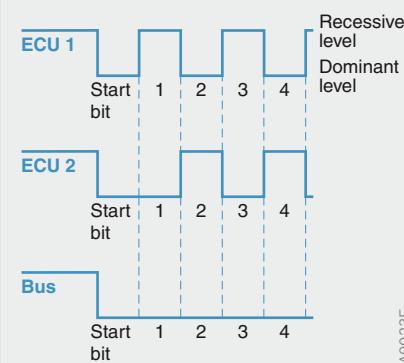
If a falling edge was detected, the electronic control unit delays sending its own data until the bus has returned to idle state and the P2min time period has again elapsed. Only then may it attempt to send its data again.

Collision detection

A collision cannot be prevented if two electronic control units send at the same time. In this event, the collision still has to be detected. This is possible by monitoring the K line. To this end, the electronic control units reread their own transmitted byte through the serial interface.

In Figure 8, two electronic control units (ECU 1 and ECU 2) send a start bit at the same time. ECU 2 then sends a 0 (dominant bit) and ECU 1 a 1 (recessive bit). The 0 appears on the K line. With bit 2, the exact opposite is true. Here, ECU 1 persists with the dominant level. When the complete byte has been transferred, the electronic control units compare the byte sent on the K line with the byte they have read. In the example, a recessive level was overwritten by a dominant level for both electronic control units. From this, both electronic control units deem the transfer to have been faulty and retreat from sending.

8 Sequence of unsuccessful communication



To prevent cyclical collisions, the electronic control unit reports after a time of P2min, which is calculated in accordance with a specific algorithm.

Let us assume that ECU 2 in the previous example sent a 1 as its first data bit. ECU 1 would then receive nothing from the data transfer of ECU 2 and would continue to establish communication. ECU 2, however, detects the collision in bit 2 and retreats. The collision in this event is non-destructive.

Communication on the CAN

Physical layout

A CAN interface containing the necessary hardware already exists in many electronic control units. This bus system can also be used for diagnostics. It is possible to communicate with the electronic control units using a diagnostic tester connected to the bus by a diagnosis connector.

Baud rates

The typical baud rates associated with CAN are 500 kBd or 1 MBd.

Addressing and message types

Tester communication on the CAN bus is defined by ISO 15765. The communication services of ISO 15765-3 or 14 229-1 are defined in a similar way to those of ISO 14230-3 (Fig. 2).

The fundamental difference between these protocols is in the format of messages and how they are transmitted. While up to 255 data bytes can be transmitted in a message on the K line, only eight data bytes are possible with the CAN. To reproduce the service messages on the CAN, the contents (data bytes) of the header and the checksum are separated and embedded in a similar, new message format. The address of the electronic control unit is now the CAN identifier itself. With different CAN identifiers, it is now possible to support functional or physical addressing.

Different addressing methods exist: normal and extended addressing, whereby normal addressing is the regular addressing method and is analyzed below. Longer messages that do not fit into a CAN frame (seven bytes with normal addressing, six with extended addressing) are segmented into several CAN frames and recomposed by the recipient.

Unsegmented messages

If the quantity of data bytes to be transmitted is no greater than seven bytes (for normal addressing), a single frame is transferred. The first byte to be sent is the PCI (Protocol Control Information). It identifies the frame in the high nibble (top four bits of a byte) as being a single frame and indicates the data length (DL) in the low nibble (Fig. 9).

The data is sent by the transmitter to the receiver in a single message (Fig. 10a).

Segmented messages

If more than seven data bytes (for normal addressing) are to be delivered, the diagnosis tester (client) sends a first frame on the CAN data bus (Fig. 10b) first of all. The electronic control unit (server) acknowledges with a flow-control frame. The other data is subsequently sent in consecutive frames.

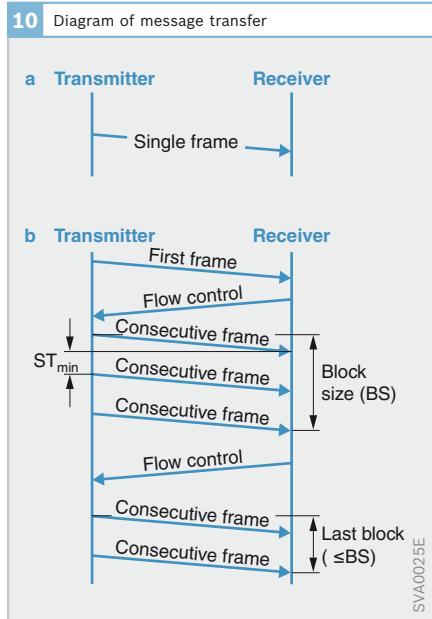
9 Format of an unsegmented message

0	1	2	3	4	5	6	7
PCI	Data						
7	6	5	4	3	2	1	0
0	0	0	0	DL			
Single frame							

The first frame contains the PCI, an additional length byte (DL, Data Length) and the first six data bytes (Fig. 11). In the high nibble, the PCI contains the information for identifying the frames as the first frame. The extended data length (XDL) is stored in the low nibble. Together with the DL, a 12-bit data word is formed with which it is possible to express values from 0 to 4,095.

The transmitter sends the first frame and waits for an acknowledgment of receipt from the receiver in the form of a flow-control frame. This message again contains the PCI, the high nibble of which identifies the frame as the flow-control frame. The low nibble contains the flow status (FS), which can be used to authorize or delay the sending of further frames (consecutive frames).

BlockSize (BS) indicates how many consecutive frames can be received without the need for a further flow-control frame to be sent. The values expressed in ST_{min}



(Separation Time) specifies the interval to be maintained between consecutive frames.

The flow-control frame is followed by the consecutive frames. Again, the PCI in the high nibble contains the identifier, while the low nibble contains the sequence number (SN). With the first consecutive frame to be sent, the SN is set to 1. The SN is incremented with each subsequent consecutive frame; after 15, the SN restarts from 0. By evaluating the sequence number, the receiver is able to detect whether or not all the frames have arrived.

11 Format of a segmented message

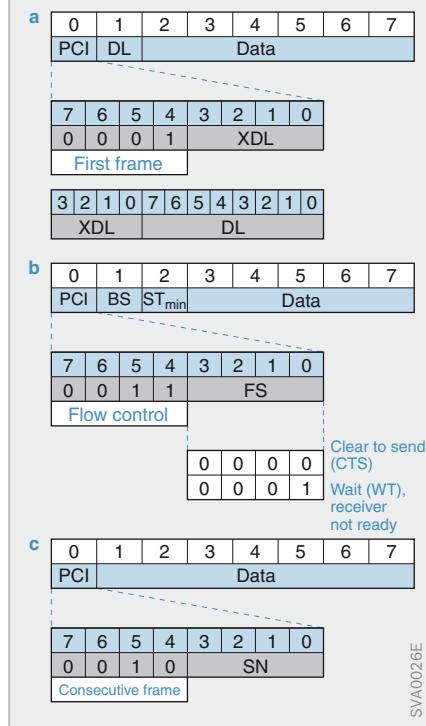


Fig. 10
a Message transfer in a single frame (unsegmented message)
b Segmented message transfer

Fig. 11
a First frame
b Flow control
c Consecutive frame

Architecture of electronic systems

Over the last three decades, tremendous progress has been made in automotive engineering. Modern injection and exhaust-gas treatment systems drastically reduced pollutants in the exhaust gas, while occupant-protection and vehicle stabilization systems improved safety on the road. Much of this success is due to the introduction of electronically-controlled systems. The proportion of these systems used in cars increased continuously. The requirements of safety and environmental compatibility, but also the demand for comfort and convenience functions, will increase yet further and this will in no small part be achieved through the use of electronics. Up to around 90 % of innovations in the motor vehicle will be realized by electronics and microprocessor-controlled systems. The networking of these electronics creates the prerequisite for having this wide variety of electronic systems integrated within the complete vehicle system to form a whole. However, this results in a complexity that can only be overcome at considerable expense.

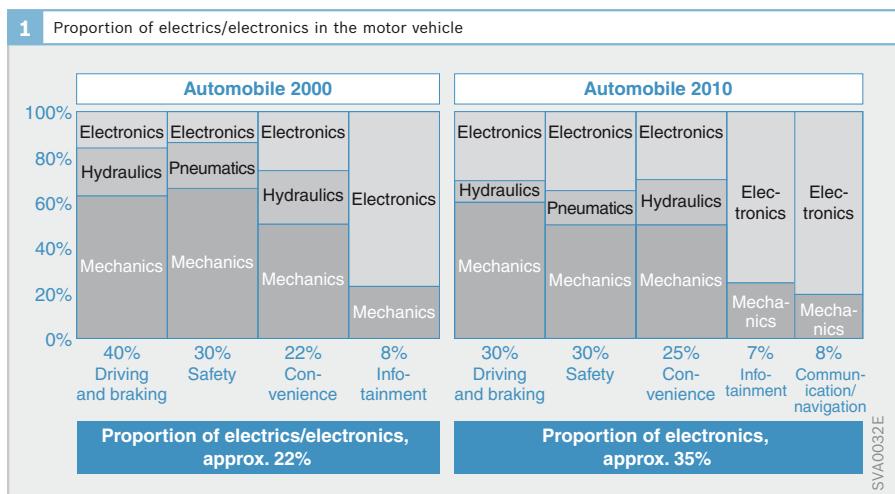
Overview

History

The on-board electrical network of a car around the year 1950 comprised approx. 40 lines. Essentially, cables were only required for the battery, starter, ignition and the lighting and signaling systems.

With the first electronic injection and ignition systems, cabling complexity began to increase fast. Sensors fitted in the engine compartment (e.g. speed sensor, engine-temperature sensor) had to deliver signals to the engine control unit, while the fuel injectors required their triggering signals from the electronic control unit.

A further increase in cabling complexity resulted from the introduction and rapid widespread adoption of the antilock brake system (ABS). Meanwhile, comfort and convenience systems, e.g. electrical power-window units, would also form part of the standard equipment. All these systems require additional connecting lines for the connection of sensors, control elements and actuators to the control unit.

**Fig. 1**

Source:
Mercer management consulting

Technology of the present day

In the 1990s the cabling work in a luxury class vehicle amounted to around 3 km. This figure clearly demonstrates how complex the vehicle system has become. The growth of the proportion of electronics in the motor vehicle (Fig. 1) can mainly be attributed to the growth in microelectronics and sensor technology.

At first, many of the new systems were integrated into the vehicle by means of their own dedicated electronic control unit. For the most part, the individual electronic control units operated in mutual independence. All the same, connecting lines became increasingly necessary between electronic control units to enable the exchange of data by means of PWM signals, for example. Depending on the vehicle class, there are between 20 and 80 electronic control units fitted in today's vehicles. They control such equipment as the engine, antilock brake system or the airbags. The number of microcontrollers in the vehicle has therefore risen continuously in recent years (Fig. 2).

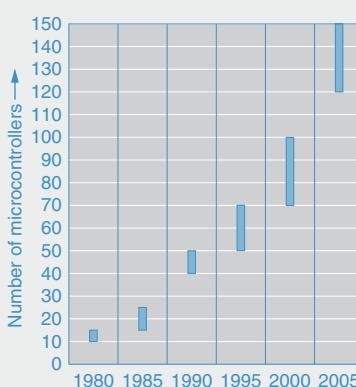
The components of the individual systems are optimally matched to each other. The systems may originate from different manufacturers that use previously agreed, albeit still their own, interfaces. The rain sensor, for example, "speaks" in a different way to the sensors for the engine management. The following example demonstrates just how networked the functions in a modern vehicle are: the radar sensor of the adaptive cruise control system (ACC) measures the distance to the vehicle traveling in front. If this distance is shorter than a specified minimum distance, the ACC electronic control unit sends this information to the engine management, the ESP electronic control unit and the airbag electronic control unit. The engine management reduces torque and thus driving speed. If this is not sufficient, the electronic stability program (ESP) must also generate brake pressure to decelerate the vehicle. If the distance continues to

shorten, the airbag and seat-belt pretensioners are set to emergency standby.

The communication between the electronic control units cannot take more than fractions of a second. The more electronic control units interact in the one complete system, the more difficult it becomes for them to communicate undisturbed.

With the number of electronic control units and the associated need for mutual communication, the costs of developing the systems rose as did the adaptation costs for making interfaces compatible. With the CAN bus (Controller Area Network) developed by Bosch, a powerful and widely used data bus system has become commonplace in vehicles for the first time. The data line of the CAN bus makes it possible for the electronic control units to exchange specific and relevant items of information with each other. At the start, the network only comprised a few electronic control units, such as the engine-management system, the electronic stability program and the transmission control. Gradually, further systems would expand this network, especially in the areas of comfort and convenience and infotainment. The CAN bus has gradually evolved into the standard for networking systems in the motor vehicle. Today it is the standard for communication between elec-

2 Number of microcontrollers in the motor vehicle



tronic control units within different areas of the electronics (drivetrain, suspension, body electronics and infotainment) and forms a powerful backbone for networking these areas with each other. Additional bus systems (e.g. LIN bus, MOST bus) are used as subbuses or for transmitting at high data rates with comparatively low realtime requirements in the motor vehicle.

Development trends

The proportion of electrics and electronics in the motor vehicle will continue to increase. In the drivetrain, the number of components in the exhaust line (e.g. exhaust-gas sensors) is increasing due to stricter exhaust-emissions legislation. While the demands for reductions in fuel consumption can, for example, be fulfilled by means of new valve-gear concepts, even this requires additional electronic components. A further increase in the proportion of electronics results mainly from the growth of electronic systems in the areas of safety, comfort and convenience, and infotainment.

Objectives

Drivers demand a high level of reliability from a car. The vehicle manufacturer and the supplier of assemblies, meanwhile, are constrained by other requirements such as minimization of manufacturing costs, space restrictions and the weight of components. An opportunity to fulfill these requirements in the face of the increasing complexity of the "vehicle" system is seen in the shift of the traditional implementation technologies of mechanics, hydraulics and electrics towards microprocessor-controlled, electronic systems. For this reason, the development of software will continue to gain in importance in future.

The current situation in the electrical and electronic architecture of motor vehicles is characterized by an increase in functionality and an increasingly strained costs situation. To achieve both of these

objectives simultaneously, development partners are more frequently tapping into resources that are already available in subsystems. These can be sensors or actuators as well as realized functions that are available to different systems over the communications network. For new systems and functions, manufacturers strive to get by on a minimum of additional resources.

In the meantime, engineers are faced with a new challenge in the form of "networked" thinking and subsystem integration, especially when the assemblies for the subsystems originate from different development partners (suppliers).

Complaints in the field (i.e. with series-production vehicles) due to electrical or electronic failures could be the consequence of not having taken the interactions of the subsystems into consideration. The causes - unmanageable behavior of functionality spread among networked systems, and their integration - are avoidable through the logical application of certified development processes as early as in the specification phase. Furthermore, modeling and tools for authoring a formal description of architectures are gaining ever more in importance.

Broadened requirements for a complete motor vehicle system in the future are leading to increased networking of vehicle components and subsystems. In this regard, new functions are being developed that go beyond the frontiers of traditional applications - and this is without additional expenditure on hardware wherever possible.

New development methods and technologies are required to make this achievable. With a top-down approach, new functions are viewed from the perspective of the complete vehicle. This means that, in accordance with the method of systems engineering, functional requirements and non-functional requirements (e.g. quality objectives, safety, costs, etc.) are set for the vehicle as a whole and derived as specifi-

cations for the subordinate subsystems. These requirements are formulated as a model and can thus be used as a specification for the subsystems and the creation of test cases. This is what is known as an “executable specification”, which makes it possible to prove the completeness and the traceability of the requirements, for example, or to identify the requirements for interaction and communication between subsystems. In this way, it is possible to form an optimized architecture for the complete vehicle and its subsystems and components. The functional relationships between the complete motor vehicle system and the subordinate subsystems can be surveyed in different levels of detail and suitable interfaces can be defined for the functions. This approach supports an expanding networking of functions. Synergies are exploited between vehicle areas (domains such as the drivetrain, interior, infotainment) that were hitherto considered in isolation and resources are spared.

As an element of the development process that works in the opposite direction, the generation of new functions from available resources and existing systems (bottom up) should also be taken into consideration to minimize innovation risks. This is how new functions are integrated into existing systems, for example. Examples of this approach are measures to avert the consequences of an accident by “preparing” subsystems for an imminent crash (closing windows, closing the sliding sunroof, activating the airbag, etc.) or the assistance of the driver in emergency braking situations in ESP in future. In this way, it is possible to reduce the number of electronic control units and counteract rising system costs.

The development process described characterizes the CARTRONIC® concept that Bosch developed in the 1980s. The results of this concept are being incorporated into the Autosar Initiative (see Autosar Initiative)

Vehicle system architecture

Architecture

The architecture of a system represents its “construction plan”. It describes the structural and dynamic system characteristics as a whole. The architecture is usually specified in a description language. Special draft mechanisms are used for specific requirements. With architecture being a construction plan for different realization technologies and a means of proving that functional and nonfunctional requirements have been fulfilled in the system draft, different views of the system architecture are required. Examples of this include:

- ▶ Hardware architecture
- ▶ Software architecture
- ▶ Network architecture in the area of realization technologies
- ▶ Cost and resource consumption in the area of economical analysis and
- ▶ For the area of social requirements, aspects such as safety, availability and legal conformity

The problems that arise in the integration of differently structured subsystems can be reduced by means of an architecture.

Functional structure

The domain of vehicle motion has the task of ensuring the controlled movement of the vehicle as well as its directional stability. This task can be subdivided into various levels (Fig. 3).

The navigation level is home to the planning tools for the driving route. These are merely informational in nature and have no interventional influence on vehicle motion.

At vehicle guidance level, the decisions of the driver are implemented by means of the steering wheel and accelerator pedal but also various assistance systems for vehicle handling (e.g. ACC, course stability

systems). At this level, the driver is able to overrule the assistance systems at any time.

At the stability level, there are the subsystems that are able to correct the decisions taken at handling level if these happen to be outside the range of safe reference variables (e.g. ABS, ESP). This may be the case when cornering or on wet road surfaces, for example.

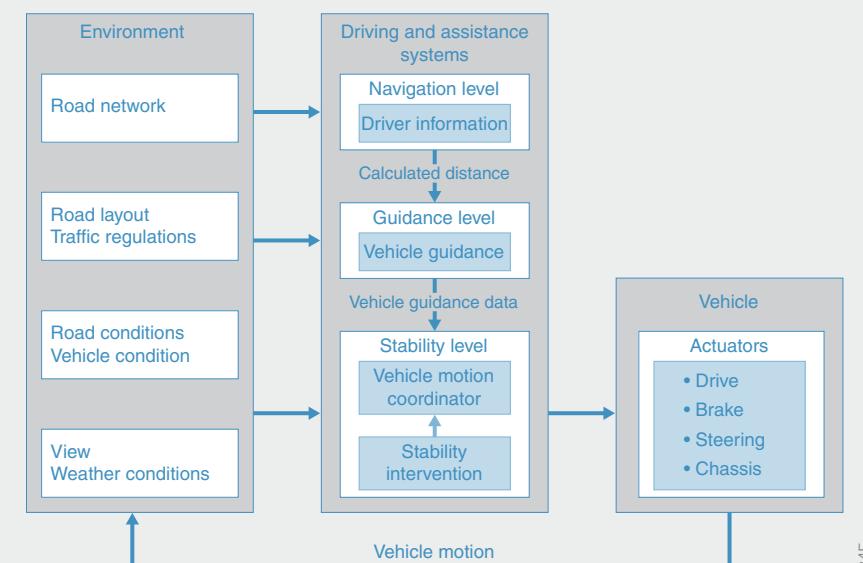
At stabilization level, correcting variables for implementation by the vehicle's actuators are determined. Information about the environment (e.g. road condition, air temperature, rain sensor signal) is still required at the various levels for the implementation of the relevant tasks.

These tasks can be assigned to functional components, which are the architectural elements of the functional architecture. In this way, the driver information

functional component represents the tasks of the navigation level, which are to inform the driver of the driving route determined by means of a mapping system (Fig. 3). Vehicle guidance represents the guidance level, and stability intervention the tasks of the stabilization level. The vehicle motion coordinator determines the correcting variables for the actuators, e.g. of the drive and electronic stability program (ESP), from the information input by vehicle guidance and stability intervention.

Figure 4 shows how the functional components of guidance level, stabilization level and vehicle actuators are related in a hierarchical structure within vehicle motion. Communication relationships between the components and interactions with other domains, e.g. body and interior, are also featured in the model.

3 Levels in the vehicle motion domain



In the same way as Vehicle motion is refined, these functional components require further detailing until the refined components represent manageable, clearly delimited tasks that make flexible, modular implementation possible through different realization technologies. Defined interfaces between the components enable communication and the exchange of data. For example, the transmission control issues a request through the engine-management system for a specific reduction in torque during a gearshift. This value is exchanged as a physical variable via the interface.

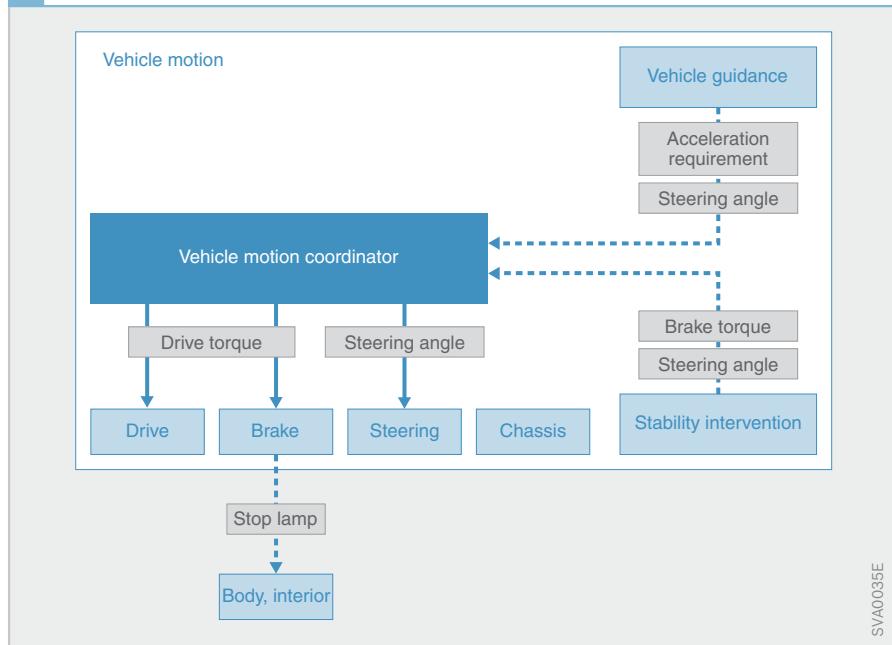
With its integration into a suitable procedural model, the functional structure is the starting point for subsequent stages in the development process.

Systematic creation of EE system architectures

The increasing amount of networking in traditional vehicle domains for the realization of new functions can be illustrated using the ACC (Adaptive Cruise Control) driver-assistance system as an example. Adaptive, same-lane driving is made possible by the networking of a combined cruise and distance control system with the engine-management system, brake system, transmission and cockpit. Here, subsystems from the drivetrain, chassis and infotainment (interaction with the driver) domains are used to realize the new function with minimal cost.

The decision as to whether a function (e.g. ACC) is realized in a dedicated logic close to the sensor or in one of the existing, subscriber electronic control units has no bearing on the function itself.

4 Example of a functional structure for the vehicle motion domain



Rather, the decision is affected by non-functional requirements such as safety, availability, costs or resource availability. In addition to the functional requirements, these requirements mainly determine how the function is realized. The "how" is described by the architecture of the system. Different requirements result in different system architectures.

CARTRONIC® concept

With the CARTRONIC® architecture concept, all closed and open-loop control tasks in the vehicle have been structured in accordance with logical, functional viewpoints and modeled in the form of a functional architecture. Delimited functions (and their dependencies) that implement specific functional requirements have been represented by defined architectural elements. The functional structure, i.e. the structural description, represented a hierarchical decomposition of the subsystems down to manageable size. Interactions between elements of the functional structure have been described by communication relationships. Since the use of the architecture concept could have led to different functional structures, it was essential to reach agreement on the tasks and interfaces. It was necessary to choose interfaces that were based on physical variables and thus supported aspects such as reusability and interchangeability.

The motor vehicle system with all its open and closed-loop tasks was dismantled into subsystems that implement clearly defined tasks. These subsystems include the engine management, brake system, transmission control, ACC, lighting management, etc. Different levels of functional structure detail can be assigned to the system and subsystem levels (Fig. 5). It was therefore possible to create development frameworks for selected functional components and component groups on which to base implementation in the form of partial real-

ization stages. This required a decoupled development process and the exploitation of synergies between subsystems. The development frameworks took into consideration the dependencies and interface contents within the individual domains and with the rest of the vehicle, as is the case with a networked system such as ACC, for example.

Bosch introduced this concept to the Auto-sar Initiative (Workpackage 10.x).

Software architecture

The independence of the functional structure, or architecture, from the later realization stage results in a decoupling of functionality and technology and thus forms the first stage of a model-based development process. The functional structure can be used on several occasions and expanded as the foundation for drafting system architectures. This architecture is characterized by architecture drivers (specific criteria of the architecture) that are essentially the product of nonfunctional requirements (e.g. costs, quality, reusability, relocatability).

Further precision of the development frameworks devised from the functional structure, and of their interfaces in particular, is required if it is to be possible to evaluate an electronic control unit for the relocatability of functions and the integration of software, which is a contribution of various participants in the project. While retaining the realization-independent information from the functional structure – such as an agreed torque interface – the frameworks are supplemented by realization-specific information such as data type, quantization, runtime properties or resource requirements.

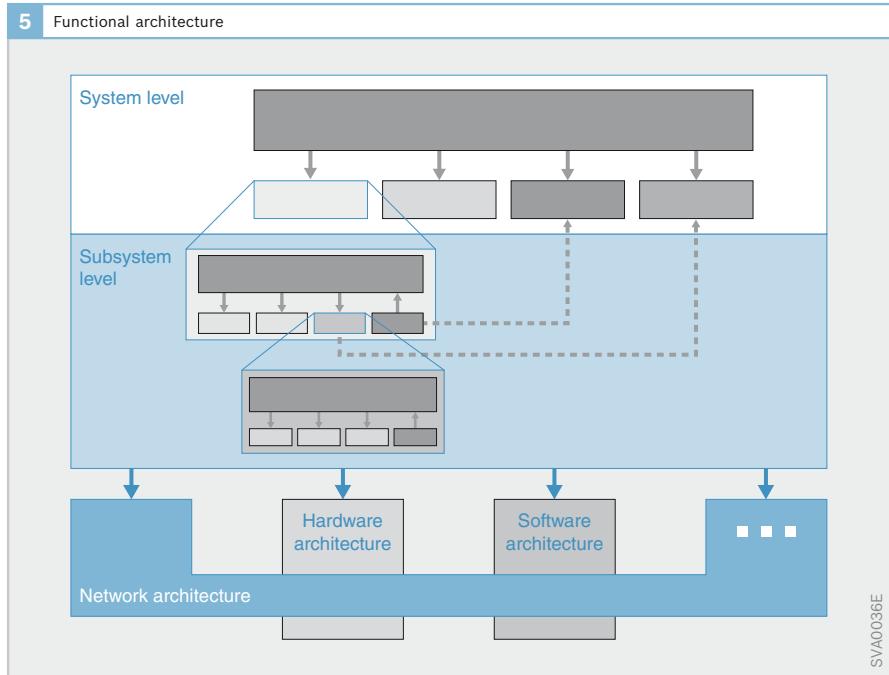
In the same way as hardware and disciplines such as mechanics or hydraulics, software can also be classified as a realization technology. Product-line or platform approaches have long been a familiar

feature of mechanical development or production optimization and their application is virtually universal. The trend towards software-based system solutions and improvements, in conjunction with the vastly expanding software scopes of electronic systems, has given rise to the demand for this strategy to also be transferred to software-intensive systems.

The decisive challenge faced today is less to do with technical feasibility but more about how to develop methods further and apply already developed methods and processes in product developments in a systematic and disciplined way, and to anchor them within the organization. The product-line approach in software development was transferred to the motor vehicle domain with the participation of Bosch with methodological support from the software engineering institute (SEI). The method will be used systematically in future Bosch product generations.

Network architecture

With the spreading of open standards such as the CAN bus, the integration of functions into application-specific electronic control units, and satellites linked by subnetworks, network architecture has become the synonym for the complexity management of distributed systems. Extensions and “attachment solutions” are easily integrated until the limits of network capacity are reached. If these possibilities were to be exploited without checking the system draft, this would result in unmanageable increases in complexities and integration conflicts. Biological systems solve these unmanageable complexities through specialization and the creation of subnetworks with new forms of organization. Their objectives are stability and the ability to survive. This model has, to a certain extent, evolved on its own in motor vehicle systems through assignment to traditional



fields of application or domains and the comparatively slow growth of networking within these domains.

Bus systems for the individual domains are becoming more specialized due to their plainly different requirements. With the CAN in the drivetrain as the point of origin, new bus systems such as the LIN sub-bus have begun to infiltrate the area of body electronics or FlexRay in the case of safety-relevant x-by-wire systems. In the multimedia field, where demands for high data rates but low safety requirements prevail, bus systems such as Bluetooth have started to make an appearance.

Breaking through these traditional domains with ever more applications leads to known consequences, e.g. dramatic increase in complexity, high start-up costs, increasing integration times and costs, and more demanding work in customer service as a consequence of diagnostics no longer being manageable. A solution for these multidimensional optimization tasks has in the past been sought in the software field. In the case of technical systems in particular, the paradigm is still king, especially in software realizations, because the absence of physical boundaries supports unlimited growth.

Autosar Initiative

The Autosar Initiative (AUTomotive Open Systems ARchitecture) was founded in July 2003 by several vehicle manufacturers and suppliers - Bosch among them. Their global objective is the joint development of an open system architecture for future automotive applications. The aims of the partnership include the standardization of fundamental system functions (basic software) and function interfaces; they will replace the company-specific, individual solutions used to date. Model-based concepts and methods ought to reduce complexity in spite of an expanding range of functions. The demands for quality and reliability are fulfilled by the multiple use

of proven standards. Autosar concerns itself with all vehicle domains.

Based on the uniform electronics platform, which primarily consists of standard software modules, each vehicle manufacturer is then free to build its own specific content. They enable integration into the electronics network. These software functions permit differentiation between the competition.

Not only does software have to conform to the Autosar standard. The electronic control units must be built in such a way that the Autosar software is able to run on them. The Autosar members are hoping that the new development methods yield such benefits as shorter development times and lower development costs.

Until now, it was often the case that dedicated electronic control units would be developed and fitted for new functions (e.g. electronic transmission control, antilock brake system, air conditioning). The number of electronic control units fitted in the vehicle grew continuously; today's generation of vehicles are equipped with between 20 and 80 electronic control units. In future vehicle generations, it is intended that all functions be covered by a network of 10 to 20 electronic control units. Some of these will function a little like main computers that will bundle the important function groups together. These include the drivetrain, suspension management system, body and interior and the multimedia/telematics domain. On data buses, sensors with integrated electronics output processed and verified signals, while the buses carry the relevant control commands to actuators with integrated triggering electronics.

In future, new functions will often be able to use the existing computer architecture up to its performance limit and will be widely realized in the form of a software add-on. This would therefore render unnecessary the additional electronic control unit that would have been required

today. The system only needs to be supplemented by the sensors and actuators required.

Software will no longer be an inevitable component of hardware, but will increasingly become a stand-alone product. The first examples of business and collaborative models between supplier and manufacturer have already been put into practice at Bosch, e.g. in drivetrain management.

Examples

Individually-controlled drive components at each of the wheels with different wheel positions and wheel loads permit optimum use of tire force potential. This results in increased driving dynamics and safety while at the same time reducing consumption, wear and emissions. For this to be possible, all active elements in the drivetrain, suspension and steering must be networked.

One example of superordinate functions realized by networking is the ASIS (Active Shift Strategy) drive strategy for automatic transmissions in passenger cars. Based on the evaluation of various control elements (e.g. accelerator pedal) and conclusions drawn from the information of other systems (e.g. cornering detection from the wheel speeds), this strategy is able to control the gearshift in such a way as to meet the driver's real-time demand for agility and power through selection of the appropriate gear. In future, the telematics will be able to support additional, improved driving strategies, e.g. through the use of GPS signals for transmission control in terms of predictive driving.

Outlook

Increasingly greater demands for safety, reliability and availability are being placed on the network architecture of modern vehicles. This is where energy network architecture will play a key role. In the face of vehicle functions increasingly being realized electronically, this architecture makes for a reliable supply of power to systems and thus forms the basis for the reliability and safety of future systems. One possible future technology is the transmission of power and information on the supply line. The following benefits arise from the powerline communication (PLC) concept used in the public grid:

- ▶ Weight and cost reductions as well as space savings from the discontinuation of data lines
- ▶ Easier retrofitting for retrofit systems (spare parts trade)
- ▶ Reduction in complexity of the wiring harness in respect of manufacture and installation
- ▶ Increase in system safety, especially in mechanically stressed zones (e.g. door, mirror) that are characterized by premature aging of lines and increased risk of failure
- ▶ Powerline as a redundancy path for systems relevant to safety
- ▶ Simultaneous, or parallel, implementation of several bus systems or services, e.g. diagnostics

Mechatronics

The term “mechatronics” came about as a made-up word from mechanics and electronics, where electronics means “hardware” and “software”, and mechanics is the generic term for the disciplines of “mechanical engineering” and “hydraulics”. It is not a question of replacing mechanical engineering by “electronification”, but of a synergistic approach and design methodology. The aim is to achieve a synergistic optimization of mechanical engineering, electronic hardware and software in order to project more functions at lower cost, less weight and installation space, and better quality. The successful use of mechatronics in a problem solution is dependent upon an overall examination of disciplines that were previously kept separate.

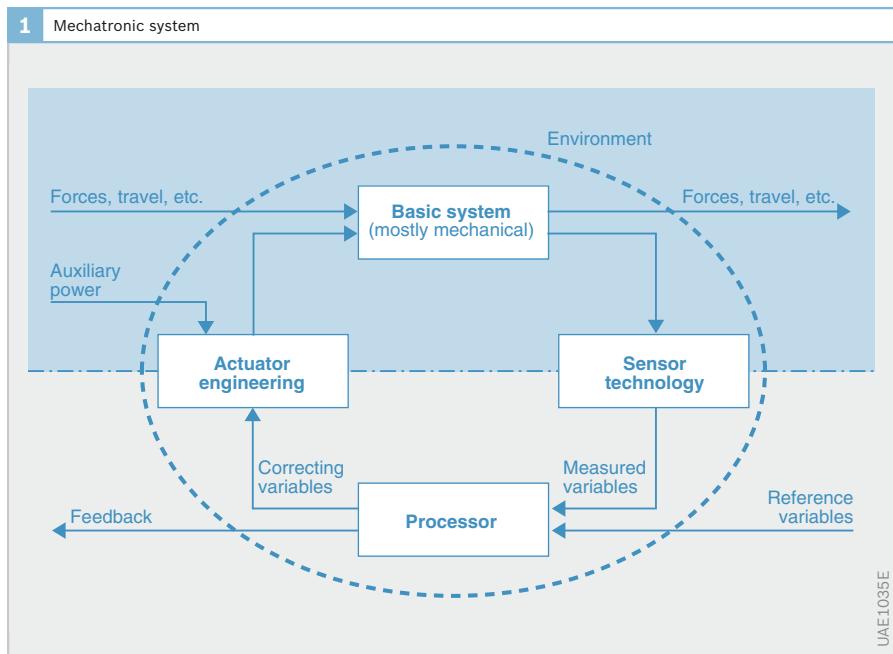
Mechatronic systems and components

Applications

Mechatronic systems and components are now present throughout almost the entire vehicle: starting with engine-management systems and injection systems for gasoline and diesel engines to transmission control systems, electrical and thermal energy management systems, through to a wide variety of brake and driving dynamics systems. It even includes communication and information systems, with many different requirements when it comes to operability. Besides systems and components, mechatronics are also playing an increasingly vital role in the field of micromechanics.

Examples at system level

A general trend is emerging in the further development of systems for fully automatic vehicle handling and steering: more and more mechanical systems will be replaced by “X-by-wire” systems in future.



A system that was implemented long ago is the “Drive-by-wire” system, i.e. electronic throttle control.

“Brake-by-wire” replaces the hydromechanical connection between the brake pedal and the wheel brake. Sensors record the driver’s braking request and transmit this information to an electronic control unit. The unit then generates the required braking effect at the wheels by means of actuators.

One implementation option for “Brake-by-wire” is the electrohydraulic brake (SBC, Sensotronic Brake Control). When the brake is operated or in the event of brake stabilization intervention by the electronic stability program (ESP), the SBC electronic control unit calculates the required brake pressure setpoints at the individual wheels. Since the unit calculates the required braking pressures separately for each wheel and collects the actual values separately, it can also regulate the brake pressure to each wheel via the wheel-pressure modulators. The four pressure modulators each consist of an inlet and an outlet valve controlled by electronic output stages which together produce a finely metered pressure regulation.

Pressure generation and injection are decoupled in the Common Rail System. A high-pressure rail, i.e. the common rail, serves as a high-pressure accumulator, constantly providing the fuel pressure required for each of the engine’s operating states. A solenoid-controlled injector with a built-in injection nozzle injects fuel directly into the combustion chamber for each cylinder. The engine electronics request data on accelerator pedal position, rotational speed, operating temperature, fresh-air intake flow, and rail pressure in order to optimize the control of fuel metering as a function of the operating conditions.

Examples at component level

Fuel injectors are crucial components in determining the future potential of Diesel-engine technology. Common-rail injectors are an excellent example of the fact that an extremely high degree of functionality and, ultimately, customer utility can only be achieved by controlling all the physical domains (electrodynamics, mechanical engineering, fluid dynamics) to which these components are subjected.

In-vehicle CD drives are exposed to particularly tough conditions. Apart from wide temperature ranges, they must in particular withstand vibrations that have a critical impact on such precision-engineered systems.

In order to keep vehicle vibration away from the actual player during mobile deployment, the drives normally have a spring damping system. Considerations about reducing the weight and installation space of CD drives immediately raise questions concerning these spring-damper systems. In CD drives without a damper system, the emphasis is on designing a mechanical system with zero clearances and producing additional reinforcement for the focus and tracking controllers at high frequencies.

Only by combining both measures mechatronically is it possible to achieve good vibration resistance in driving mode. As well as reducing the weight by approx. 15 %, the overall height is also reduced by approx. 20 %.

The new mechatronic system for electrically operated refrigerant motors is based on brushless, electronically commutated DC motors (BLDC’s). Initially, they are more expensive (motor with electronics) than previous DC motors equipped with brushes. However, the overall optimization approach brings benefits: BLDC motors can be used as “wet rotors” with a much simpler design. The number of separate parts is therefore reduced by approx. 60 %.

In terms of comparable cost, this more robust design doubles the service life, reduces the weight by almost half and reduces the overall length by approx. 40 %.

Examples in the field of micromechanics

Another application for mechatronics is the area of micromechanical sensor systems, with noteworthy examples such as hot-film air-mass meters and yaw-rate sensors.

Because the subsystems are so closely coupled, microsystems design also requires an interdisciplinary procedure that takes the individual disciplines of mechanical components, electrostatics and possibly fluid dynamics and electronics into consideration.

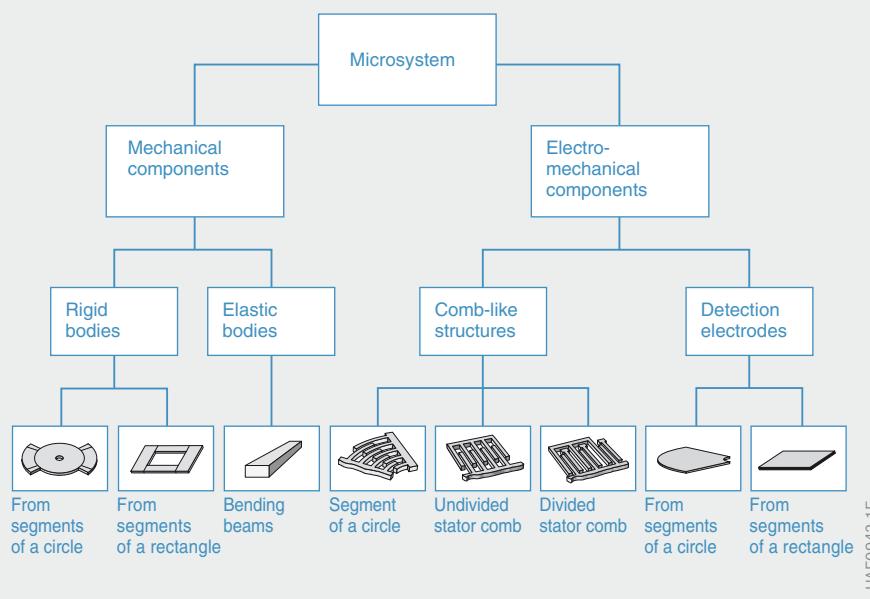
Development methods

Simulation

The special challenges that designers face when developing mechatronic systems are the ever shorter development times and the increasing complexity of the systems. At the same time, it is vital to ensure that the developments will result in useful products.

Complex mechatronic systems consist of a large number of components from different physical domains: hydraulic components, mechanical components and electronic components. The interaction between these domains is a decisive factor governing the function and performance of the overall system. Simulation models are required to review key design decisions, especially in the early development stages when there is no prototype available.

2 Model library for a micromechanical yaw-rate sensor



Basic issues can often be clarified by producing relatively simple models of the components. If more detail is required, more refined component models are needed. The detailed models focus mainly on a specific physical domain:

- ▶ This means that detailed hydraulic models of common rail injectors are available, for example. These can be simulated using special programs with numeric calculation methods that are exactly tailored to hydraulic systems. Cavitation phenomena have to be taken into consideration, among other things.
- ▶ Detailed models are also needed to design the power electronics that trigger the injector. Again, this involves the use of simulation tools which must be developed specifically to design electronic circuits.
- ▶ The development and simulation of the software that controls the high-pressure pump and the power electronics in the control unit with the aid of the sensor signals also takes place using tools that are specially designed for this area of the overall system.

As the components in the overall system interact with each other, it is not sufficient to consider specific detailed models of the components in isolation. The optimum solution is also to take into account the models of other system components. In most cases, these components can be represented by simpler models. For example, the system simulation that is focussed on the hydraulic components only requires a simple model of the power electronics.

The application of various domain-specific simulation tools during the development of mechatronic systems is only efficient if there is some sort of support for exchanging models and parameters between the simulation tools. The direct exchange of models is highly problematic due to the specific languages used for describing the models of each of the tools.

However, an analysis of the typical components in mechatronic systems shows that they can be composed of a few simple elements specific to the domains. These standard elements are, for example:

- ▶ In the hydraulic system: throttle, valve or electric line
- ▶ In the electronic system: resistor, capacitor or transistor
- ▶ In the mechanical system: ground with friction, transmission or clutch (or the equivalent for micromechanics)

The preferable solution is that these elements should be stored in a central standard model library that is also decentrally accessible to product development. The essence of the standard model library is a documentation of all the standard elements. For each element, this comprises:

- ▶ Description of physical behavior in words
- ▶ The physical equations, parameters (e.g. conductivity or permeability), state variables (e.g. current, voltage, magnetic flux, pressure) and
- ▶ The description of the associated interfaces

In addition, a major part of the environment is a reference model written in a modeling language that is independent of the tool. Overall, the library includes reference models from the mechanical, hydraulic, electronic, electrodynamic and software areas.

V model

The dependencies of the different product development phases are illustrated in the “V model”: from requirement analysis to development, implementation, testing and system deployment. A project passes through three “top-down” levels during the development stage:

- ▶ Customer-specific functions
- ▶ Systems and
- ▶ Components

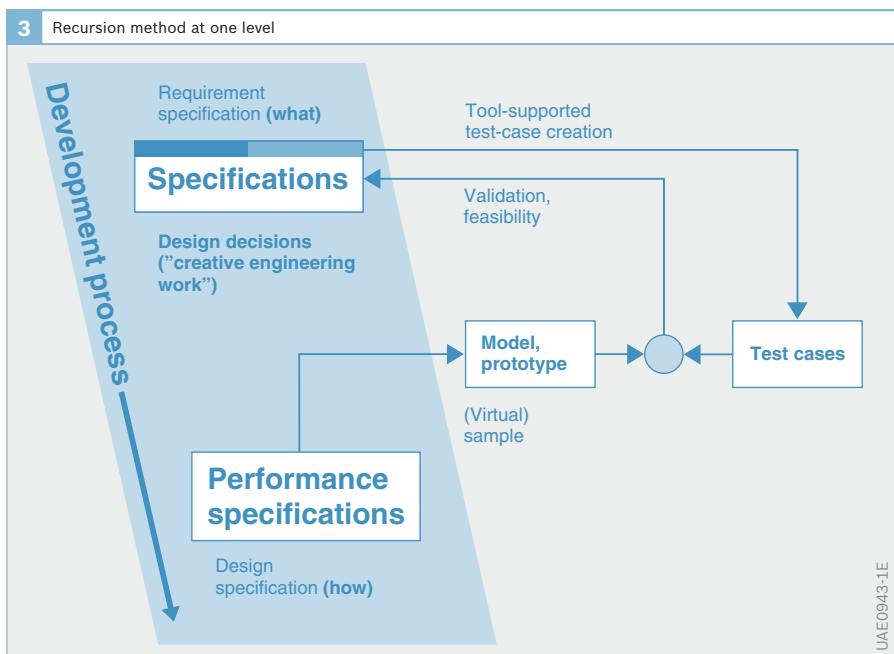
A requirements specification (what) must first be produced at each level in the form of specifications. This results in the design specification, which is drawn up on the basis of design decisions (the actual creative engineering work). The performance specifications describe how a requirement can be met. The performance specs form the basis for a model description which allows a review (i.e. validation) of the correctness of each design stage together with previously defined test cases. This procedure

passes through each of three stages, and, depending on the technologies applied, for each of the associated domains (mechanical engineering, hydraulics, fluid dynamics, electrics, electronics, and software).

Recursions at each of the design levels shorten the development stages significantly. Simulations, rapid prototyping, and simultaneous engineering are tools that allow rapid verification, and they create the conditions for shortening product cycles.

Outlook

The major driving force behind mechatronics is continuous progress in the field of microelectronics. Mechatronics benefits from computer technology in the form of ever more powerful integrated computers in standard applications. Accordingly, there is a huge potential for further increases in safety and convenience in motor vehicles, accompanied by further



reductions in exhaust emissions and fuel consumption. On the other hand, new challenges are emerging with regard to the technical mastery of these systems.

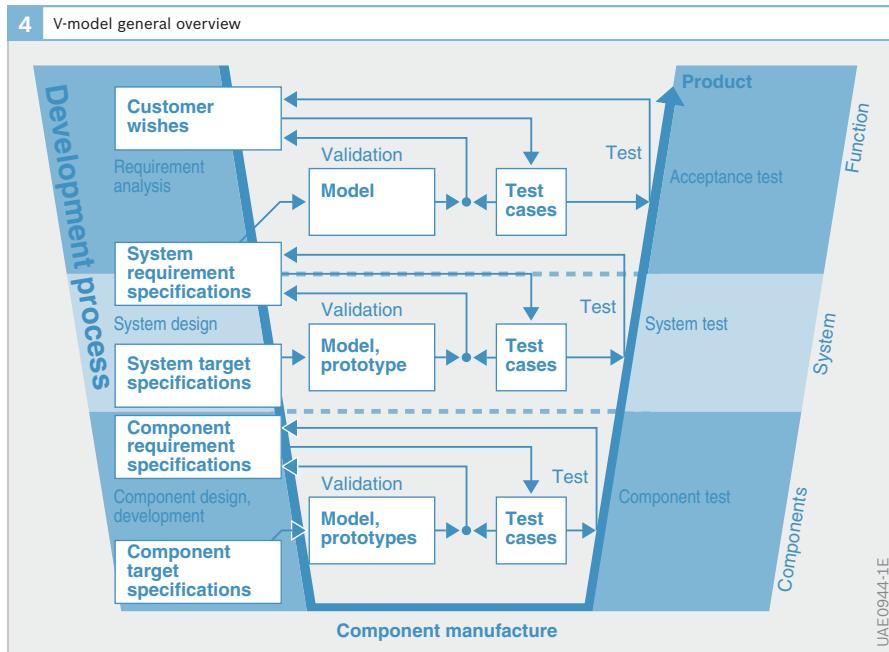
However, future “X-by-wire” systems without the mechanical/hydraulic fall-back level must also provide the prescribed functionality in the event of a problem. The condition for their implementation is a high-reliability and high-availability mechatronic architecture which requires a “simple” proof of safety. This affects both single components as well as energy and signal transmissions.

As well as “X-by-wire” systems, driver-assistance systems and the associated man/machine interfaces represent another area in which the consistent implementation of mechatronic systems could achieve significant progress for both users and vehicle manufacturers.

The design approaches of mechatronic systems should strive toward continuity in several aspects:

- ▶ Vertical:
“Top-down” from system simulation, with the objective of overall optimization, through to finite element simulation to achieve a detailed understanding, and “bottom-up” design engineering from component testing through to system testing
- ▶ Horizontal:
“Simultaneous engineering” across several disciplines in order to deal with all product-related aspects at the same time
- ▶ Beyond company boundaries:
Step by step, the idea a “virtual sample” is nearing our grasp

Another challenge is training in order to further an interdisciplinary mindset and develop suitable SE processes and forms of organization and communication.



Electronic components in the vehicle

Since the first electronic systems were introduced in the passenger car sector in the 1960s, the proportion of electronics in the motor vehicle has steadily risen. The first components to be used were predominantly discrete components. However, manufacturing costs and high requirements for system reliability soon inhibited their further development. Technological advances in microelectronics promoted the increasingly dense integration of components. This gave rise to compact, dependable electronic systems for automotive applications.

Electronic systems in the motor vehicle make high engine output possible with low exhaust emissions and low fuel consumption, they improve safety on the road and make for high levels of comfort and convenience. The basis for these modern systems are the electronic components fitted in the electronic equipment concerned. They consist of a multitude of different passive and active components, the latter characterized by semiconductor components.

Table 1

Basic principles of semiconductor technology

Electric conductivity

The specific suitability of different materials for the conduction of electricity is determined by the number and mobility of free charge carriers that they contain. The electric conductivity of solids at room temperature has a range of 24 powers of ten.

Solids are subdivided into three classes of material according to their electrical conductivity (Table 1).

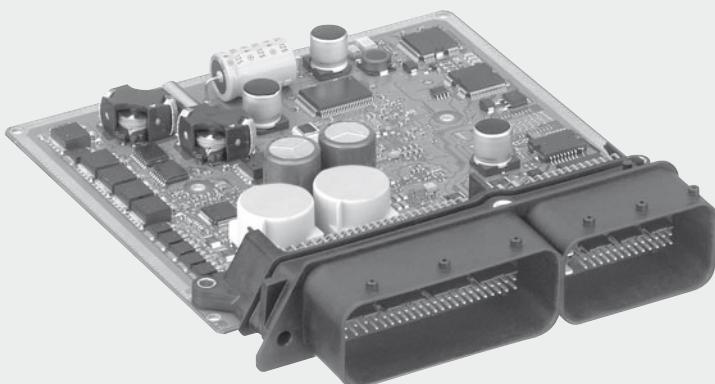
1 Classes of material based on conductivity with examples		
Conductor (metals)	Non-conductors (insulators)	Semiconductors
Silver	Teflon	Silicon
Copper	Silica glass	Germanium
Aluminum	Aluminium oxide	Gallium arsenide

All solids contain around 10^{22} atoms per cubic centimeter which are held together by electrical forces.

Conductors (metals)

In metals, the number of free charge carriers is very large (one or two free electrons per atom). Their level of mobility is moder-

1 Electronic control unit equipped with electronic components



UAE0985-1Y

ate. Due to the high number of free electrons, the electric conductivity of metals is high: good conductors have a conductivity of 10^6 siemens/cm.

Non-conductors (insulators)

In insulators, the number of free charge carriers is practically zero and consequently the electric conductivity virtually non-existent. The conductivity of good insulators is of the order of 10^{-18} siemens/cm.

Semiconductors

The electric conductivity of semiconductors is somewhere between that of conductors and insulators. Under normal conditions, semiconductors also have a very small number of free charge carriers, but that number can be substantially increased by energy input from an external source. Their conductivity, therefore - in contrast to that of metals and insulators - is heavily dependent on:

- ▶ Pressure (affects the mobility of charge carriers)
- ▶ Temperature (affects the number and mobility of charge carriers)
- ▶ Exposure to light (affects the number of charge carriers) and
- ▶ Added impurities (doping, affects the number and type of charge carriers)

By doping the semiconductor material, it is possible to alter the properties of the semiconductor component in a specific manner. The sensitivity of semiconductors to pressure, temperature and light also makes them suitable for use as sensors.

Doping of semiconductor materials

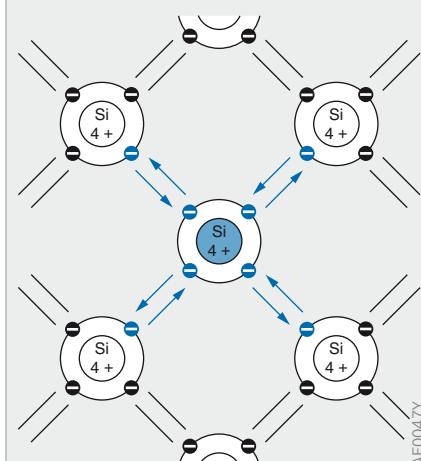
The basis of semiconductor technology is the ability to accurately modify and localize the conductivity of semiconductors by the controlled introduction of impurities that affect electrical properties (doping). The electric conductivity that can reliably be brought about in silicon by doping ranges from 10^4 to 10^{-2} siemens/cm.

As silicon is by far the most commonly used semiconductor material, the explanations that follow will restrict themselves exclusively to that material. When solid, silicon consists of a crystal lattice in which each silicon atom is linked to four, equally spaced adjacent atoms. Every silicon atom has four outer electrons (Fig. 2), with two shared electrons forming the bond with the contiguous atoms. In this ideal state, silicon has no free charge carriers; thus it is not conductive. This condition is changed by the introduction of suitable additives (doping), and energy input from an external source.

n-type doping

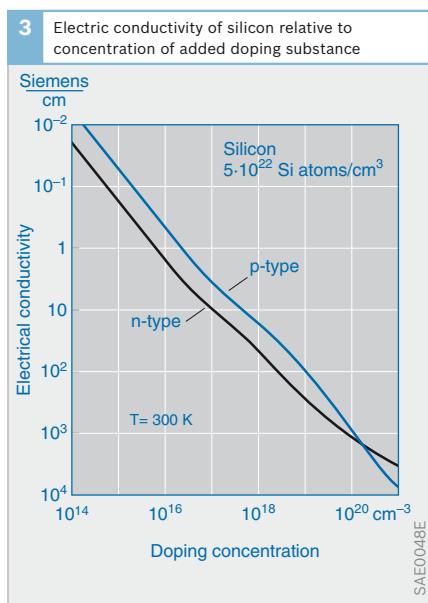
The addition of foreign atoms with five outer electrons (e.g. phosphorus) introduces free electrons because only four are required to bind each atom within the silicon crystal lattice. Thus, each additional phosphorus atom will provide a free, negatively charged electron. The silicon becomes n-conductive (negatively conductive). It is then referred to as n-type silicon (Fig. 3).

2 Structure of a perfect, pure-silicon crystal lattice with four outer electrons (valence electrons •)



p-type doping

The addition of foreign atoms with three outer electrons (e.g. boron) creates electron gaps. The boron atom is one electron short of the number required to properly bind it within the silicon crystal lattice. This gap in the bonding pattern is also called a hole.



Holes are mobile in the silicon. In an electrical field, they behave like positively charged charge carriers. They are able to drift by allowing the electrons of adjacent atoms to fill the gaps. It appears as if positive charge carriers are drifting through the silicon and positive current is flowing.

Each boron atom introduced therefore supplies one free, positively charged hole. The silicon becomes positively conductive and is therefore referred to as p-type silicon (Fig. 3).

The conductivity of n-type silicon is around 10 times as high as that of p-type silicon assuming other factors are equal (e.g. number of foreign atoms, temperature, pressure) because electrons can move about 10 times as fast as holes.

Intrinsic conductivity

In response to external energy input, such as heat or incident light radiation, electrons may break free from the crystal structure. Conductivity, therefore, can even be generated in undoped silicon. Electron-hole pairs produce intrinsic conductivity in the semiconductor. Such conductivity is generally low compared with that produced by doping. Increases in temperature induce an exponential rise in the

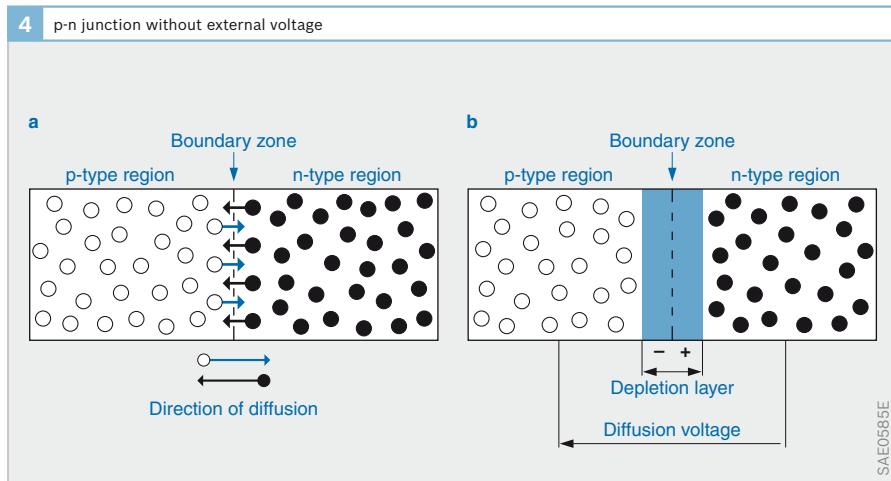


Fig. 4

- a Diffusion of mobile charge carriers
- b Formation of depletion layer (space-charge region)
- Holes
- Electrons

number of electron-hole pairs, ultimately obviating the electrical differences between the p and n regions produced by doping. Consequently, there are maximum limits for the operating temperatures of semiconductor components (Table 2).

2 Maximal permissible operating temperatures	
Material	Max. operating temperature
Germanium	90 to 100 °C
Silicon	150 to 200 °C
Gallium arsenide	300 to 350 °C

An n-type semiconductor always has some holes due to thermal fluctuations even at room temperature and a p-type semiconductor always has some free electrons. Such minority charge carriers are fundamental to the way in which almost all semiconductor components work, as will be demonstrated later.

p-n junction

The boundary between the p-type and n-type regions of the same semiconductor crystal is called the p-n junction. Its properties are fundamental to almost all semiconductor components.

p-n junction without external voltage

In the p-type region (Fig. 4a), there are a large number of holes (○) and few free

electrons (●). In the n-type region, by contrast, there are very few holes and a large number of free electrons. Due to the concentration differentials, the mobile charge carriers in each region diffuse into the other region in each case (diffusion currents).

As a result, the p-type region is negatively charged and the n-type region positively charged. A potential difference (diffusion voltage) is created between the p-type region and the n-type region which counteracts migration of the charge carriers (Fig. 4b), ultimately bringing the exchange of holes and electrons to a halt. A region that is deficient in mobile charge carriers, and therefore has poor electric conductivity, is created at the p-n junction. It is referred to as the depletion layer or space-charge region. Due to the diffusion voltage, the depletion layer has a strong electric field.

p-n junction with external voltage

If an external voltage is applied to a p-n junction, it produces the effects explained below (Fig. 5).

Forward direction

If the positive terminal is connected to the p-type region and the negative terminal to the n-type region, the charge carriers are

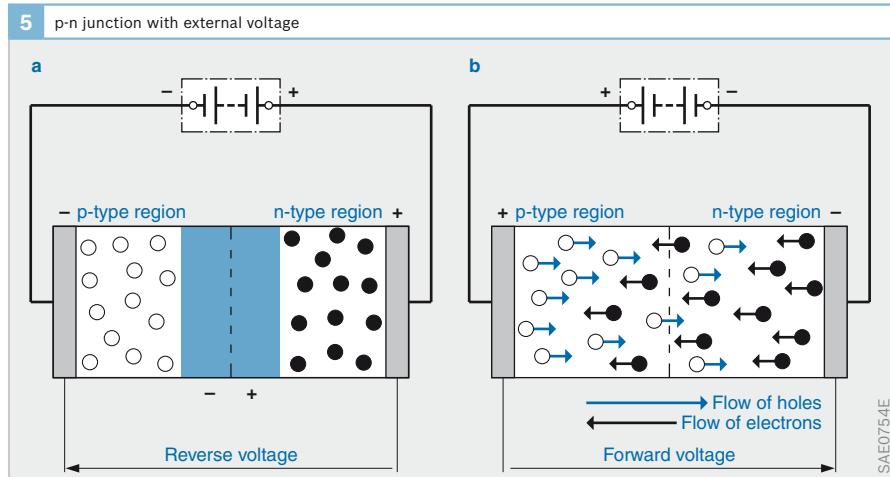


Table 2

Fig. 5

- a Reverse direction
- b Forward direction

forced into the space-charge region. The depletion layer collapses. When the diffusion voltage is exceeded, the charge carriers flood the p-n junction and a large current flows in a forward direction.

Reverse direction

If the negative terminal is connected to the p-type region and the positive terminal to the n-type region, the width of the space-charge region increases. Consequently, current flow is largely inhibited apart from a very small residual current (reverse current) produced by minority charge carriers.

Breakdown voltage

Above a specific voltage in the reverse direction (breakdown voltage), there is a sharp rise in reverse current.

The cause of this effect is the release of bound electrons from the crystal lattice in the space-charge region due to the high field strength (Zener breakdown) or due to surges of accelerated electrons. The accelerated electrons strike other electrons, breaking them free of their bonds and starting an avalanche-like increase in the number of charge carriers ("avalanche breakdown" or "first breakdown"). Both effects are reversible. The nature of the breakdown and the level of the breakdown voltage are dependent on the doping concentration profile.

A second breakdown occurs if there is localized heating of a semiconductor component caused by current constriction so that the area concerned becomes more conductive. This results in a self-accelerating increase in current and leads to the destruction of the semiconductor component. The reason for this is the increasing intrinsic conductivity. Semiconductors have a very high positive temperature coefficient, i.e. on heating, their conductivity increases fast (exponentially). Due to the stronger flow of current, this produces even further heating (positive feedback).

Passive components

Ohmic resistors

Design

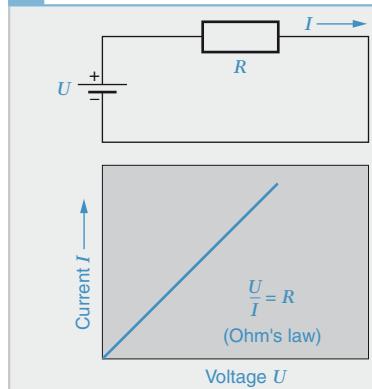
Ohmic resistors are generally made of materials with conductive properties similar to metals, e.g. carbon (carbon-film resistors) or special metal alloys (metal-film resistors). Through the suitable selection of metal alloys, they are constructed so as to reduce to the absolute minimum the effect of voltage, current and temperature on their electrical resistance. The conductor has a very small cross-section relative to its length, this being achieved either by the application of thin films to insulators or by winding wire into a coil.

The resistor in a circuit is referred to by the symbol R and its unit of measurement is the ohm (Ω).

Applications

In electronic circuits, resistors limit the current flow I or produce a voltage drop U proportional to the current. The resulting energy loss is converted into heat. Resistors can be used, for example, to adjust the operating point of a transistor.

1 Characteristic curve of an ohmic resistor



Capacitors

Design

A capacitor comprises two metal plates that are electrically insulated from each other. If a negative voltage is applied to one of these plates, its electrons will gather together and it becomes negatively charged. Since like charges repel, the electrons in the other plate are forced away and the plate becomes positively charged. The area and separation of the plates as well as the medium insulating them (dielectric) determine the quantity of charge carriers that can be stored by the capacitor (its capacitance). Using a vacuum (or air) as the dielectric offers the lowest capacitance. The capacitance can be substantially increased by the use of other insulating materials. The ratio to capacitance with air as the dielectric is known as the relative dielectric constant ϵ_r .

Since, in many cases, very large plates would be required for the commonly used capacitance values, capacitors are normally made by winding long strips into a coil (wound capacitors) or packing a large number of small plates together (multi-layer capacitors).

The electrolytic capacitor uses a thin layer of oxide as the dielectric. This method makes it possible to manufacture small capacitors with large capacitances.

The capacitance in a circuit is referred to by the symbol C and its unit of measurement is the farad (F).

Applications

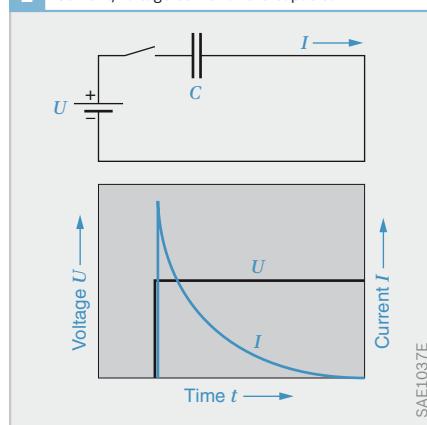
If a DC voltage is applied to both plates, the capacitor stores up charge until the limit of its capacitance is reached (Fig. 2). A charging current I starts to flow that drops as the charge stored in the capacitor increases. No more current can flow once the capacitor is fully charged. Thus – for a certain period at least – the capacitor stores electrical energy that is then available even if the external power supply fails.

This effect is utilized for triggering the airbag, for instance. Even if the wires connecting the airbag triggering unit to the vehicle's battery are severed in a serious accident, the capacitor still holds sufficient electrical energy to ensure that the airbag is deployed. In alternating-current circuits, a capacitor has a similar effect to a resistor.

In alternating-current circuits, the capacitor is repeatedly charged and discharged. A charge and discharge current therefore flows in permanent alternation. The current intensity depends on the capacity of the capacitor and the frequency of the alternating current. A capacitor therefore acts like a resistor, the resistance of which depends on capacity and frequency. As the frequency decreases, the resistance increases. If the frequency is reduced to its lowest limit, i.e. zero (direct current), the resistance is equal to infinity and, consequently, no current flow is possible.

This relationship is utilized, for example, by frequency filters in speaker systems in order to filter out high-frequency sounds from lower-frequency sounds (high-pass).

2 Current/voltage curve for the capacitor



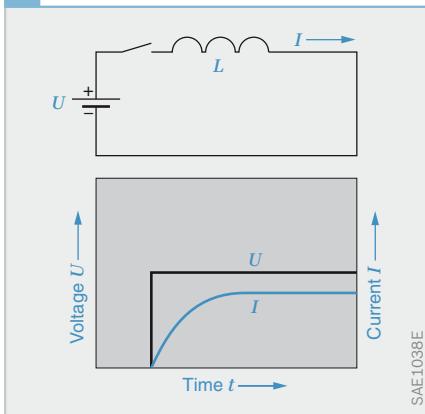
Coils

When an electric current flows through a coil, a magnetic field is created. Its strength depends on the current intensity, the number of windings in the coil and the properties of the coil core (generally a ferrite or iron core). Inductance, as a characteristic of a coil, indicates the amount of magnetic energy the coil can store for a given current intensity.

The magnetic field changes with a change in current intensity. A voltage is consequently induced in the coil which resists the generation of the magnetic field (Lenz's law). To clarify: if a DC voltage is applied to the coil, a current flows that generates a magnetic field. A change in magnetic field produces an induced voltage that is in the opposite direction to the applied DC voltage and thus resists the increase in current. The current intensity increases, but the increase becomes smaller and smaller. The change in magnetic flux and thus the induced voltage are also commensurately reduced. The current seeks to achieve its maximum value as determined by the ohmic resistance of the coil winding. The magnetic field of the coil through which the direct current is flowing has then reached its maximum strength; no more voltage is induced due to the magnetic flux being constant. The coil does not therefore represent a lasting hindrance to the flow of direct current.

The inductance in a circuit is referred to by the symbol L and its unit of measurement is the henry (H).

3 Current/voltage curve for a coil



SAE1038E

Applications

The generation of a magnetic field in a coil and the force that it exerts on a permanent magnet is used for applications such as fuel injectors. In the energized valve, the magnetic field causes the valve needle to lift away from its seat and allow the valve to exhaust gas.

In an alternating-current circuit, because of the constant generation and collapse of the magnetic field (and the energy contained within it), the coil acts as a frequency-dependent resistor, the characteristics of which, however, are exactly the opposite of those of a capacitor. In this case, the higher the frequency, the greater is the resistance. This means that the coil can be used for frequency filters in loudspeaker systems to filter out the low-frequency sounds from the higher-frequency sounds (low-pass).

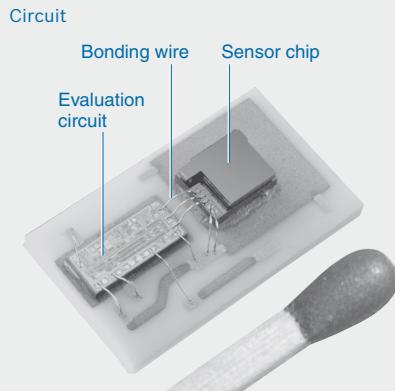
► Miniaturization

Thanks to micromechanics it has become possible to locate sensor functions in the smallest possible space. Typically, the mechanical dimensions are in the micrometer range. Silicon, with its particular characteristics, has proved to be a highly suitable material for the production of the very small, and often very intricate mechanical structures required. With its elasticity and electrical properties, silicon is practically ideal for the production of sensors. Using processes derived from the field of semiconductor technology, mechanical and electronic sensor functions can be

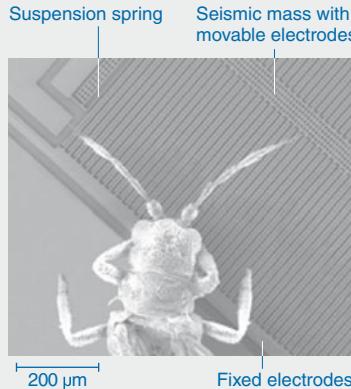
integrated with each other on a single chip or by using other methods.

Bosch was the first to introduce a product with a micromechanical measurement cell for automotive applications. This was an intake-pressure sensor for measuring load, and went into series production in 1994. More recent examples of miniaturization are micromechanical acceleration and yaw-rate sensors in vehicle safety systems for occupant protection and electronic stability program control. The illustrations below show quite clearly just how small such components really are.

▼ Micromechanical acceleration sensor

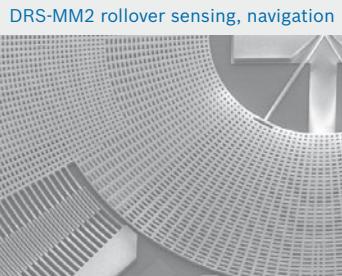
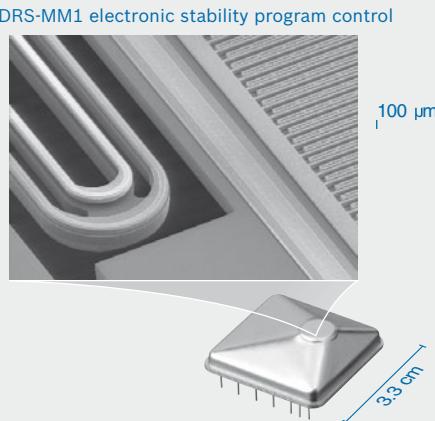


Comb-like structure compared with an insect



UAE0787E

▼ Micromechanical yaw-rate sensors



UAE0788E

Semiconductor components

Semiconductor components are active components that are generally subdivided into four categories (Table 1). These categories are further subdivided according to the way in which the components are manufactured and their function.

Discrete semiconductor components are stand-alone, “distinguishable” components. According to the currently accepted definition, semiconductor components with fewer than 100 transistor functions are referred to as discrete.

Monolithic integrated circuits (ICs, Integrated Circuits) are active components which incorporate more than 100 transistors (monolithic literally meaning “made from a single stone” from the Greek “monolithos” meaning “single stone”).

1	Semiconductor components (examples)
Discrete active components	
–	Diodes
–	Semiconductor resistors
–	Transistors
–	Thyristors
Integrated circuits (ICs)	
–	Analog circuits
–	Digital circuits
–	Mixed-signal circuits
Optoelectronic components	
–	Photoresistor
–	Photodiode
–	Photovoltaic cell
–	Laser diode
–	Phototransistor
–	Charge-coupled device
Micromechanical sensors (examples)	
–	Pressure sensors
–	Acceleration sensors
–	Yaw rate/angular-position sensors
–	Flow sensors
–	Temperature sensors
–	Position/angular-position sensors (Hall-effect sensors)
–	Gas sensors

A complex electronic circuit is therefore integrated on a single, silicon chip.

Optoelectronic components emit or receive light. They are fundamentally different from the traditional active components in terms of their method of manufacture and usage. For example, most light emitting components (light emitting diodes, laser diodes) are not made of silicon, but from III/IV semiconductors (alloy made from semiconductor materials with three and five outer electrons, e.g. GaAs).

Micromechanical sensors contain mechanical and electrical functions of sensors on a semiconductor chip (MST, MicroSystem Technology, or MEMS, MicroElectro-Mechanical System). In this case too, the methods of production and the type of use differ substantially from conventional active components. Normally, they are for the most part made from silicon but they contain some very complex, mechanical structures, e.g. “fingers”, that are free to oscillate and are thus suitable as a means of measuring acceleration. Pressure sensors contain fine membranes that can be used to measure pressure differentials.

Diodes

Diodes are semiconductor components with a p-n junction and two electrical connections – one to the p-type region and one to the n-type region. The pattern of doping impurity concentration within the crystal determines the specific characteristics of diodes.

Diodes designed for a forward current of more than 1 A are referred to as power diodes.

Rectifier diode

A rectifier diode allows current to pass in one direction (positive terminal connected to p-type region) but not in the other (positive terminal connected to n-type region). It acts like a flow control valve and is suit-

able for rectifying alternating current (Figs. 1a and 2).

The current flowing in reverse direction (reverse current) is much smaller than the forward current (approx. 1/100 the strength). It rises rapidly in response to increases in temperature.

Application: amongst other applications, rectifier diodes are used in automotive alternators to convert the alternating current into direct current. Because of the high ambient temperatures to which the alternator is subjected, the reverse current is a critical factor and has to be taken into account in the design of the diodes used.

Zener diode

The Zener diode is a semiconductor diode in which the reverse current rises abruptly upwards of a certain voltage as a result of Zener breakdown and/or avalanche effects. This diode is rated for different breakdown voltages depending on the particular application. They are designed for continuous operation at the breakdown voltage.

Even though large numbers of these diodes depend for their function on avalanche effects rather than Zener breakdown, they are still referred to as Zener diodes.

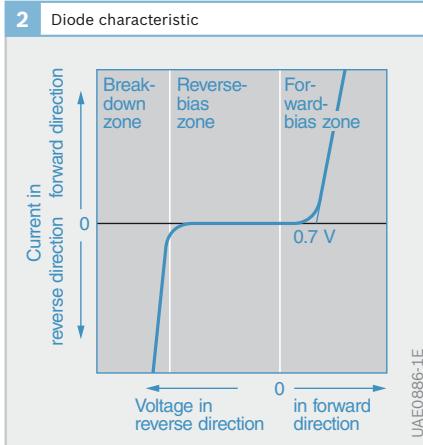


Fig. 2
Width of the depletion layer depending on the diode (e.g. rectifier diode, Zener diode)

Application: the Zener diode is mainly used for voltage limitation (e.g. damping diodes against overvoltage) and for producing a voltage reference (e.g. in the voltage regulator).

Variable-capacitance diode

With the variable-capacitance diode, it is possible to exploit the property whereby the space-charge region at the p-n junction of a diode acts like a capacitor; the semiconductor material (depletion layer) freed of charge carriers acts as the dielectric. An increase in the applied voltage increases the width of the depletion layer

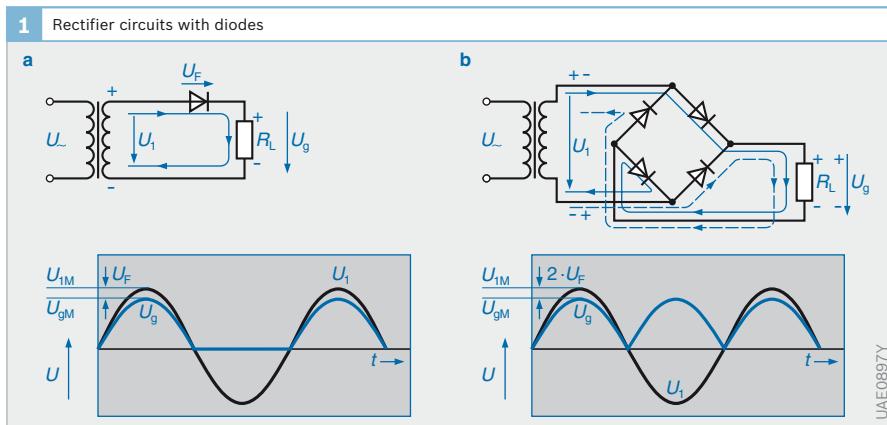


Fig. 1
a Half-wave rectification
b Bridge rectification
 U_F Diffusion voltage (approx. 0.7 V)
 U_{\sim} Alternating voltage
 U_1 Transformed alternating voltage
 U_g Rectified alternating voltage
 R_L Load resistor
 U_{1M} Amplitude of U_1
 U_{gM} Amplitude of U_g
 t Time

and reduces the capacitance; reducing the voltage increases the capacitance.

Application: variable-capacitance diodes are used mainly for resonant circuit tuning and frequency multiplication (e.g. in tuners).

Semiconductor resistors

The electrical resistance of these components is, by contrast with ohmic resistors, dependent on voltage, current intensity and temperature. They generally consist of polycrystalline semiconductor materials. The effects that occur in semiconductor resistors are based in part on depletion-layer properties that appear at the crystallite boundaries.

NTC resistors (thermistors)

NTC (Negative Temperature Coefficient) resistors (thermistors or thermal resistors) are, as the name suggests, resistors that have a marked negative temperature coefficient. The electrical resistance decreases with increasing temperature due to the increasing number of free charge carriers.

NTC resistors are made from polycrystalline metal oxides such as Fe_2O_3 , ZnTiO_4 or MgCr_2O_4 by a process of compression and sintering. Their temperature coefficients can be as much as $-6\text{ \%}/\text{K}$.

Application: NTC resistors are used as temperature sensors, for example.

PTC resistors (thermistors)

PTC (Positive Temperature Coefficient) resistors are resistors that have a positive temperature coefficient (of the order of a few $\text{ \%}/\text{K}$). The resistance value increases with rising temperature because lattice oscillations impede the flow of free electrons. PTC resistors conduct electricity better at low temperatures than at high temperatures. Most metals are thermal resistors with very low temperature coefficients.

The term PTC resistors refers to resistors made of semiconductor materials. PTC resistors made of ferroelectric ceramic material (e.g. polycrystalline barium titanate) have a relatively narrow temperature range and a very high positive temperature coefficient ($+6$ to $60\text{ \%}/\text{K}$). PTC resistors made of silicon have a wide temperature range and a virtually constant positive temperature coefficient (approx. $+0.8\text{ \%}/\text{K}$).

Uses: e.g. heater elements for auxiliary heaters or diesel-fuel preheating. The greater resistance with increasing temperature allows heat output to be adjusted automatically.

Transistors

Transistors can be used to control a large, electrical current with a small current. Consequently, these semiconductor components can be used as switches or amplifiers. "Transistor" is a contraction of the term "transfer resistor".

Figure 3 illustrates the classification of the transistor family.

Application: transistors are hardly ever used as individual components any more. They are integrated into integrated circuits (IC) on a semiconductor chip, e.g. as circuit breakers in output-stage modules or as component parts of densely-integrated logic chips or microcontrollers.

Bipolar transistors

Bipolar transistors consist of three zones of varying conductivity, the configuration being either pnp or npn. These regions (and their connections) are called the emitter (E), the base (B) and the collector (C).

Depending on their applications, transistors are categorized as low-signal transistors (up to 1 watt power loss), power transistors, switching transistors, low-frequency transistors, high-frequency

transistors, microwave transistors, photo-transistors, etc.

Such transistors are called bipolar because they make use of charge carriers of both polarities (holes and electrons). In an n-p-n transistor, positive charge carriers (holes) in the base current control roughly 100 times as many negative charge carriers (electrons) flowing between the emitter and the collector. This corresponds to a current amplification factor of around 100. Bipolar transistors are thus controlled by means of the base current.

Operating principle of a bipolar transistor (with reference to an n-p-n transistor): The emitter-base junction (EB) is poled for forward bias (Fig. 4). This means that electrons migrate into the base region. The base-collector junction (BC) is polarized for reverse bias. This induces the formation of a space-charge region with a strong electric field.

Coupling (transistor effect) occurs if the two p-n junctions are very close to one another (less than 10 µm apart in silicon).

The electrons injected at the EB then diffuse through the base to the collector.

As soon as they come within the range of the electrical field at the BC junction, they are accelerated into the collector region and travel onwards as the collector current. Thus, the concentration gradient in the base is retained, and additional electrons continue to migrate from the emitter to the collector.

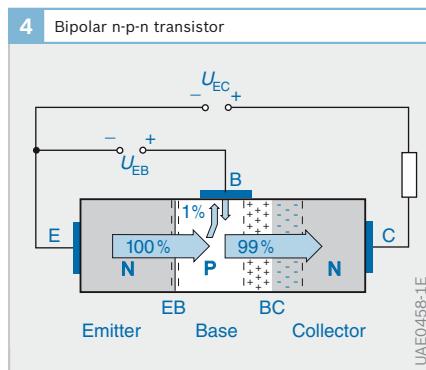


Fig. 4

N n-type silicon
P p-type silicon
E Emitter
B Base
C Collector

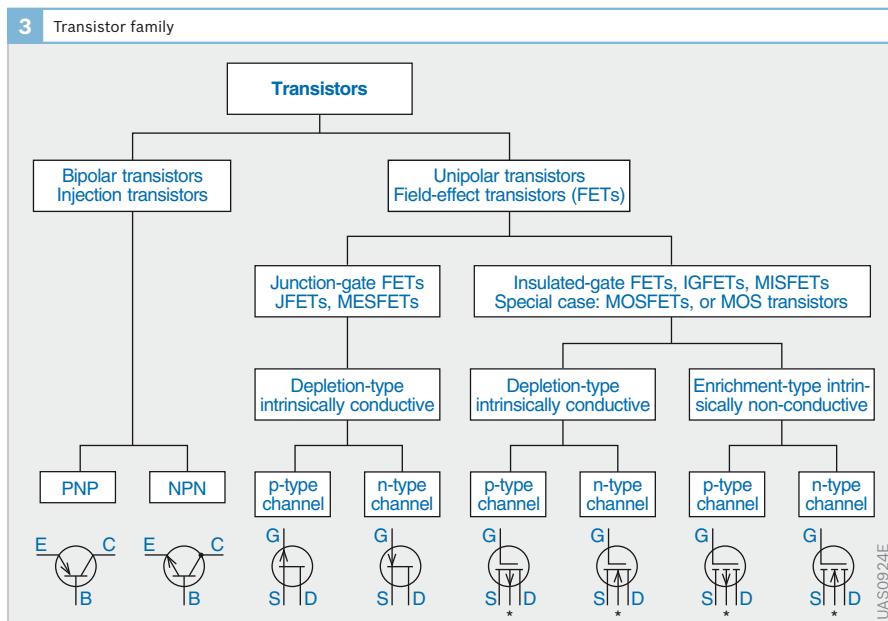


Fig. 3

B Base
E Emitter
C Collector
S Source
G Gate
D Drain
* Bulk connections (connection designation not usual)

In standard transistors, 99 % or more of all the electrons emanating from the emitter reach the BC space-charge region and become collector current. The few missing electrons are caught in the electron holes while traversing the p-doped base. If no base current is flowing, the base would become negatively charged and the resulting repulsion forces would completely prevent the continued flow of electrons within an extremely short space of time (50 ns). Such a build-up of negative charge can be partially or entirely counteracted by a low base current of positive charge carriers

(holes). Small changes in the base current bring about large variations in the emitter-collector current.

With moderate base currents, the n-p-n transistor acts as a bipolar, current-controlled, amplifying semiconductor component. If the base current alternates abruptly between very low and very high levels, the transistor acts as a switch.

Field-effect transistors (FET)

In a field-effect transistor (Fig. 5a), the current flowing from the source (S) to the drain (D) through a conductive channel is controlled by an electric field. That field is generated by a voltage applied via a control electrode known as the gate (G). Between the electrode and the semiconductor is an electrically insulating layer. The use of an electric field to control the current flow explains the origin of the term "field-effect transistor" (FET).

Field-effect transistors differ from their bipolar counterparts in utilizing only a single type of charge carrier (electrons or holes), giving rise to the alternate designation "unipolar transistor". Field-effect transistors are subdivided into:

- ▶ Junction field-effect transistors (junction FET, JFET)
- ▶ Insulated-gate field-effect transistors, in particular metal-oxide semiconductor (MOS) field-effect transistors (MOS-FETs).

MOS-FETs are not only suitable for extremely densely integrated circuits, they have also penetrated most other fields of application for bipolar transistors.

Operating principle of a junction FET (with reference to a JFET with an n-type channel, Fig. 5a):

DC voltage is applied to the ends of an n-type crystal. Electrons flow from the source (S) to the drain (D). Two p-type regions at the sides forming the gate (G), and the negative voltage applied to them,

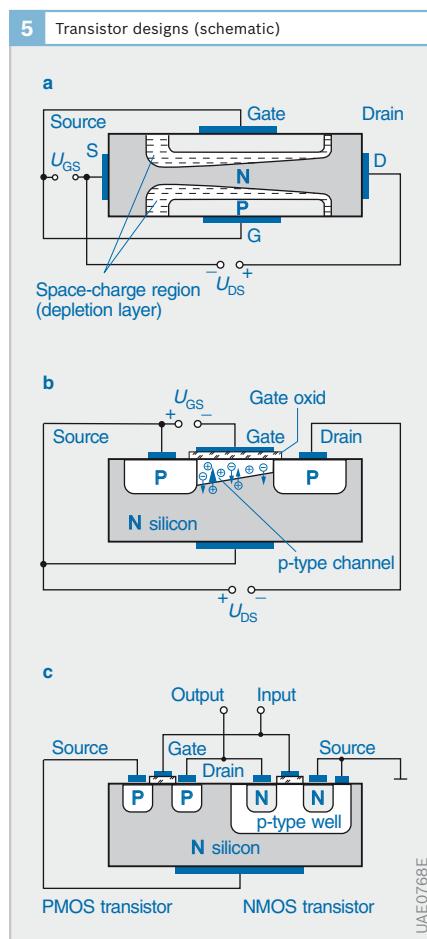


Fig. 5

a Depletion layer, field-effect transistor

b PMOS transistor

c CMOS transistor pair

N n-type silicon

P p-type silicon

S Source

G Gate

D Drain

U_{GS} Voltage between G and S

U_{DS} Voltage between D and S

determine the width of the channel. If the negative gate voltage is increased, the space-charge regions extend further into the channel and constrict the current path. The voltage at the control electrode (G) thus controls the current between the source (S) and the drain (D). The FET only requires charge carriers of a single polarity in order to function. The current is controlled virtually without power consumption. Thus, the junction FET is a unipolar, voltage-controlled component.

Operating principle of a MOS FET (with reference to a MOS FET with p-type enrichment):

While no voltage is present at the gate electrode, no current flows between the source (S) and the drain (D); the p-n junctions are reverse biased (Fig. 5b). A negative voltage applied to the gate (G) forces the electrons in the n-type region below the electrode into the center of the crystal and draws holes – which are always present as minority charge carriers even in n-type silicon – to the surface. A narrow p-type layer forms beneath the surface: This is called the p-type channel. Current can now flow between the two p-type regions (source and drain). This current consists exclusively of holes. As the gate voltage acts through an insulating oxide layer, no current flows in the control circuit and control therefore requires no power. The MOS transistor is a unipolar, voltage-controlled component.

PMOS, NMOS, CMOS transistors

If a p-type channel MOS transistor (PMOS transistor) is doped with a donor impurity rather than an acceptor impurity, it becomes an NMOS transistor. Due to the higher mobility of electrons, NMOS transistors switch faster than PMOS transistors, which are easier to manufacture and were therefore available first.

If PMOS and NMOS transistors are created in pairs on the same silicon chip, they are referred to as CMOS transistors (Complementary MOS transistors, Fig. 5c).

Their advantages are:

- ▶ Very low power loss
- ▶ High interference resistance
- ▶ Low supply voltage
- ▶ Suitability for analog-signal processing

The low power consumption, particularly for digital circuits, is due to the fact that current only flows when a digital cell changes (e.g. switches from “0” to “1”). Retention of the information requires only that a voltage is applied. Since, with this configuration, the PMOS and NMOS transistors are connected to one another, the leakage flow is minimal as a consequence of the intrinsic conductivity of the silicon.

More than 80 % of all integrated circuits are now manufactured using CMOS technology.

Monolithic integrated circuits (ICs)

Monolithic integration

The planar process is a method of manufacturing semiconductors which makes it possible to produce all components of a circuit (resistors, capacitors, diodes, transistors) and the conductive connections between them on a single silicon wafer (chip) in a single production process. It involves the creation of multiple p- and n-type layers in a multi-stage doping process using a pre-doped single-crystal silicon disk or wafer.

An integrated circuit (IC) no longer contains any discrete components but rather switching elements or functional elements (e.g. gates, capacitors). In this way, it is possible to manufacture complex circuits of small size and still meet reliability requirements.

Integration level

The integration level is defined as the number of functional elements, transistors or gates on a single chip. The following categories are defined based on the integration level (and chip surface area):

- ▶ SSI (Small-Scale Integration): up to around 1,000 elements per chip, average chip area 3 mm² (varies considerably depending on level of power loss)

- ▶ MSI (Medium-Scale Integration): up to around 10,000 elements per chip, average chip area 8 mm²
- ▶ LSI (Large-Scale Integration): fewer than 100,000 elements per chip, average chip area 20 mm²
- ▶ VLSI (Very-Large-Scale Integration): more than 100,000 elements per chip, average chip area 30 mm²

The number of functional elements on VLSI chips is constantly increasing. Microprocessors can now have up to 10 million transistor functions per chip. Memory modules (DRAM) can even contain several hundred million transistors. The number of such functional elements is subject to an exponential increase over time. This was first recognized by Gordon Moore (co-founder of the company Intel) and is formulated in "Gordon Moore's Law": Elements per chip = $2^{(\text{year} - 1956) \cdot 2/3}$

This means that the number of functions per chip will double every 18 months. This is because the structural units (width of transistors) in which the chips are manufactured are becoming smaller at the same time as the number of functional elements is growing. Thus the structural units of new processors have shrunk from 10 µm in 1972 to 0.18 µm in the year 2000.

The integration levels of LSI and VLSI demand methods such as CAD (Computer-Aided Design). VLSI circuits in particular can only be created with the help of highly sophisticated programs which can convert entire function blocks into corresponding circuit subdivisions. Such program systems are called HDL (Hardware Description Language) or VHDL (Visual HDL).

IC classification

There are various systems for classifying ICs, although there are invariably mixed categories within each system.

- ▶ *Classification by production technique:*
bipolar/unipolar (MOS)
mixed form: e.g. BiCMOS, BCD
- ▶ *Classification by function:*
analog/digital;
mixed form: e.g. mixed signal IC
- ▶ *Classification by application:*
standard IC - ASIC;
mixed form: e.g. ASSP

ASIC modules (Application-Specific IC) are developed and produced specifically for a particular application and exclusively for a particular customer.

ASSP modules (Application-Specific Standard Product) are developed and produced specifically for a particular application for a particular customer but also sold to others for the same type of application.

Apart from a few exceptions, ICs are now always based on MOS or combined technologies.

Analog circuits

Analog circuits are required wherever electronic equipment has to communicate with the outside world – which is always analog. This applies in particular to the areas of the preparation and processing of input and output signals. Examples of analog-input signal generators are microphones (telephone) and knock sensors. Examples of analog-output signal receivers are speakers or actuators.

Applications

- ▶ Standardized functions:
stabilized-voltage sources, stabilized-current sources, differential amplifier stages, coupling circuits, potential shifting circuits, output stages
- ▶ Application-based ICs:
operational amplifiers, voltage regulators, comparators, timers, transducers, interface circuits

- ▶ Specialized ICs:
voltage references, broadband amplifiers, analog multipliers, function generators, phase-locking circuits, analog filters and switches

Digital circuits

Digital circuits are used wherever large volumes of data have to be processed within a short space of time. High-performance microcontrollers can perform several hundred million computational operations per second. They are thus able to utilize the input data to provide the required, highly accurate response at the output of the particular device. With digital circuits, even large amounts of data can be coded and carried on a line, e.g. data transfer on the CAN bus.

Since inputs and outputs for communication with the real world always have to be analog, analog-digital converters (ADCs) are used at the inputs of a digital circuit, and digital-analog converters (DACs) at the outputs.

Digital systems make use of a whole series of recurring basic circuits and variations of them. The scale ranges from simple gates (e.g. inverters, logical AND operator) to memories, microprocessors and microcontrollers.

Digital modules can only be connected up to form a complete system if power supply voltages, logic signal levels, switching speeds and signal transmission times are compatible. This requirement is met within a family of circuits (e.g. CMOS). Since, apart from a few specialized applications mainly in military systems, CMOS logic circuits are the only type now used, compatibility is guaranteed.

Mixed-signal ICs

ICs in which both analog and digital circuit elements are combined are called mixed-signal ICs. They are an essential requirement for the combination of entire electronic systems on a single piece of silicon (SoC: System-on-a-Chip).

With ever-increasing degrees of integration, this type of circuit is becoming more and more attractive and the number of applications is continually increasing. It is now conceivable, for example, to integrate the entire electronic circuitry for a mobile phone or a complete Internet-access system on a single chip. For commercial reasons, however, two or three chips are used. Where only low levels of electrical power are required, such as in the examples quoted, both the analog and the digital sections are manufactured using CMOS technology.

Mixed-signal ICs for applications involving higher electrical power levels, such as are frequently encountered in automotive systems, are typically hybrids of various technologies. An example would be BCD systems (where B stands for bipolar, C for CMOS and D for DMOS). Here, three different technologies are used on the one chip because they each have properties that are complex and expensive to implement using the other technologies. Bipolar structures are particularly suitable for handling very small currents without distortion. The bipolar section is therefore responsible for converting the analog inputs. CMOS is the optimal choice for the rapid processing of digital data. The CMOS section of the IC performs the logical processing operations using digital technology. Using DMOS technology, it is possible to produce high currents; the DMOS section is therefore able to provide high analog output levels.

Optoelectronic components

Photoresistor

A photoresistor is a resistor whose resistance decreases when exposed to light (Fig. 6). Light (photons) generates free charge carriers in the semiconductor when the energy of a photon is equal to the energy holding a charge carrier within the lattice (internal photoelectric effect). The photoresistor is thus not equally sensitive to all wavelengths. Its maximum sensitivity lies within a narrow wavelength band that

is specific to the material used. Photoresistors generally consist of polycrystalline semiconductor materials in which the effect is particularly marked such as CdS, CdSe, PbS, PbSe, CdTe, ZnO, Se, InSb, InAs, Ge or Si.

Application: photoresistors are mainly suitable for use as light meters, e.g. in cameras.

Photodiode

A photodiode is a semiconductor diode which utilizes the depletion-layer photoelectric effect. Reverse voltage is present at the p-n junction. Incident light releases electrons from the crystal lattice (photoelectric current). This also results in free electrons and holes that increase the reverse current in proportion to the light intensity (Fig. 7).

Application: in the automotive sector, photodiodes are used in the rain sensor, for example. They measure the scattered light emitted by a light emitting diode and reflected by the raindrops.

Photovoltaic cell

Like the photodiode, the photovoltaic cell releases charge carriers when exposed to light; no external voltage is applied to the p-n junction, however. If the electrons and

holes reach the p-n junction, they are separated by the internal field of the space-charge region. A photoelectric voltage is generated which produces a photoelectric current in the external electric circuit. Light energy is thus converted into electrical energy.

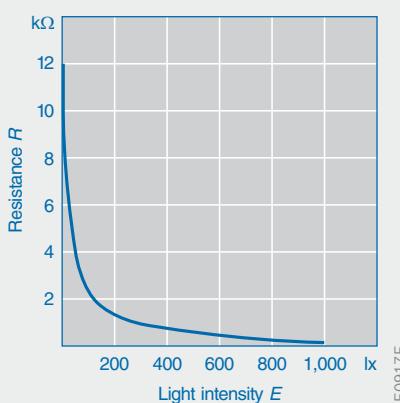
Application: photovoltaic cells are used to measure light intensity and to generate electricity with solar cells.

LED (Light Emitting Diode)

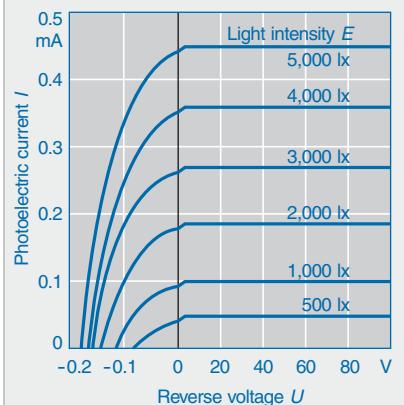
The light emitting diode (LED) emits light when a forward current is passed through it. This signifies a reversal of the photodiode effect: free electrons and holes are recombined, the energy that is freed in the process is emitted as light. The wavelength of the light emitted by an LED depends on the semiconductor material used. There is now a choice of colors available (e.g. red, yellow, green, blue, and even infrared). The semiconductor materials most commonly used are gallium compounds such as gallium arsenide (GaAs) and gallium phosphide (GaP).

Application: LEDs are used for numerical and alphabetical displays as well as for warning lamps. Since it has been possible more recently to produce many colors and substantially increase the amount of light

6 Photoresistor (characteristic curve)



7 Photodiode (characteristics curve)



emitted, LEDs are now also used in motor vehicles to illuminate the instrument cluster or as a third stop lamp. They have the advantage of a relatively high light yield and, as a result of their very long service life, do not have to be replaced. Another example of the LED's application in the motor vehicle is its use in the rain and dirt sensor.

Laser diode

With the appropriate mechanical/optical design, LEDs can also be made to emit laser light as laser diodes, i.e. the light emitted is extremely concentrated and coherent. These properties are required for the optical data transfer in optical fiber cables (e.g. MOST bus).

Phototransistor

A similar effect to that used by the photodiode also occurs in the phototransistor. Incident light striking the base alters the conductivity between the emitter and the collector in such a way that a current proportional to the intensity of the light is able to flow through the transistor. Phototransistors can thus be used as a switch for photoelectric light barriers, for instance.

There are also phototransistors that operate in the opposite way, i.e. they emit light. The advantage compared to the LED is that the phototransistor can be switched on and off very quickly.

Application: the phototransistor is particularly suited for use in the manufacture of flatscreen displays. A transistor is used for each pixel. As the areas required are generally substantially larger than the silicon slices available, and such slices would, in any case, be too expensive, flatscreen displays are manufactured using thin-film techniques (vapor deposition through a screen).

The advantages of flat screens over cathode-ray tubes are their slimmer dimensions, lower power consumption and the absence of flicker.

Charge-coupled device (CCD)

CCDs are integrated circuits arranged as a charge-coupled array. They are used to record images in video and digital cameras as well as in scanners. Thousands or even millions of individual elements laid out in a matrix form (array) are created on a chip. Digital cameras can record up to 3.5 million pixels with a color resolution of 16 million colors. CCDs are capable of converting the visual information into electrical information and storing it until it can be transferred to an external storage medium. To do this, the output register is scanned at a high rate.

Micromechanical sensors

It has been possible for a number of years now to manufacture different sensor elements using silicon technology. Since a large number of sensors can be produced simultaneously on a single silicon wafer, the unit costs are substantially lower than in the production of conventional sensors. In addition, there is the possibility of having the sensor signals directly processed on the chip electronically. Examples of micromechanical sensors are:

- ▶ **Magnetic-field sensors:** these use the hall effect to measure an external magnetic field purely electrically (e.g. position sensor for camshaft position)
- ▶ **Acceleration and pressure sensors:** these use capacitance changes, caused by mechanical deformation of individual regions in the silicon, to record measured variables (yaw-rate sensor, intake-manifold pressure sensor)
- ▶ **Temperature sensors:** these make use of the relationship between a semiconductor material's temperature and its conductivity (e.g. engine-temperature sensor)
- ▶ **Chemical sensors:** these make use of the effect whereby certain gases and liquids specifically alter the electric conductivity of some semiconductor materials

Manufacture of semiconductor components and circuits

Semiconductor components

Most semiconductor components are made of silicon. The basic material is SiO₂ (quartz). It occurs naturally in large quantities as rock crystal or silica sand. It is used to make monocrystalline cylinders or rods of extremely pure silicon with a diameter of between approx. 50 mm (2") and 300 mm (12"). The most common diameters are 150 mm (6") and 200 mm (8").

During the production of the crystals, they are given a precise conductivity specified by the component manufacturer (basic doping). The silicone is sawn into disks approximately 0.3 to 0.7 mm thick which are then smoothed and polished. These disks are referred to as wafers. The process of producing the semiconductor components starts with these silicon wafers (Fig. 1).

A large number of identical circuits (chips) is created on each wafer and those circuits are then separated by mechanical means. Before separation, however, the chips are tested to make sure they match the electrical specifications. Any unusable chips are marked and then discarded once separated. Only the usable chips are assembled, i.e. cemented, bonded, encased, sealed and passed on for final testing (Figs. 1 and 3).

The value-added chain from the basic material to the end product shows an extremely high rate of value growth: silicon in the form of silica sand costs about 1 euro per kg, but by the time it has become a finished microprocessor it can be worth up to 1 million euro per kg.

Doping processes

Doping involves introducing impurities with a known electrical effect into the semiconductor crystal at specific points

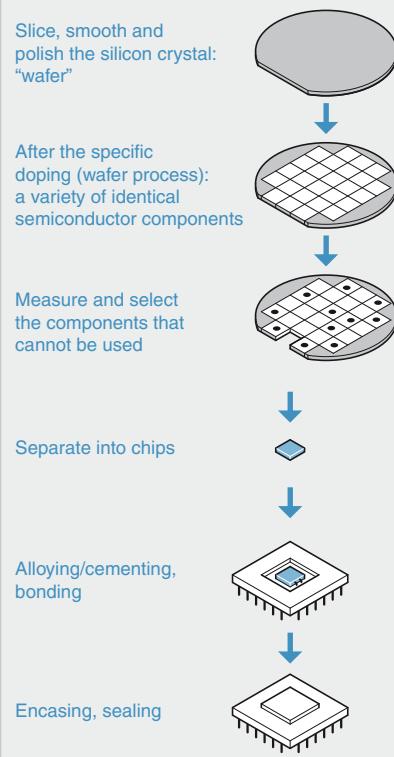
and in precisely defined concentration levels using reproducible methods. Doping techniques are the fundamental processes of semiconductor technology. The target parameters are concentration profile, penetration depth, surface concentration and planar parallelity of the doping-material fronts.

Manufacturers of semiconductor components buy their basic material in the form of wafers that have been pre-doped according to precisely defined specifications.

Doping during crystal growth

While the crystal is being grown, phosphorus, for example, can be added to

1 Stages in the manufacture of semiconductor components



the silicon melt as a doping material. As the monocrystal forms, the phosphorus atoms become fixed in the silicon crystal, giving it n-type conducting properties.

Doping by diffusion

At high temperatures, the doping material can diffuse into the silicon crystal. The effect is achieved by creating a specific concentration level of the impurity atoms at the surface of the wafer. The difference in concentration causes the doping atoms to diffuse into the silicon wafer.

The process involves placing between 50 and 200 wafers in a kiln together and exposing them to boron or phosphorus compound vapors at temperatures of around 1,000 °C. Boron produces p-type regions while the effect of phosphorus is to create n-type regions. Surface concentration levels, temperature, and time, are the parameters that determine the penetration depth of the doping material.

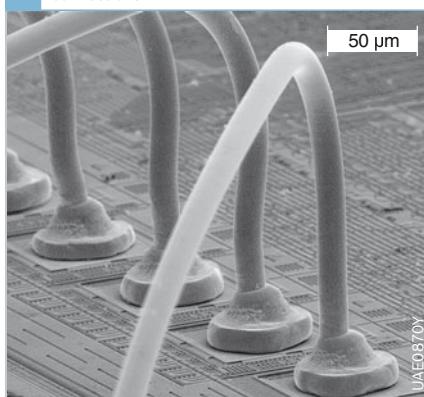
Doping by ion implantation

In this process, the atoms of a gaseous doping material are first of all ionized in a vacuum and then accelerated by a high voltage (up to 300 kV) so as to be “catapulted” into the semiconductor. This allows particularly precise control of concentration level and localization of the implanted material. Embedding of the implanted atoms and restoration of the crystal lattice requires thermal aftertreatment.

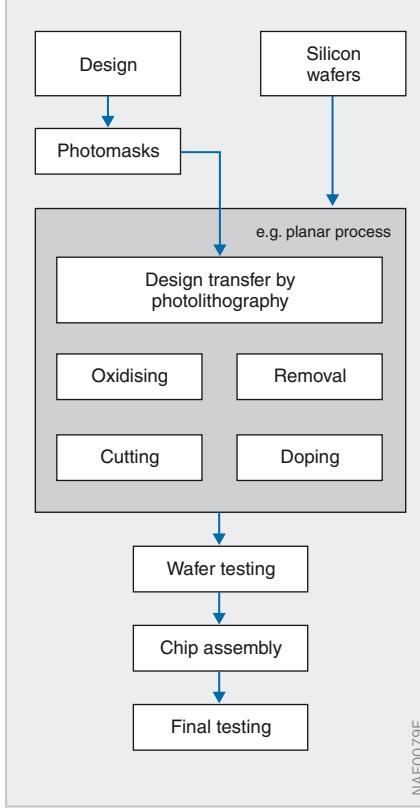
Epitaxy

This doping method creates a monocrystalline doped semiconductor layer a few micrometers thick on a monocrystalline substrate. If gaseous silicon tetrachloride and hydrogen are passed over silicon wafers heated to around 1,200 °C in a quartz tube, the vapor breaks down and silicon is deposited at a rate of about 1 µm/min, forming a monocrystalline layer. If a defined quantity of doping material is added to the gas flow, an epitaxial layer is created

2 Microstructure of a microchip with bonded connections



3 Process of manufacturing semiconductor components



whose electric conductivity and conduction type differs fundamentally and abruptly from that of the substrate.

Photolithography and the planar process
 In the *photolithography* process, the pattern of the component design is transferred to the wafer by using metal screens. When the design has been completed, the data for producing the screen is saved onto a storage medium (e.g. magnetic tape or CD). It is then used to control a photographic exposure device that transfers the design pattern to photographic plates. This pattern is subsequently reduced by optical means to the actual size for the application and copied onto metal screens on which it is repeated as many times as will fit onto the size wafer being used.

As this method can be used for structures that are many times smaller than the wavelength of the light used, it will continue to be used in the future. The size of the smallest achievable structures depends on the wavelength of the light source. Using UV lasers, photolithographic methods have already produced structures as small as 0.08 µm in the laboratory (by comparison, a human hair has a diameter of 40 to 60 µm).

Even smaller structures can be produced using other methods such as X-ray or electron-beam lithography. However, those methods are substantially more expensive as they can only “expose” one IC at a time on the wafer. For that reason, they are only used in special cases.

Silicon wafers are easily oxidized by oxygen or water vapor. The oxide layer thus created prevents penetration by the doping atoms in the doping process. In the planar process, holes are created in the oxide layer so that when the wafer is doped, localized areas of n-type and p-type material are created.

The wafer is coated with a special lacquer and then exposed to light while covered with a metal screen. Following developing, the areas of the lacquer coating previously covered by the screen and the oxide layer below it can be removed by etching. The position, size and shape of the holes thus created match precisely the specified design. In the subsequent doping process in the diffusion furnace or during ion implantation, materials which have a known electrical effect such as boron or phosphorus pass through those holes in the oxide layer into the silicon, thereby creating n-type or p-type regions in the desired locations. Afterwards, the oxide layer is removed again and the wafer is ready for the next stage in the manufacturing process.

The photolithographic process and the doping process are repeated as many times as required to produce the desired number of layers of different conductivity on the semiconductor component. In the case of complex integrated circuits, this may involve as many as 20 or more separate manufacturing stages. In order to provide the electrical connections between the resulting functional elements, the wafers are coated with aluminum or copper and then the metal conductor track patterns are formed. This process too may involve the formation of multiple layers of metal, one above the other.

Completion of the wafer production process is followed by electrical testing (preliminary testing) of the individual chips on the wafer. Chips that do not meet the specifications are marked with colored dots. The wafers are then cut up into individual chips using a diamond cutting tool. The functional chips are subsequently placed in metal or plastic casings and fitted with external connections. After being hermetically sealed or encased in plastic, they go through a final testing stage.

Conventional printed-circuit boards

The printed-circuit board has become an electronic component in its own right. It has to have precisely defined electrical and mechanical properties. In a motor vehicle, for example, the printed-circuit board may be required to withstand temperatures ranging from -40 to +145 °C. The standards required in respect of EMC (ElectroMagnetic Compatibility), maximum current capacity and complexity are becoming more and more demanding. And at the same time, the circuit boards are continually expected to become smaller and cheaper – in spite of shorter product life cycles.

The basic material – the baseboard – is glass fiber. It can be rigid or semi-flexible. The conductor tracks on the surface are made from a thin layer of copper 12 to 70 µm thick (copper base). To protect it against corrosion, the surface of the copper is protected by a coating of tin alloy, gold or an organic surface sealant, depending on the type of application.

Designs

Depending on the complexity of the circuit, printed-circuit boards may be made up of a number of layers (Fig. 4).

Circuit boards used in motor vehicles consist of between two and eight layers.

Single-sided circuit board

The pattern of conductor tracks and solder eyes is only on one side of the baseboard (a).

Double-sided, non-interconnected circuit board

There are patterns of conductor tracks on both sides of the circuit board. However, those two circuits are not connected to each other (b).

Double-sided, interconnected circuit board

The conductor track patterns on each side are interconnected via copper layers on the walls of the holes (c).

Multilayer circuit board

In addition to the two outer layers (in this case the 1st and 4th layers), there are additional conductor layers on the inside of the baseboard (internal layers). Those conductor layers may be electrically connected with one another. To that end, the relevant conductor layers are connected to the copper layers applied to walls of the holes (d).

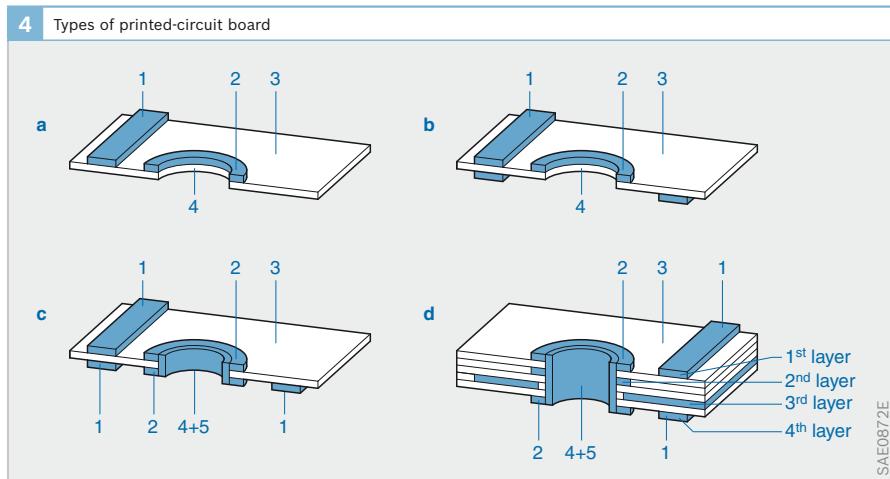


Fig. 4

Schematic diagram:
conductor tracks
are shown as raised
surfaces

- a Single-sided
- b Double-sided,
non-interconnected
- c Double-sided,
interconnected
- d Multilayer

- 1 Conductor track
- 2 Solder eye
- 3 Baseboard
- 4 Bore
- 5 Copper layer
(metallized hole)

Production process

For the manufacture of printed-circuit boards, pattern plating and panel plating

have become the accepted methods (Fig. 5).

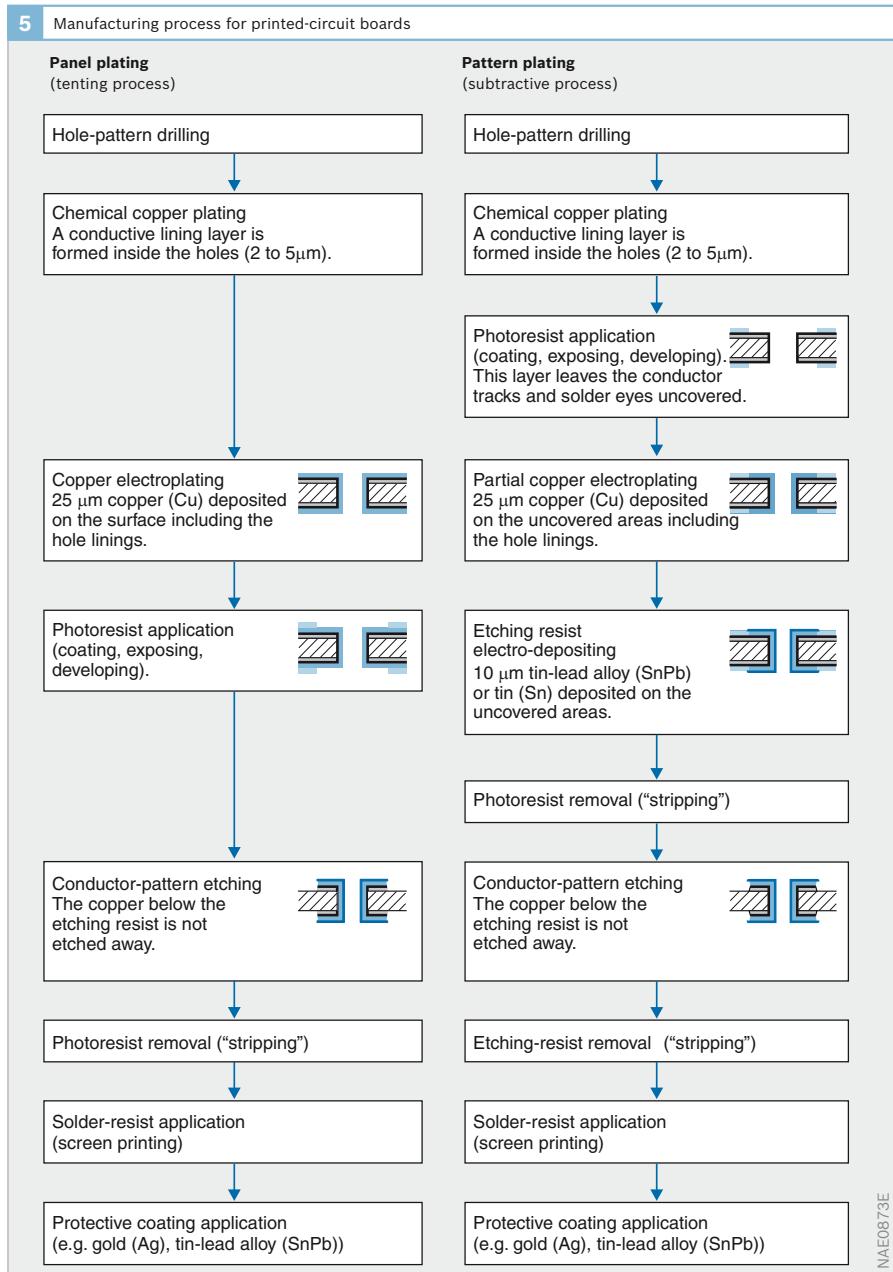


Fig. 5

The base material is coated on both sides with a copper film

Subsequent processing

Once the circuit board has been produced, it has to be fitted with the circuit components. These subsequent processing stages also demand the highest quality standards so as to ensure reliable operation of the electronic control units in the motor vehicle. There are two methods of mounting the components on the board - through-fitting and surface mounting.

Through-fitting method

With the through-fitting method, the component connections are passed through the holes in the circuit board and then soldered (Fig. 6a).

Surface-mounting method

The surface-mounting method (SMT, Surface Mount Technology) uses special electronic components whose connections lie flat against the surface of the printed-circuit board (Fig. 6b). Such components are referred to as SMDs (Surface Mounted Devices). Another advantage - in addition to the greater component density achievable - is that the components can be fitted to the circuit board fully automatically. The surface-mounting method is therefore becoming more and more widely adopted. The component fitting machines used can achieve work rates of more than 60,000 components an hour, increasing productivity as a result.

Production process

There are various production systems that are used in the subsequent processing of circuit boards and they are classified according to the spatial arrangement of the production equipment and the work areas. The categories distinguished are job-shop, pool, flow-shop and flow production.

The process described below is an example of flow production (Fig. 7). According to the standard DIN 33 415, flow production involves "a process of operations organized according to the flow principle in a rigid

sequence, aligned to a particular spatial arrangement and tied to a cyclic timing pattern". Furthermore, the flow principle involves "arrangement of the individual work stations according to the order in which the operations are performed". This method is also referred to as line production.

Production of the electronic circuitry in an ECU involves the following stages:

Basic materials

The basic materials consist of printed-circuit boards that have not yet been fitted with circuit components. Usually, one or more circuit boards are arranged on a "panel" of standardized dimensions.

After being fitted with components, the circuit boards are cut out of the panel.

6 Component fitting methods for printed-circuit boards

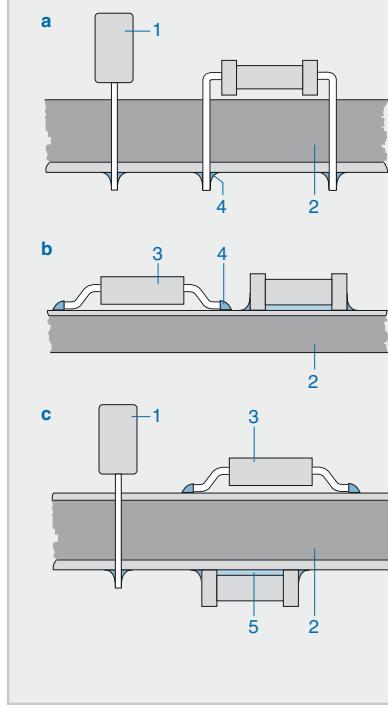


Fig. 6

- a Through-fitting method
 - b Surface-mounting method
 - c Mixed mounting
- 1 Wired component
2 PCB
3 SMD
4 Soldering point
5 Cement

Solder-paste application

The first stage is the application of solder paste for the SMDs using a screen printing process. The solder paste is a mixture of metal powder, flux and other organic additives. It is pressed through the spaces in the printing screen onto the panel with the aid of a "scraper".

SMD fitting (reflow side)

An automatic component-insertion machine presses the SMDs into the solder paste applied to the panel.

Reflow oven

The panel passes through a reflow oven on a conveyor belt. The heat of the oven melts the solder paste on the panel. This electrically and mechanically connects the components to the pcb.

Position checking

The camera of a vision system examines the soldering points (position checking system). Depending on the findings of the examination, the panel may be diverted to a repair station or automatically moved on to the next stage of production.

Turning the circuit board

In this stage, the circuit board is rotated through 180 degrees so that the "underside" is facing upwards and can be worked on. The way in which components are fitted to the underside depends on the nature of the second soldering stage: if the *reflow/reflow method* is used, the process described above is repeated.

If the alternative *reflow/wave method* is used, the SMDs are fitted differently since the soldering method is different from that used for the top side. A cementing station is required.

Cementing station (underside)

At the cementing station, cement spots are first of all applied in those positions on the "underside" where SMDs are to be placed. This is done using pipettes or a screen,

similar to the procedure for applying the solder paste. The cement holds the SMDs in position until they are soldered.

SMD fitting (underside)

At succeeding stations along the assembly line, the circuit components are placed on the cement spots. On this side of the circuit board, only components that are suitable for attachment by wave soldering can be fitted.

Cement hardening

In order to prevent the components falling off during the soldering process, the cement is hardened in an oven.

Component checking

A visual component checking system checks that all components are present and that they are in the correct positions on the underside of the circuit board. Non-compliant products are indicated and diverted to a repair station.

Turning the circuit board

The circuit board is once again rotated through 180 degrees, returning it to its original orientation.

Wired-component fitting

This stage involves fitting components that have leads attached (e.g. large coils and connectors) into ready-made holes (solder eyes) in the panel using the through-fitting method. They are subsequently soldered on the reverse side at the same time as the SMDs during the wave-soldering stage. The wired-component fitting stage completes the component fitting operations.

Wave soldering

The wave soldering stage solders the connections of all SMDs and "wired components" on the underside. The wave soldering process involves the following three stages:

1. Flux is applied to the underside of the panel.

2. The panel then passes through a pre-heating zone so that the components are not damaged by the sudden increase in temperature in the subsequent stage.
3. The panel passes over a wave of liquid solder created by a jet. As it does so, the solder is deposited on the soldering points (pads) on the panel. The solder resist on the circuit board prevents the solder attaching itself in the wrong places.

Position checking

Another camera checks the finished soldered connections. If faults are found, the panel is passed to a repair station.

In-circuit testing (ICT)

The purpose of the in-circuit test is to check that the electrical circuit is working properly. The testing equipment is connected via a testing adaptor and checks the components for correct function and electrical readings.

Cutting

The individual circuit boards are now cut out of the panel using computer-controlled machines.

Final assembly

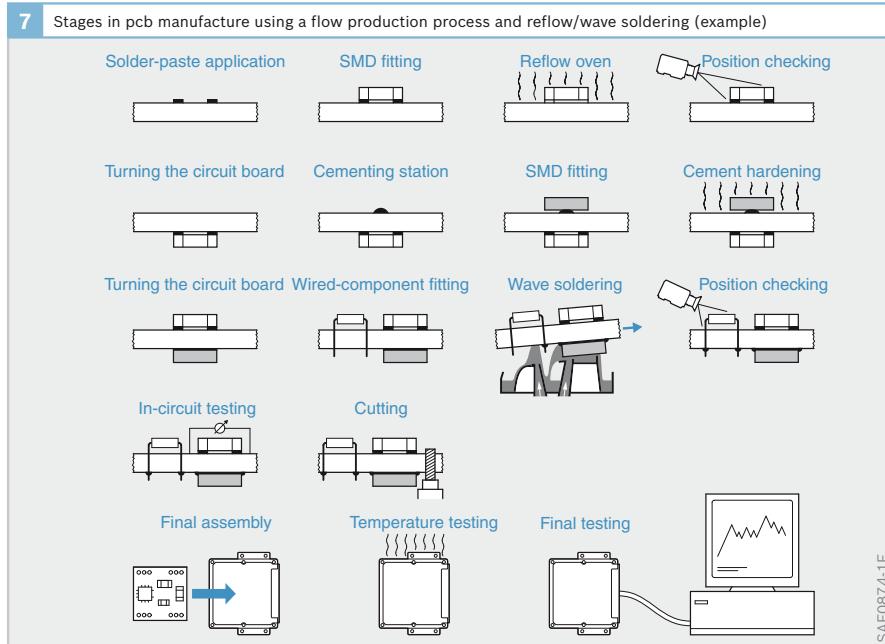
In the final-assembly stage, the circuit board is fitted into the ECU casing.

Temperature testing

The finished ECU is subjected to very high temperatures in order to test its ability to withstand extreme conditions. The tests simulate operation of the ECU in real conditions. In this way, component or soldering faults can be detected.

Final testing

Before the ECU leaves the production line, it is subjected to a final test to make sure it performs the functions required of it in actual use.



Film and hybrid circuits

Film circuits

In film integrated circuits, passive circuit elements (usually conductor tracks, insulation and resistors, but also capacitors and inductors) are applied to a base (substrate) in layers. The advantages of these circuits are:

- ▶ Fine structures (up to approx. 10 µm) with high element density and
- ▶ Good high-frequency characteristics

Those advantages are counterbalanced by relatively high production costs.

The thickness of the films originally gave rise to the terms “thin-film circuit” and “thick-film circuit”. Classification is now based on differences in the method of production.

Thin-film circuits

With thin-film circuits, the films are applied to glass or ceramic substrates using vacuum-coating processes.

Thick-film circuits

For thick-film circuits, the preferred method of manufacture involves application of the films to ceramic substrates by screen printing processes followed by firing.

Ceramic multilayer substrates

Ceramic multilayer substrates are made up of unfired ceramic foils onto which conductor tracks are applied by screen printing methods. Several such foils are then combined to form a multilayer laminate which is fired at 850 to 1,600 °C so that it becomes a solid ceramic element with integral conductor channels. A typical hybrid substrate consists of four or five layers. Particularly high wiring densities can be achieved with LTCC-line substrates (Low-Temperature Cofired Ceramic, i.e. ceramic material which has been fired (sintered) together (simultaneously) at low temperature).

To make the electrical connections between layers, holes are punched in the individual films and filled with metal paste. Those holes are referred to as so-called “vias”. Special materials are employed to integrate resistors and capacitors in the device. The wiring densities of these circuits is considerably greater than with thick-film circuits.

Hybrid circuits

Hybrid circuits are integrated film circuits with additional discrete components such as capacitors and integrated semiconductor circuits (ICs) that are attached by soldering or cementing. The use of unpacked semiconductor chips, to which the connections are made by bonding, or SMDs makes high component densities possible. A ceramic multilayer substrate can be used to achieve extremely small hybrid control devices (microhybrids). The advantages of these circuits are:

- ▶ High permissible installation temperatures due to their good heat dissipation properties
- ▶ Compactness of design combined with good vibration resistance and
- ▶ Good resistance against media

Hybrid circuits are therefore particularly suited to use in telecommunications systems and for automotive applications where they can be found in ABS, traction control, ESP, transmission control and engine-management control units (engine-mounted).

Figure 8 shows the basic stages in the production of a hybrid-circuit substrate. The tapes are punched with holes for the vias for each separate wiring layer and the holes are then filled with silver paste (Fig. 9a). Screen-printing stations print the conductor tracks onto the film. The various layers are aligned with each other, laminated and then fired at 890 °C. A specially controlled sintering process keeps the ceramic firing to within toler-

ance limits of approximately 0.03 %. This is important for the packing density. The circuit's resistors are printed on the reverse side of the substrate and fired (Fig. 9b).

For bonding on the top side, the surfaces are finished using a plating process adapted to the LTCC. The spacing of the microcontroller contacts (bond land grid on the substrate) ranges from 450 to 260 µm. The component connections are bonded using 32-µm gold wire and 200-µm aluminum wire.

Alongside the electrical vias, there are also thermal vias with a diameter of 300 µm for optimum cooling of ICs with high power losses. The effective thermal conductivity of the substrate can thus be increased from approx. 3 W/mK to 20 W/mK.

All components are cemented with conductive cement. There are two methods used for final assembly of the finished hybrid.

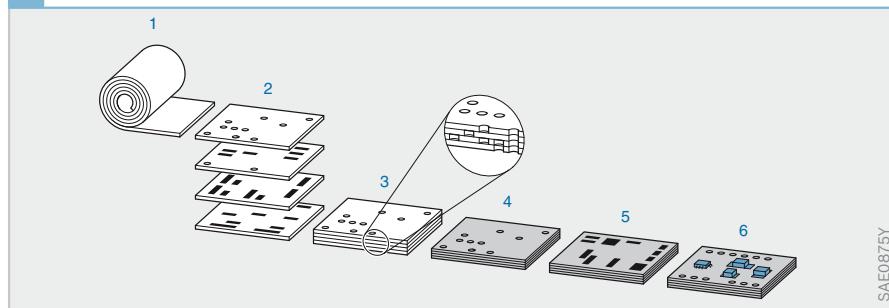
Method 1: the finished hybrid is cemented to the steel panel of the casing using heat-conducting cement and connected to the glass feed-through for the connector using bonded 200-µm aluminum wire. The casing is then hermetically sealed.

Method 2: the finished hybrid is cemented to the aluminum casing using heat-conducting cement and connected to the plastic-encased connector pins using bonded gold or 300-µm aluminum wire. Before the cover is cemented in place, a gel is applied to protect the circuit.

Fig. 8

- 1 Unfired glass-ceramic film
- 2 Punching of holes, filling with conductor paste and printing of conductor tracks
- 3 Aligning and stacking (laminating)
- 4 Sintering
- 5 Printing of resistors (reverse side), firing and plating of the bonding pads (top side)
- 6 Fitting of components and wire bonding

8 Production process of a microhybrid substrate



9 Example of a hybrid circuit (sections)

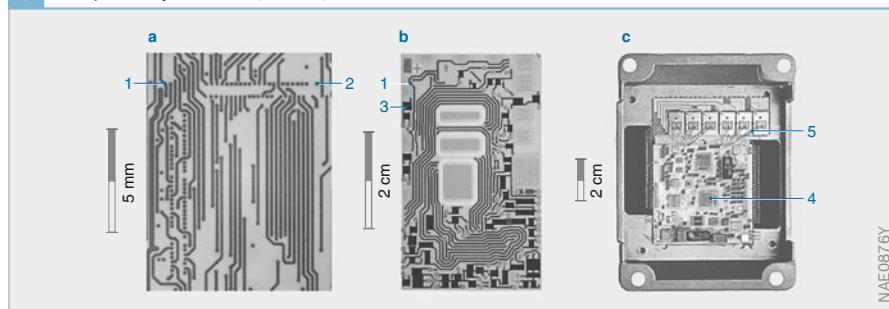


Fig. 9

- | | |
|---|-----------------------------|
| a | Inner layer |
| b | Reverse side with resistors |
| c | Top side inside ECU |
-
- 1 Conductor track
 - 2 Via
 - 3 Resistor
 - 4 Microcontroller
 - 5 Bonded wire

Control units

Digital technology furnishes an extensive array of options for open and closed-loop control of automotive electronic systems. A large number of parameters can be included in the process to support optimal operation of various systems. The control unit receives the electrical signals from the sensors, evaluates them, and then calculates the triggering signals for the actuators. The control program, the “software”, is stored in a special memory and implemented by a microcontroller. The control unit and its components are referred to as hardware. The Motronic control unit contains all of the algorithms for open and closed-loop control needed to govern the engine-management processes (ignition, induction and mixture formation, etc.).

Operating conditions

The control unit is subjected to very high demands with respect to

- ▶ Extreme ambient temperatures (in normal vehicle operation from -40 to +60...+125 °C)
- ▶ Extreme temperature changes
- ▶ Indirect materials and supplies (oil, fuel etc.)
- ▶ The effects of moisture and
- ▶ Mechanical stress such as vibration from the engine

The control unit must operate reliably when the vehicle is started with a weak battery (e.g. cold start) and with high charge voltages (vehicle electrical system fluctuations).

Other requirements arise from the need for EMC (ElectroMagnetic Compatibility). The requirements regarding immunity to electromagnetic interference and limitation of high-frequency interference signal emission are extremely stringent.

Design

The printed circuit board with the electrical components (Fig. 1) is installed in a housing of plastic or metal. A multiple plug connects the control unit to the sensors, actuators and electrical power supply. The high-performance driver circuits that provide direct control of the actuators are specially integrated within the housing to ensure effective heat transfer to the housing and the surrounding air.

The majority of the electrical components are of the surface-mounted device technology type. This concept provides extremely efficient use of space in low-weight packages. Only a few power components and the connectors use push-through assembly technology.

Hybrid versions combining compact dimensions with extreme resistance to thermal attack are available for mounting directly on the engine.

Data processing

Input signals

In their role as peripheral components, the actuators and the sensors represent the interface between the vehicle and the control unit in its role as the processing unit. The electrical signals of the sensors are routed to the control unit via a wiring harness and the connector plug. These signals can be of the following type:

Analog input signals

Within a given range, analog input signals can assume practically any voltage value. Examples of physical quantities which are available as analog measured values are intake-air mass, battery voltage, intake-manifold and boost pressure, coolant and intake-air temperature. They are converted into digital values by an analog-digital converter in the microcontroller

of the control unit and used for calculations by the microcontroller CPU. The maximum resolution of these analog signals is 5 mV. This translates into roughly 1,000 incremental graduations based on an overall measuring range of 0 to 5 V.

Digital input signals

Digital input signals only have two states. They are either “high” or “low” (logical 1 and logical 0 respectively). Examples of digital input signals are on/off switching signals, or digital sensor signals such as the rotational-speed pulses from a Hall generator or a magnetoresistive sensor. Such signals are processed directly by the microcontroller.

Pulse-type input signals

The pulse-shaped input signals from inductive-type sensors containing information on rotational speed and reference mark are conditioned in their own control unit stage. Here, spurious pulses are suppressed and the pulse-shaped signals converted into digital rectangular signals.

Signal conditioning

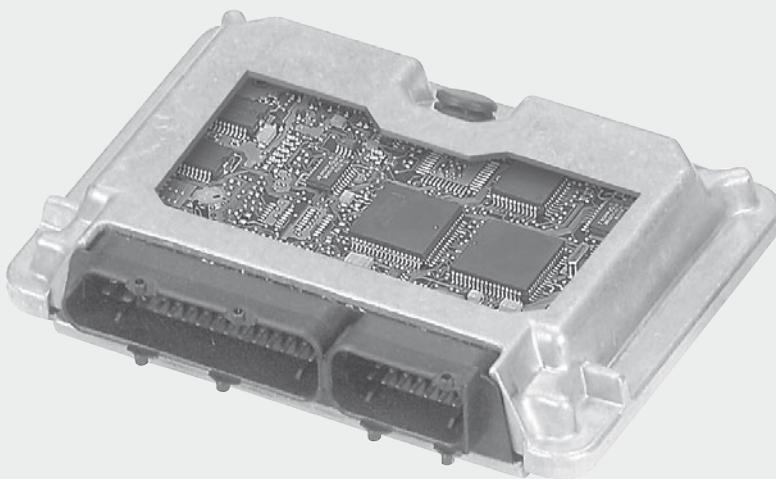
Protective circuits limit the voltages of input signals to levels suitable for processing. Filters separate the useful signal from most interference signals. When necessary, the signals are then amplified to the input voltage required by the microcontroller (0 to 5 V).

Signal conditioning can take place completely or partially in the sensor depending upon the sensor's level of integration.

Signal processing

The control unit is the switching center governing all of the functions and sequences regulated by the engine-management system. The closed and open-loop control functions are executed in the microcontroller. The input signals that are provided by the sensors and the interfaces to other systems (such as the CAN bus) are used as input variables, and are subjected to a further plausibility check in the computer. The control unit program supports generation of the output signals used to control the actuators.

1 Design of a control unit using the example of an ME Motronic (sectional view through housing cover)



UAE0992Y

Output signals

The microcontroller uses the output signals to control output stages that usually provide enough power for connecting the actuators directly. It is also possible to activate certain output stage relays for consumers that use up a great deal of power (e.g. motor fans).

The output stages are proof against short circuits to ground or battery voltage, as well as against destruction due to electrical or thermal overload. Such malfunctions, together with open-circuit lines or sensor faults are identified by the output-stage IC as an error and reported to the microcontroller.

Switching signals

Actuators can be switched on and off using the switching signals (e.g. motor fans).

PWM signals

Digital output signals can be in the form of PWM (Pulse-Width Modulated) signals. These are constant-frequency rectangular signals with variable on-times (Fig. 2). Various actuators can be moved to various operating positions using these signals (e.g. exhaust-gas recirculation valve, boost-pressure actuator).

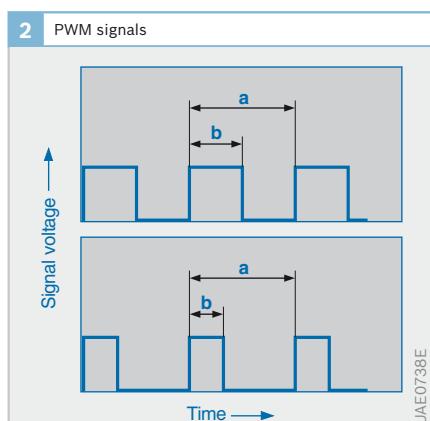


Fig. 2

- a Period duration (fixed or variable)
- b Variable on-time

Control unit-internal communication

In order to be able to support the microcontroller in its work, the peripheral components must communicate with it. This takes place using an address/data bus which, The microcontroller outputs the RAM address whose contents are to be read (for example) via the address bus. The data bus is then used to transmit the relevant data. For former automotive applications, an 8-bit bus topology sufficed. This meant that the data bus comprised 8 lines which together could transmit 256 values simultaneously. 65,536 addresses can be accessed using the 16-bit address bus in this system. Presently, more complex systems demand 16 bits, or even 32 bits, for the data bus. In order to save on pins at the components, the data and address buses can be combined in a multiplex system, i.e. addresses and data are dispatched through the same lines but offset from each other with respect to time.

Serial interfaces with just a single data line are used for data that does not have to be transmitted extremely quickly (e.g. fault memory data).

EoL programming

The extensive variety of vehicle variants with differing control programs and data records, makes it imperative to have a system which reduces the number of control unit types needed by a given manufacturer. To this end, the Flash-EPROM's complete memory area can be programmed at the end of production with the program and the variant-specific data record (this is the so-called End-of-Line, or EoL, programming).

Another way of reducing the type diversity is to store several data variants (e.g. transmission variants) in the memory, which are then selected using coding at the end of the production line. This coding is stored in an EEPROM.

► Performance of electronic control units

The performance of electronic control units goes hand-in-hand with advances achieved in the field of microelectronics. The first gasoline injection systems were still analog – with limited flexibility in the implementation of control functions. These functions were constrained by the hardware.

Progress advanced in quantum leaps with the arrival of digital technology and the microcontroller. The entire engine management system was taken over by the universally applicable semiconductor microchip. The actual control logic in microcontroller-controlled systems is in a programmable semiconductor memory.

From systems that initially simply controlled fuel injection, complex engine-management systems were then developed. They controlled not only fuel injection but also the ignition system including knock control, exhaust-gas recirculation and a whole variety of other systems. This continuous process of development is bound to continue in a similar vein over the next decade as well. The integration of functions and, above all, their complexity are constantly increasing. This pattern of development is only possible because the microcontrollers used are also undergoing a similar process of improvement.

Microcontrollers in the Intel 8051 family were used quite some time until they were replaced with 80515 derivatives with additional

I/O facilities for timer-controlled signals and an integrated analog-digital converter at the end of the 1980's. It was then possible to create relatively powerful systems. Figure 3 shows a comparison between the performance of a fuel-injection system (LH3.2) and an ignition system (EZ129K) – equipped with 80C515 controllers – and that of the succeeding Motronic systems. The ME7 has approximately 40 times the performance capability of the LH/EZ combination with a clock frequency of 40 MHz. With the benefit of a new generation of microcontrollers and a further increase in clock frequency on the ME9, this figure will increase to a factor of well over 50.

In the foreseeable future microcontrollers will process more than just digital control sequences. Signal processors are integrated that can also directly process the signals provided by knock sensors, for example.

Advances in the development of semiconductor memory chips are also worthy of note. Complex control programs require an enormous amount of memory space. The capacity of memory chips at the start of the 1980s was still only 8 kilobytes. The ME7 now uses 1-megabyte chips and soon memory capacities of 2 megabytes will be required. Figure 3 shows this pattern of development and likely future trends.

3 Development of electronic control units

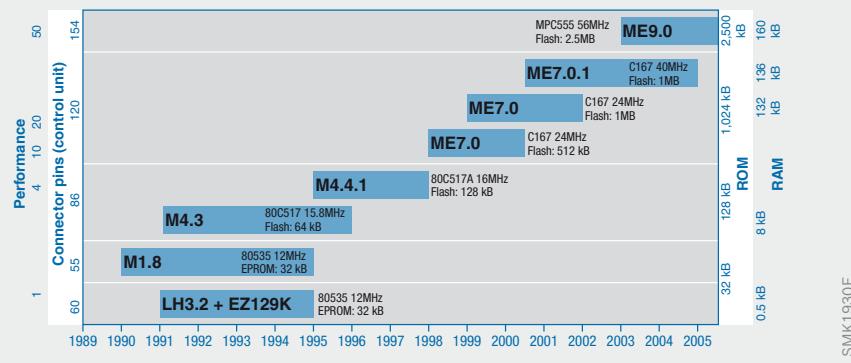


Fig. 3

Chart illustrating

- ▶ Performance capability of engine-management systems
- ▶ Number of control unit connector pins
- ▶ Program memory capacity
- ▶ Data memory capacity (RAM)

By way of comparison:

The performance capability of a state-of-the-art engine-management system far exceeds that of Apollo 13.

Digital modules in the control unit

Microcontroller

Structure

A microcontroller consists of the following interacting components (Fig. 1):

- ▶ **Central processing unit (CPU)**: this contains the control unit and the arithmetic and logic unit. The control unit executes the instructions from the program memory, whereas the arithmetic and logic unit performs arithmetical and logical operations.
- ▶ **Input and output devices (I/O, Input/Output)**, which handle the exchange of data with peripheral devices. Peripheral devices include input and output devices and external data storage media.
- ▶ **Program memory**, in which the operating program (user program) is permanently stored (ROM, PROM, EPROM or flash EPROM).
- ▶ **Data memory**, which is accessed for reading and writing (RAM). This contains the data that is currently being processed. Non-volatile memory (EEPROM) is used for data that must not be deleted when the supply voltage is switched off.
- ▶ The **bus system** connects the individual elements of the microcontroller.
- ▶ A **clock generator** (oscillator) ensures that all operations in the microcontroller take place within a defined timing pattern.
- ▶ **Logic circuits** are modules with specialized tasks such as program interrupts. They are integrated in individual I/O units.

The chief components of a microcomputer are generally separate modules connected to one another on a printed-circuit board. The microprocessor within such as system - the CPU - is not functional on its own: it is always part of a microcomputer.

In a microcomputer, however, the above-mentioned functions are integrated on a silicon wafer (system-on-a-chip). This is not functional on its own (standalone) and is therefore referred to as a single chip microcomputer.

The microcontroller is used to control self-regulating systems such as an engine-management system. Depending on the application, they may also have expansion modules connected to them (e.g. additional memory for data and program code).

The user program is fixed and is not replaced for different applications. This is the difference between a microcontroller system, for example, and a PC.

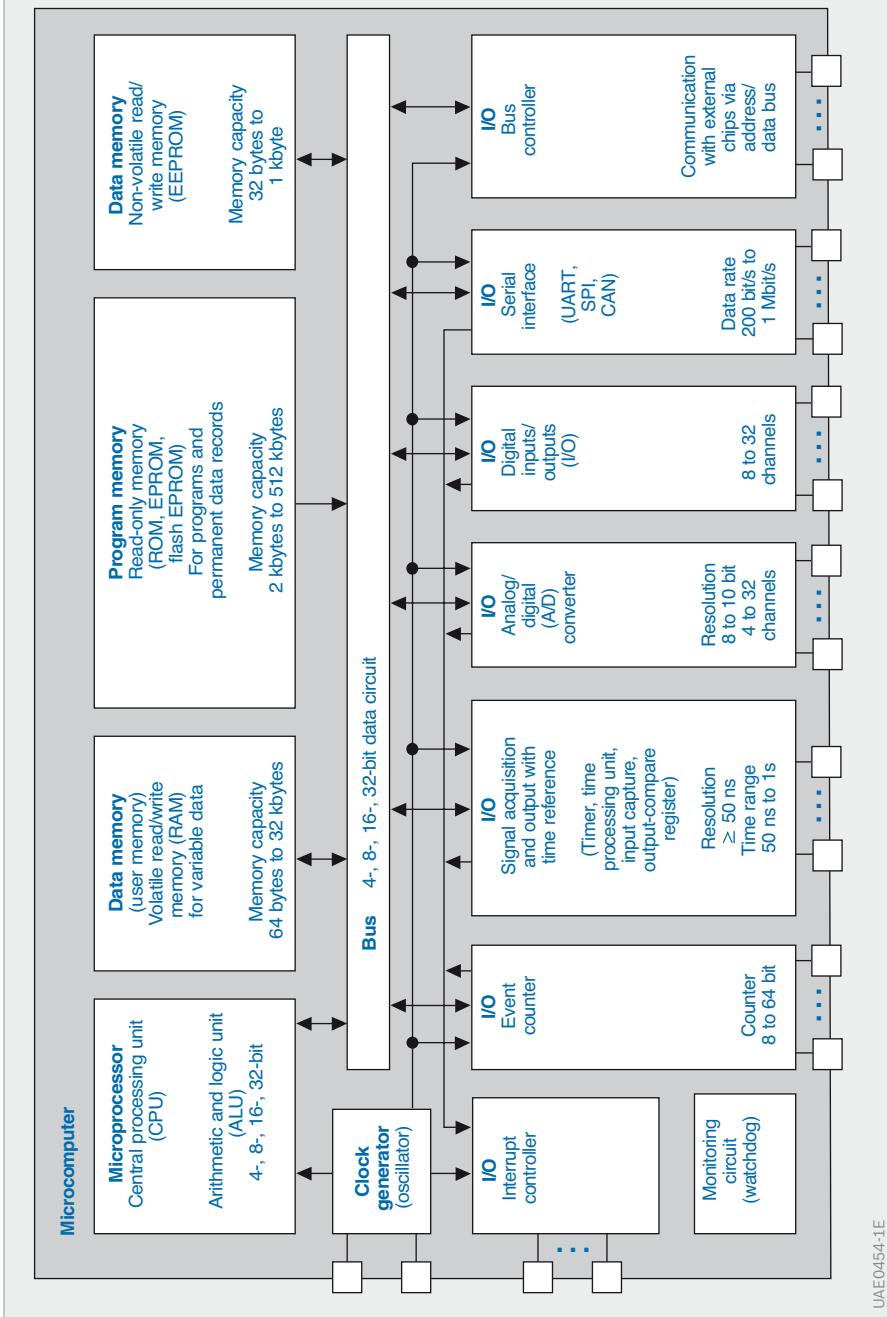
Programming

The only command form capable of direct interpretation by a microprocessor is a bit pattern, i.e. the binary representation of a number. Since, however, this form of instruction is not easy to work with for a programmer, and is therefore susceptible to errors, easily memorable abbreviations (mnemonics) are used. These are automatically translated by an assembler program into bit patterns (machine code) that can be understood by the microprocessor.

For more complex systems and programs, high-level programming languages such as C are needed, as otherwise it would be impossible to keep extensive programs manageable and free of errors. Such languages require sophisticated translation programs (compilers) which convert the text of the high-level language into a form that can be processed by the microcontroller.

The machine code is stored in the program memory, where it remains permanently. The CPU accesses these components via the bus system, reads the numerically coded commands and executes them.

1 Microcontroller



Semiconductor memories

Applications

- Memories are used to store large volumes of
- ▶ Digital signals representing data (I/O data, statuses, intermediate results involving frequent and rapid reading and writing)
 - ▶ Program code (usually permanently stored) and
 - ▶ Constants (permanently stored)

Storage involves

- ▶ Recording (writing)
- ▶ Permanent retention (actual storage) and
- ▶ Location and retrieval (reading) of information

Memories utilize physical effects that clearly and easily create and show two different states (e.g. conducting/non-conducting or loaded/not loaded). The information that is to be stored must therefore exist in binary form, i.e. encoded as a series of “yes or no” statuses (logical “1” or logical “0”). Such a “yes or no” unit of information is called a bit (binary digit).

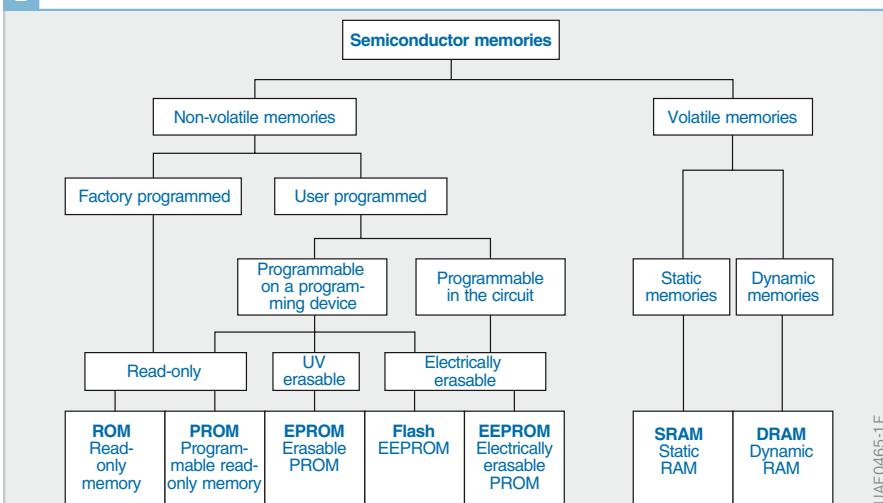
Memory modules are organized on a bit or word basis, depending on the application. A “word” is a group of bits that can be processed as a single unit. The word length is equal to the number of bits processed as a single unit. Eight bits are referred to as a byte.

Memories can be organized on the basis of different word lengths. An 8 M x 8-RAM, for example, has a memory capacity of 8 million times 8 bits (64 Mbits). The data is organized into bytes (8 bits), making the memory capacity 8 Mbytes.

Word lengths of 4, 8, 16 and 32 bits are common in microcontroller systems. The word length is one of the factors that determines the performance capability of the system. The word length that is used depends on the performance capability requirements of the system.

The most important terms are explained below according to their standardized definitions, where applicable, or their most common usage (see Figure 2 for overview).

2 Overview of semiconductor memories



Random-access memory (RAM)

Random-access memory or RAM is a short-term memory that allows direct access to any storage location. Information can be read/written from/to the memory any number of times.

Static RAM (SRAM)

Static RAMs use bistable switching elements as the data storage cells. Their functionality is similar to that of a flip-flop, a simple circuit with two transistors, of which either the one (logical "1") or the other (logical "0") conducts at any one time. In SRAM, the information remains stored until the storage cell concerned is addressed and overwritten, or the operating voltage is switched off. SRAM is therefore volatile memory.

Dynamic RAM (DRAM)

Unlike SRAM, the information is stored as an electrical charge in the gate capacity of a CMOS transistor in dynamic RAM (DRAM). As such capacitors are susceptible to leakage, the charge is gradually lost. In order to retain the information, the charge has to be refreshed at regular intervals (every few ms).

Read-only memory

Read-only memory (ROM) is permanent-storage memory that allows any memory location to be accessed directly but - as the name indicates - allows the information only to be read and not modified.

A ROM is nonvolatile memory, i.e. the information it contains is retained even when the operating voltage is switched off. It is usually used to store program code (control programs) and fixed data (function tables, encoding rules, engine characteristic data maps) that need to be retrievable at any time. The information may be indelibly entered in the memory by the manufacturer or the user by means of appropriate programming of specially prepared memories (PROMs or programmable ROMs).

Erasable ROM

There are also ROMs whose contents can be erased and reprogrammed as outlined below.

EPROM (Erasable PROM)

This type of erasable read-only memory can have its contents completely wiped by irradiation with UV light and can then be reprogrammed using a programming device.

EEPROM (Electrical EPROM)

The EEPROM (also known as E²PROM) can be electrically erased and reprogrammed. Every storage cell of an EEPROM can be individually overwritten. For that reason, this type of memory module can also be used as nonvolatile data memory (e.g. for learned information in engine management systems).

Flash EEPROM

A more sophisticated variant of the EPROM and EEPROM is flash EEPROM. In this case, electrical flash pulses are used to erase specific storage areas or the entire contents of the memory. The erased areas can subsequently be reprogrammed.

The flash memory can be reprogrammed on a programming station. However, the advantage of flash EEPROM is that it can also be reprogrammed while still inside the sealed control unit.

Flash EEPROM is used in cases where relatively large quantities of data need to be stored, but must also be modifiable (e.g. program memory in vehicle control units).

Control unit software

Real-time capability

One of the requirements on electronic systems is real-time capability. This means that control procedures must react to input signals within an extremely short time. For example, a wheel that has a tendency to lock must be detected so quickly that the ABS control algorithm in the control unit can reduce the brake pressure via the hydraulic modulator quickly enough. Brake slip is therefore reduced before the wheel can lock. Engine management systems make considerable real-time capability demands so that crankshaft angles can be adhered to with extreme accuracy at fast engine speeds for injection and ignition timing purposes.

The complexity of an electronic system therefore makes extremely high demands of the software that is developed. The software structure is explained below on the basis of an example.

Software structure

The microcontroller in the control unit executes commands sequentially. The command code is obtained from the program memory. The time taken to read in and

execute the command depends on the microcontroller that is used and the clock frequency. The microcontrollers that are currently used in vehicles can execute up to 1 million commands per second.

Because of the limited speed at which the program can be executed, a software structure with which time-critical functions can be processed with high priority is required.

The engine-management program has to react extremely quickly to signals from the speed sensor, which records the engine speed and the crankshaft position. These signals arrive at short intervals that can be a matter of milliseconds, depending on the engine speed. The control unit program has to evaluate these signals with high priority. Other functions such as reading in the engine temperature are not as urgent, since the physical variable only changes extremely slowly in this case.

Interrupt control

As soon as an event occurs that requires an extremely rapid response (e.g. speed sensor pulse), the program that is currently running must be interrupted. This can be done using the microcontroller's interrupt control facility. Events can trigger a program execution interrupt, whereupon the program jumps and executes the "interrupt routine". When this routine has been executed, the program resumes at the point at which it was interrupted (Fig. 1).

An interrupt can be triggered by an external signal, for example. Other interrupt sources are timers integrated in the microcontroller, with which timed output signals can be generated (e.g. ignition signal: microcontroller ignition output is switched at a point in time that is calculated beforehand). However, the timer can also generate internal time frames.

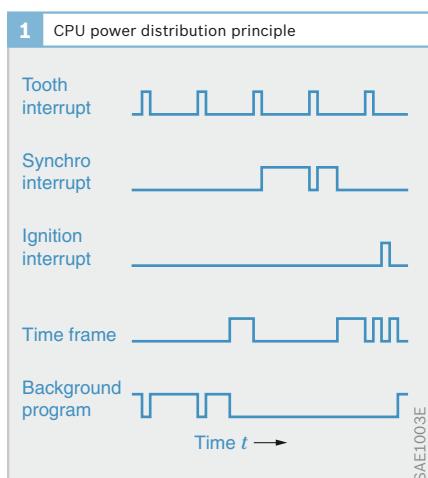


Fig. 1

Depiction of several program levels on the example of the software from a Motronic system

SAE1003E

The control unit program reacts to several of these interrupts. An interrupt source can therefore request an interrupt while another interrupt routine is currently being executed. Every interrupt source therefore has a fixed priority assigned to it. The priority controller decides which interrupt is allowed to interrupt another interrupt.

Tooth interrupt

The crankshaft is equipped with a pulse wheel (Fig. 2a) that has a certain number of teeth on its circumference. The teeth are scanned by the speed sensor. This allows the crankshaft position to be recorded. The typical distance between a pair of teeth on the crankshaft sensor wheel is 6°. In order to determine the crankshaft position, the control unit program must execute certain routines as each tooth is detected. At 6,000 rpm the detection time between two teeth is approximately 300 µs. Every command in these routines must be executed within this time. This requires a rapid response

to the speed sensor signal. For this purpose, the engine-speed signal is connected to a microcontroller interrupt input. Every falling signal edge at this input interrupts the current calculations that are in progress and forces a branch to the interrupt routine. After executing the commands in the interrupt routine, the program continues execution at its point of origin.

In order to perform certain operations the control unit program requires the time taken for the crankshaft to travel between one tooth and the next. This calculation is performed by an internal timer. This is a freewheeling 16-bit counter (Fig. 3) that increments at a certain rate, depending on the microcontroller oscillator clock cycle. This time frame amounts to about 0.5 µs. When the falling tooth flank occurs, the current counter status is recorded. The difference (and therefore the tooth interval) is calculated using the stored counter from the previous tooth.

Example: crankshaft position calculation
The engine-management system (Motronic for gasoline engines, EDC for diesel engines) must know the crankshaft position at any given point in time. This is a prerequisite for injecting into the right cylinder at the right time and ensuring that ignition takes place at the calculated ignition angle (Motronic systems). In order to

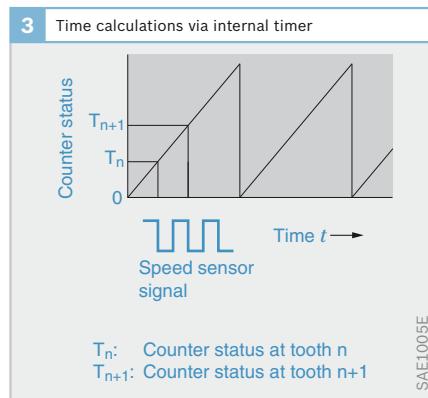
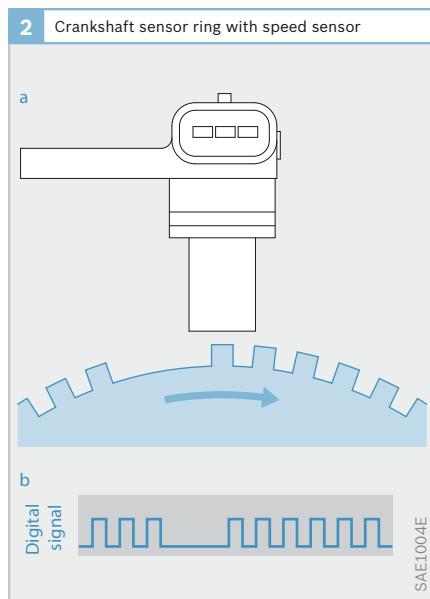


Fig. 2
a Design
b Speed sensor signal

detect the engine position and the engine speed, the control unit evaluates the speed sensor signal (Fig. 2b).

There is gap in the crankshaft sensor wheel in which two teeth are missing. The tooth space has a defined position in relation to the top dead center (TDC) of cylinder no. 1. The control unit program has to synchronize itself with this tooth space. This is done by measuring the times between two consecutive falling tooth flanks. The time for the tooth space is considerably greater than the time before and after the gap. Following a “short – long – short” sequence the last thing to be scanned was the falling flank of the second tooth after the space.

The crankshaft has rotated by 6° for each falling tooth flank that has been detected by the control unit program. This is how the control unit program knows the crankshaft position within this time frame.

Since cylinder no. 1 is in the tooth space position in the vicinity of top dead center (TDC) or bottom dead center (BDC), an additional signal is required to determine the position. The camshaft sensor provides a different voltage level in both cases. The control unit is therefore able to uniquely assign the crankshaft and camshaft positions.

Combustion-synchronous interrupt

Some calculations have to be performed for every combustion cycle. For example, the ignition angle and the injection have to be recalculated synchronously with combustion for each cylinder. The program does this by branching to the “synchronization program” after certain teeth (Fig. 4). This interrupt takes place after every 30 teeth (ignition interval) for a four-cylinder engine, and after every 20 teeth for a six-cylinder engine.

The synchronization program is fixed to a certain tooth position and has to be executed with high priority. For this reason it is activated via an interrupt (triggered by a command in the tooth interrupt routine).

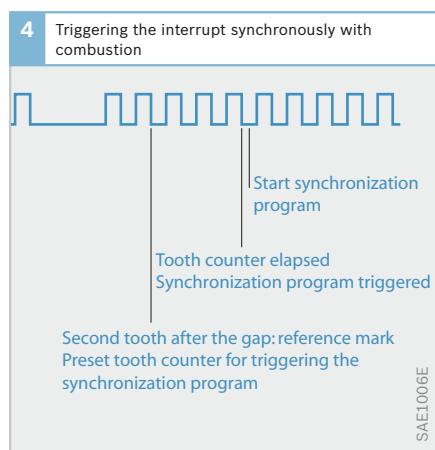
Since the synchronization program runs over several teeth at fast engine speeds, it has to be interrupted by the tooth interrupt. The tooth interrupt is given higher priority than the synchronization program.

Ignition interrupt

The ignition output takes place within a certain crankshaft range, depending on the value from the ignition map. Since the specified ignition angle has to be adhered to exactly, the ignition output is controlled by an interrupt. Like the synchronization program, the ignition interrupt is also called up once per combustion cycle.

The control unit program is aware of the crankshaft position in the 6° framework. However, this framework is not accurate enough for ignition angle output. For this reason, accurate ignition output between two teeth must take place as well as this approximate counting for the last 0 to 6° crankshaft degrees. This is done using a timer (Fig. 5). Ignition angle output that was purely timer-controller would lead to an ignition angle output error at high engine speed dynamics.

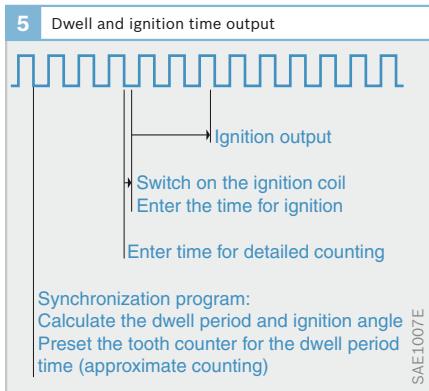
Firstly, the ignition coil must be enabled for a defined time (the so-called dwell period). In order to do this, the program calculates the switch-on time by calculating



backwards from the ignition angle at which the ignition coil has to be switched off. This makes it possible to calculate the tooth after which the ignition coil has to be switched off (approximate counting in 6° time frame). The remaining angle (detailed counting 0 to 6°) is converted into an output time using the current engine speed. As soon as the specified tooth position has been reached using approximate counting, a time is loaded with the output value from the detailed counting. When this time period expires, the timer triggers an interrupt. The commands that switch on the ignition coil are programmed in this interrupt routine. Then the timer is preset to the dwell period value, which causes an interrupt to be triggered when the timer elapses, switching off the ignition coil and therefore initiating ignition.

Time frame

Many control algorithms have to run within a certain time frame. Lambda control, for example, has to be processed within a fixed time frame (e.g. 10 ms) so that the correcting variables are calculated quickly enough.



Background program

All other activities that do not run in an interrupt routine or a time frame are processed in the background program. At fast engine speeds, the synchronization program and the tooth interrupt are called frequently, leaving little CPU time for the background program. The time taken for a complete run-through of the background program therefore increases rapidly with the increasing engine speed. The background program must therefore only contain low-priority functions.

Automotive sensors

The term sensor has become common, as in the past 20 to 40 years measuring gages have also come into use in consumer applications (e.g. motor vehicle and domestic appliance technology). Sensors – another term for measuring detectors or measuring sensors – convert a physical or chemical (generally non-electrical) variable Φ into an electrical variable E ; this process often also takes place over further, non-electrical intermediate stages.

Table 1 summarizes and compares the various fields of application for sensors. Figure 1 provides an overview of the abundance of electronic vehicle systems already on the market. Undoubtedly, this number will increase immensely in the years to come.

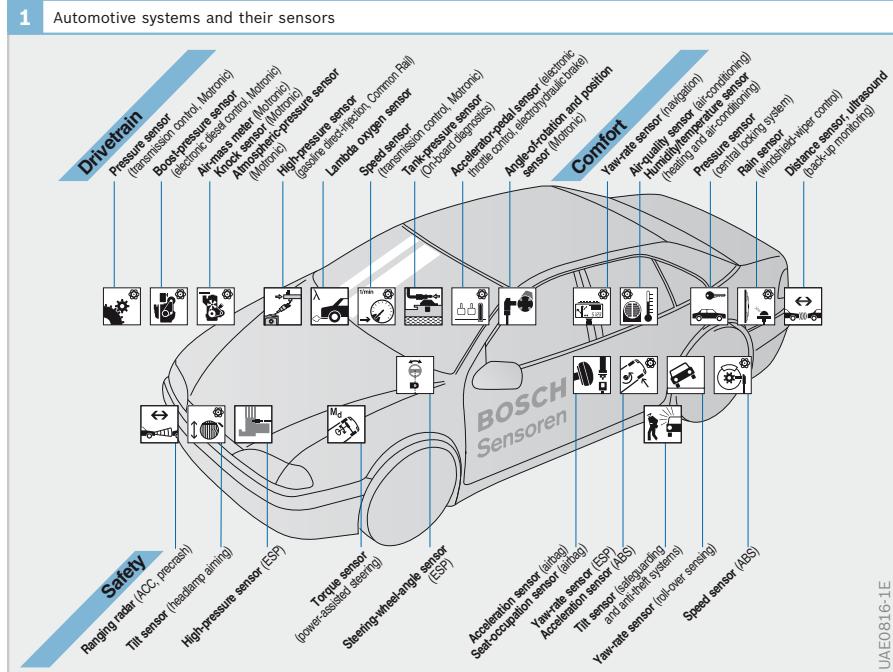
Basics and overview

Sensor, term/definition

The electrical sensor outputs are not only provided in the form of current and voltage, but are also available as current or voltage amplitudes, frequency, phases, pulse widths, and cycles or periods of an electrical oscillation, or as the electrical parameters, resistance, capacitance, and inductance. A sensor can be defined using the following equation:

- (1) $E = f(\Phi, Y_1, Y_2\dots)$
Sensor output signal
- (2) $\Phi = g(E, Y_1, Y_2\dots)$
Required measured variable

If the functions f or g are known, then they represent a sensor model with the help of which the measured variable sought may be mathematically calculated from output



1 Areas in which sensors are used

Typical features	Primary standards	Precision measurement	Industrial measurement	Consumer technology
Accuracy	10^{-11} to 10^{-7}	2 to $5 \cdot 10^{-4}$	2 to $5 \cdot 10^{-3}$	2 to $5 \cdot 10^{-2}$
Costs	€ 100 k to € 1 m	€ thousands	€ hundreds	€ 1 to 10
Units/year	single figures	approximately 10	100 to 1 k	10 k to 10 m
Use	<ul style="list-style-type: none"> – Research – Testing of secondary normals 	<ul style="list-style-type: none"> – Calibration 	<ul style="list-style-type: none"> – Process instrumentation – In-process measurement 	<ul style="list-style-type: none"> – Auto electronics – Building services

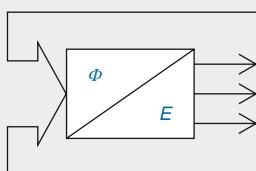
Table 1

signal E and the influencing variables Y_i practically without error (intelligent or smart sensors).

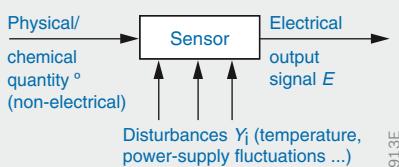
Adjustment

In the real world, the sensor model always includes some free parameters, with which the model can be adapted to the actual properties of the individual sensor type in a kind of calibration process (Fig. 4a).

These model parameters are generally stored in a programmable, nonvolatile memory component (PROM) in the digital processes for conditioning sensor signals, which have come to dominate. By contrast with the conventional analog compensation of influencing variables, it is not only roughly linearly acting influences that may be corrected, but also strongly nonlinear processes that can be corrected successfully. Another significant advantage of this type of calibration, which takes place over a purely electrical connection, is that each sensor may easily be maintained under operating conditions during the calibration phase.

2 Sensor symbol

UAE0815Y

3 Basic sensor function

LAE0913E

▶ Concept of the smart sensor

In a somewhat more general form, intelligent sensors (smart sensors) may be defined as follows: Intelligent sensors, sometimes also known as integrated sensors or sensors with specific local electronics, allow the (static or dynamic) accuracy inherent in a sensor to be exploited by means of (generally digital) microelectronics to a far greater degree than conventional sensors. The sensor information, in particular the complex information from multisensor structures, can be compressed here by local further processing, that is to say, it is brought to a higher level (than the sensor alone is capable of providing), without a multiplicity of external connections being required for this.

There is no clear determination as to whether sensors can already incorporate a part of the signal processing or not; it is, however, recommended, for example, that no distinction be made between elementary sensor, sensor cell, or the like and an integrated sensor.

Calibration process

A smart sensor is generally programmed or calibrated, in a procedure corresponding to the calibration of conventional analog sensors, in three steps with the help of an external computer (host) (Fig. 4):

Recording the actual value

The host computer systematically varies both the measured variable x_e and the influencing variable(s) y , thus setting a specific number of relevant and representative operating points. The smart sensor here passes the, still uncorrected, raw signals x_a^* to it. The host, however, is simultaneously also receiving the “true” values x_e and y from substantially more accurate reference sensors. The host computes the correction values necessary by comparing these two values, and interpolates these corrections over the full measuring range.

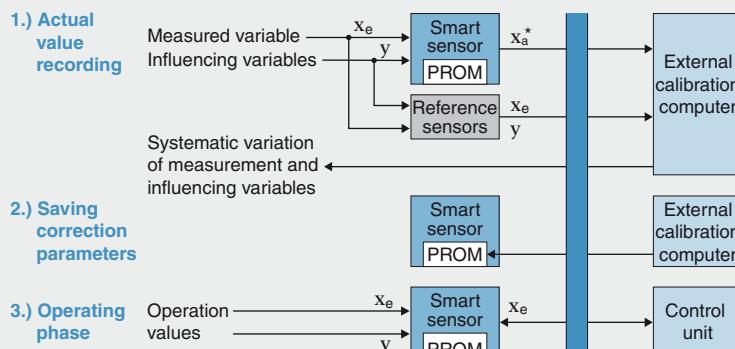
Saving the correction parameters

The host computer calculates the model parameters for the specific sensor from the data previously obtained, e.g. for a linear characteristic curve course and saves these to the smart sensor's PROM. In a check process, these parameters may also first be emulated in the host computer's RAM before they are finally burnt to the nonvolatile memory of the smart sensor. If characteristic curves are approximated with higher order polynomials, program maps (look-up tables) may also be stored in the smart sensor to avoid protracted computing processes. The solution of saving a coarse-grid program map in conjunction with a simple linear interpolation between the data points (example shown in Fig. 5) has also proved very useful.

Operating phase

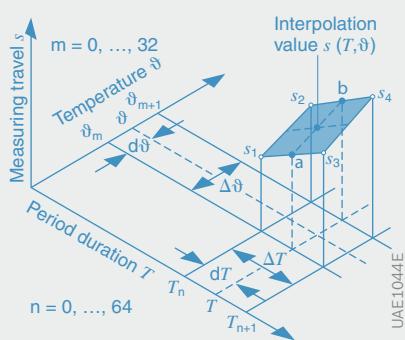
The smart sensor is now disconnected from the host computer and is in a position itself to calculate the measured variable x_e with very low error thanks to the stored model data. It is able to pass the value to a connected control unit in digital, bit-serial or even analog form (e.g. pulse-width modulated), for instance. The measured variable can also be distributed digitally to further control units via a bus interface.

4 Trimming/calibration of a smart sensor



This process can, unlike conventional laser calibration, in principle also be repeated if an erasable PROM is used. This is a particular advantage in the development phase of sensors.

5 Interpolation of measured values using a data point program map



Example: two-dimensional data point-program map $s(T_n, \vartheta_m)$ of a smart sensor for the measurement of a travel s :

For a highly precise evaluation of the sensor, which operates using variable inductance, its natural characteristic curve and its temperature sensitivity are each approximated using 5th order polynomials. As a frequency-determining element of a very simple oscillator circuit, it outputs the period duration T as an uncorrected output signal. As a sensor model for the measuring travel s , an outline program map embracing only $32 \times 64 = 2,048$ example-specific values $s_{n,m}$ is stored (in the PROM) along with a simple interpolation algorithm (in the ROM), instead of a total of 36 polynomial coefficients and a protracted polynomial evaluation. Should a signal T occur between these data points T_n and T_{n+1} in conjunction with a temperature ϑ between data points ϑ_m and ϑ_{m+1} , the sensor interpolates in two dimensions in accordance with the mapping between the benchmark figures stored "error-free" s_1, \dots, s_4 and thus the measured value $s(T, \vartheta)$ sought is determined as the result of interpolation.

Automotive applications

With increasing demands on all functions in the vehicle, the control and regulation functions that were previously implemented by mechanical means have, in the last 40 years, successively been replaced by electronic devices (ECU, Electronic Control Unit). This has necessarily caused high demand for sensors and actuators with which these electronic control units can, on the one hand determine the relevant states of the vehicle, and on the other hand, actually influence these states. Over this period, the vehicle industry has become one of the, previously unprecedented, drivers of the development of sensors that could be manufactured in large numbers.

Whereas at the outset they still were of electromechanical or macromechanical form of some sort, the trend beginning in the 80s was for miniaturization, with

6 Milestones in the development of sensors for automotive applications

1950	Lambda oxygen sensor
1960	Electromechanical pressure sensor Piezoelectric knock sensor
1970	First integrated Hall sensor Strain-gage acceleration sensor for airbag First pressure sensor on silicone base
1980	Hot-wire air mass meter Thick-film air mass meter Integrated pressure sensor
1990	Micromechanical acceleration sensor for airbag Piezoelectric yaw-rate sensor for ESP Micromechanical air-mass meter Micromechanical yaw-rate sensor
2000	Yaw-rate sensor for roll-over sensing

UAE1045E

semiconductor methods (batch processing) for the manufacture of high use sensors. For a while, sensors originating in hybrid technology using thick-film techniques also played a not insubstantial role. These are still used in isolated applications today, e.g. in the wafer-shaped Lambda oxygen sensors and high temperature sensors for the exhaust line.

Where temperature and magnetic field sensors could initially still be designed in circuit-like structures and be produced in batches, this trend increased as it became possible also to micromachine silicon in many various ways in two or three dimensions, and also to connect them, even in multiple layers, very soundly by very efficient methods.

Where the technologies of the electronic semiconductor circuits were practically exclusively based on silicon as the base material, quite different materials and technologies played a not insignificant role in sensors. Thus, for instance, quartz can also be micromachined using anisotropic etching techniques, and unlike silicon also has very advantageous piezoelectric properties. III-V semiconductors such as

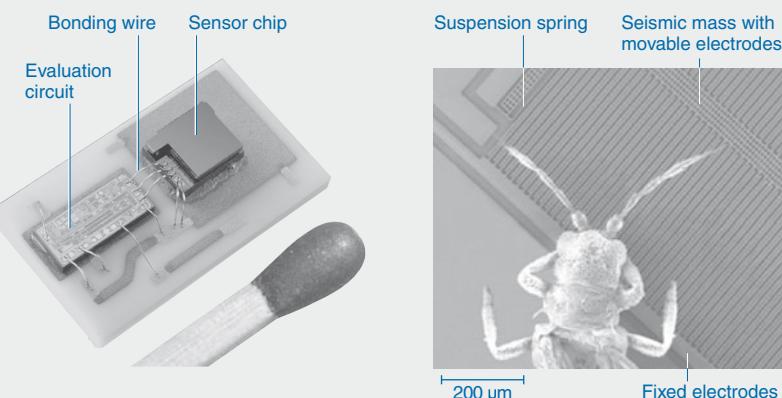
gallium arsenide (GaAs) have a substantially greater operating temperature range than silicon, which can be very advantageous, particularly at some points in the vehicle. Thin metallic layers are very well suited to the manufacture of precise strain-gage resistors, accurate temperature sensors and magnetic field-dependent resistors.

With silicon it is possible also to integrate the electronics monolithically with the sensor. This technique has lost its importance – with a few exceptions (e.g. Hall-effect ICs) – because of the generally very different number and type of process steps and also because of the inflexibility associated with this. Hybrid integration technologies in the tightest of spaces as a rule lead to substantially more cost effective, but functionally equivalent solutions (Fig. 7).

While the development of sensors was initially concentrated almost exclusively on systems inside the detail in the drivetrain, the suspension and the body and driving safety, the direction of sensing of newer developments is increasingly named to the outside and to the area close to and further from the vehicle:

7

Hybrid integration of sensor and electronics: surface micromechanical acceleration sensor on a microhybrid circuit



UAE0787E

- ▶ Ultrasound sensors detect obstacles on parking and will even allow automatic parking in the foreseeable future, perhaps in combination with other sensors.
- ▶ Near-range radar scans the area around the vehicle to detect objects that probably could cause a collision, to gain time and to prime safety systems before a collision occurs (precrash sensors).
- ▶ Imaging sensors can not only detect traffic signs and send them to the driver's display, but also detect the edge of the carriageway, warn the driver of any hazardous deviation and, where required, in the long term also permit automatic driving. In combination with infrared beams and a screen in the driver's field of vision, IR-sensitive imaging sensors could permit long distance observation of the carriageway, even at night (night vision) or in foggy conditions.
- ▶ Long-range radar sensors observe the carriageway for 150 m in front of the vehicle, even in poor visibility, to adapt the driving speed to vehicles ahead and, in the longer term, also to support automatic driving.

As part of the vehicle's peripheral equipment, the sensors and actuators form the vehicle's interface to its complex drive, braking, chassis, and bodywork functions, as well as to the vehicle guidance and navigation functions and the (usually digital) ECUs which operate as the processing units (Fig. 8). An adapter circuit is generally used to convert the sensor's signals into the standardized form (measuring chain, measured-data registration system) required by the control unit.

8 Sensors in the multilayered process that is the motor vehicle

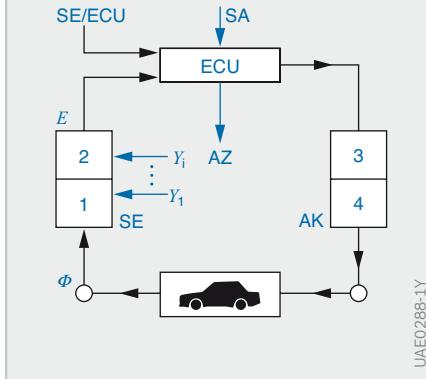


Fig. 8

1	Measuring sensors
2	Adapter circuit
3	Driver circuit
4	Actuators
AK	Actuator
AZ	Display
SA	Operating switch
SE	Sensors
ECU	Control unit
Φ	Physical variable
E	Electrical variable
Y _i	Disturbances

These customer-specific adapter circuits are tailor-made for specific sensors and are available in integrated design and in a wide variety of versions. They represent a quite substantial and very valuable addition to the sensors described here, without which their use would not be possible, and the measuring accuracy of which is, properly said, only defined in conjunction with these.

The vehicle can be regarded as a highly complex process, or control loop, which can be influenced by the sensor information from other processing units (control units), as well as from the driver using his/her controls. Display units keep the driver informed about the status and the process as a whole.

Details of the sensor market

The proportion of value added in the vehicle by the electrical and electronics systems is currently around 26 %. Nowadays, practically every second sensor is fitted into a vehicle, with annual growth rates that are, in some cases, still in double digits. Since the start of the 90s, micromechanical and microsystem-technical sensors have taken a rapidly increasing share, which amounted to about a third in 2005.

Unlike in the general sensor market, Europe has a current market share of 41 % of the vehicle sensors sector and Bosch is the world market leader; this has clearly surpassed America with a share of just 34 %. The total market for sensors for vehicle applications is expected to grow from US\$ 8.88 billion in 2005 to around US\$ 11.35 billion in 2010, that is by a total of 28 % (Fig. 9).

There are typically three groups of companies offering sensors for use in the vehicle:

- ▶ The semiconductor industry: sensors from semiconductor manufacturing originated here with the application of a few process steps. They serve the entire sensor market including the automotive industry and have an efficiently-operating sales and distribution system.

Micromechanical processes for the production of sensors are continually being further developed in tandem with the semiconductor processes. These companies do not, however, have any specific know-how in the area of specification, testing and packaging appropriately for the vehicle.

- ▶ Special, generally medium-sized manufacturers of sensors, who do not produce semiconductor circuits, but have generally chosen just a few sensor types as their product to supply the sensor market overall or just preferred branches such as the vehicle market.
- ▶ Large vehicle suppliers and system manufacturers (e.g. Bosch) or large subsidiaries of vehicle manufacturer who have specialized in the needs and support of their mother groups. Here too, since the introduction of electronics in the vehicle, experience has been gained with the production of semiconductor and hybrid circuits in close collaboration with semiconductor manufacturers (process development, licensing). Thanks to system knowledge, it has been possible to build up comprehensive know-how in the field of specification, testing and packaging techniques appropriate for the vehicle.

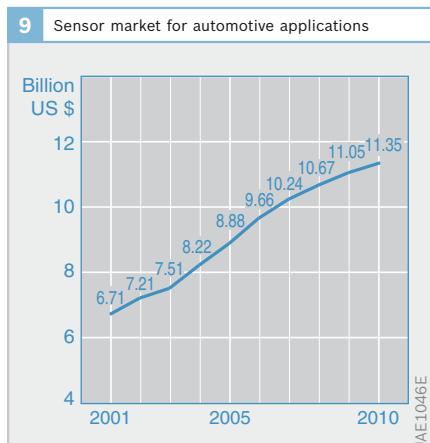


Fig. 9

Source: Bosch

Features of vehicle sensors

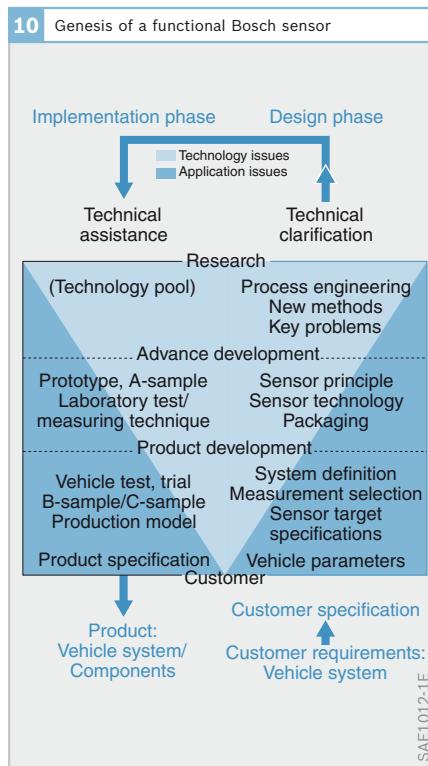
While general sensors are being developed for as broad a user group as possible and in multilevel measuring ranges, often without the manufacturer being aware of the application, vehicle sensors are generally specified and optimized for a special application. They are a part of a system and are often not freely available in the trade. Their development generally takes longer than that of commercial sensors, not just because of the enhanced demands. It is more associated with the development of the system and generally lasts as long as this does, as the sensor specification can still change until system development is concluded.

The high thrust for innovation in systems for the automobile sector very often compels the development of new sensor technologies or significant extensions of their specifications. Figure 10 shows the typical development phases that vehicle sensors pass through at the supplier's works.

The development process begins naturally with the system idea originating with the vehicle manufacturer or the supplier. The initial task is to make a selection of the measured variables required, whether or not this is feasible. At this stage, the sensors' function is simulated as part of the systems engineers' usual system simulation and a first specification is drawn up. If it is possible to use a sensor technology that has already been introduced, the wishes in respect of the sensors are passed directly to product development or the producing area. If no technology is immediately available, sensor and technology experts are gradually increasingly included in research and advanced development. Here, the first laboratory prototypes can be created using known technologies and often with the help of external partners. These prototypes can then be passed to product development for initial tests.

If no sensor principles are known for the requirements set, research will be started into new procedures and methods for the measurement of the variables required. This phase is one of fundamental research which ultimately also supplies new initial technology prototypes. This procedure can repeat itself recursively until a promising solution which can then be sent to product development is found. It is not unusual for this development loop to be repeated again in full or for a different measured variable to be selected.

Five distinct phases are generally identified in the development of a completely new sensor, just as with other electronic products (Table 2). While prototypes and A-samples mostly still come from advance development or research, B- and C-samples



are products of product development. If recursions are necessary in difficult developments, there can easily also be a number of B- or C-sample phases (B1, B2, C1, C2).

2 Sensor prototyping phases leading to series production		
Prototype phase	Function/specifications	Manufacture
Prototype	restricted	Prototype construction without tools
A	restricted	Prototype construction without tools
B (possibly B1, B2)	full	Prototype construction without tools (construction identical to C)
C (possibly C1, C2)	full	Prototype construction with series production tools
D	full	Pilot series, partially manual
Series production	full	Automated

Table 2

Sensor classification

Sensors can be classified and grouped according to very different points of view. They can be classified with regard to their use in the vehicle as follows:

Assignment and application

- ▶ Functional sensors (pressure, air-mass flow) mainly used for open and closed-loop control assignments
- ▶ Sensors for safety (passenger protection: airbag, ESP) and security (theft-deterrence feature)
- ▶ Sensors for vehicle monitoring (on-board diagnosis (OBD), fuel-consumption and wear parameters) and for driver/passenger information

Characteristic curve type

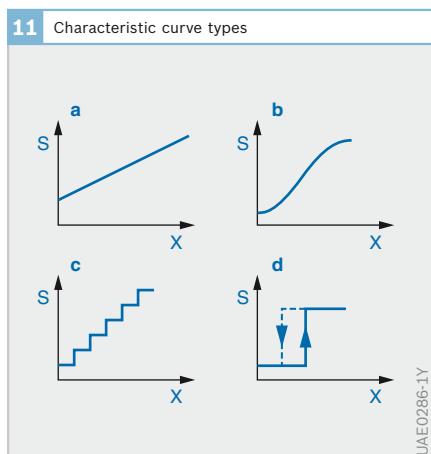
- ▶ Continuous linear curves (Fig. 11a) are used mainly for control assignments covering a wide measuring range. Linear curves are also distinguished by uncomplicated testing and calibration.
- ▶ Continuous nonlinear curves (Fig. 11b) are often used for the closed-loop control of a measured variable across a very restricted measuring range (e.g. exhaust-gas control to $\lambda = 1$, vehicle spring-deflection level). When, for instance, a constant permissible deviation relative to the measured value is demanded throughout the complete measuring range (e.g. HFM air-mass meter), curves which feature both pronounced non-linearity and a special shape (e.g. logarithmic) are at an advantage.
- ▶ Such two-step curves (possibly even featuring hysteresis, Fig. 11d) are used for limit-value monitoring in such cases where remedial measures are easy to apply when the limits are reached. If remedial measures are more difficult, then multi-step curves (Fig. 11c) can be used for an earlier warning.

Fig. 11

S Output signal

X Measured variable

- a Continuous, linear
- b Continuous, nonlinear
- c Discontinuous, multi-step
- d Discontinuous, two-step (with hysteresis)



Type of output signal

Sensors may also be distinguished by the type of output signals (Fig. 13):

Analog signals

- ▶ Current/voltage, or a corresponding amplitude
- ▶ Frequency/period duration
- ▶ Pulse duration/pulse duty factor

Discrete output signal

- ▶ Two-step (binary coded).
- ▶ Multi-step, with irregular steps (analog coded).
- ▶ Multi-step equidistant (analog or digital coded).
- ▶ A further distinction must be made, as is represented in Figure 12 in a systematic overview of the determined, i.e. nonrandom (stochastic) signal - as to whether the signal is permanently available at the sensor output (continuously) or only at discrete intervals (discontinuously).

For instance, the signal is bound to be discontinuous if it is digital and issued in bit-serial form.

13 Signal shapes

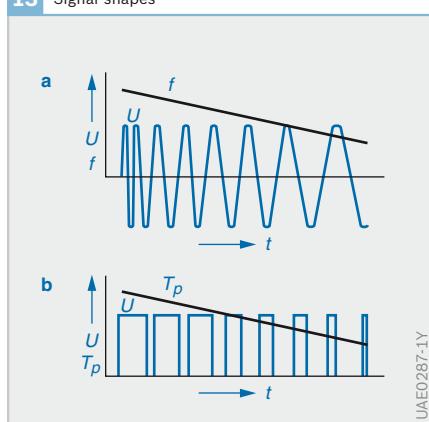
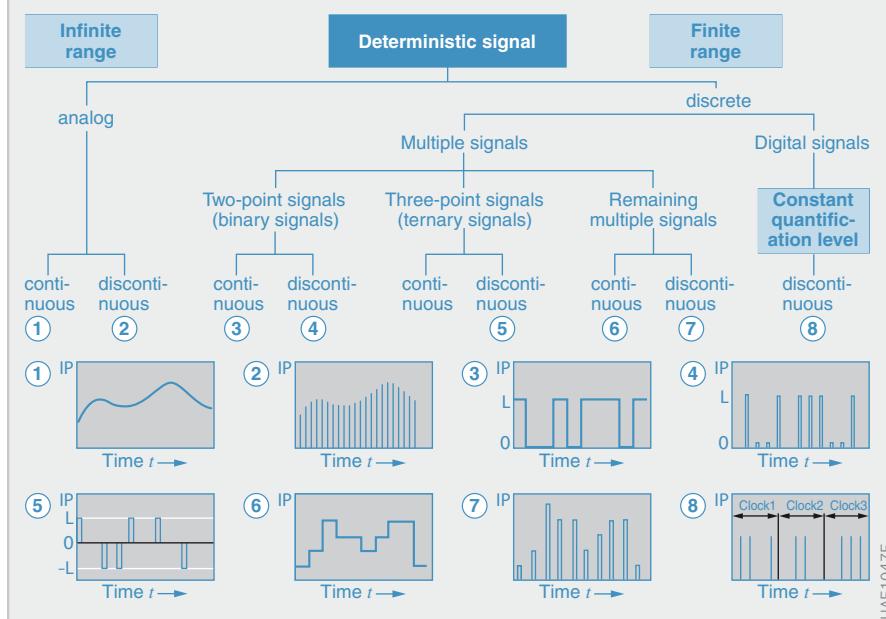


Fig. 13

- a Output signal U , information parameter: frequency f
b Output signal U , information parameter: pulse duration T_p

12 Classification of the determined signals according to the information parameter (IP) with examples



UAE1047E

Error types and tolerance requirements

The deviation of the actual characteristic curve of a sensor from its reference characteristic curve is designated its error e . It is best stated in relation to the input range y (measured variable) and not to the output range x (output signal):

$$(3) \quad e = y_{\text{indicated}} - y_{\text{true}}$$

$y_{\text{indicated}}$ = indicated value for the measured variable

y_{true} = ideal value, setpoint value for the measured variable (is determined with a measuring sensor that is more accurate than the sensor being examined by at least 1 class)

The amount of the deviation represents, as shown in Figure 14, the absolute error e_{abs} (unit as measured variable). Related to the true measured value y_{true} , this becomes the relative error (% of read-

14 Characteristic curves and error graph for a sensor

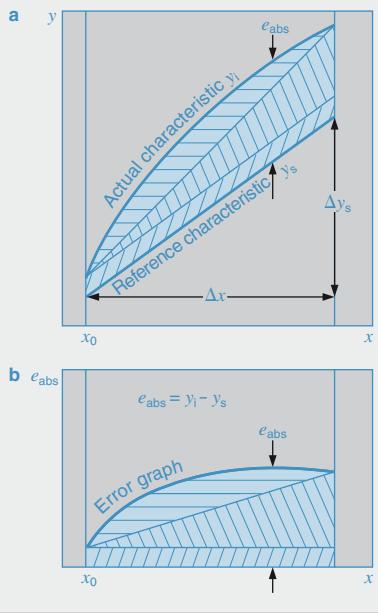


Fig. 14

- a Actual and reference characteristic
- b Error graph

y Measured variable
x Output signal
 Δx Measuring range
e Error (deviation)

Fig. 15

- a Zero error
- b Gradient error
- c Linearity error

y Measured variable
x Output signal
 Δx Measuring range
e Error

ing), related to the measuring-range final value it becomes the percentage error of the range.

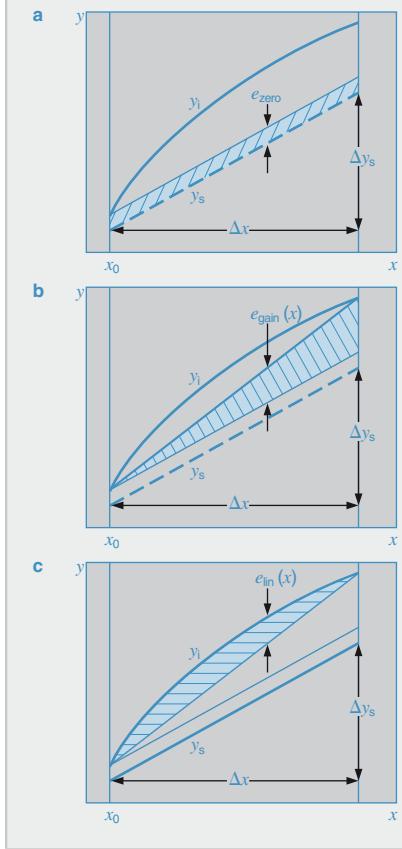
Assuming a generally desirable linear characteristic, the absolute deviation e_{abs} can be classified into three categories (Fig. 15):

- Zero offset (offset error) e_{zero}
- Gradient deviation (gain error) e_{gain}
- Linearity deviation e_{lin}

The causes of these errors are predominantly found in

- The production scatter of the characteristic curve

15 Subdivision of the total error



UAE1048E
UAE1049E

- ▶ The temperature sensitivity of the characteristic curve and
- ▶ The production scatter of the temperature sensitivity

The deviations referred to are exclusively systematic or terministic errors which, by contrast with random (stochastic) errors, are well defined, foreseeable and, as a matter of principle, correctable, and are also for the most part more or less precisely corrected.

- The non-correctable, stochastic errors include, for example,
- ▶ Drift (low and higher frequency background noise) and
 - ▶ The effects of aging

In the specification of a sensor, the total error in the new state and after aging is prescribed in the product specifications by means of a tolerance graph (Fig. 16). Sometimes, however the permissible individual error shares such as the offset, lead and linear deviations are additionally specified.

Strict metrology teaching states that, in the case of systematic errors, the sum of the amounts of the individual errors must be assumed as the total error (these may add up in the worst case). In the case of stochastic errors, statistical addition, which calculates the total error as the root of the sum of squares of the individual errors, is permitted. As statistical addition leads to a smaller total error, it is, however, often also applied to the systematic errors in a less strict configuration:

$$(4) \quad e_{\text{total}} = \sum_1^n |e_i| \quad \text{Sum of } n \text{ systematic errors}$$

$$(5) \quad e_{\text{total}} = \sqrt{\sum_1^n e_i^2} \quad \text{Sum of } n \text{ stochastic errors} \\ (\text{statistical addition})$$

Reliability

Failure rate

The operational reliability of a sensor is a purely statistical quantity and, as with any component, is characterized by its failure rate λ , which is stated in 1/h, %/h or ppm/h. Here λ is determined with a large number of parts. If it is wished to determine approximately the failure rate with a number N (< 40) of sensors that is not too great, it is possible to observe the failure behavior of this random sample under operating conditions until, after a finite time, all the parts have failed. If this observation is begun at time t_0 and if the inventory of parts remaining at a later point t_i is identified as $B(t_i)$, then a good approximation for the failure rate λ is the failure quota q :

$$(6) \quad q(\Delta t_i, t_i) = \frac{B(t_i) - B(t_{i+1})}{\Delta t_i \cdot B(t_i)} \quad \text{Failure quota}$$

where $\Delta t_i = t_{i+1} - t_i$

t_i are the times at which individual parts or several parts fail (Fig. 17). The ratio of the current inventory of parts to the initial inventory of parts is known as the relative inventory B_R :

16 Tolerance graph for a sensor

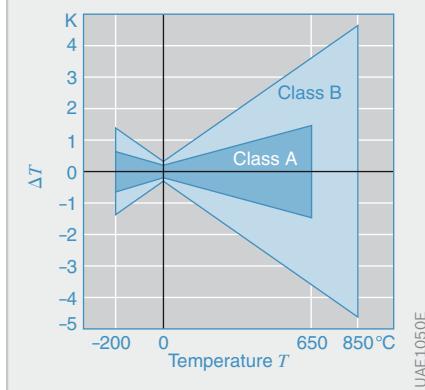


Fig. 16
Tolerance graph illustrated on the example of a resistance temperature sensor

$$(7) \quad B_R(t_i, t_0) = \frac{B(t_i)}{B(t_0)}$$

With a very large number of parts ($N \rightarrow \infty$), the survival probability $R(t)$ at the, here now continuously variable, time t corresponds to this. The failure rate is calculated here for a large number (in practice $N \approx 2,000$) of sensors as a percentage variation on the survival probability R per unit of time dt as:

$$(8) \quad \lambda(t) = -\frac{1}{R(t)} \cdot \frac{dR}{dt} \quad \text{Failure rate}$$

Reliability is the inverse value of the failure rate:

$$(9) \quad z = \frac{1}{\lambda} \quad \text{Reliability}$$

The definition of the failure rate λ demands a failure criterion:

- ▶ Complete failure
- ▶ Partial failure
- ▶ Sudden failure (sudden change in a feature)
- ▶ Drift failure (gradual change in a feature)

Furthermore, it is absolutely necessary to specify under what operating conditions the failure rate defined in this way is to be understood. Particularly in the case of electrical components such as sensors, for example, it is necessary to distinguish between true, active operating time (switched on state) and service life in the sense of pure storage time. Any statement of a failure rate without this additional information is worthless.

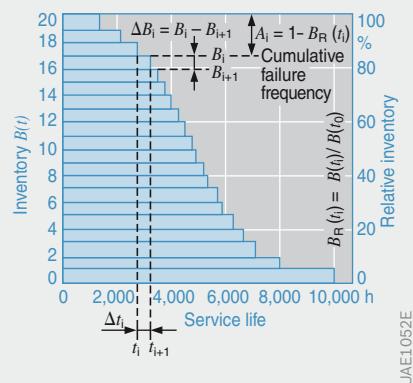
Failure rates are normally determined using time-acceleration methods. The acceleration factors are achieved here by exposing the sensors to enhanced operating conditions. A high degree of experience is required in order to apply time acceleration methods that truly reflect reality.

The concept of the mean service life T_M is also used to characterize the reliability of a sensor. A good approximation of this can be calculated from a random sample from the sum of the individual service lives T_i :

$$(10) \quad T_M \approx \frac{1}{N} \cdot \sum_{i=1}^N T_i \quad \text{or}$$

$$(11) \quad T_M = \int_0^\infty R(t) dt \quad \text{for a very large number of parts}$$

17 Inventory and relative inventory of a random sample



18 Distribution of the failure rate $\lambda(t)$ over time

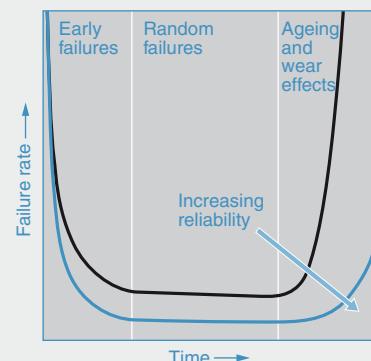


Fig. 17

Observation for a random sample of $N = 20$ sensors; mean service life $T_M = 4,965$ h.

The failure rate of a product against time shows a typical bathtub pattern (Fig. 18). At the start, the failure rate is somewhat high because of early failures, it then passes through a longer, relatively low, horizontal range, then to rise again drastically towards the end of the service life (aging and wear effects). In the case of sensors for which high operational reliability is demanded, an attempt is made to avoid the elevated failure rate at the start by sorting out the early failures by pre-aging, e.g. by conditioning at higher temperatures ("burn in"). Early failures are basically nothing other than manufacturing defects that have not been recognized.

Table 3 gives some examples of permissible failure rates λ applicable for the vehicle. The ppm values given relate to a period of 10 years, or alternatively to 150,000 km covered, if this is completed in a shorter time. Where the sensors are given a flat rate of <10 ppm, it means that in 10 years, only fewer than 10 of 1 million sensors may fail. This value is, however, substantially lower for sensors in passenger protection systems.

Measures for increasing reliability

The best method for ensuring high quality is to engineer and design in reliability. This means that as soon as the sensor is designed, correspondingly durable materials must be selected and solid protective measures against anticipated mechanical,

chemical and electrical ambient influences must be provided (packaging, passivation). On the other hand, it is costly to create operational reliability by testing alone, i.e. to exclude early failures through premature aging.

It is sensible to provide monitoring and diagnostic capabilities for sensors in complex systems (e.g. signal range check, etc.), so that any failures occurring may be detected in good time. In an emergency, a sensor's function can here be temporarily replaced by other measured variables or sensible fixed values (limp-home mode, back up). A sensor-less, often purely mechanical, limp-home mode can be provided. For instance, should the accelerator-pedal sensor failure in a diesel vehicle, the vehicle could also be driven slowly back for repair under regulated idle speed alone (limp home).

Where reliability must be guaranteed with a probability bordering on certainty (e.g. for sensors in electronic braking and steering systems), redundancy, i.e. provision of multiple parts, is generally the means used. Dual redundancy of a sensor of the same type here only permits failure detection for the case of widely different indications, whereas triple redundancy with 2 from 3 analysis additionally still supplies a correct measured value. Here, however, care must be taken to provide not only redundancy of the sensors, but other essential parts such as the power supply, signal evaluation and transmission means must also be correspondingly redundant, as otherwise the probability of simultaneous failure increases. It is often also advisable to provide sensor redundancy in different technologies.

3 Reliability requirements on vehicle systems	
Warranty target: 150,000 km/10 years	
→ ECU failure rate (field)	< 50 ppm
→ ECU failure rate (0 km)	< 15 ppm
→ Failure rate of modules and sensors	< 10 ppm
→ ASIC failure rate	< 3 ppm
→ IC failure rate	<< 1 ppm
→ Failure rate of discrete components	< 0.5 ppm
→ <i>For comparison</i>	~ 5,000 ppm Mobile phone

Table 3

Main requirements, trends

The vehicle sensors customized for the requirements of special electronic systems in the vehicle, unlike the universal sensors conventionally found in the market, are subject to five strict requirements (Fig. 19) which must be fulfilled by development and to which the most important development trends also correspond.

Low manufacturing costs

Electronic systems in modern vehicles easily contain up to 150 sensors. Compared to other sectors of sensor application, this abundance forces a radical reduction in costs. The target costs – typically in the range of € 1 to 30 – are here often less than one-hundredth of those for conventional sensors of the same capacity. Naturally, particularly with the introduction of a new technique or technology, the costs are generally on a falling learning curve, starting from a higher level.

Development trends

Automated manufacturing processes (Fig. 20) are largely used, working to a high yield. This means that each process step is always executed for a large number of

sensors simultaneously. For example, semiconductor sensors are manufactured using “batch processing” in which 100 to 1,000 sensors are manufactured simultaneously on a single silicon wafer. On the other hand, such manufacturing equipment is only an economic proposition when correspondingly large numbers of sensors are produced. These quantities sometimes exceed an automotive-industry supplier’s own in-house requirements, and can typically often be between 1 and 10 million per year. Here, the high numbers of sensors needed by the automotive industry has played an unprecedented and revolutionary role, and set completely new standards.

High reliability

In accordance with their assignments, vehicle sensors are subdivided into the following reliability classes, given in descending order of severity:

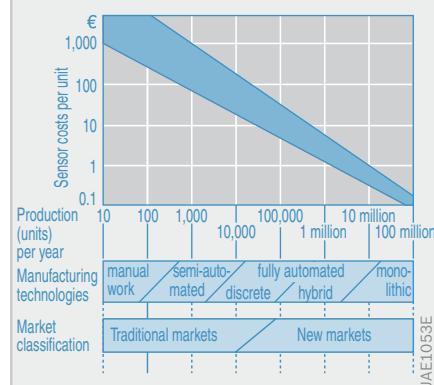
- ▶ Steering, brake, passenger protection
- ▶ Engine/drivetrain, chassis/tires
- ▶ Comfort and convenience, diagnosis, information, and theft deterrence

19 Principal requirements on vehicle sensors

Vehicle sensor	Requirement
Rational mass production	⇨ Low costs
Robust, proven technology	⇨ High reliability
Appropriate packaging technology	⇨ Extremely severe operating conditions
Appropriate miniaturization technologies	⇨ Low space requirement
On-site error compensation	⇨ High-level accuracy

SAE1013-1E

20 Relationship of costs, markets, technologies



UAE1053E

The requirements on the highest class here easily match the high reliability values familiar from aerospace applications. To some extent, they demand similar measures, such as the use of the best materials, redundant componentry, self-monitoring, (short-term) backup power supply, multiple programming of critical decision-making algorithms.

Development trends

Appropriate design measures guarantee built-in reliability. For instance, this necessitates the use of reliable, top-quality components and materials, coupled with rugged and well-proven techniques and engineering. And furthermore, an effort is made to achieve a consistent integration of the systems to avoid connection points which may be broken and are at risk of failure. This is possible, for instance, with radio-scanned sensors based on the antenna-coupled SAW (Surface Acoustic Wave) elements which do without wiring completely. If necessary, redundant sensor systems are fitted.

Severe operating conditions

Like practically no other breed, vehicle sensors are exposed to extreme stresses because of where they are fitted, and must withstand all kinds of aggression:

- ▶ Mechanical (vibration, shock)
- ▶ Climatic (temperature, moisture)
- ▶ Chemical (e.g. spray, salt mist, fuel, engine oil, electrolyte)
- ▶ Electromagnetic (incident radiation, line-conducted interference, excess voltages, polarity reversal)

It is this trend of using sensors immediately at the site where the measurement is taken in order to exploit the benefits associated with this which itself leads to the requirements being considerably cranked up.

Figure 21 illustrates this problem on a sensor which might be a temperature sensor, speed sensor, flow-rate sensor or concentration sensor, for instance. The sensor cannot in the slightest always be surrounded by a hermetic protective sleeve. Although this sleeve could offer rough protection for the measurement of temperature, flow rates and concentration, it must permit more or less direct contact between the sensor and the generally very aggressive monitored medium (an exception to this is the inertia sensor). Sometimes, thin, but very resistant, passivation layers are permissible on the sensor.

The form of the plug-in sensor demands a permanently tight seat of the mount in the associated wall (inner packaging, outer packaging). The sensor may be connected to the electronic control units via a fixed housing plug or also, as illustrated, via a plug flexibly fastened with a length of cable (e.g. ABS wheel-speed sensor). Here, too, there are three critical electrical connection points that absolutely must be protected against conductive fluids and corrosion: the inner connection on the sensor element, the connection of the cable tail and finally the connection of the partially shielded cable to the external plug connection.

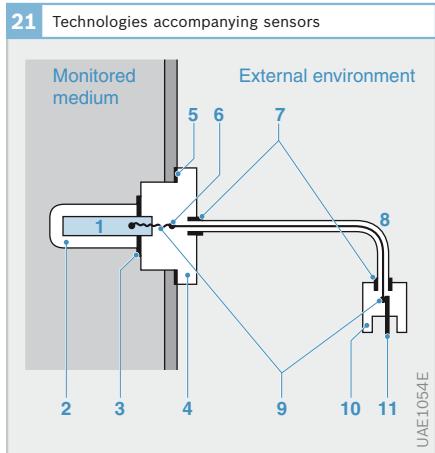


Fig. 21

- 1 Sensor
- 2 Protective sleeve (coating)
- 3 Seal
- 4 Mount
- 5 Seal, mounting
- 6 Data point
- 7 Seal, strain relief
- 8 Insulation (flexible)
- 9 Contacts
- 10 Plug housing
- 11 Plug contact

tion. If the seal is inadequate, it is only a question of time until corrosive liquids penetrate from the external connector into the internal sensor terminal.

The plug-in connection itself must be sufficiently tight overall, that no shunts form at the sensor output. The cable itself must retain its flexibility and tightness after many years of operation under the most adverse conditions.

Separable plug-in connections in the vehicle unfortunately still represent one of the most frequent causes of failure. Wireless signal connections (e.g. infrared light or radio) could break down this problem in the longer view, especially if the sensors were even to have a wireless energy supply (autonomous sensors).

The costs of these indispensable technologies complementary to the sensors often exceed those of the actual sensor element many times over. They constitute the actual value of an vehicle sensor, not only in terms of cost, but also in terms of function.

Development trends

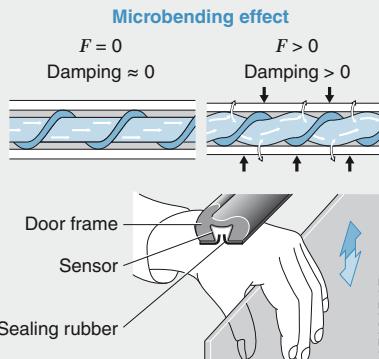
Protective measures against the stresses referred to demand a very high degree of specific know-how in the field of sensor packaging.

- ▶ Passivation and connecting techniques
- ▶ Sealing and joining techniques
- ▶ EMC protective measures
- ▶ Low-vibration installation
- ▶ Service life test and simulation methods
- ▶ The use of resistant materials, etc.

Furthermore, detailed knowledge of the loading to which the sensor will be subjected at the particular installation point is required. It is often forgotten that the quality of a sensor stands and falls with total competence with these protective measures.

Fiber-optic sensors in which the light guided in the optical fibers (glass, plastic) can be modified as a function of the measured variable are particularly immune to electromagnetic interference. This is true up to the point at which the optical signals are converted back into electrical signals. If these are to be used in the future, there will have to be some development work the provision of low-priced measuring elements and the accompanying technologies. There are, for example, some very interesting applications for this in the field of force measurement, e.g. in providing finger protection for electric power windows and sliding sunroofs (Fig. 22). Sensors of this type have also been tested very successfully as very early responding, distributed sensors in the door and frontal area of the vehicle for triggering passenger and pedestrian protective systems.

22 Microbending effect



SAE1015-1E

Example of development: fiber-optic finger protection in power windows based on the microbending effect:
The light in an optical fiber is attenuated proportionally to the corrugated distortion on the application of a force F transverse to the fiber (large measuring effect), independently of where the force is applied or whether it is acting at a point or is distributed (distributed sensors).

Low-volume design

On the one side, the number of electronic systems in the vehicle continues to climb steadily. On the other, today's vehicles are becoming more and more compact. These facts, together with the need to retain the high level of passenger-compartment comfort forces development to concentrate on an extreme space-saving designs. Furthermore, the increasing demand for further improvements in fuel economy mean that minimization of the vehicle's weight is of prime importance.

Development trends

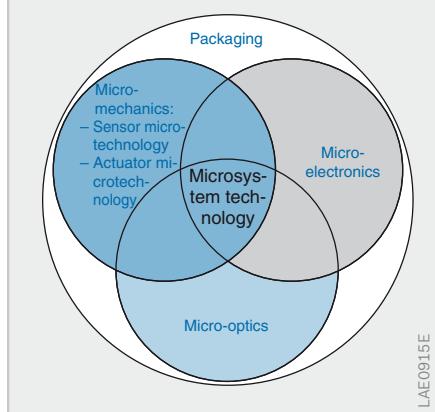
Widespread use is made of the familiar technologies applied in circuit engineering for the miniaturization of electronic components:

- ▶ Film and hybrid technologies (deformation-dependent resistors, thermistors, and magnetoresistors); nanotechnology
- ▶ Semiconductor technology (Hall-effect and temperature sensors)
- ▶ Surface and bulk micromechanical techniques (silicon pressure, acceleration and yaw-rate sensors)
- ▶ Microsystem technology (combination of two and more microtechnologies such as microelectronics and micro-mechanics, Fig. 23)

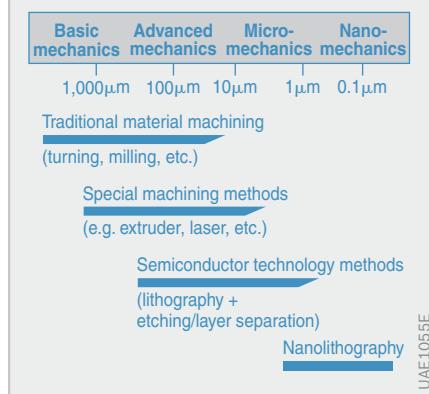
Micromechanical manufacturing refers on the one hand to dimensions in the μm -range and tolerances in the sub- μm -range which cannot be achieved with conventional machining methods. On the other hand, sensors which may have dimensions in the mm-range, but which are manufactured with micromechanical methods, are also considered micromechanical (Fig. 24). We will discuss only the most well-known and important method here, anisotropic etching of silicon. This is the most important method, because silicon is manufactured to high precision and in large numbers at low cost, and is the most thoroughly researched and best known material. In addition, it offers the capability of the monolithic integration of sensor and evaluation electronics. The etching speeds, which can be very different depending on the crystal axis (1:100), with which suitable etching fluids such as KOH attack are exploited in this method (Fig. 25); some crystal faces, for instance, remain practically untouched, while other faces are rapidly etched to some depth.

In conjunction with appropriate etching stop methods (doping, depletion layer) practically three-dimensional structures of the smallest dimensions may be produced to the highest precision in this way

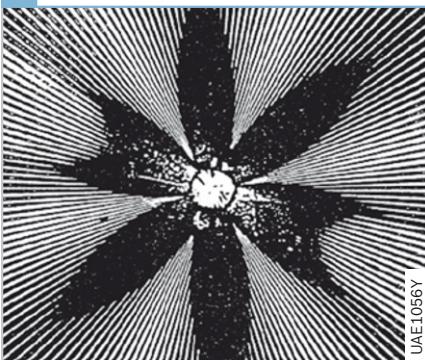
23 Microsystem technology



24 Micromechanics and mechanical structures



25 Test structure for the determination of the etching rate in Si monocrystal wafers



26 Anisotropic etching of (100) silicon

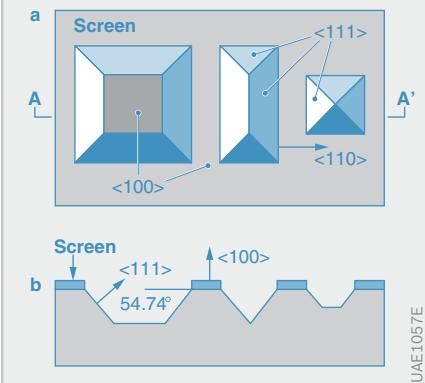


Fig. 26

- a Top view
- b Cross-section at AA'

(Figs. 26, 27). The problem of simultaneously integrating the circuit can be considered largely resolved, even if currently the associated circuit for most sensors is still made separately for reasons of a better yield and higher flexibility.

By contrast with bulk micromechanics, anisotropic, often wafer-depth etching has no role to play in surface micromechanics (SMM). SMM sensor structures are mostly built up additively on the surface of a silicon substrate (Figs. 28 and 29). Even if the dimensions of bulk silicon sensors are still mostly in the mm-range, those of SMM structures are an order of magnitude smaller (typically 100 µm), as a rule.

Frequently, mechanically indispensable parts are used simultaneously for enclosing and even cooling (associated) sensors and electronics (e.g. microhybrid control unit as an add-on control unit to the VP44 diesel distributor-type injection pump). This fusion of mechanical and electronic components known as mechatronics is coming more and more to the forefront in the search for cost and space savings. In the foreseeable future, practically all systems will operate on this basis.

27 First attempts at three-dimensional forming of silicon

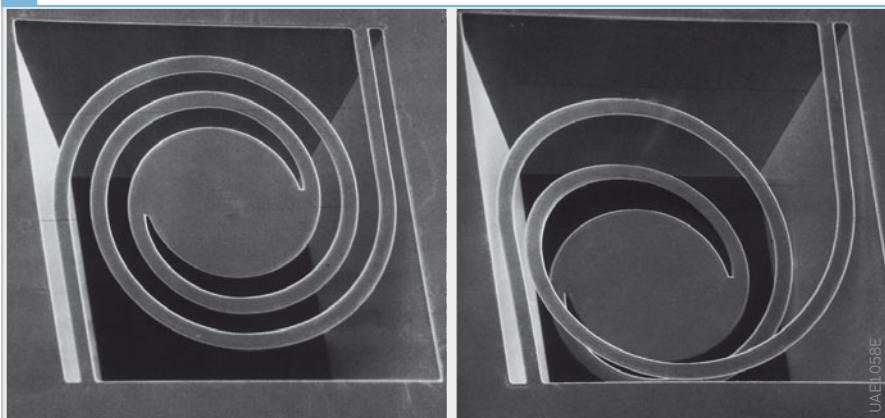
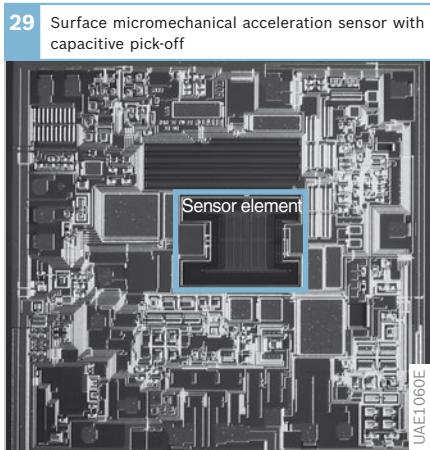
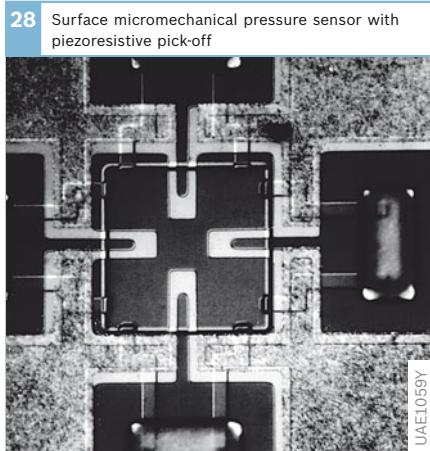


Fig. 27

Source:
Prof. Heuberger,
Fhg Berlin

Another example of a mechatronic sensor system is the Hall-effect speed sensor built into the wheel bearing. The addition of magnetic particles to the indispensable shaft seal in the bearing can even give it the function of the rotor or a pole wheel. The sensor benefits from excellent protection and encapsulation along with the high precision of the roller bearing and no longer needs any bias magnets. Finally, the Hall-effect technology allows installation in the very tight space within the bearing.



High accuracy

By comparison with the probes and sensors used for instance in the processing industry, with only a few exceptions (e.g. the air-mass meter), the accuracy demands made on vehicle sensors are relatively modest. Generally, the permissible deviations are $> 1\%$ of the measuring-range final value. This applies in particular when considering the unavoidable effects of aging. The permissible deviations are normally achieved by the application of complex techniques to compensate for manufacturing tolerances, and to balance the effective compensation measures used against interference. Particularly since the above-mentioned requirements have for the most part become achievable, continually more demanding and sophisticated systems are imposing higher and higher demands in this sector.

Development trends

Initially, a tightening up of the tolerances in manufacture, and refinement of the calibration and compensation techniques help in this regard. An important step forward here is the hybrid or monolithic integration of the sensor and signal electronics directly at the measuring point, even embracing complex digital circuits such as analog/digital converters and microcomputers (Fig. 30).

Such microsystems are also known as "intelligent sensors". They take full advantage of the sensor's inherent accuracy, and offer the following features:

- ▶ Reduction of the load on the control unit
- ▶ Uniform, flexible, bus-compatible interface
- ▶ Multiple application of sensors
- ▶ Use of smaller measuring effects and of radio frequency measuring effects (local amplification and demodulation)
- ▶ Correction of sensor deviations at the measuring point, and mutual calibration and compensation of sensor and electronics, is simplified and improved by

Fig. 28

Source:
Prof. Guckel,
University of Michigan
in Madison, USA

storing the individual correction information in the PROM

While simultaneously detecting and digitizing interfering factors, intelligent sensors can calculate the required measured variable practically without error by applying the mathematical sensor model (Fig. 31). Here, the item-specific model parameters are defined in a preceding process which is equivalent to the calibration as previously performed, and stored in a PROM integrated within the sensor. In this manner, it is possible to considerably improve the sensor's static and dy-

namic characteristics (evaluation of the differential equation which defines the dynamic performance).

Local electronic circuitry necessitates the use of multi-sensor structures which use a number of identical sensors, or a number of different sensors, to register a variety of highly complex facts and locally reduce these to their basic information content, as necessary. This applies in particular to imaging sensors which in future will play an ever increasing role in detecting the situation inside and outside the vehicle.

30 Sensor integration levels

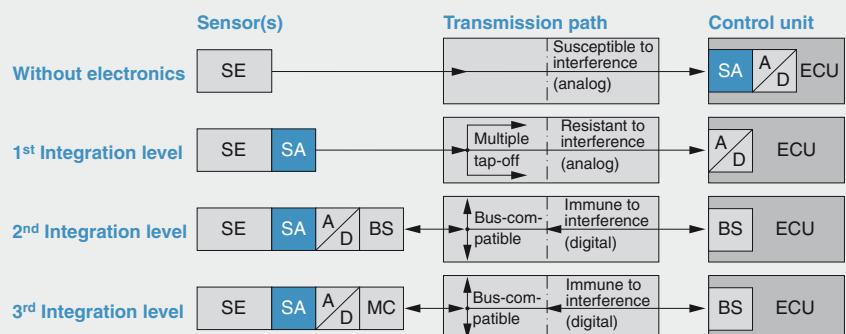
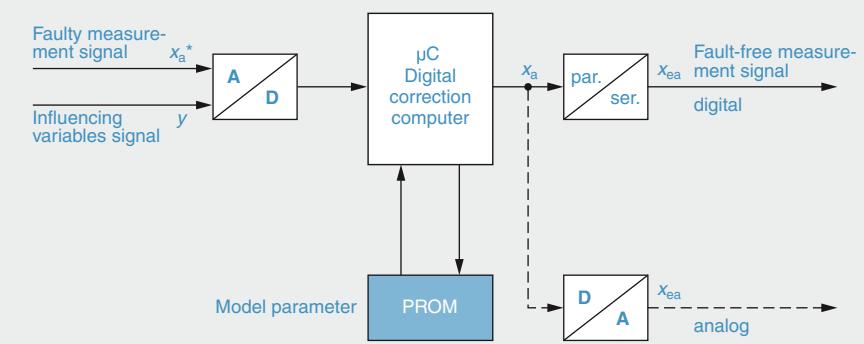


Fig. 30

SE Sensors
 SA Signal conditioning
 A/D Analog-Digital converter
 ECU Control unit
 MC Microcontroller
 BS Bus interface

31 Correction model for a smart sensor



UAE0018-1E

With a number of pressure sensors integrated into the smallest of spaces, it is possible to not only increase the reliability of measurement, but also to reduce the, generally irregular, aging drift by applying mean-value generation. Individual failures (outliers) can also be detected and eliminated. Sensors of this type are also known as soft sensors. If the individual sensor elements are designed for differing measuring ranges – and at the same time feature high overload capabilities (e.g. capacitive) – such a sensor can be used to considerably extend the high-accuracy measuring range. Sensor structures of this type were developed and tested years ago, but have not been converted into marketable products.

Overview of the physical effects for sensors

The sensors are not classified here into a systematic structure by measuring effect, but rather by measured variable. For this reason, the most important physical effects or measuring concepts exploitable for measuring purposes will here only be reproduced in the form of a rough overview, which cannot be complete. Overlaps between the various categories are unavoidable. The table specifies principally the actual electrical effects; mechanical or fluid prior stages, such as strain springs (force), membranes (pressure), spring-mass systems (acceleration), vibration systems (tuning forks) or even fixed and rotating turbine blades (flow) are not included.

4 Physical effects for sensors		
Physical effect	Example	As at: Series production ¹ /Development ²
Resistive effects (dependency of electrical resistance):		
Influence of temperature on metallic and semiconductor materials	NTC and thin-film resistors for the measurement of air and engine temperature	S
Length or angular proportionality of resistors (potentiometer sensors)	Accelerator-pedal and throttle-valve sensor, fuel tank level	S
(In plane) tension or pressure dependency (piezoresistive): strain-gage resistors	High-pressure sensors (e.g. common rail, ABS): metal diaphragm, low pressure sensors (silicon diaphragm), force sensor	S S D
Vertical pressure dependency (out of plane)		
Magnetic field dependency (magnetoresistive): semiconductor (magnetoresistors), AMR ³ thin metal layers (e.g. NiFe, also in barber's pole form), GMR ⁴ sensors (nanolayers)	Speed or delivery angle measurement in diesel distributor injection pumps	S/D
Light dependency: semiconductor photoresistors	Rain sensor, dirt sensor for headlight cleaning, automatic headlights, automatic low-beams	S
Inductive effects (effects of Faraday's law)		
Induction voltage sensors (alternator): movement in the magnetic field	Wheel, camshaft, engine speed, needle lift (injection nozzle)	S
Wiegand effect	Speed of rotation	D
Variation in inductance as a result of the positional change of a ferromagnetic coil core	Solenoid armature sensor	D
Variation in inductance as a result of field limiting conductive elements (eddy current)	Semi-differential short-circuiting ring sensor (diesel pump load sensor)	S

Table 4

¹⁾ Series production, at RB or competitors, possible also already phased out

²⁾ Development possibly also concluded and in reserve

³⁾ AMR = Anisotrop Magneto Resistive

⁴⁾ GMR = Giant Magneto Resistive

Physical effect	Example	As at: Series production/ Development ¹⁾
Variation of the degree of transformer coupling (by electrical or magnetic conductive elements)	Full-differential short-circuiting ring sensor	D
Variation in inductance or of the degree of transformer coupling (by magneto-elastic conductive elements)	Load bolt (Hitchtronik), braking force	S D
Saturation core sensors (e.g. Foerster probe)	Compass sensor	S
Capacitive effects (influence)		
Capacitance change as a result in change in plate distance and degree of cover	Micromechanical acceleration sensors, e.g. for airbag, ESP, MM2 yaw-rate sensor, pressure sensor	S S D
Capacitance change as a result in a change of the relative dielectric constant	Oil grade, humidity sensors	S S
Capacitance change as a result of change in the electrolyte level with a dielectric medium	Fuel tank level	D
Charge generating effects		
Piezoelectric effect (quartz, piezoceramics)	Knock sensor, airbag sensor, DRS1 yaw-rate sensor	S S
Pyroelectric effect	IR sensor (dynamic)	D
Photoelectric charge generation	CCD and CMOS imaging sensor (also IR range)	D
Voltage generating, galvanic effects		
Hall effect (out-of-plane sensitivity, semiconductor material)	Hall switch (ignition), wheel and engine speed, acceleration sensor (ABS, 2g), front passenger weight detection (iBolt™), ARS1,2 (accelerator pedal etc.)	S
Pseudo-Hall effect (in-plane sensitivity, metal thin film)	LWS2 and LSW4 steering-angle sensors	S
Electrolytic diffusion probes (doped Zr oxide ceramic)	Lambda oxygen sensors	S
Thermocouple, thermopile	IR sensor (Bolometer)	D
Photoelectric and fiber-optic effects		
Photoelectric cells, photodiodes, phototransistors (also in IR range)	Rain sensor, dirt sensor for headlight cleaning, automatic headlights, automatic low beams	S
Medium-dependent absorption	Soot particles, humidity	D
Extrinsic and intrinsic fiber-optic effects: influence on intensity, interference (influencing phases), influencing polarization; e.g. microbending effect	Finger protection (windows, sliding sunroof), D pedal force, collision	D
Thermal effects		
Resistor cooling as a function of the flow velocity, of the medium, of the density or fill level of a medium	HFM air-mass sensors, analysis, concentration, fill level (fuel tank)	S – – D
Wave propagation effects		
Sound waves: propagation time effects (echo sounding), superposition with medium speed, Doppler effect (moving source/receiver)	Parking-aid assistant, volume flow rate, speed over the ground	S D S
Light waves: total reflection on boundary surfaces, optical resonators (color analysis)	Rain sensor, fiber-optic finger protection, fill level (analog, limit value), D yaw rate: fiber and laser gyroscopes, lidar (light wave radar)	S D D D
Sagnac effect Propagation time	ACC distance sensor	S
Electromagnetic radiation: Doppler, FMCW, propagation time radar		

- 1) Series production, at RB or competitors, possible also already phased out
- 2) Development possibly also concluded and in reserve
- 3) AMR = Anisotropic Magneto Resistive
- 4) GMR = Giant Magneto Resistive

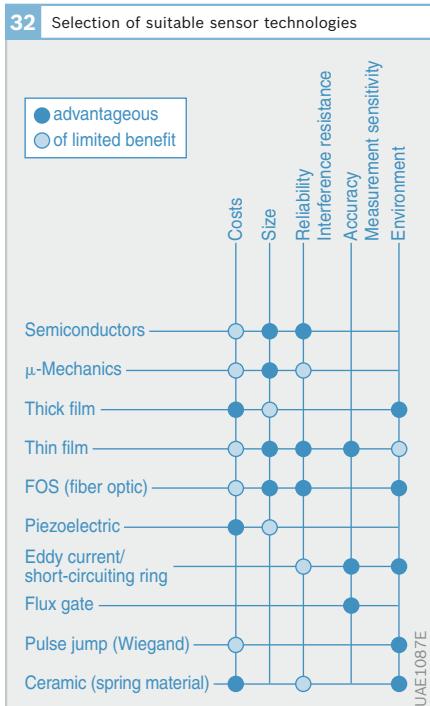
Overview and selection of sensor technologies

The various sensor technologies for exploiting the measuring effects described are naturally closely bound up with the measuring principle. They will first be listed here in a rough overview:

- ▶ Wound or photolithographically produced induction coils (with electrical or magnetic conductors), eddy current and short-circuiting ring sensors
- ▶ Flux gate probes (Metglas etc.) for magnetic field measurement
- ▶ Pulse jump sensors (Wiegand)
- ▶ Wire-wound (inductance-free) resistors
- ▶ Foil resistors (laminated onto plastic substrates)
- ▶ Sintered ceramic resistors
- ▶ Thin- and thick-film techniques (especially resistors and capacitances)

- ▶ Semiconductor technology (mono-crystalline or polycrystalline resistors, depletion layers, charge-storage cells, etc.), electronics for signal conditioning: Si (bipolar, CMOS, BICMOS, EEPROM), GaAs
- ▶ Micromechanics (silicon and other materials, e.g. quartz, metal (LIGA technology) etc.)
- ▶ Piezoceramics
- ▶ Piezofoil
- ▶ Isolating ceramics as a spring material (e.g. as a pressure-sensor diaphragm)
- ▶ Ceramic solid electrolytes (e.g. as Lambda oxygen sensor)
- ▶ Quartz and other piezoelectric crystals
- ▶ Glass or plastic optical fibers or plates

The matrix shown in Figure 32 indicates which sensor technologies should be taken for various requirements. Thus, for instance, should the requirements be for small size, high reliability and high accuracy, it is appropriate to take a thin-film technology, as long as this is available for the measured variable desired.



Sensor measuring principles

There is a great number of sensors at work in motor vehicles. They act as the sensory organs of the vehicle and convert input variables into electrical signals. These signals are used in control and regulation functions by the control units in the engine-management, safety and comfort and convenience systems. Various measuring concepts are applied, depending on the task.

1 Travel/angular positions as direct measured variables	
Measured variable	Measuring range
Throttle-valve position in the gasoline engine	90°
Accelerator/brake-pedal position	30°
Seat, headlamp, rear-view mirror position	
Control-rack travel and position for diesel in-line fuel-injection pump	21 mm
Angular setting of the injected fuel-quantity actuator on the diesel distributor-type injection pump	60°
Fill level in the fuel tank	20 to 50 cm
Travel of clutch servo unit	50 mm
Distance: vehicle – vehicle or vehicle – obstacle	150 m
Steering (wheel) angle	±2 · 360° (±2 revolutions)
Tilt angle	15°
Angle of direction of travel	360°
GPS (Global Positioning System)	360° geographical latitude/longitude, geographical altitude
Near-range distance (US parking-aid assistant)	1.5 m
Near-range radar (precrash)	10 m
External video	40 m
Long-range and near-range infrared viewer	100 m

Table 1

2 Travel/angular positions as indirect measured variables	
Measured variable	Measuring range
Spring compression travel (headlamp range, vehicle tilt)	25 cm
Torsion angle (torque)	1 to 4°
Deflection of a sensor plate (volume flow)	30 to 90°
Deflection of a spring-mass system (acceleration)	1 to 500 µm
Deflection of pressure sensor diaphragm	1 to 20 µm
Deflection of force measuring spring (passenger weight)	10 to 500 µm

Table 2

Position sensors

Characteristics

Position sensors record the widest variety of one or multi-dimensional travel and angular positions (translational and rotary variables) in a wide variety of areas. This includes gaps, distances, displacements, fill levels and even minor expansions – anything that can be measured in meters and angles.

In this sector of applications, activities have long since been directed at changing over to proximity or non-contacting sensor principles. Such sensors are wear-free and thus have a longer service life as well as being more reliable. The costs involved though often force vehicle manufacturers to retain contacting sensor principles which still perform sufficiently well for many measuring tasks.

Position variables are “extensive measured variables”, the measured variable or the measuring range of which has a substantial influence on the size of the sensor (by comparison, pressure and temperature, for instance are “intensive measured variables”). Sensors for extensive variables of medium and larger ranges (e.g. including flow and force sensors) are thus initially less suited to miniaturization on the one hand and hence for cost-effective mass production. As this is much less true for angular-position sensors than for travel sensors (the angle value does not depend on the radius or leg length), the former are clearly preferred in the vehicle.

In the case of extensive variables with a large measuring range, however, the extensive variable can often be first converted into an intensive variable, which can then be measured using microsensors. Thus flow measurement can be converted to flow density measurement (partial flow measurement, e.g. 1:100) where the flow tube becomes an essential component part of the sensor without which the sensor ultimately cannot be tested and calibrated.

In the case of travel and distance sensors covering large measuring ranges (approximately 0.1 to 150 m), the measured variable is converted into a pulse or phase runtime by means of wave-propagation sensors (sound and electromagnetic waves), which can then be easily measured electronically with relatively small transceiver devices. In the case of sound waves, however, the propagation medium (e.g. air) and the time standard required (e.g. oscillating crystal) then become an essential component part of the measuring device.

Measured variables: overview

There are a large number of applications in which position represents the actual measured variable. This is shown in overview by Table 1. In other cases, the measured position or angle represents a different measured variable (Table 2).

In practice, “incremental sensor systems” are also often referred to as angular-position sensors, even when they are used for measuring rotational speed. These sensors are in reality not angular-position sensors. Since the increments (steps with which a given variable increases) which have to be measured with these sensors in order to measure the deflection angle must be counted with the correct preceding sign (in other words, added), these sensors are in reality not angular-position sensors. Due to the danger of the counter being falsified as a consequence of spurious pulses, such angular-position measuring systems are only in limited use. Fixed, directly locatable reference marks only provide very little help in this dilemma. Another disadvantage of such angular-position measuring systems is the fact that the absolute position is lost when the operating voltage is switched off. It is no use storing the final position in a non-volatile memory here, since most angular positions can also change mechanically when power has been removed (e.g. steering angle).

Although it is true that the crankshaft angle is measured incrementally, this measurement is recalibrated after each revolution by a reference mark detectable by the sensor, in case a fault should have occurred. This also takes account of the fact that the position of the crankshaft when the engine is started is not known; it must first rotate almost one revolution under the starter so that the reference mark is passed at least once. The crankshaft’s motion is also very monotonous, i.e. it is always pretty much smooth (without sudden changes) and is always in one direction (that is forwards, and never backwards). You can thus be sure that no more than one revolution will ever be incorrectly detected.

The disadvantages of an incremental angle measurement cannot be absorbed if the engine starts immediately without being cranked, as is also the case with the steering angle. An absolute angle sensor for a 360° range is required here.

Potentiometer sensors

The wiper potentiometer, usually made as an angular-position sensor (Fig. 1), uses the correspondence between the length of a wire or film resistor (Cermet or conductive plastic) and its resistance value for measuring purposes. At present, this is still the lowest-priced travel/angular-position sensor. The shape of the characteristic curve can be influenced by shaping the measuring track (or just sections of it).

1 Principle of the wiper potentiometer

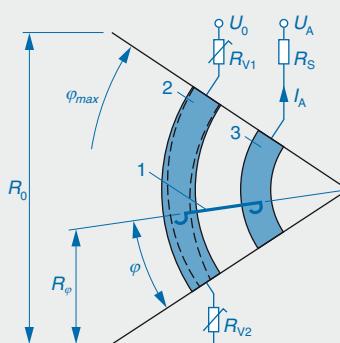


Fig. 1

- 1 Wiper
- 2 Resistance track
- 3 Contact conductor track
- I_A Wiper current
- U_0 Supply voltage
- U_A Measurement voltage
- R Resistance
- φ_{max} Maximum angle of rotation
- φ Measured angle

The standard wiper connection is furnished by a second contact track consisting of the same material mounted on a low-resistance substrate. The supply voltage is usually applied to the measurement track through low-resistance series resistors R_V (these can also be used for calibration of zero point and progression-rate adjustment) to protect against overloading.

Wear and falsification of measured values can be kept to a minimum by keeping the electrical loading of the pick-off as low as possible ($I_A < 1 \text{ mA}$) and by dust-proof and waterproof encapsulation. One of the prerequisites for low wear is the optimal friction pairing between the wiper and the conductor track. To this end, wipers can be of "spoon" or "scraper-shape" design, and one or more can be mounted. Brush-shaped wipers are also in use.

A whole range of clear advantages are faced by a considerable number of serious disadvantages:

Advantages of the potentiometer-type sensors

- ▶ Low costs
- ▶ Simple design
- ▶ Very broad measuring effect (measuring range \approx supply voltage)
- ▶ No electronic circuitry required
- ▶ High level of interference immunity
- ▶ Broad temperature range (up to 250°C)
- ▶ High precision (better than 1 % of measuring range)
- ▶ Wide measuring range (almost 360° is possible)
- ▶ No difficulty with redundant design
- ▶ Calibration possible (laser ablation etc.)
- ▶ Flexible characteristic curve (variable conductor track width)
- ▶ Flexible assembly (on curved as well as on flat surfaces)
- ▶ Numerous manufacturers
- ▶ Rapid prototyping

Disadvantages of potentiometer-type sensors

- ▶ Mechanical wear from abrasion

- ▶ Measuring errors due to abraded particles
- ▶ Problematic when operated in fluids
- ▶ Variable contact resistance between wiper and measurement track
- ▶ Strong acceleration or vibration can result in wiper lift-off
- ▶ Testing is costly
- ▶ Limited possibilities for miniaturization
- ▶ Background noise

Applications

Examples of potentiometer-type sensors:

- ▶ Accelerator-pedal sensor or accelerator-pedal module for capturing the torque request for the engine-management system
- ▶ Fuel-level sensor (Fig. 2)
- ▶ Sensor plate potentiometer (KE-Jetronic and L-Jetronic) for detecting the air mass taken in by the engine
- ▶ Throttle-valve position sensor for capturing the position of the throttle valve in a gasoline engine (Figs. 3 to 5)

2 Potentiometric fuel-level sensor (design)

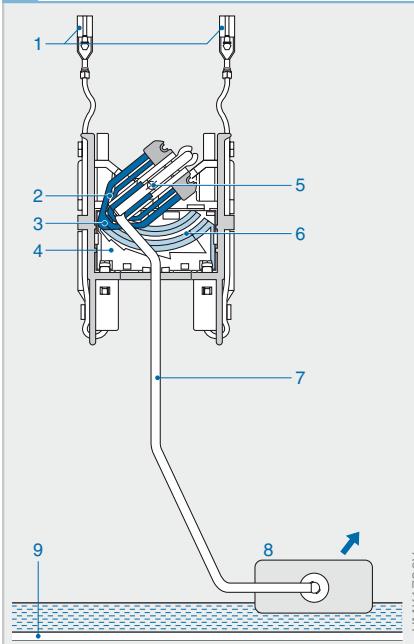
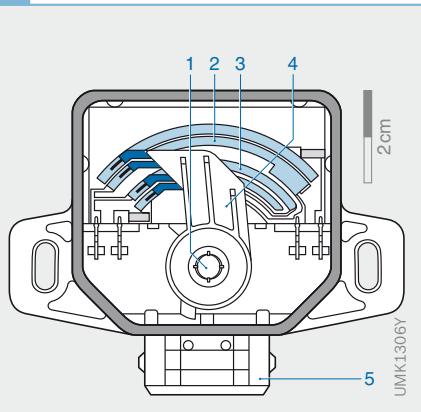
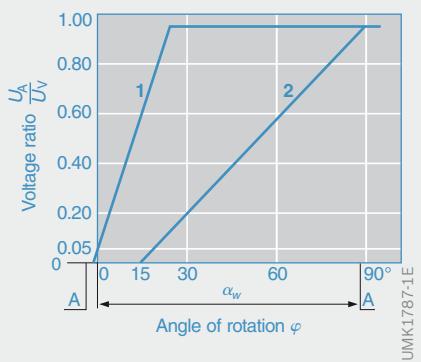
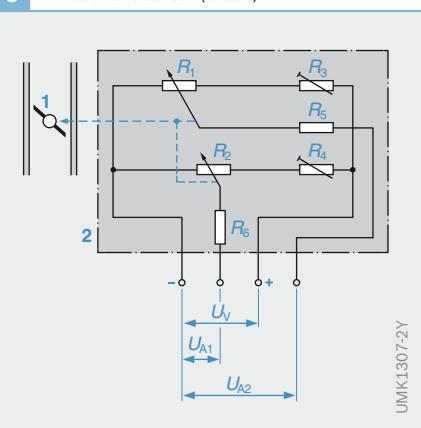


Fig. 2

- 1 Electrical connections
- 2 Wiper spring
- 3 Contact rivet
- 4 Resistor board
- 5 Bearing pin
- 6 Twin contact
- 7 Float lever
- 8 Float
- 9 Fuel-tank floor

UMK1790Y

3 Throttle-valve sensor (design)**4** Throttle valve sensor with two characteristic curves**5** Throttle-valve sensor (circuit)

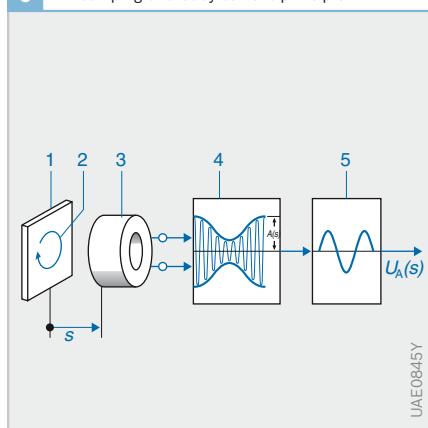
Magnetically inductive-type sensors

Of all the sensors using proximity and non-contacting principles for position measurement, the magnetic sensors have proved to be the most rugged and most insensitive to interference. This applies in particular to those principles relying on alternating current, in other words magnetically inductive principles. Compared with a micromechanical sensor though, the coil configuration needed here requires far more space. This means therefore that there is no favorable possibility of redundant (parallel measurement) design. Furthermore, coil contacting is less favorable from the costs and reliability viewpoint.

Although there are a multitude of different principles in use for this form of sensor, only two have come to the forefront for motor vehicle applications. Regarding their operating concepts, these are very similar to each other. However, they are currently no longer in use in their macro-mechanical version by Bosch for motor vehicle sensors in new developments.

Eddy-current sensors

When an electrically conductive flat or curved disk (for instance in aluminum or copper) approaches a coil (usually nonferrous) (Fig. 6) to which high-frequency alternating current has been applied, it has

6 HF damping and eddy-current principle**Fig. 3**

- 1 Throttle-valve shaft
- 2 Resistance track 1
- 3 Resistance track 2
- 4 Wiper arm with wipers
- 5 Electrical connection

Fig. 4

- A Internal stop
- 1 Characteristic curve for high resolution in the angular range 0° to 23°
- 2 Characteristic curve for angular range 15° to 88°
- U_A Measurement voltage
- U_V Operating voltage
- α_w Effective measured angle

Fig. 5

- 1 Throttle valve
- 2 Throttle-valve sensor
- U_A Measurement voltages
- U_V Operating voltage
- R_1, R_2 Resistance tracks 1 and 2
- R_3, R_4 Calibration resistors
- R_5, R_6 Protective resistors

Fig. 6

- 1 Spoiler
- 2 Eddy currents
- 3 Air-core inductor
- 4 Variable-damping oscillator
- 5 Demodulator
- s Measured travel
- $A(s)$ Oscillator voltage
- $U_A(s)$ Output voltage

an effect upon the coil's equivalent resistance and its inductance. This is the result of the eddy currents generated in the disk (otherwise known as a spoiler) due to the increasing magnetic coupling. The disk's position represents the measured travel s .

The generally high operating frequency (MHz range, a low-frequency feed would take up too much current because of the low coil inductance) requires a direct assignment of the electronics to the sensor or a shielded supply lead, which is generally not supportable. In order to convert the measuring effect into an electrical output voltage, either the damping effect (equivalent resistance) or field-displacement (inductance) principles can be applied. In the first (damping effect) case, a variable-amplitude oscillator can be used, and in the second a variable-frequency oscillator or a constant frequency supplied inductive voltage divider (differential configuration).

There are many ways to adapt the eddy-current principle to the measuring assignment. It is suited both to the determination of large travel distances or angles (there has already been a practically production-ready development, for the throttle valve and accelerator pedal, for instance) and also for very small variables (e.g. torque sensors). Because of the self-compensating properties, this principle generally only shows a very small temperature sensitivity over a wide temperature range. As eddy-current sensors can also, in principle, be manufactured micromechanically, their future use cannot be completely written off given their very advantageous properties. This sensor principle is, however, widely used in manufacturing and quality measuring technology, for the precise detection of the smallest displacements/distances or for the measurement of layer thicknesses in the μm range.

Short-circuiting ring sensors

In contrast to eddy-current sensors, the coil of a short-circuiting ring sensor has a straight or curved U or E-shaped soft magnetic, (usually) laminated iron core (Fig. 7). The coil, or short-circuiting ring ("spoiler"), is of highly conductive material (copper or aluminum), and is located around one (or all) of the core limbs. Because of their iron core, sensors of this type have a far greater inductance than eddy-current sensors and also a far higher measuring effect because of the guidance of the magnetic flux being very concentrated by the iron core. This means that they operate well at low frequencies, and do not necessarily require their signal electronics locally, that is directly on the sensor itself. The iron core also protects the measuring space between the core limbs very well against external interference fields.

The alternating field generated in and around the iron core by the coil current I cannot pass through the short-circuiting ring, since the eddy currents in the short-circuiting ring reduce it practically to zero.

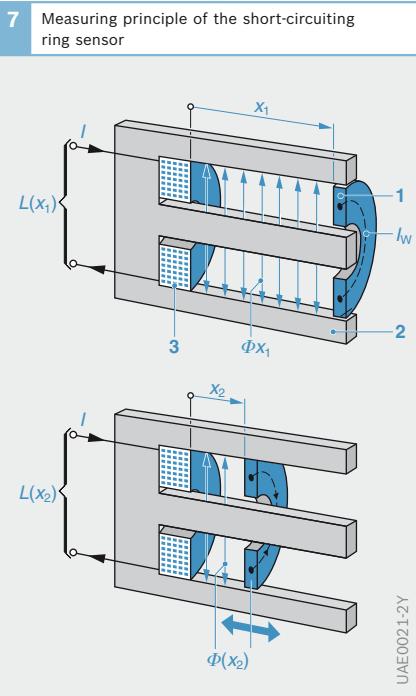


Fig. 7

Imaging for two different measured travels

- 1 Short-circuiting ring
- 2 Soft-magnetic core
- 3 Coil
- I Current
- I_w Eddy current
- $L(s)$ Inductance and
- $\Phi(s)$ Magnetic flux in measured travel s

UAE0021-2Y

Thus the eddy currents in the short-circuiting ring, acting like a “magnetic insulator”, limit the extent of the magnetic flux Φ to the space between the coil and the short-circuiting ring (hence its other name of shading ring) and thus make it dependent on the position x of the short-circuiting ring $\Phi = \Phi(x)$.

Inductance is defined:

$$(1) \quad L = \Phi/I$$

This is why the position x of the short-circuiting ring also directly influences the inductance L of the field coil. The relationship $L = L(x)$ shows a good linear fit over a wide domain. Virtually the entire length of the sensor can be utilized for measurement purposes. A mechanically tight guidance of the short-circuiting ring is not necessary here.

The mass of the short-circuiting ring to be moved is very low. Shaping the gap between coil and limb has an effect upon the shape of the characteristic curve: reducing the gap towards the end of the measuring range further improves the already good linearity. Operation is generally in the 5 to 50 kHz range, depending on material and structural shape. This sensor can also be used in very severe operating conditions, for instance on diesel fuel-injection pumps (Figs. 8, 9).

The cores are laminated (e.g. NiFe plate, 0.2 mm thick) because of the alternating current operation. To achieve the necessary mechanical stability to keep the plates stable under acceleration, they are not only bonded in the normal fashion, but also interlocked. To this end, the laminations are given a knobbly pattern in selected places which can be used to connect the laminations together before soft annealing (just like LEGO bricks), without interfering too much in the electrical insulation from lamination to lamination required.

The half-differential short-circuiting ring sensor with moving measuring ring and fixed reference short-circuiting ring is very accurate; it is evaluated as an inductive voltage divider (analysis of the inductances L_1/L_2 , or $(L_1 - L_2)/(L_1 + L_2)$) or even as a frequency determining element of an oscillating circuit to generate a frequency analog signal (very interference-proof, easy to digitize).

Applications

Examples for short-circuiting ring sensors:

- Rack-travel sensors for measuring the control-rack setting on in-line diesel injection pumps (Fig. 8)
- Angular-position sensors in the injected-fuel quantity actuators of diesel distributor-type injection pumps

Fig. 8

Design of the rack-travel sensor (RWG) for diesel in-line injection pumps

- | | |
|---|---------------------------------|
| 1 | Soft-magnetic core |
| 2 | Reference coil (L_2) |
| 3 | Reference short-circuiting ring |
| 4 | Control rack |
| 5 | Measuring coil (L_1) |
| 6 | Measuring short-circuiting ring |
| s | Control-rack travel |

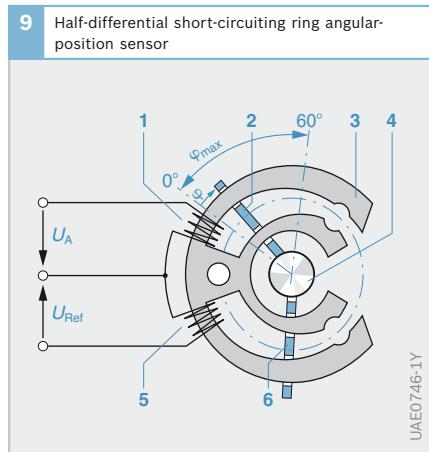
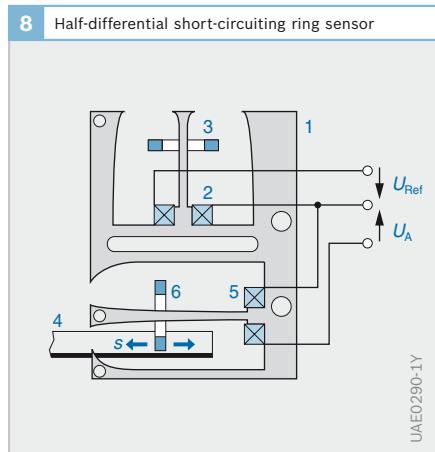


Fig. 9

- | | |
|------------------|---|
| 1 | Measuring coil |
| 2 | Measuring short-circuiting ring |
| 3 | Soft-magnetic core |
| 4 | Control-collar shaft |
| 5 | Reference coil |
| 6 | Reference short-circuiting ring |
| φ | Measured angle |
| φ_{\max} | Adjustment angle range for the control-collar shaft |

Figure 9 shows the structure of a half-differential short-circuiting ring sensor for diesel distributor-type injection pumps. A sensor coil and a reference coil are attached to each leg of the core. As the alternating current transmitted by the control unit permeates the coils, it creates alternating magnetic fields. The reference short-circuiting ring is fixed in position, while the measuring short-circuiting ring is attached to the control-collar shaft (adjustment angle φ). As the measuring short-circuiting ring moves along, the magnetic flux changes and the voltage across the coil changes, since the control unit maintains the current constant (load-independent current). An evaluation circuit forms the ratio of the output voltage U_A to the measurement inductance L_1 to the reference voltage U_{Ref} to the reference inductance L_2 . It is proportional to the excursion of the measuring short-circuiting ring and can be evaluated by the control unit.

Sensors with rotatable alternating fields

With alternating current of the angular frequency ω fed coils or coil-like structures (such as meandering conductor track structures) two or more pole alternating fields structures may be arranged either in a circle or even linearly. These pole structures with fixed pole pitch can be displaced in relation to a generally fixed set of receiver coils having the same pole pitch, by the movement of the system to be measured, either rotary or translatorial. Here the amplitudes of the receiver signals are changing with the movement in a sinusoidal pattern. If the receiver coils are displaced in relation to one another by a specific portion of the pole pitch T (e.g. $T/4$ or $T/3$), the sinus pattern is phase-shifted by a corresponding angle (e.g. by 90° or 120°).

The voltages arise as follows, for instance:

$$(1) \quad u_1 = U \cdot \sin \varphi \cdot \sin \omega t$$

$$(2) \quad u_2 = U \cdot \sin(\varphi - 90^\circ) \cdot \sin \omega t$$

$$= U \cdot \cos \varphi \cdot \sin \omega t$$

or even:

$$(3) \quad u_1 = U \cdot \sin \varphi$$

$$(4) \quad u_1 = U \cdot \sin(\varphi - 120^\circ) \cdot \sin \omega t$$

$$(5) \quad u_1 = U \cdot \sin(\varphi - 240^\circ) \cdot \sin \omega t$$

After rectification, the angle of rotation α can be calculated with great precision from these voltages. The sensors preferably made as angle sensors identified in classical measurement technology as synchro, resolver or even inductosyn methods function in this way.

In a simple bipolar arrangement, the electrical phase angle also corresponds directly to the mechanical angle of rotation α . In an arrangement of n pole pairs, the mechanical full angle $\varphi = 2\pi$ is converted into a phase displacement of $\alpha = n \cdot 2\pi$, so that the phase displacement α corresponds only to an angle of rotation of only φ/n , which considerably enhances the resolution of the measuring system. However, the unambiguity of the measurement signal must be restored by additional means, such as a simple angle sensor, in larger measuring ranges.

Hella sensor

The angle sensor developed by Hella does not strictly correspond to any of the methods described above, but is most similar to the inductosyn method. Figure 10 shows the sensor with a 6-count pole structure ($n = 6$), for instance, which electrically seen converts an angle of rotation of $\varphi = 60^\circ$ into a phase displacement of the signal amplitudes of $\alpha = 360^\circ$. All the conductor track structures required are applied to multi-layer PCBs, at least in the case of the fixed part (stator). The rotor may also be made as a stamped part, if necessary, either free-standing or mounted on a plastic carrier (hot stamping).

A circular conductor track loop is made on the stator. This induces an eddy current at an operating frequency of 20 MHz in a closed meandering loop of approximately the same outside diameter on the rotor.

This eddy current naturally generates a secondary magnetic field, just like the exciter loop, which superimposes itself on the excitation field in the sense that it attempts to eliminate it. If the track on the rotor were only a circular conductor track congruent with the stator loop instead of the meandering track, this would simply extinguish the primary field to a great extent. The meandering structure, however, creates a resulting multi-pole field which can be rotated with the rotor and the total flux of which is also naturally almost zero.

This multi-pole alternating field is sensed by concentric receiver coils or meanders of practically identical form located on the stator. These are offset within a pole pitch (of e.g. 60°) by $1/3$ each, i.e. electrically in their signal amplitude by 120° each (Fig. 10b). The receiver coils however extend over all n pole pairs (series circuit) and use the sum of all pole fields.

According to Figure 10c, the receiver coils are connected in a star connection. Their signals are sent to an ASIC to determine the electrical phase angle α or mechanical angle of rotation φ , which applies the necessary (sign-correct) rectification, selection and ratio formation. A version ASIC 1 receives the digital control signals necessary for this from an adjacent microcontroller. A different version ASIC 2 is, however, also capable of operating the sensor completely independently (on a stand-alone basis). The ASICS also permit an end-of-line calibration of the mechanical and electrical tolerances in production. It is also possible, for the case of applications with elevated safety requirements, to construct a redundant system with two galvanically isolated signal paths and two ASICS. The sensor principle can also very advantageously be applied in a “cut-open” form as a travel sensor.

The principal benefits of this sensor, in addition to the very favorable production costs, are as follows:

- ▶ Contact-free and wear-and-tear resistant measuring principle
- ▶ Complete temperature independence (up to 150°C)
- ▶ High accuracy (up to approximately $\pm 0.09^\circ$ for a measuring range of 360°)
- ▶ Flexibility (may be adapted to any angular range up to 360°)
- ▶ Possibility of a redundant design
- ▶ High EMC
- ▶ Use of standard materials exclusively, without the need for ferromagnetic components
- ▶ Flat design

The sensor can therefore be used at very many points on the motor vehicle.

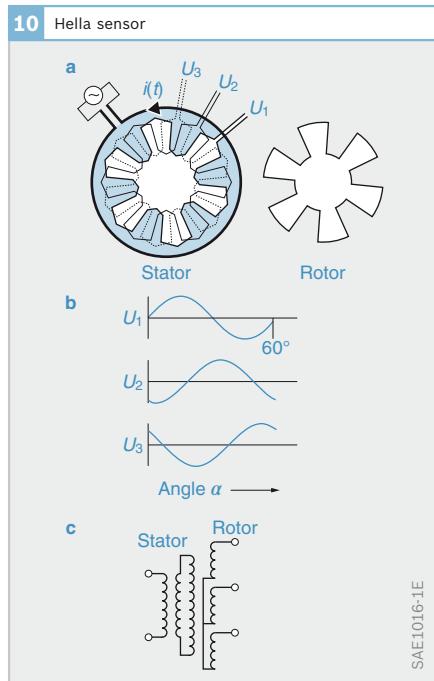


Fig. 10
a Schematic design
b Circuit
c Output signals

Magnetostatic sensors

Magnetostatic sensors measure a DC magnetic field. In contrast to the magnetically inductive (coil) sensors they are far more suitable for miniaturization and can be manufactured at reasonable cost using microsystem technology. As DC magnetic fields can reach through housing walls of plastic, and also those of non-ferromagnetic metals, with no problems, magnetostatic sensors have the advantage that the sensitive, generally fixed part, can be well encapsulated and protected against the rotating part, generally a permanent magnet or soft-magnetic conductive element – and also against the environment. Above all, galvanomagnetic effects (Hall and Gaussian effects) are used, as well as anisotropic magnetoresistive (AMR and GMR) effects.

Galvanomagnetic effect

The Hall effect is normally evaluated using thin semiconductor wafers. If a current-carrying wafer of this type is exposed perpendicularly to magnetic induction B , the charge carriers are deflected by the Lorentz force normal to the field and the

current I by angle φ from their otherwise straight path (Fig. 11). This means that it is possible to pick-off a voltage U_H transverse to the current direction between two opposing points on the edge of the wafer proportional to the field B and to the current I (Hall effect).

$$(2) \quad U_H = R_H \cdot I \cdot B/d$$

where R_H = Hall coefficient,
 d = wafer thickness

At the same time, the series impedance of the wafer increases independently of the field direction following an approximately parabolic characteristic curve (Gaussian effect, magnetoresistor).

The coefficient determinant for the measurement sensitivity of the wafer R_H is only comparatively small for silicon. As the wafer thickness d can be made extremely thin, however, using diffusion techniques, the Hall voltage amounts to a technically usable magnitude. When silicon is used as the basic material, a signal-conditioning circuit can be integrated on the wafer. Sensors using these principles are very cheap to produce. Silicon is by a long way not the best semiconductor material for Hall sensors in respect of measurement sensitivity and temperature sensitivity. III-V semiconductors such as gallium arsenide or indium antimonide possess better characteristics.

Hall switch

In the most simple case, the Hall voltage is applied to an electronic threshold circuit (Schmitt trigger) built into the sensor which then outputs a digital signal. If the magnetic induction B applied at the sensor is below a given minimum threshold level, the Schmitt trigger's output value corresponds to a logical "0" (release status); if it exceeds a given upper threshold the output value corresponds to a logical "1" (operate status). Since this behavior is guaranteed across the complete operating-temperature range and for all sensors of a given type, the two threshold values are relatively far apart (approx. 50 mT). A con-

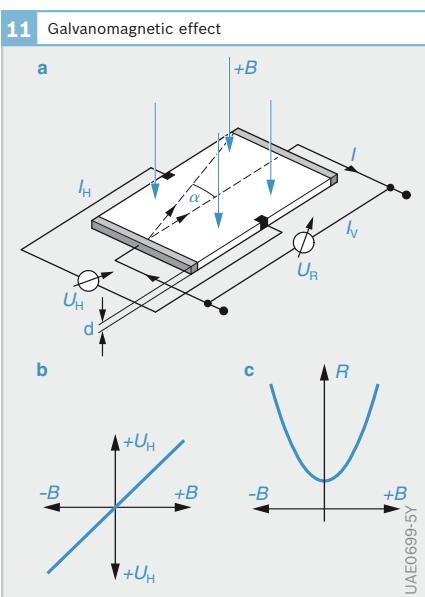


Fig. 11

- a Circuit
- b Curve of Hall voltage U_H
- c Increase of wafer resistance R (Gaussian effect)
- B Magnetic induction
- I Wafer current
- I_H Hall current
- I_V Supply current
- U_H Longitudinal voltage
- α Deviation of the electrons due to the magnetic field

siderable induction jump ΔB is thus necessary to operate the Hall switch.

Sensors manufactured in bipolar technology of this kind were used in Hall vane switches (Fig. 12), for example, as were fitted in the housing for the camshaft-driven ignition distributor. This Hall vane switch contains the sensor and a permanent magnet and soft magnetic conductive elements. The magnetic circuit is U-shaped or forked in such a way that a trigger wheel vane in soft magnetic material can travel through the open end, alternately screening or exposing the magnet, thereby toggling the Hall sensor between the operate and release states.

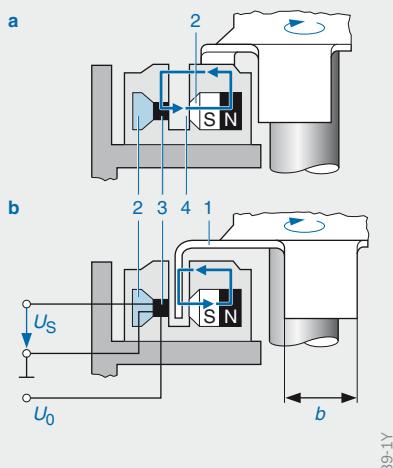
A further application is the digital steering-angle sensor LWS3 (see "Hall angular-position sensor up to 360°").

Hall sensors of this type may be very cost-effective, but are really only good for a switch operation and are too imprecise for capturing analog values.

Hall sensors using the spinning-current principle

A disadvantage of a simple silicon Hall sensor is its simultaneous sensitivity to the unavoidable mechanical strain (piezoelectric effect) resulting from the packaging, leading to unfavorable temperature sensitivity of the offset. Applying the spinning current principle (Fig. 13), associated with a transition to CMOS technology, has overcome this disadvantage. Although the piezoelectric effect still occurs here, it is compensated for by the signal being averaged over time, as it occurs with a different sign in the case of very rapid electronically controlled interchange (rotation) of the electrodes. If it is wished to save the expense of the complex electronics for switching the electrodes, it is possible also to integrate several Hall sensors (two, four or eight) with correspondingly different alignment of the current paths in close proximity and to add their signals together in order to form an average. It is only in this way that the Hall ICs were also suitable for analog sensor applications.

12 Hall vane switch



13 Hall sensor using the spinning-current principle

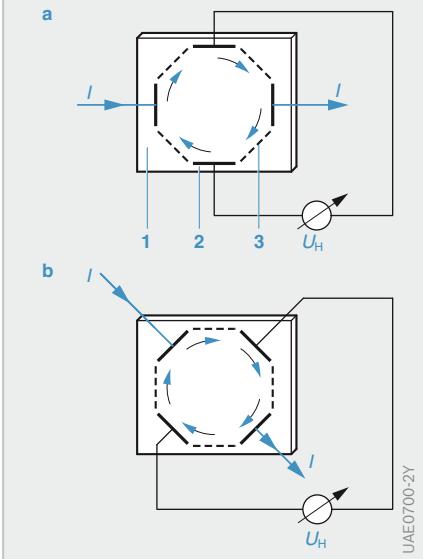


Fig. 12

- a Magnetic flux: unhindered
- b Magnetic flux: short-circuited
- 1 Vane width b
- 2 Soft-magnetic conductive element
- 3 Hall IC
- 4 Air gap
- U_0 Supply voltage
- U_S Sensor voltage

Fig. 13

- a Rotary phase φ_1
- b Rotary phase $\varphi_2 = \varphi_1 + 45^\circ$
- 1 Semiconductor wafer
- 2 Active electrode
- 3 Passive electrode
- I Supply current
- U_H Hall voltage

These measures though did not result in a reduction of the considerable effects of temperature on the sensor's measurement sensitivity.

Such Hall ICs are suitable above all for the measurement of small travel distances, in which they register the fluctuating field strength of a permanent magnet as it approaches (e.g. the iBolt force sensor detects the passenger's weight to optimize the triggering of the airbag). Similarly good results could, in the past, only be achieved by using single Hall elements, e.g. in III-V compounds with a hybrid amplifier connected downstream (e.g. Hall acceleration sensor).

Differential Hall sensors

The double Hall sensor (differential Hall sensor, Fig. 14) has two complete Hall systems mounted at a specified distance from each other on one chip. The associated electronics evaluate the difference between the two Hall voltages. The advantage of such sensors lies in the fact that their output signal is for the most part independent of the absolute value of the magnetic-field strength, and as differential sensors they only register the magnetic induction's change in space, in other words the field gradient (thus the common name of gradient sensor).

Since the polarity of their output signal is independent of the air gap between rotor and sensor, these sensors are mostly used for rotational-speed measurement. If a simple Hall sensor only is used to sense a gear wheel, this will not be able to distinguish whether the magnetic flux has modified as a result of the gear wheel turning further or by a change in distance (e.g. vibrations, mounting tolerances). Significant scanning errors occur, as the signal must be passed to a threshold detector. The situation with differential sensors is fundamentally different. Here only the difference in the signals from two Hall sensors located at a suitable distance from one another around the circumference is evaluated. If the signal difference is positive, for example, it is possible to freely vary the distance between the gear wheel and sensor; the difference will remain positive, even if the absolute value may become smaller. The sign can only change if the rotor turns further. A threshold detector connected afterwards thus has no problem in distinguishing between a change in distance and rotation.

Usually, in order to achieve as high an output signal as possible, the distance between the two Hall sensors each located on the edge of the (elongated) chip is then selected to correspond to about half an increment (half the rotor tooth interval).

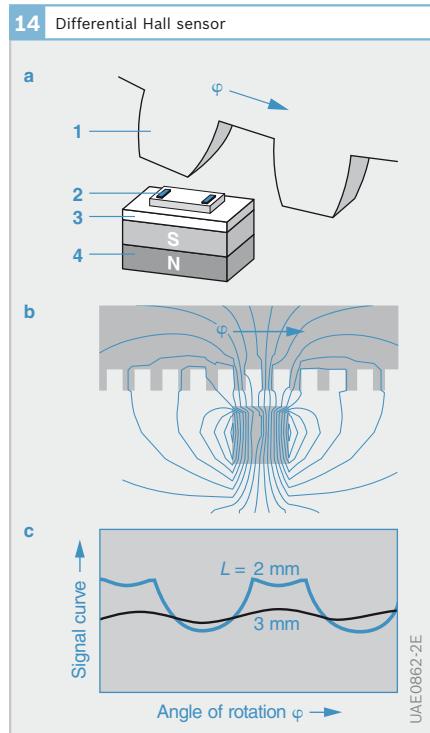


Fig. 14

- a Design
- b Field-strength distribution (1.5 times increment spacing)
- c Signal curve for air-gap widths L
- 1 Ring gear
- 2 Differential Hall IC
- 3 Homogenizing wafer (soft iron)
- 4 Permanent magnet

The signal maximum is very wide and covers a broad range of variation of the increment spacing. More pronounced variations in the spacing necessitate a highly complex redesign of the sensor.

A gradient sensor must be precisely aligned to the rotor's direction of rotation.

Hall angular-position sensors in the range up to approximately 180°

Using a rotatable magnetic ring ("movable magnet"), together with a number of fixed soft-magnetic conductive elements, a linear output signal can be generated for a larger angular range without conversion being necessary (Fig. 15). In this case, the bipolar field of the magnet ring is passed through a Hall sensor arranged between semicircular conductive elements. The effective magnetic flux through the Hall sensor is dependent on the angle of rotation φ .

This principle is used on accelerator-pedal sensors.

The Type ARS1 Hall angular-position sensor with a measuring range of approx. 90° (Fig. 16) is derived from the basic "movable magnet" principle. The magnetic flux from a practically semicircular permanent-magnet disk is returned to the magnets through a pole shoe, two additional conductive elements each of which contains a Hall sensor in its magnetic path, and the shaft which is also ferromagnetic. Depending upon the angular setting, the flux is led to a greater or lesser degree through the two conductive elements, in the magnetic path of which a Hall sensor is also situated. Using this principle, it is possible to achieve a practically linear characteristic in the measuring range.

The Type ARS2 is a simplified version which does without soft magnetic conductive elements (Fig. 17). In this version, the magnet moves around the Hall sensor in a circular arc. Only a relatively small section of the resulting sinusoidal characteristic curve features good linearity. If the Hall

sensor is located slightly outside the center of the circular arc, the characteristic curve increasingly deviates from the sinusoidal. It now features a short measuring range of just 90°, and a longer measuring range of a little over 180° with good linear-

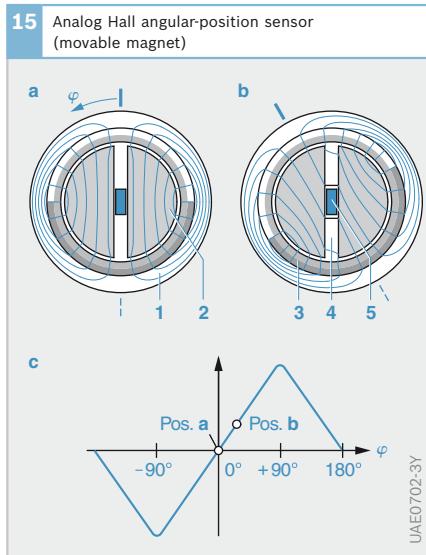


Fig. 15
Linear characteristic curve for angles up to 180°
a Position a
b Position b
c Output signal
1 Magnetic yoke
2 Stator (1, 2 soft iron)
3 Rotor (permanent magnet)
4 Air gap
5 Hall sensor
φ Angle of rotation

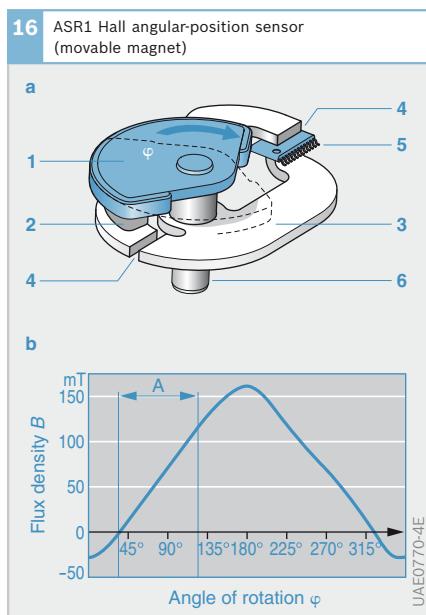


Fig. 16
Linear characteristic curve for angles up to approximately 90°
a Design
b Characteristic curve with working range A
1 Rotor disk (permanent-magnetic)
2 Pole shoe
3 Conductive element
4 Air gap
5 Hall sensor
6 Shaft (soft magnetic)

ity. A great disadvantage though is the low level of shielding against external fields, as well as the remaining dependence on the geometric tolerances of the magnetic circuit, and the intensity fluctuations of the magnetic flux in the permanent magnet as a function of temperature and age.

Angular-position sensors in the range up to approximately 360°

Figure 18 shows an analog Hall angular-position sensor with a measuring range of up to 360°. As shown, a permanent magnet rotates in front of two Hall sensors arranged at right angles to each other. The magnet should be adequate in size in order not to be too dependent on the tolerance of the magnet positioning. The two Hall sensors, fitted as closely together as possible, should here be positioned in the homogeneously directed portion of the permanent-magnet's stray field, which also represents the angular location φ of the permanent magnet. They are perpendicular to one another and parallel with the permanent magnet's rotary axis. When the field-strength vector B rotates past the sensors, therefore, they register its x and y components:

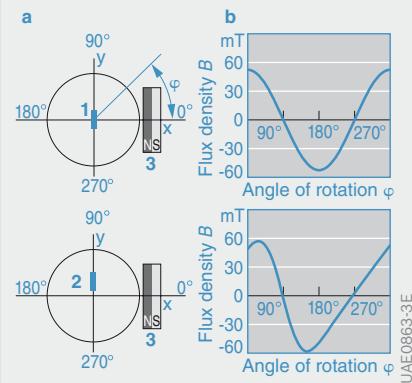
Fig. 18

- a Constructed from discrete Hall ICs
- b Constructed from planar integrated Hall ICs
- 1 Signal electronics
- 2 Camshaft
- 3 Control magnet
- B Induction
- I Current
- U Voltage
- U_A Output voltage
- φ Angle of rotation

Fig. 17

- Linear characteristic curve for angles over 180°
- a Principle of operation
 - b Characteristic curve
 - 1 Hall IC positioned in the mid-point of the circular path
 - 2 Hall IC located outside the mid-point (linearization)
 - 3 Magnet

17 ASR2 Hall angular-position sensor (movable magnet)



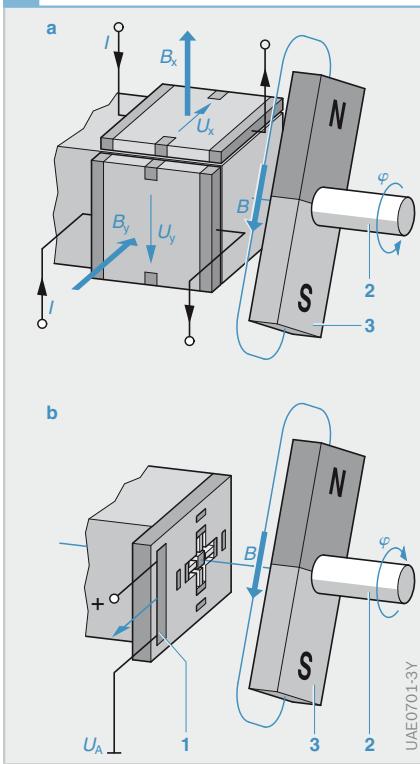
$$(3) \quad U_{H1} = U_x = B \cdot \sin \varphi$$

$$(4) \quad U_{H2} = U_y = B \cdot \cos \varphi$$

The angle φ can be calculated using the trigonometric relationship $\varphi = \arctan (U_{H1}/U_{H2})$ in an associated commercially available evaluator chip, through which the sensor signal is generally digitized.

A Hall sensor arrangement of this nature can also, in principle, be integrated vertically with VHD (Vertical Hall Devices) in the manner illustrated so that the plane of the sensor chip is arranged perpendicular to the axis of rotation and the sensor has an in-plane sensitivity, unlike a normal, planar Hall sensor (Fig. 19). Monolithic integration also guarantees the high level of precision as required for the 90° arrangement and for the requisite compact form of

18 Analog Hall sensor for 360° angles

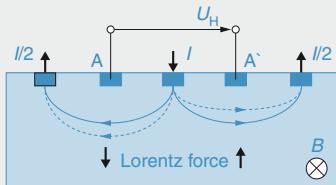


the two Hall systems. Sentron (Melexis), of Switzerland, is working on sensors of this type and will be launching them on the market in the near future.

Angular-position sensors of this type were considered for the implementation of starter motor-less instant start systems, for instance, where the absolute rotary position of the camshaft over a 360° range must be measured.

There is, however, another trick which can be exploited to (additionally) give a normal, planar Hall sensor an in-plane sensitivity. In the case of the method used by Melexis-Sentron (e.g. its 2SA-10 model) a total of four Hall sensors are mounted on the same chip in a tight circle offset at 90° to one another to measure the field components B_x and B_y . To be accurate, in this case, each sensor consists of a Hall element pair rotated at 90° for its part for the reasons described above. Once the chip has been manufactured, a circular disk of soft magnetic material approximately $200\ \mu\text{m}$ in diameter is fixed on its surface as shown in Figure 20 in such a way that the Hall sensors are located exactly beneath the edge of the disk.

19 Vertical hall device (VHD) in cross-section

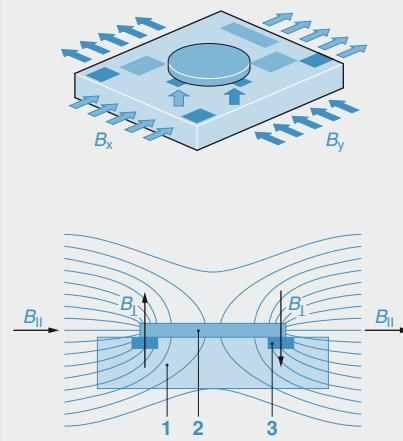


SAE1019-1E

Because of its high relative permeability, this disk acts as a flow concentrator (IMC, Integrated Magnetic Concentrator) and compels all field lines to pass into its surface vertically. This compels the field lines running horizontally (in-plane) without the concentrator into a vertical direction (out-of-plane) at the position of the Hall sensors and thus they can modulate the Hall elements. As diagonally opposing elements each see fields in opposite directions, the difference between the two Hall voltages running in opposite directions is evaluated to capture a field component. At the same time, any vertical components B_z present have their effect eliminated.

The additional evaluation of the summary signals of diagonally opposing elements can also be used to capture the vertical field component B_z , since these are not in the slightest influenced by the ferromagnetic disk and the elements have also not lost their normal out-of-plane sensitivity

20 Set of 4 x 2 planar Hall sensors with IMC



SAE1020-1V

Fig. 19
The currents directed into the interior of the chip I are deflected by the Lorentz forces because of the flux density aligned in parallel with the chip surface B in such a way that a Hall voltage U_H can be picked off between the surface electrodes A and A'.

Fig. 20
IMCIntegrated Magnetic Concentrator

The Hall elements are mounted on the chip in pairs at an angle of 90° to one another directly below the edge of a small ferromagnetic disk. Thanks to their presence, the flux density parallel to the chip surface (induction) $B_{||}$ is converted into a flux density B_\perp that can be sensed by the Hall sensors.

by it being fitted. In this way, an arrangement of this type can be used to capture the induction vector B acting at the sensor location in all its three components. The electronic circuitry required to evaluate the signals including a μC-based digital signal processor (DSP) used to calculate the arctan function and the means required for an end-of-line calibration of the sensor (EEPROM) can cost-effectively also be integrated on the same chip as the sensor.

In order to measure angles up to 360° , it is sufficient to rotate a permanent magnet, preferably round in form, magnetized in parallel with the chip over the sensor chip with its integrated signal processing (Fig. 21). As the angle of rotation is formed using the arctan function from the ratio of the two sine wave and cosine wave sensor signals, the strength of the magnetic field, and thus also the aging of the magnets, its temperature sensitivity and its distance from the sensor surface are largely irrelevant. Only the rotary position of its magnetization is detected.

21 Angular position measurement over 360° with quadruple Hall sensor

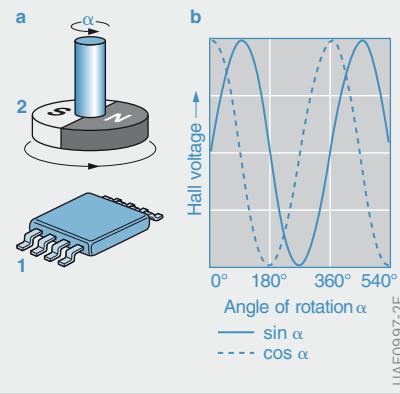


Fig.

21

Angle measurement at the end of a rotating shaft with quadruple Hall sensor manufactured by Melexis containing an integrated flux concentrator (IMC) to deflect the field.
a Measuring system
b Primary output signals

The maximum intrinsic deviation of the sensor is stated by the manufacturer as $\pm 2^\circ$ for a measuring range of 360° . This can, however, be significantly improved by the user on calibration of the sensor. The accuracy of the digital output is then 10 bits at a resolution of 12 bits. Because of the processing time required by the signal processor, the smallest sampling rate for a timing frequency of 20 MHz is 200 μ s. A pulse width modulated signal output can be operated at a maximum of 1 kHz. The sensor needs 15 ms for initialization. A typical value for the measuring field strength is approximately 40 mT. The concentrator plate becomes saturated at flux densities in excess of 0.7 T.

The sensor may also be programmed for any measurement range $< 360^\circ$ to good effect, with additional functional security being generated by out-of-range monitoring (application as an angular-position sensor in the accelerator-pedal module FPM2.3, for instance).

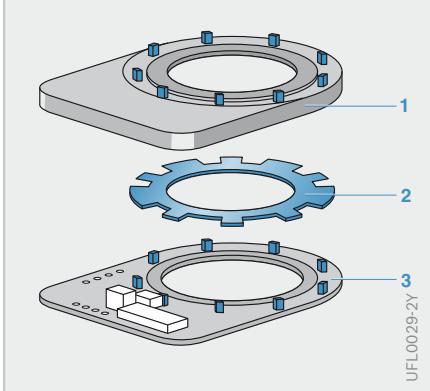
Angles up to 360° may also be captured with the simplest Hall ICs (Hall switches), as are also used for the measurement of speeds of rotation, (application in LWS type steering-angle sensors). Here, in order to obtain an n-bit resolution, n Hall switches are arranged on a single track in a circle at equal distances from each other (Fig. 22). Depending upon its position, a rotatable soft-magnetic code disk blocks the magnetic field of the individual permanent magnets located above each Hall switch, or opens it when it rotates further so that one after another the Hall switches generate n different code words (serial code). It is useful to apply the Gray code (cyclic binary code) to prevent errors in transitional states of the code disk. The Gray code is designed such that the code words for two adjacent positions – unlike the case with a natural binary code, for instance – differ in only 1 bit, and so any incorrect indication on the transition from

one position to the next can never amount to more than one angular step.

For the practical implementation of a steering-wheel-angle sensor, the code disk is connected to the steering shaft for instance, and the sensor's non-moving parts to the vehicle body. One difficulty with this sensor is that the soft magnetic code disk must be mounted floating on the steering column for reasons associated with assembly tolerances, which leads to a complex and space-hungry mounting there using a flexible plastic sleeve. Because of the air gap tolerance between the upper disk and the lower disk, the code disk is mounted in a wiping guide. A corresponding optoelectronic solution using photoelectric light barriers is less satisfactory, as it is not possible to construct an adequate encapsulation against the susceptibility of the sensor elements to dirt.

In order to measure a number of complete rotations, an additional 3-bit configuration can be used in which the code disk is rotated by a reduction gear. The resolution of such configurations is mostly no better than 2.5° .

22 Digital Hall angular-position sensor with n Hall switches



Magnetoresistive sensors

In addition to the transversally directed Hall effect, there is also a longitudinal resistance effect, also known as the Gaussian effect, that occurs on semiconductor wafers. Elements that exploit this effect are known as magnetoresistors which are manufactured from a III-V-semiconductor (crystalline indium antimonide (InSb)). In contrast with the Hall sensors, the optimum wafer shape for a magnetoresistor tends to be shorter and squatter, and represents a very low electrical resistance. In order to arrive at technically applicable values in the $\text{k}\Omega$ range, it is therefore necessary to connect a large number of these wafers in series. This problem is solved elegantly by adding microscopically fine, highly conductive nickel-antimonide needles to the semiconductor crystal. These are located transverse to the current direction. A further measure is to apply meander techniques to the semiconductor resistor (Figs. 23 and 24).

The dependence of the resistance on the magnetic induction B follows a square-law function up to inductances of approx. 0.3 T, and above this point it is increasingly linear. There is no upper limit to the control range, and dynamic response can be regarded as practically free from lag, as is the case for the Hall sensor.

23 Magnetoresistive differential sensor

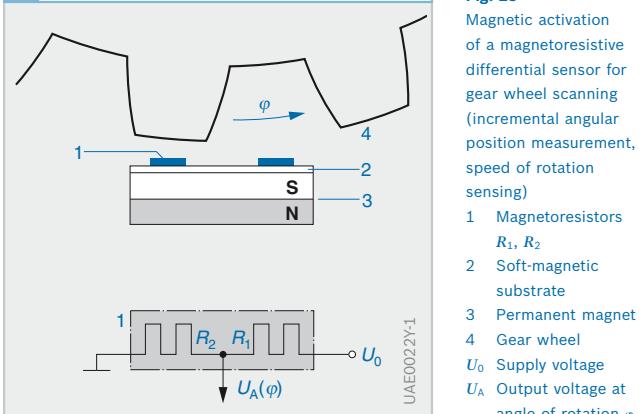


Fig. 22
Angular position measurement up to 360° with a circular, equidistant arrangement of simple Hall switches

1 Housing with permanent magnets

2 Code disk

3 Printed-circuit board with Hall switches

Fig. 23
Magnetic activation of a magnetoresistive differential sensor for gear wheel scanning (incremental angular position measurement, speed of rotation sensing)

1 Magnetoresistors R_1, R_2

2 Soft-magnetic substrate

3 Permanent magnet

4 Gear wheel

U_0 Supply voltage

U_A Output voltage at angle of rotation φ

Since their temperature sensitivity has a pronounced effect upon their resistance value (approximately 50 % reduction over 100 K), they are usually supplied only in the dual-configuration form in voltage-divider circuits (differential magnetoresistors). For the particular application, each of the two resistor sections must then be magnetically triggered (as far as possible with opposite polarities). Notwithstanding the high temperature coefficients of the individual resistors, the voltage-divider circuit guarantees good stability of the working-point (that point at which both resistor sections have the same value).

In order to achieve good measurement sensitivity, it is best to operate the magnetoresistors at a magnetic working point between 0.1 to 0.3 T. Generally, the required magnetic initial-tension is supplied by a small permanent magnet, the effects of which can be increased by using a small magnetic return plate.

The magnetoresistor's main advantage is its high signal level which is usually in the volts range. This means that amplification is unnecessary, as well as the local electronic circuitry and the associated protec-

tive measures which would otherwise be needed. Furthermore, in their role as passive, resistive components they are highly insensitive to electromagnetic interference and, as a result of their high bias voltage, practically immune against external magnetic fields.

Because of its pronounced temperature sensitivity, the magnetoresistor is used almost exclusively in incremental angular-position and rev counters, or in binary limit-value sensors (with switching characteristic).

Anisotropic magnetoresistive sensors (AMR)

Thin, only around 30 to 50 nm thick NiFe alloy coatings show an electromagnetically anisotropic behavior. That means that their electrical resistance changes under the influence of a magnetic field. Resistance structures of this type are thus known as anisotropic magnetoresistive (abbreviated to AMR) elements or sensors. The metal alloy generally used is also known as permalloy.

Technology and designs

Even without an external control fields, a small spontaneous magnetization M_S will arise in the longitudinal direction of a conductor track on a long slim resistor strip, as shown in Figure 25a, (shape anisotropy). In order to give this magnetization a clearly defined direction - theoretically, it could be in the other direction - AMR sensors are often provided with weak bias magnets, as shown. In this state, the longitudinal resistance achieves its greatest value R_{\parallel} . If the magnetization vector is rotated under the effect of an additional external field H_y through angle ϑ , the longitudinal resistance declines gradually until it reaches its minimum value R_{\perp} at $\vartheta = 0$. Here, the resistance depends only on the angle ϑ which is enclosed by the resulting magnetization M_S and the current I , and it has an approximately cosine wave pattern as a function of ϑ :

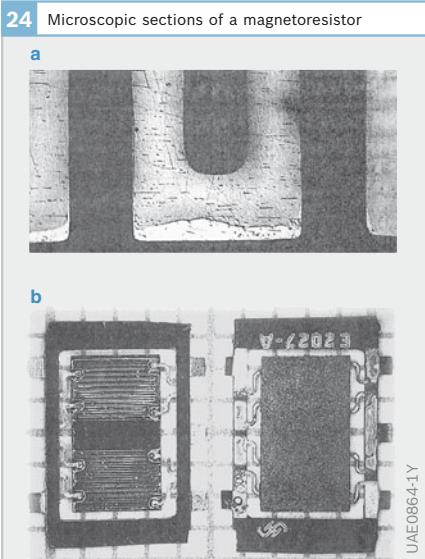


Fig. 24

- a Microscopic section (cut-out)
- b On ferrite substrate in Super-8 film carrier

$$(5) \quad R = R_0 \cdot (1 + \beta \cdot \cos^2 \vartheta)$$

where $R_{\parallel} = R_0 \cdot (1 + \beta)$; $R_{\perp} = R_0$

The coefficient β here identifies the maximum possible variation in resistance. It amounts to around 3 %. If the external field is much stronger than the spontaneous generated magnetization, and this is usually the case when permanent magnets are used, the effective angle ϑ is almost completely a function of the direction of the external field. The field strength is now irrelevant, and in other words the sensor is now operating in a "saturated state".

Highly-conductive short-circuiting strips (for instance of gold) on the AMR film force the current to flow at an angle of below 45° to the spontaneous magnetization (longitudinal direction) without the application of an external field. As a result of this "trick", the sensor curve shifts by 45° compared to that of the simple resistor. This results in the so-called "Barber Pole" sensor. This means, therefore, that even with the external field strength $H_y = 0$, the curve is at the point of maximum measurement sensitivity (reversal point).

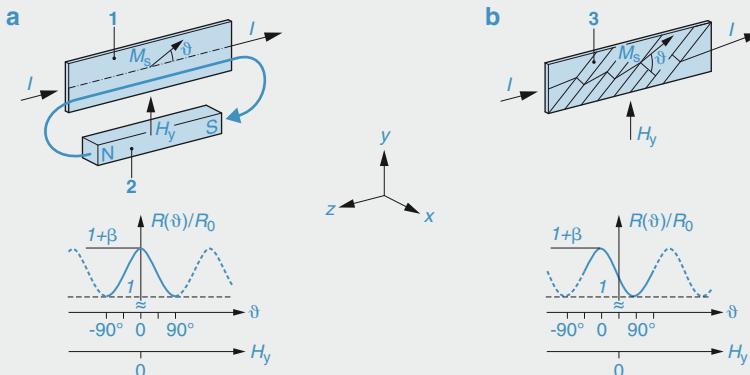
The striping of two resistors in opposite directions causes them to change their resistances in opposite directions under the effects of the same field. In other words,

one of the resistances increases while the other drops.

Being thin-film sensors, anisotropic magnetoresistive sensors also have the advantage that they can be calibrated to a setpoint value (e.g. zero point) by laser trimming, for instance. In principle, the oxidized silicon wafers which serve as the substrate material can also incorporate the electronic circuitry for signal conditioning. At present, cost considerations dictate that sensor chips and electronic chips are for the most part manufactured separately, and then mounted and packaged in a common leadframe, for instance. The controlling magnetic field B is usually generated by magnets moving side-to-side or rotating over the sensor.

In addition to the simple, two-pole AMR elements, there are also pseudo Hall sensors in practically square NiFe thin layer structures, that have four connections, just like normal Hall sensors. Two for the current path and two transverse to this for pick-off a (pseudo) Hall voltage (Fig. 26). Unlike a normal Hall sensor, the pseudo Hall sensor, however, has its sensitivity to magnetic fields in the plane of the layer, and not perpendicular to it. In addition, it does not have a proportional characteristic curve, but a sinusoidal one with very a

25 AMR basic principle, barber pole structure



high match to a sinus wave, which does not depend in the least on the strength of the control field and the temperature. The output voltage vanishes for a field parallel to the current path, to describe a sinus half-period on rotation up to an angle $\varphi = 90^\circ$. The sinus voltage obtained in this way combined with the amplitude results in:

$$(6) \quad U_H = u_H \cdot \sin 2\varphi$$

If the external control field is rotated once through $\varphi = 360^\circ$, the output voltage follows two full sinus periods. The amplitude u_H is, however, very dependent on the temperature and the air gap between the sensor and the control magnet; it declines with increasing temperature and widening air gap.

The measurement sensitivity of the pseudo Hall element can be considerably increased (without excessively falsifying the sinusoidal shape) by “hollowing out” the element from the inside so that only the “frame” remains (Fig. 26b). This modification converts the pseudo Hall sensor to a full bridge consisting of four AMR resistors from its geometrical form (Fig. 26c). Even when the bridge resistors are meander-shaped, provided a given minimum conductor width is not fallen below, this still has negligible effect of the signal’s sinusoidal shape.

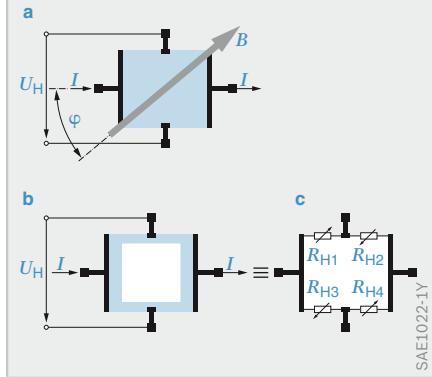
Fig. 27

- Measuring range up to $\pm 15^\circ$
- a Measuring principle
- b Characteristic curve
- 1 Permalloy resistors
- 2 Rotatable permanent magnet with control induction B
- 3 Lower operating temperature
- 4 Higher operating temperature
- a Linear measuring range
- b Effective measuring range
- U_A Output voltage
- U_0 Supply voltage (5V)
- φ Angle of rotation

Fig. 26

- a Full area basic form
- b Modified form with hollowed surface
- c Electrical circuit diagram for form b

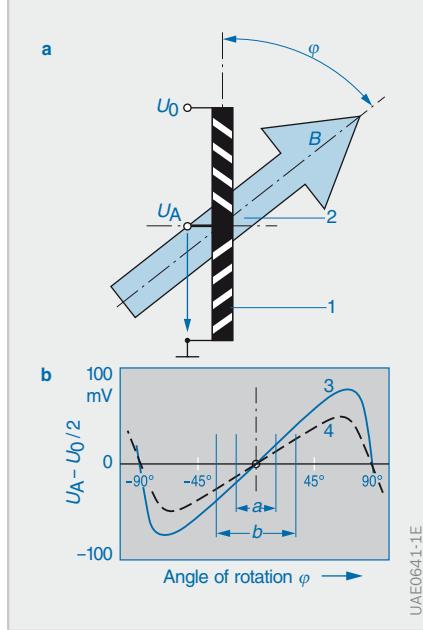
26 Pseudo Hall sensor



Simple AMR angular-position sensor for ranges $< 30^\circ$

Barber pole AMR angular-position sensors with limited precision and a restricted measuring range (max. $\pm 15^\circ$) rely on the unbalance of a magnetoresistive voltage divider (differential sensor) comprising an elongated (possibly meander-shaped) permalloy resistor with highly conductive gold cross strips (Fig. 27). Although in the case of such sensors the zero point is largely independent of the distance between the magnet and the sensor, the gradient of the characteristic curve, which for its part also depends on the temperature (TK around $-3 \cdot 10^{-3}/K$), on the other hand, is not. A sensor of this kind has already been considered as a cost-effective alternative for the measurement of pedal positions.

27 Magnetoresistive angular-position sensor (barber pole)



UAE0641-IE

Simple AMR travel sensor for the mm range

Similarly, the simplest travel sensors may be built for a measuring range of typically just a few millimeters using individual differential barber pole sensors (Fig. 28). However, to achieve good accuracy a constant distance (air gap) is required between sensor and moving magnets, which embodies the position s to be measured. The temperature sensitivity of the characteristic curve gradient can be very easily compensated for by means of an additional metallic thin-film sensor on the same carrier substrate, which will have around the same temperature coefficient, but with the inverse sign (e.g. Pt, Ti, Ni).

High-precision multisensor (POMUX) for travel ranges > 10 mm

There is an intelligent multisensor for position measurement built up with simple barber poles also known under the commercial name of POMUX (position multiplexer). It exploits the fact that the symmetry point or zero point of the characteristic curve is absolutely independent of the air gap and the temperature. Its use in a motor vehicle has often been considered (e.g. shock absorber stroke measurement), as it has excellent characteristics, in particular, extreme accuracy. It has also already been illustrated in a circular form as an angular-

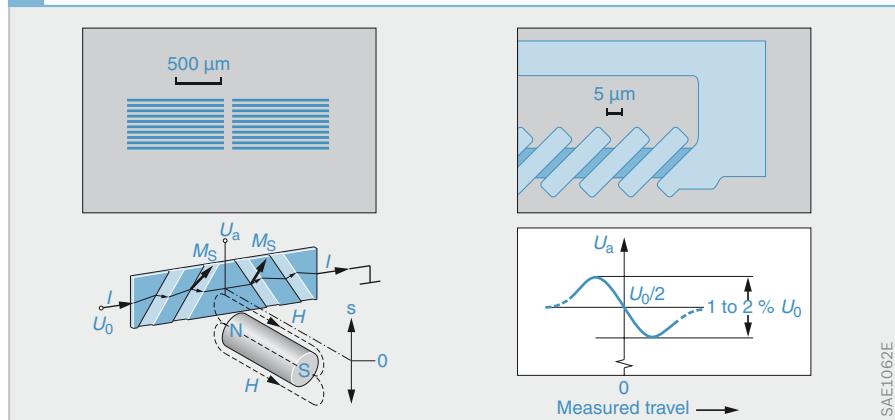
position sensor for scanning a gravity pendulum tuned to a low frequency, which may be used to determine the orientation of the vehicle (tilt sensor).

Angular-position sensor for 180°

Pseudo Hall type magnetoresistive angular-position sensors utilize the practically 100 % sinusoidal signal that is picked off from the output terminals of the four-pole, sensor structure, whereby two complete periods of the electrical output signal correspond to a mechanical rotation of the magnets through 360°. Using a second element, which has been turned through 45°, a cosine signal is also generated (Fig. 30). By applying the arc-tangent function for instance, the relationship between the two signal voltages can be applied to determine the measured angle using a microcontroller or an ASIC. This applies throughout a range of 180°, and is for the most part independent of temperature changes and magnetic-field fluctuations (distance, aging).

A further prerequisite for the high accuracy of this sensor principle is that the field at both bridges is at least in the same direction (above a given magnitude, field strength is irrelevant). This can only be guaranteed when both bridges are directly above one another. A design was drawn up

28 Displacement measurement with single differential barber pole



in which the two bridges, which are at 45° to each other, were interwoven so that they can be regarded as being at the same point and quasi "on top of one another" (Fig. 29).

Fig. 29 Nested AMR bridges

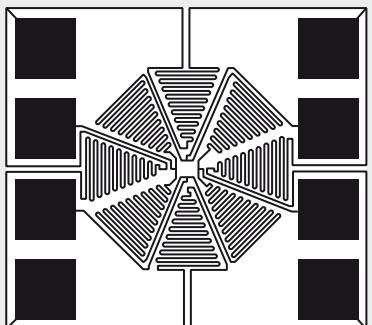


Fig. 29
"Nested" design of
two AMR bridges offset
from each other by 45°

Fig. 30
Measuring range
up to 180°

- a Measuring principle
- b Output signals
- 1 Thin NiFe film (AMR sensor)
- 2 Rotatable permanent magnet with control induction B
- I_y Supply current
- U_{H1}, U_{H2} Measurement voltages
- φ Angle of rotation

Fig. 31
a Bridge circuit

- b Output signals from the Wheatstone bridges
- c Influence of the auxiliary field on the signals
- d Output signal from the evaluation circuit

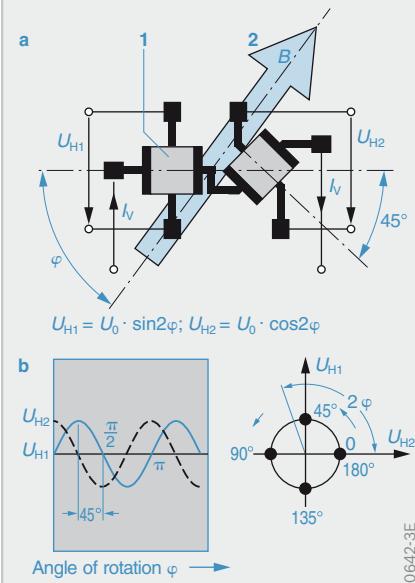
B_H Control induction

U_{DD} , U_{SS} Supply voltages

U_{SIN}, U_{COS} Measurement voltages

φ Angle of rotation

Fig. 30 Magnetoresistive angular-position sensor (pseudo Hall)

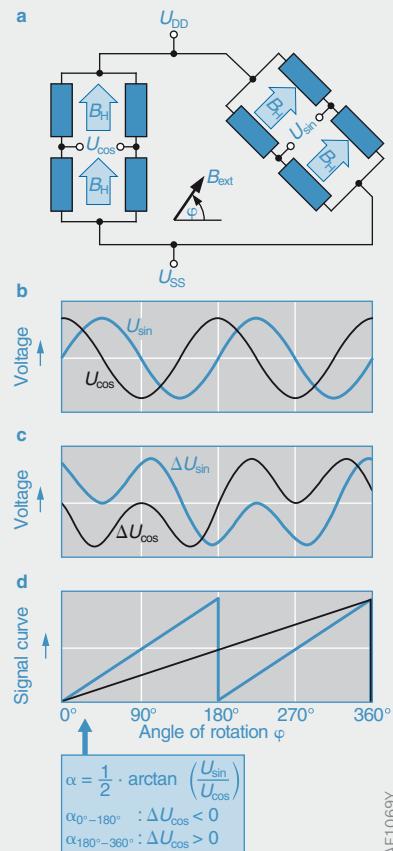


UAE0642-3E

Angular-position sensor for 360°

One serious drawback with AMR angular-position sensors is their natural limitation to an unambiguously detectable angular range of 0° to 180° . This restriction may, however, be overcome by modulation of the direction of the magnetic flux by means of an alternating auxiliary magnetic field (Fig. 31). By contrast with the COS and SIN sensor signals which have a 180° periodicity, these modulation signals have a 360° periodicity. In this way it is possible, simply by evaluating the sign of the modulation signals, to distinguish between

31 AMR360 angular-position sensor with field direction modulation



$$\alpha = \frac{1}{2} \cdot \arctan \left(\frac{U_{\sin}}{U_{\cos}} \right)$$

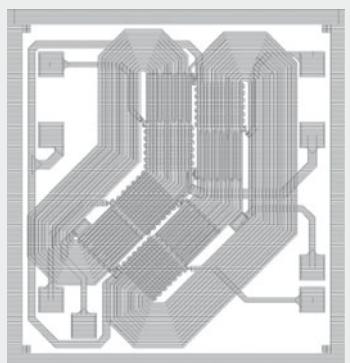
$$\alpha_{0^\circ-180^\circ} : \Delta U_{\cos} < 0$$

$$\alpha_{180^\circ-360^\circ} : \Delta U_{\cos} > 0$$

UAE1069Y

ranges and thus to double the unambiguous range to 0° to 360° . Because the auxiliary field for the two sensor bridges points in different directions, the two modulation signals are phase-shifted. It is sufficient, to distinguish ranges, merely to determine the sign of the largest modulation signal in respect of absolute value. This means that distinguishing the ranges is not critical and is still easily possible, even in the case of modulation signals suffering heavy interference.

32 Layout of the AMR360 angular-position sensor



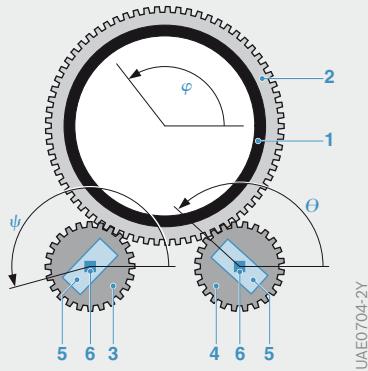
The magnetic auxiliary field is generated by a planar coil integrated into the AMR sensor chip (Fig. 32). The auxiliary field in the COS and SIN full bridge points in directions 45° apart in accordance with the wiring arrangement.

The sensor can only be mounted on the end of a rotating shaft. Thus, for instance, it may be mounted on the end of the steering spindle for measuring the steering-wheel angle in systems that do not require determination of multiple rotations of the steering spindle (e.g. steering-angle sensor LWS4).

Angular-position sensor in the range over 360° (multiple revolutions)

A dual-configuration of pseudo-Hall angular-position sensors (180°) can be used to measure multiple rotations of a rotating component, a steering spindle, for instance. The two associated permanent magnets are rotated via a step-up gear train (Fig. 33). As the two driving smaller gear wheels, which carry the control magnets, differ by one tooth (number of teeth m and $m+1$), their mutual phase difference (difference of the angle of rotation: $\Psi - \Theta$) is an unambiguous measure of the absolute angle position φ of the steering spindle. The system is designed such that this phase difference does not exceed 360° in a total of four revolutions of the steering spindle and so the unambiguity of the measurement is ensured. Each individual sensor also offers an indeterminate fine resolution of the angle of rotation. Using such a configuration, it is possible for instance to determine the complete steering-wheel range with a resolution of better than 1° (this principle is applied in the LWS3 steering-angle sensor).

33 Arrangement for angular measurement in excess of 360°



Giant magnetoresistive sensors (GMR)

GMR sensor technology was developed a few years ago and is now finding its first applications in sensing angles and speeds of rotation in the automotive sector.

Fig. 33

- 1 Steering axle
- 2 Ring gear
- 3 Gear wheel with m teeth
- 4 Gear wheel with $m+1$ teeth
- 5 Magnets
- 6 AMR sensor elements
- φ Angle of rotation of the steering column
- Ψ, Θ Angle of rotation of the small gear wheels

The essential advantages of GMR sensors over AMR sensors are the natural 360° range of unambiguity in angle sensing and the higher sensitivity to magnetic fields in speed of rotation sensing.

Unlike AMR sensors, GMR sensors consist not of one magnetic function layer, but rather of a complex system of layers. A distinction is essentially made between two systems: on the one hand the GMR multilayers with typically some 20 sequences of alternating soft magnetic and nonmagnetic individual layers (Fig. 34a). On the other hand there are the GMR spin valves, comprising antiferromagnetic, ferromagnetic and nonmagnetic functional layers. The individual layer thickness are in the 1 to 5 nm range in both systems, and so embrace just a few atomic layers.

The electrical resistance of a GMR layer stack depends on the angle between the magnetizations of adjacent ferromagnetic individual layers. It is at its maximum when they are aligned antiparallel and at their minimum when in parallel alignment (Fig. 34b).

Fig. 34

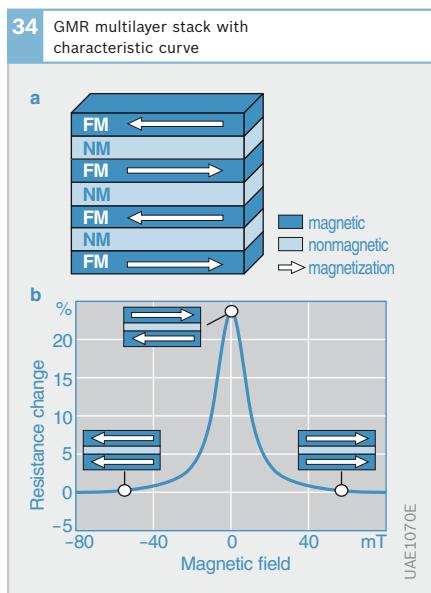
- a Design
- b Change in resistance as a function of the magnetic field

FM Soft magnetic single layer
NM Nonmagnetic single layer

Fig. 35

- a Design
- b Change in resistance as a function of the magnetization angle

FM1 Antiferromagnetic single layer
FM2 Ferromagnetic single layer
NM Nonmagnetic single layer

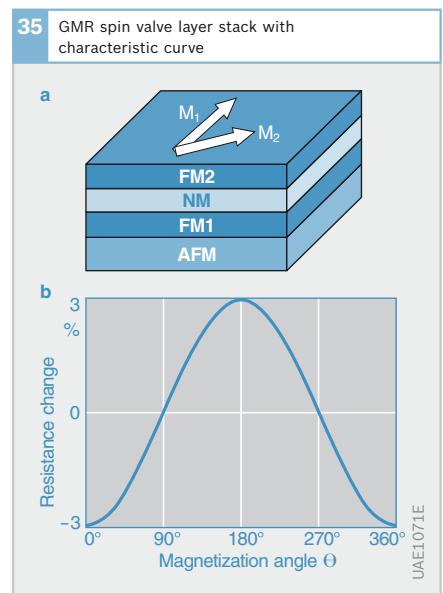


GMR multilayers

The relative change in resistance (GMR effect) is 20 to 30 % in GMR multilayers and thus is ten times greater than the AMR effect. A hard magnetic layer, which will provide a suitable bias field, is added to the GMR multilayer stack to place the working point in a sensitive region of the characteristic curve. The sensitivity of a sensor structure of this nature is about 0.8 %/mT which makes it greater than that of AMR barber pole structures.

GMR spin valves

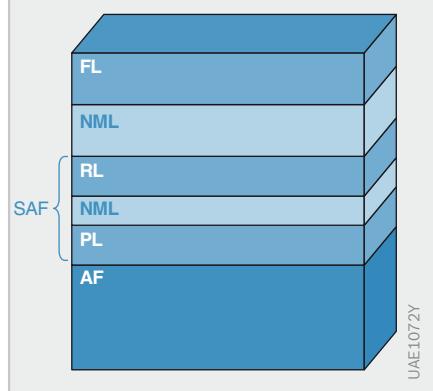
Unlike the GMR multilayers, GMR spin valves (Fig. 35) are also suitable for angle sensing. In the latter, the reference magnetization necessary for angle sensing is generated, fixing (pinning) the direction of magnetization of one of the ferromagnetic layers (FM1) through interaction with an adjacent antiferromagnetic (AFM) layer. This is hence also known as the “pinned layer”. On the other hand, the magnetization of the second ferromagnetic layer (FM2), which is largely magnetically de-



coupled by a nonmagnetic interlayer, is freely rotatable using the external magnetic field. This is accordingly known as the “free layer”.

The resistance varies with a cosine-wave dependency on the angle φ between the external field direction and the reference direction. The stability of the reference magnetization against the action of the external field is decisive for the accuracy of the angular measurement. This stability is significantly enhanced by the use of an additional synthetic antiferromagnet (SAF) (Fig. 36). This consists of two ferromagnetic layers which are strongly antiferromagnetically coupled through a very thin nonmagnetic interlayer and, because of their vanishing magnetic net moment, no longer offer a face for the external field to attack. The magnetization of one of these two layers is pinned by the adjacent natural antiferromagnet (AFM). Both the interaction of the pinned layer (PL) with the natural antiferromagnet (AFM) and the mutual interaction of the ferromagnetic layers (PL, RL, FL) are determinant for the magnetic behavior of the layer system. Furthermore, the behavior is also formed by the shape and crystal anisotropy of the individual magnetic layers, as is the case with the AMR sensor structures.

36 GMR spin valve with synthetic antiferromagnets



Like an AMR angular-position sensor, a GMR angular-position sensor consists of two full bridges, of which one supplies a cosine-wave signal and the other a sine signal depending on the external field direction (Fig. 37). The different magnetic reference directions M_R required to achieve this are generated by local heating of the individual bridge resistors and cooling in the presence of a magnetic field of suitable orientation. The field direction can be unambiguously determined over the full angular range from 0° to 360° by an arctan function.

37 GMR angular-position sensor with two full bridges

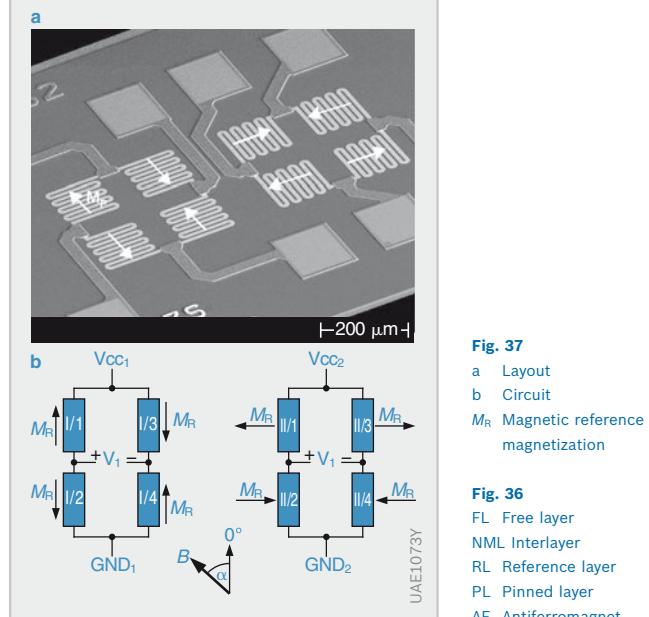


Fig. 37
a Layout
b Circuit
 M_R Magnetic reference magnetization

Fig. 36
FL Free layer
NML Interlayer
RL Reference layer
PL Pinned layer
AF Antiferromagnet

Wave-propagation sensors

In recent years, the development of sensors has concentrated increasingly on sensors that acquire information from the immediate and more distant surroundings of the vehicle. This refers to the distance to other vehicles or traffic and to obstacles (all-round view from the vehicle, Fig. 38), either for safety reasons or to support vehicle handling.

Ultrasound propagation time methods are suitable for vehicle-to-vehicle distance monitoring in the near range (0.5 to 5 m), with optical propagation time or triangulation methods using short-wave infrared light (lidar) for the medium range up to 50 m, and electromagnetic radar for the long range (up to 150 m).

Ultrasound sensors

According to the echo-sounding process (Fig. 39), the sensors transmit ultrasound pulses with a frequency of around 40 kHz, and record the time taken for the echo pulses to arrive having been reflected back from an obstacle. The distance l from the send/receive head to the closest obstacle is determined from the propagation time t of the first echo pulse to arrive and the velocity of sound c thus:

$$(8) \quad l = 0.5 \cdot t \cdot c$$

In practice, the geometrical distance a between an obstacle and the front of the vehicle is determined using the triangulation method from the results measured (distance b and c) from two ultrasound heads, mounted at a distance d from one another (Fig. 40):

Long-range radar sensors (76 GHz)

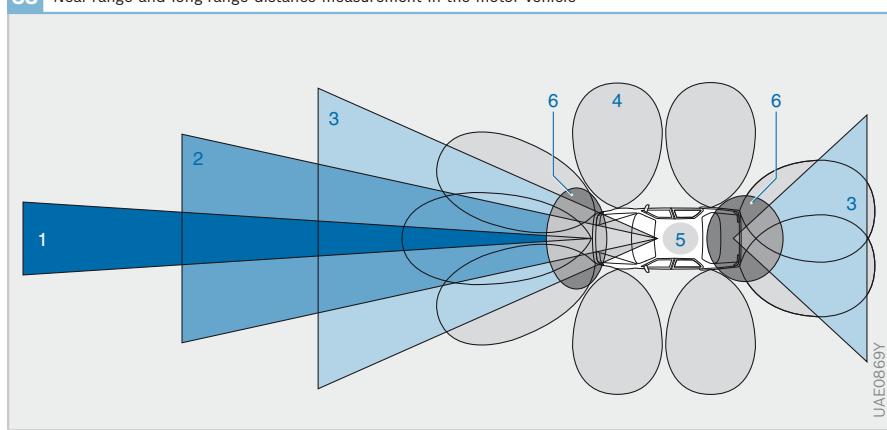
The vehicle-speed controller fitted is equipped with a long-range radar sensor (adaptive cruise control, ACC) for the automatic detection of vehicles preceding in the same lane or even of obstacles in the path of the vehicle which could eventually necessitate application of the brakes. The working frequency is 76 GHz (wavelength $\lambda \approx 3.8$ mm) permits the relatively compact construction as needed for automotive applications. A Gunn oscillator (Gunn diode in the cavity resonator) feeds in parallel four adjacently arranged patch antennas, for example, which also serve to receive the reflected signals. Referred to the vehicle axis, a plastic Fresnel lens in front concentrates the transmit and receive beam horizontally and vertically.

The signal reflected on the vehicle in front is delayed in accordance with the propagation time. The distance a to the vehicle in front and its relative speed Δv

Fig. 38

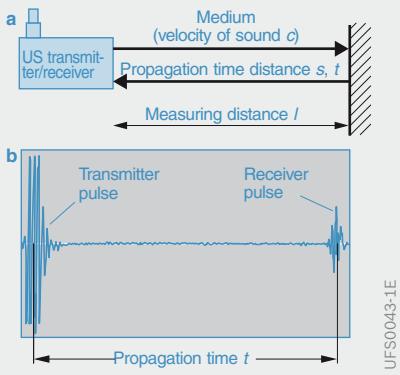
- 1 Long-range radar (77 GHz)
- 2 Long-range/near-range infrared viewer (night vision)
- 3 Video coverage of the vehicle's immediate vicinity (night vision, mid-range < 40 m)
- 4 Near-range radar 24 GHz (near range < 10m)
- 5 Passenger-compartment video
- 6 Ultrasound (near range < 1.5 m)

38 Near-range and long-range distance measurement in the motor vehicle

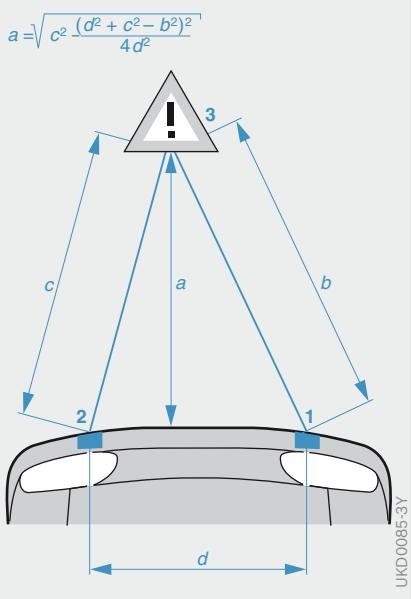


can be determined from the signal propagation time and/or the phase displacement between the signal transmitted and the echo received. The antenna transmit/receive characteristics are aligned in four different directions due to the antennas

39 Principle of distance measurement using ultrasound



40 Calculation of the distance (triangulation)



being off-set from the center. The angle of the vehicle in front can be determined from this (see "Adaptive cruise control" chapter).

GPS position and distance traveled measurement

GPS is the global positioning system using satellites, originally created by the USA for military purposes, but later increasingly released for civil uses. The system was completed in 1993 with a total of 24 satellites (21 operating, three in reserve) orbiting at an altitude of 20,183 km. More than 24 satellites have since become available. They are distributed over six circular orbital planes taking 12 hours to complete an orbit and so at least four (but generally up to eight) satellites are always visible above the horizon from any point on the earth's surface. They transmit continuous (digital) signals on a carrier frequency of 1.57542 GHz. The signals contain the following message, in simple terms:

- ▶ Satellite identification code
- ▶ Satellite position
- ▶ Message transmit time (including date)

The satellites each have two cesium and two rubidium clocks which vary by less than 20 to 30 ns on board for the high-pre-

Fig. 40

- a Distance between the fender and the obstacle
- b Distance sensor 1 to obstacle
- c Distance sensor 2 to obstacle
- d Distance sensor 1 to sensor 1
- 1 Transmit/receive sensor
- 2 Receive sensor
- 3 Obstacle

41 Position determination by GPS

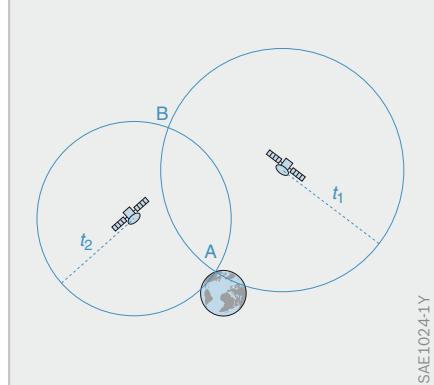


Fig. 41

Representation in the plane (two-dimensional): If the position of the satellites is known, with the propagation times t_1 and t_2 measured, the possible reception locations are on two circles around the satellites which intersect at points A and B. The point A located on the earth's surface is the location of the receiver.

cision determination of the transmission time. The signal propagation times to be evaluated lie typically in the region of approximately 70 ms. The satellites identify themselves with a pseudo random code (PRN number), which is 1,023 bits long and is continuously repeated after every 1 ms. It is imposed on the carrier as a phase modulation.

A user can take this information and determine his own geographical position in three dimensions with a GPS receiver, consisting of a GPS antenna, signal receiver, precision clock and microprocessor. Theoretically, the signals from three satellites are sufficient for this; since, however, the less complex clocks in mobile receivers (oscillating crystals) can differ from the satellite clocks somewhat, a fourth satellite must be used in order to eliminate the errors generated by the receiver clock. The distance to the visible satellites is derived from the signal propagation times measured.

The position is determined - as shown in Figure 40 in only two dimensions of space - then using a trilateration method, according to which there is precisely one point in space which satisfies the three dimensional conditions. In addition, there is also the knowledge that the motor vehicle, unlike an airplane, must be located on the earth's surface. Measurement is thus the more accurate the more satellites can be received and the better, i.e. the more uniformly the satellites are distributed across

the sky. Since complete release for civil use (May 2000), the accuracies obtainable are about 3 to 5 m in a plane and around 10 to 20 m in the determination of altitude. With the refinement of GPS to DGPS (Differential GPS, mobile stations have contact with a fixed reference station via long wave signals) this variance can be significantly further reduced. This makes resolutions down to less than 1 m possible today.

These accuracies are the more astonishing when the manifold sources of error in this very complex measurement is taken into consideration. For example, simply the fact that at the very high velocity of approximately 12,000 km/h at which the satellites move, relativistic effects already play a not insignificant role. Time passes more slowly at high velocities than it does on earth. Time passes more quickly in weaker gravitational fields as are found in the orbital planes. The high velocity causes a dilation of time of 7.2 μ s per day, the lower gravitation even has an effect that is around six times greater, although opposing.

If the position data determined in a close sequence is stored in a mobile GPS receiver, such as in a motor vehicle, it is easy to derive the orbit and the speed of the vehicle. If the data is sent via a radio module to a fixed central location, the location of the vehicle at any time will then be known.

Speed and rpm sensors

Measured variables

Speed and rpm sensors measure the number of revolutions or the distance traveled per unit of time. When automotive applications are concerned, these are in both cases measured variables which occur between two components or with respect to the road surface or another vehicle. However, it is also necessary to measure the absolute rotational speed in space or about the vehicle axes (yaw rate). For instance, for the electronic stability program, the yaw rate about the vehicle's vertical (or yaw) axis must be determined by sensing.

Measuring principles

Conventional sensors used for rotational-speed measurement are based on pronounced measuring effects (e.g. inductive). They are therefore for the most part electrically passive. That is, they are usually not provided with any form of local/on-site electronics. With the newer sensors, however, measurement is based on less-pronounced measuring effects (for instance, the Hall effect), and these sensors thus need local, integrated electronics for signal conditioning. They belong in the broadest sense to the category of "intelligent sensors" (which are also often referred to as "active" sensors).

In fact, the sensors used for measuring absolute rotating speed (yaw rate) need highly complex electronic circuitry directly at the sensor since the measuring effects used here are not only particularly small, but also require complex signal conditioning.

Incremental rotational-speed measurement takes advantage of a wide variety of different physical effects (some of which can be applied in sensors at a very reasonably price). Optical and capacitive sensors, though, are highly unsuitable for the rough operating conditions encountered in the motor vehicle. Here, magnetic-effect sensors are used almost exclusively.

Relative rpm and speed measurement

Examples of relative rotational speed are as follows:

- ▶ Crankshaft and camshaft rotational speed
- ▶ Wheel speeds (for ABS/TCS/ESP)
- ▶ Rotational speed of the diesel fuel-injection pump

Here, measurement usually makes use of an incremental pick-up system comprised of a gear wheel and a speed sensor.

In the detection of relative yaw rate, depending upon the number and size of the scanned rotational rotor markings, a distinction is made between the following types of sensor (Fig. 1):

- ▶ Increment sensor with closely spaced rotational markings. Up to a certain point, this form of sensor permits instantaneous speed to be measured at points on the circumference, or the registration of very fine angular divisions.
- ▶ Segment sensor, with only a small number of scanned rotational segments (for instance, equivalent to the number of engine cylinders).
- ▶ Simple speed sensor, with only a single scanned marking per revolution, so that only the average rotational speed can be registered.

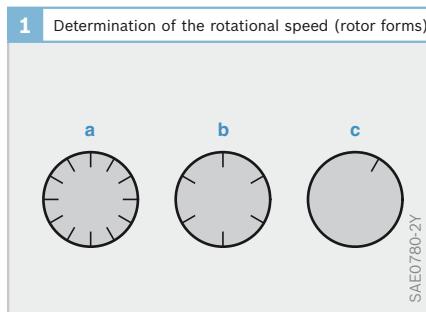


Fig. 1

- a Incremental sensor
- b Segment sensor
- c Speed sensor

Sensor shapes

The following shapes of sensor are presently in use (Fig. 2) for speed and rpm sensors:

- ▶ Rod sensor
- ▶ Fork-shaped sensor
- ▶ (Internal and external) toroidal sensor

The preferred form is the rod sensor, in which the rotor teeth approach the sensor and which is also the easiest to install. The fork-shaped sensor is permissible in some cases, and is also in service in the field. This sensor must be roughly aligned to the rotor when installed. The sensor type in which the sensor surrounds the rotor shaft in the form of a ring is practically no longer used. From time to time, an inner-ring sensor is used which is introduced into the end of a hollow shaft featuring an incremental internal structure.

Unfortunately, although it is the most widely used sensor type, the rod-shaped sensor features the lowest measurement sensitivity and is problematical when air gaps become excessive. The fork-shaped sensor on the other hand is practically impervious to axial and radial play. Regarding toroidal sensors, the most widely used form combines a large measurement signal with pronounced insensitivity to geometric tolerances.

In many respects, the traditional inductive-type sensors are highly unsatisfactory. Their output signal features an rpm-dependent amplitude, and for this reason they are unsuitable for measuring low rotational speeds. In addition, they only permit comparatively low-level air-gap tolerances and are usually unable to differentiate between air-gap fluctuations and genuine rotational-speed pulses. And at least the sensor tip must resist high temperatures when installed near hot components (e.g. the brakes). These disadvantages are behind the additional features aimed at with new, innovative sensors:

- ▶ Static detection (that is, at zero speed, or at extremely low cranking or wheel speeds)
- ▶ Larger air gap (non-aligned mounting with air gaps > 0)
- ▶ Small size
- ▶ Independence from air-gap fluctuations
- ▶ Temperature stability ($\leq 200^\circ\text{C}$)
- ▶ Detection of the sense of direction (optional for navigation)
- ▶ Reference-mark identification (ignition)

Magnetostatic sensors (Hall, magnetotransistors, AMR) are highly suitable for complying with the first two demands. And, as a rule, they also permit compliance with the second and third stipulations.

Figure 3 shows three fundamentally suitable sensor shapes, which generally are insensitive to air-gap fluctuations. Here, one must differentiate between sensors which sense radially and those which sense tangentially. This means that independent of the air gap, magnetostatic sensors are always able to differentiate between the north and south poles of a magnetically active pole wheel or rotor ring.

In the case of magnetically passive rotors, the sign of the output signal is then no longer independent of the air gap when they register the tangential-field strength (here though, the fact that the air gap is

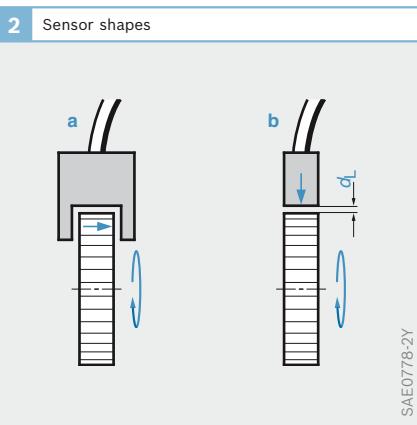


Fig. 2

- a Fork shape (vane principle)
- b Rod shape (proximity principle)
- d Air gap

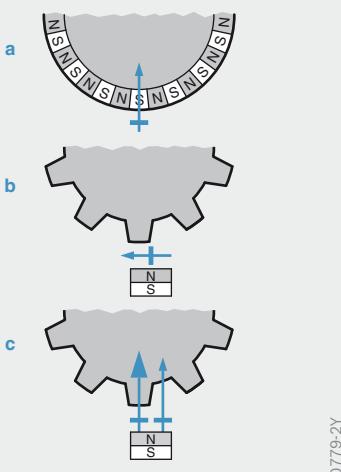
often enlarged due to the sensor is a disadvantage).

Radially measuring differential-field or gradient sensors are often used. These always register only the gradients of the radial-field components, the signs of which do not change with the air gap but only with the angle of rotation.

Rotors

The rotor is of decisive importance when measuring rotational speed. It is usually provided by the vehicle manufacturer, while the sensor itself comes from a component supplier. Previously, magnetically passive rotors have been used almost exclusively. These are made from magnetically soft materials (usually iron), and are less expensive than magnetically hard rotors (also known as pole wheels). Apart from that, since they are not magnetized they are easier to handle, and there is no danger of mutual demagnetization (for instance during storage). Unfortunately, this form of rotor is the most difficult to scan, particularly in combination with rod sensors.

3 Sensors insensitive to air-gap fluctuations



As a rule, presuming the same increment width and output signal, the pole wheel's intrinsic magnetism (a pole wheel is defined as a magnetically active rotor) permits a considerably larger air gap.

Normally, passive rotors are in the form of gear wheels. In many cases, these are already present on the engine (for instance the starter ring gear). Otherwise they must be installed in a specific position in order to generate the required signal (as needed for instance for ABS). In the latter case, both planar-toothed and axial-scan versions are in use.

An easily identifiable reference mark is required for each revolution when pick-off the crankshaft speed or position (for instance, at the starter ring gear). This reference mark must allow a fault-free and optimum timing of ignition and injection. The reference mark can be in the form of a completely (or partially) removed tooth. Due to the fact that a tooth space "takes more time", the reference mark is immediately identifiable, particularly since the engine speed can only vary gradually and never abruptly.

In addition to gear wheels, stamped perforated disks or wave-shaped metal rings have been introduced in the meantime as low-priced rotors (ABS).

It was the integration of ABS sensors in the vehicle's wheel bearing which led to the introduction of pole wheels some of which also assume the role of a sealing ring (plastic-bound magnetic powder). Small, for the most part encapsulated, tachometer sensors connected through a short flexible shaft with one of the vehicle's wheels, also use pole wheels (with only very few poles) for generating a speed signal. These are usually picked off by means of integrated Hall sensors.

Fig. 3

- a Radial-field sensor with pole wheel
- b Tangential sensor
- c Differential sensor with gear wheel

Inductive-type sensors

Inductive-type sensors were available on the market as coil versions long before the first suitable microstructure sensor versions (e.g. using the Hall principle) became available. Such inductive-type sensors use Faraday's law when measuring rotational speed. In other words, they generate a voltage U_A at their two-pole output which is proportional to the change (referred to time) of a magnetic flux Φ (w = number of turns).

$$U_A = U_{\text{ind}} = w \cdot d\Phi/dt$$

The magnetic flux Φ is also a function of the rotational position x and the air gap d_L : where

$$\Phi = \Phi(x, d_L) \text{ and } d_L = \text{const}$$

the following applies:

$$U_A = U_{\text{ind}} = w \cdot \partial\Phi/\partial x \cdot dx/dt$$

where dx/dt represents the (rotational) speed to be measured.

The inductive-type sensors' weak point is underlined though by the equation: if it is impossible to keep the air gap d_L constant (due to flutter or other forms of mechanical play), the air-gap fluctuations induce the same change of flux as does a fluctuation of speed. This effect can cause the

generation of voltage pulses which are either impossible, or at least difficult, to separate from the genuine engine-speed signals. Since the flux varies exponentially along with the change in air gap, and the air-gap fluctuations are often of the high-frequency type (e.g. brake flutter), these unwanted pulses can feature a high voltage amplitude.

Inductive-type sensors are therefore always of the dynamic type, and being as their output signal tends to fall to zero in such cases, they are in principle also unsuitable for the detection of extremely low speeds (quasi-static or static). The only exception here are the coil sensors which are powered by a carrier frequency and which are based on the eddy-current or damping principle. These though are hardly ever used in automotive applications.

Basically, inductive speed sensors are comprised of three important magnetic component parts (Fig. 4):

- ▶ Fixed coil
- ▶ Soft-iron component
- ▶ Permanent-magnet component

The change in flux needed for the generation of the output voltage results from the motion or rotation of the hard or soft-iron magnetic component. The inductive-type sensors are preferably composed of a rod magnet (1) with a soft-magnetic pole pin (2) carrying the 2-connection induction coil (3). When a ferromagnetic gear wheel (5), or some form of similar rotor, rotates in front of this sensor (pick-up) the changing magnetic flux (which varies as a function of time) induces a practically sinusoidal voltage.

For scanning very fine tooth structures, the end of the pole pin is sometimes pointed and acts as a flux-concentrating element. In other words it is shaped like a pole "blade" which usually projects through the metal or plastic housing and is adapted to the increment structure regarding shape and direction.

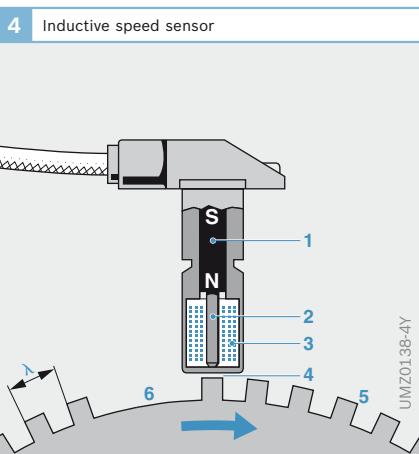


Fig. 4

- 1 Rod magnet
- 2 Soft-magnetic pole pin
- 3 Induction coil
- 4 Air gap d_L
- 5 Ferromagnetic gear-wheel (or rotor or pulse wheel)
- 6 Rotational or reference mark
- λ Tooth interval

The rotor can be provided with one or more rotational markings (6). Figure 5 shows the flux curve and the voltage induced by a single rotational or reference mark (slot, cam, or pole pin).

Normally, the steep passage through zero which takes place at the mid-point of maximum flux is utilized for the electronic

detection of such a rotational or reference mark. According to Faraday's Law, in all phases the signal's amplitude is proportional to rotational speed.

In order to ensure adequate, interference-free evaluation in the control unit, the spacing between the peaks of a double pulse (or of a periodic voltage pulse) U_{SS}

5 Flux and voltage curve for an inductive-type sensor

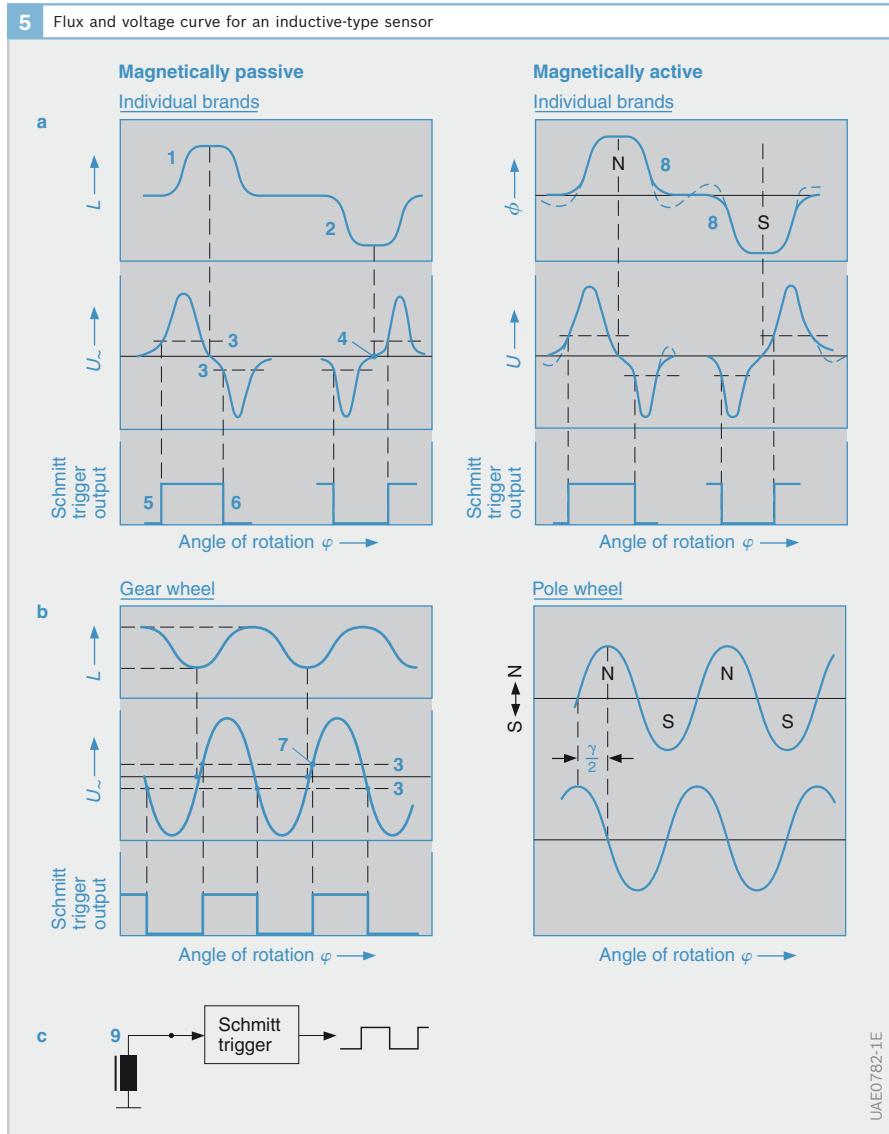


Fig. 5

- a Single magnetic marking per revolution
 - b Periodic increment structure (e.g. gear wheel, pole wheel)
 - c Evaluation circuit
 - 1 Cam
 - 2 Slot
 - 3 Switching point
 - 4 Steep passage through zero evaluated
 - 5 Priming edge
 - 6 Switching flank
 - 7 Shifting point
 - 8 Pole pin
 - 9 Sensor
 - $\gamma/2$ Phase shift
 - L Inductance
 - U Voltage
 - Φ Magnetic flux
- UAE0782-1E

should be at least 30 mV. The major disadvantage of the inductive-type sensors is the fact that at high speeds/rotational speeds their output voltages can reach levels far in excess of 100 V which are difficult to process electronically.

If Zener diodes are used to clip the high voltage peaks, the resulting changes in the sensor's load impedance lead to considerable phase-angle errors. With camshaft and crankshaft sensors this can have highly undesirable results with regard to the ignition where the correctness of phase relationship must be better than approx. 0.2°.

Normally, the prepulse generated by the magnetic return field can be ignored at low speeds. With some magnetically passive or active rotational markings though, at high speeds the prepulse voltage can increase to such an extent that it exceeds the threshold value of the downstream threshold discriminator and can cause an even greater error (Fig. 5a). For this reason, the threshold values of the control unit input circuit are adapted dynamically to the speed in question.

Provided that the air gap is not too narrow, a uniform tooth structure results in the practically sinusoidal voltage curve shown in Figure 5b. The speed is reflected in the periodic interval between the voltage's zero transition points, while the amplitude is also proportional to the rotating speed.

The signal amplitude is highly dependent (exponentially) upon air gap and tooth size. As is the case with all magnetic increment processes, up to air gaps of d_L , teeth can be efficiently detected as from half or 1/3 of a tooth interval I.

$$d_L \leq \lambda/(2...3)$$

The conventional ABS gear wheels and crankshaft applications cover air gaps of up to 0.8 and 1.5 mm respectively. The reference mark needed for the ignition and injection results from leaving out one or two teeth or by closing a tooth space.

The reference point is detected when the distance between the passages through zero changes abruptly and causes a far higher signal voltage (corresponding to an apparently larger tooth) which has a negative effect upon the previous and upon the subsequent incremental voltage - this can under certain circumstances be undesirable.

Advantages

- ▶ Low manufacturing costs
- ▶ High immunity to interference: low static internal resistance (dynamically higher)
- ▶ No local electronics (electrically passive), that need to be protected
- ▶ No problems with DC voltage drift (dynamic measuring principle)
- ▶ Broad temperature range (limited primarily by the sealing compound characteristics)

Disadvantages

- ▶ Conventional coil technology imposes limits on size reduction
- ▶ The output signal is rpm-dependent, unsuitable for quasi-static movements
- ▶ Sensitive to air-gap fluctuations

Examples of application

- ▶ Inductive engine-speed sensor (crank-shaft-speed sensor)
- ▶ Inductive wheel-speed sensor
- ▶ Inductive camshaft sensor (transistorized ignition with induction-type pulse generator TZ-I)
- ▶ Needle-motion sensor (diesel-fuel injection)

Magnetostatic sensors

Quasi-static speed measurement is best performed using magnetostatic sensors. Their output signal is independent of rotational speed and depends solely on field strength, so that at high rotational speeds their low signal voltages are easier to handle electronically. Furthermore, they not only have the advantage of imposing practically no limits with regard to miniaturization, but also signal amplification and/or signal preprocessing can be integrated locally. The fact that they are very small means that multiple systems such as differential configurations, or arrangements with integrated recognition of direction, present no difficulties.

On the other hand, such active sensors have a serious disadvantage, and that is the fact that their operating-temperature range is for the most part defined by the relevant Si (silicon) evaluation electronics which as a rule cannot withstand such high temperatures as the sensor element itself.

For some time now, active sensors have been available with the option of a two-pole current output, so that the two-core connection can no longer be regarded as a specific advantage of the inductive-type coil sensor.

Hall vane switches

If Hall Si sensors are to be used for incremental rpm measurement, pronounced manufacturing scatter together with the effects of temperature mean that they must be provided with an adequate induction jump of typically 40 to 50 mT in order to ensure reliable high-speed switching. With conventional Hall sensors and acceptable air gaps, this was only possible with the sensor in the form of a "Hall vane switch" (for instance when used as ignition-triggering sensors in the ignition distributor for earlier electronic ignition systems). The sensor and the electronic circuits for supply and signal evaluation are integrated on the sensor chip.

This "Hall IC" (using bipolar technology, for sustained temperatures $\leq 150^\circ\text{C}$, and for direct connection into the vehicle's onboard electrical supply), is located in a practically closed magnetic circuit comprising permanent magnet and pole pieces (Fig. 6). A soft-magnetic trigger wheel (driven, for instance, by the camshaft) rotates through the remaining air gap. With a trigger-wheel vane in the air gap, the magnetic flux is short-circuited past the sensor element. On the other hand though, when a trigger-wheel opening passes through the air gap the flux is unhindered on its way to the sensor. This principle ensures that the sensor also performs perfectly when the trigger-wheel varies in how far it penetrates into the air gap, or in case the air-gap shifts radially, that is vertically to the direction of rotation.

Since Hall vane switches of this type feature limited rotational resolution they are mainly used as segment sensors. If the vane slot is too narrow, it is practically impossible for the magnetic field to pass through, with the result that the required induction jump is not generated.

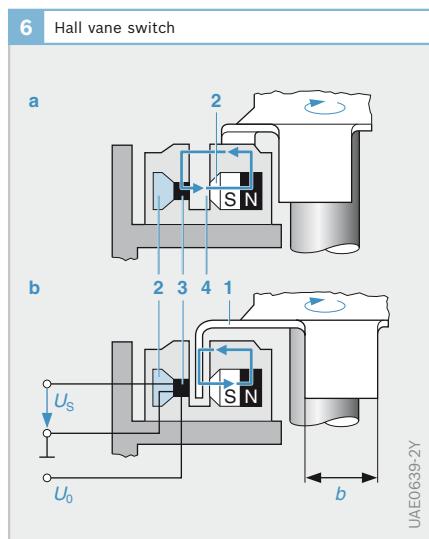


Fig. 6

- a Magnetic flux: unhindered
- b Magnetic flux: short-circuited
- 1 Vane width b
- 2 Soft-magnetic conductive element
- 3 Hall IC
- 4 Air gap
- U_0 Supply voltage
- U_s Sensor voltage

Simple Hall rod sensors

In contrast to the Hall vane switch, the working-point magnetization depends too much upon the width of the air gap, and the induction excursion is too small for reliable switching with this configuration. Simple rod-shaped Hall sensors provided with a working-point magnet are therefore unsuitable for static or quasi-static scanning of a magnetically passive rotor (Fig. 7a, gear wheel). The switching point of a downstream threshold comparator (Schmitt trigger) would have to be continually adapted to the varying working point. Such applications are only feasible when DC coupling is dispensed with and therefore also static signal evaluation. The coupling capacitors required for such configurations (also known as $\Delta\Phi$ sensors) involve high costs though and lead to reduced operational reliability.

However, simple Hall sensors are highly suitable for scanning a magnetically active rotor (Fig. 7b, pole wheel). The sensor is triggered by the pole wheel with varying polarity only in the vicinity of the magnetic zero point. Although the magnetic control excursion reduces along with increasing air gap, the position of the working point

($B = 0$) remains unchanged. Considering that the working point on new types of Hall sensors is to a great extent thermally stable, the switching points of the downstream threshold comparator can be set relatively close together. This means that relatively wide air gaps become possible. Air-gap fluctuations in this configuration cannot cause missing pulses as they do not lead to pole change. The pole change only characterizes the continuing measurement motion (rotation).

Gradient sensors

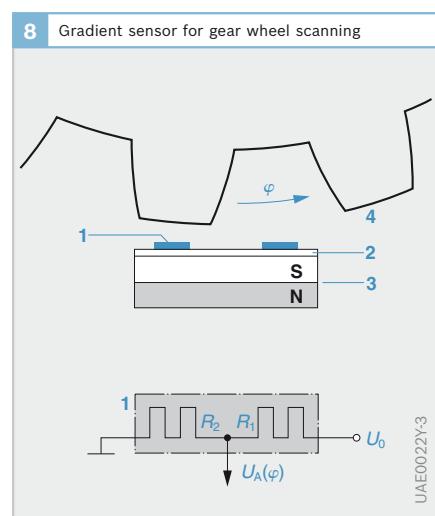
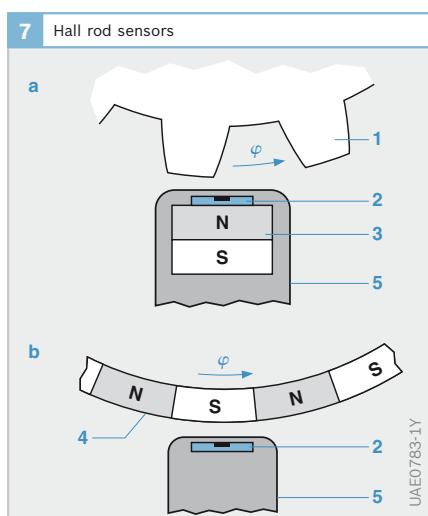
When it comes to scanning passive rotors, gradient sensors (Fig. 8) designed on the basis of differential Hall sensors or differential magnetoresistive sensors, are far better than simple Hall sensors. They are provided with a permanent magnet. This pole face of this magnet facing the gear wheel is homogenized by a thin ferromagnetic wafer (2). Two galvanomagnetic elements (generic term for Hall sensors and magnetoresistors) are located on each element's sensor tip, at a distance of roughly one half a tooth. Thus, one of the elements is always opposite a tooth space when the other is opposite a tooth. The sensor mea-

Fig. 7

- a Passive rotor
- b Active rotor
- 1 Incremental rotor
- 2 Simple Hall sensor
- 3 Permanent magnet
- 4 Pole wheel
- 5 Housing

Fig. 8

- 1 Magnetoresistors
 R_1, R_2 or Hall elements H_1, H_2
- 2 Ferromagnetic wafer
(soft-magnetic substrate)
- 3 Permanent magnet
- 4 Gear wheel
- U_0 Supply voltage
- $U_A(\varphi)$ Measuring voltage for an angle of rotation φ



sures the difference in field strength between two neighboring points on the circumference of the rotor. The output signal corresponds roughly to the field strength derived as a function of the angle at the circumference, and its sign is therefore independent of the air gap. Being as they do not alter the gradient signal's sign, air-gap fluctuations does not cause missing pulses.

For signal evaluation, it is a simple matter to connect the two magnetoresistors as a voltage divider which is supplied by a constant voltage and whose (usually unloaded) output signal is registered by the control unit. At room temperatures and with the customary air gap, this signal is in the volts range, and even at high temperatures it is suitable for transmission to the control unit without any form of preamplification.

Provided appropriate circuitry is used, measuring the loaded output current of the magnetoresistor divider instead of its open-circuit voltage permits the sensor's pronounced temperature measurement sensitivity to be compensated for to a great extent.

In the case of a gradient sensor based on the Hall effect, the current paths of both Hall elements can be connected in parallel, and their opposite-polarity output voltages in series, so that their differential voltage can be picked off directly and inputted to the downstream amplification and evaluation stages.

Tangential sensors

Tangential sensors differ from gradient sensors by their reaction to variations in polarity and intensity in the components of a magnetic field located tangentially to the periphery of the rotor. Using AMR thin-film techniques, tangential sensors are available as barber-pole or permalloy resistor types in full-bridge or half-bridge circuits (Fig. 9). In contrast to gradient sensors, they need not be matched to the particular tooth pitch of the rotor and can in fact be designed to sense practically at a

given point. Local amplification is necessary, even though their measuring effect is approximately 1 to 2 orders of magnitude larger than that of the silicon Hall sensors.

In the case of the crankshaft speed sensor integrated in the bearing (composite seal with sensor), the AMR thin-film sensor is mounted together with an evaluation IC on a common leadframe.

For the purposes of space saving and temperature protection, the evaluation IC is bent at an angle of 90° and also located further away from the sensor tip.

Giant magnetoresistive (GMR) elements

The giant magneto resistance effect (GMR effect) was first demonstrated on multilayers consisting alternately of thin soft magnetic and nonmagnetic layers (thickness of a few nanometers). An antiparallel alignment of the magnetization of adjacent ferromagnetic layers leads to the maximum electrical resistance, while parallel alignment shows a significantly lower resistance. The relative change in resistance for application-relevant layer structures thus varies in the range 20 to 30 %. This represents an increase over the AMR effect by a factor of around 10.

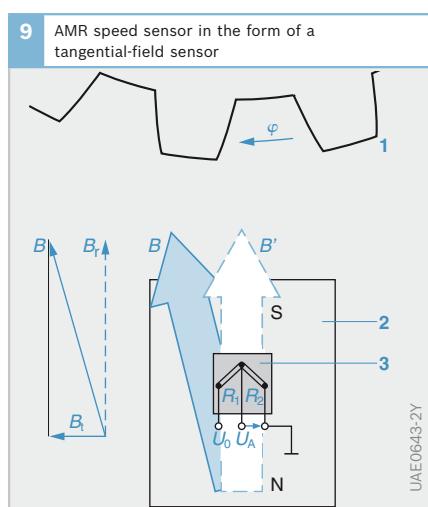


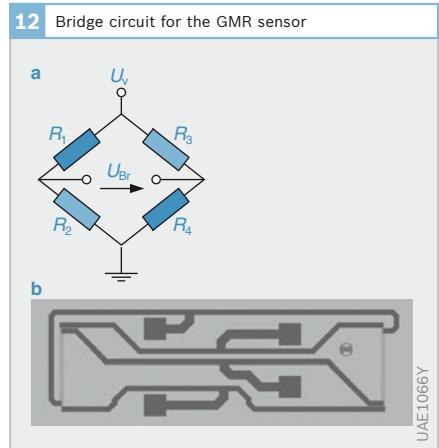
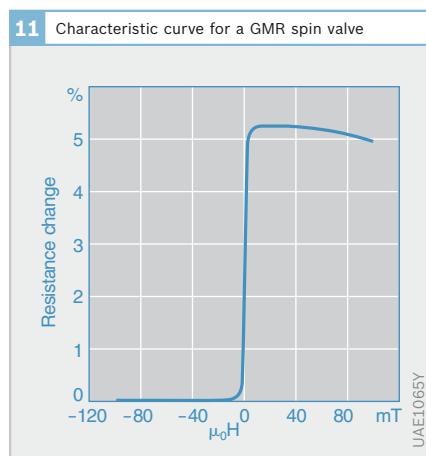
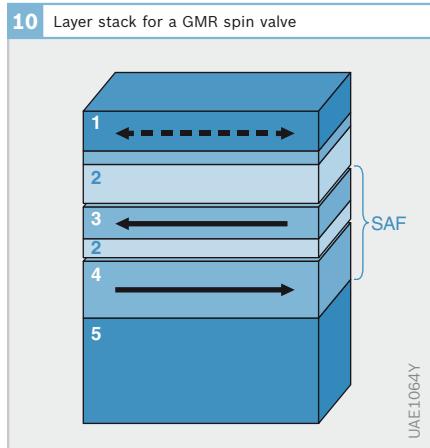
Fig. 9

- 1 Gear wheel (Fe)
- 2 Permanent magnet
- 3 Sensor
- B Control field strength with tangential components B_t and radial component B_r
- B' (B' off position, $B_t = 0$)
- R_1, R_2 permalloy thin-film resistors (AMR)
- φ Angle of rotation
- U_0 Supply voltage
- U_A Measurement voltage

The simplest layer system that shows a GMR effect is a spin valve (see also “Position sensors” section). It consists of two soft magnetic layers, separated by a non-magnetic layer. While the magnetization of the one soft magnetic layer is fixed by an antiferromagnet, the magnetization of the second soft magnetic layer can ideally follow the external field without interference (free layer). The resistance is then minimal if the magnetizations of the two layers are aligned in parallel and is increased by approximately 5 % if the magnetizations are aligned antiparallel. The interaction of the reference magnetization with the external field may be minimized with the help of a synthetic antiferromagnet (SAF), comprising two ferromagnetic layers, that are strongly antiferromagnetically coupled by means of very thin nonmagnetic interlayer (Fig. 10). This significantly increases the robustness against interference fields.

Figure 11 shows a typical characteristic curve for a spin valve. The two characteristic states for the two opposing directions of the external magnetic field can clearly be distinguished. The spin valve is sensitive to field strength changes in the transitional range between high and low resistance.

A bridge circuit of four GMR resistance elements, designed as a gradiometer is suited for the determination of rotational speed (Fig. 12). The resistors are interconnected here in such a way that only one magnetic field difference at the two bridge locations results in a signal. The effects of homogenous interference fields can be suppressed in this way. As the reference direction for all resistors is identical, local heating in the magnetic field, as is necessary to write differing reference magnetization directions in an angular-position sensor, is not required.



The maximum signal amplitude for a gradiometer sensor then results when the direction of the field at the two bridge locations is precisely 180° phase-shifted. This is the case on a multipole wheel, for instance, if the distance between the resistance elements corresponds to the distance between poles in the multipole ring (Fig. 13).

Examples of application

- ▶ Hall sensor (transistorized ignition TZ-H)
- ▶ Hall phase sensor (camshaft)
- ▶ Transmission Hall sensor (RS50, RS51)
- ▶ Active Hall speed sensor
- ▶ Active AMR speed sensor
- ▶ Magnetoresistive sensor (for diesel radial-piston distributor pumps)
- ▶ Active GMR speed sensor

Absolute rotating-speed measurement

Measuring principle of the oscillation gyroscope

Basic principles

Mechanical gyroscopes (gyros for short) utilize inertial forces in precisely measuring angular movements in space independent of the reference system. Notwithstanding their pronounced measuring effect, rotating gyroscopes as well as optical sensors based on the interferometric Sagnac effect (laser and fiber gyroscopes),

are out of the question for automotive applications due to the high costs involved.

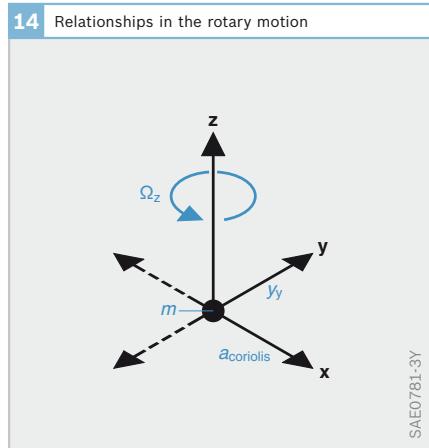
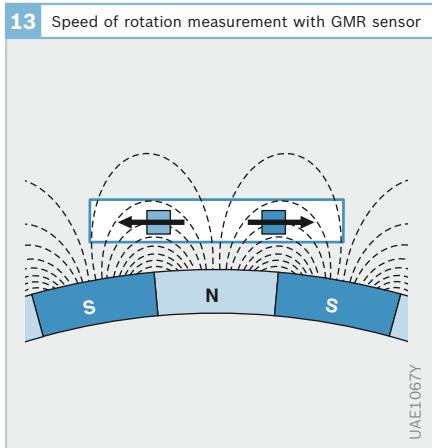
On the other hand, the slightly less-severe requirements on precision for applications in the motor vehicle can be complied with by gyroscopes manufactured using fine-mechanical and micromechanical processes. Instead of a rotational movement, these units use an equivalent, elastic, oscillatory movement to generate the measuring effect. In principle, these sensors are similar to mechanical gyroscopes. They exploit the Coriolis accelerations which are generated when rotary motions are coupled with an oscillatory component (velocity v) (Fig. 14). These sensors are known as oscillation gyroscopes and are used in the electronic stability program (ESP) for the detection of the vehicle breaking away, in navigation systems for the detection of the direction of travel and in cameras to stabilize the image against wobbling. They also comply adequately with all automotive stipulations such as freedom from maintenance, long service life, switch-on time constant etc., not to forget the question of costs.

Oscillation gyroscopes measure the yaw rate around their “sensitive” axis. The yaw rate Ω_z around the vehicle’s vertical axis (yaw axis) is measured in

Fig. 13

Field lines in the space surrounding a multipole ring

The two squares symbolize the two bridge locations
The arrows indicate the direction of the field lines at these points



the electronic stability program (ESP) and in the navigation system. Systems for triggering rollover protection systems need data on the yaw rates Ω_x and Ω_y around the vehicle's pitch and roll axes.

Measurement of the Coriolis acceleration

Oscillation gyroscopes measure the Coriolis acceleration in the following way: as the distance between a mass and the axis of rotation increases, the mass must cover a greater distance in the time of one revolution because of the greater radius. Circumference increases with the radius. The mass must, therefore, acquire a higher tangential speed and must be accelerated to achieve this. This acceleration (Coriolis acceleration) is measured by suspending the mass on springs and measuring the bend in the suspension. If the distance of the mass, now rotating at a higher tangential speed, from the axis of rotation is now reduced, the mass must now become slower, as it needs only cover a shorter distance for each revolution. The braking deceleration required is also measured via the bend of the suspension.

In the oscillation gyroscopes, the distance of one (or more) masses from the axis of rotation is periodically increased and then reduced again by an oscillation excitation (Fig. 14, movement in the y direction). This means that the mass must be accelerated and then braked again in the same cycle (movement in the x direction). The forces necessary for this depend on the amplitude of the exciting oscillation and the current yaw rate. If the exciting oscillation is kept constant, the yaw rate Ω can be determined from the acceleration forces.

The Coriolis force acts in accordance with a known vector law perpendicular to the rotary motion and speed of the mass in motion (Fig. 14). This gives the Coriolis acceleration as:

$$(1) \quad a_{\text{Coriolis}} = a_x = 2 \cdot v_y \cdot x \cdot \Omega_z$$

Whereby, the velocity v_y changes sinusoidally in accordance with the oscillatory movement:

$$(2) \quad v_y = v_y \cdot \sin \omega t$$

This means that, at a constant yaw rate Ω_z , a sinusoidal Coriolis acceleration a_{Coriolis} of the same frequency is measured. The amplitude is then:

$$(3) \quad a_{\text{Coriolis}} = 2 \cdot v_y \cdot \Omega_z$$

The acceleration a_y which is also applied to the mass m in the oscillatory direction is usually several orders of magnitude higher than the useful Coriolis acceleration.

$$(4) \quad a_y = dv_y/dt = \omega \cdot v_y \cdot \cos \omega t$$

As the Coriolis acceleration has the same frequency as the exciting function, the useful signal (the yaw rate) can be obtained by multiplication of the exciting signal and the Coriolis signal with subsequent averaging (phase-correct rectification). Interference signals of other frequencies are filtered out (principle of the lock-in amplifier). Averaging (with a low-pass filter) frees the output signal from the excitation frequency. An output voltage is generated which is proportional to the yaw rate:

$$(5) \quad U_A = \text{const} \cdot a_{\text{Coriolis}} = \text{const}' \cdot \Omega$$

Examples of application

- Piezoelectric yaw-rate sensors
- Micromechanical yaw-rate sensors

Acceleration sensors

Measured variables

Acceleration sensors are suitable

- ▶ For knock control in gasoline engines
- ▶ For triggering restraint systems (e.g. airbag and seat-belt pretensioners)
- ▶ For detecting the accelerations of the vehicle for the antilock brake system (ABS) or the electronic stability program (ESP) or
- ▶ For the evaluation of body acceleration for use by the chassis and suspension control systems

The measured variable is acceleration a , often stated as a multiple of the acceleration due to gravity g_0 ($1 \text{ g} \approx 9.81 \text{ m/s}^2$) (see Table 1 for typical values in motor vehicles).

1 Measuring range of acceleration sensors	
Applications	Measuring range
Knock control	40 g
Passenger protection	
– Airbag, seat-belt pretensioner	35 to 100 g
– Side impact, front sensing	100 to 400 g
– Roll-over detection	3 to 7 g
ESP, HHC, ABS	0.8 to 1.8 g
Chassis and suspension control	
– Design	1 g
– Axle/damper	10 to 20 g
Car alarm	1 g

Measuring principles

Acceleration sensors measure the force F executed on an inert mass m by an acceleration a :

$$(1) \quad F = m \cdot a$$

Here, just as in force measurement, systems are available to measure displacement as well as mechanical strain.

Displacement or travel-measuring systems

In the case of travel-measuring systems (Fig. 1), a mass m - the seismic mass - is elastically attached to the body, the acceleration a of which it is intended to measure.

Constant acceleration

In the event of a constant acceleration, the acceleration force is in equilibrium with the return force applied to the spring with spring constant c deflected by x :

$$(2) \quad F = m \cdot a = c \cdot x$$

The system's measurement sensitivity S is therefore:

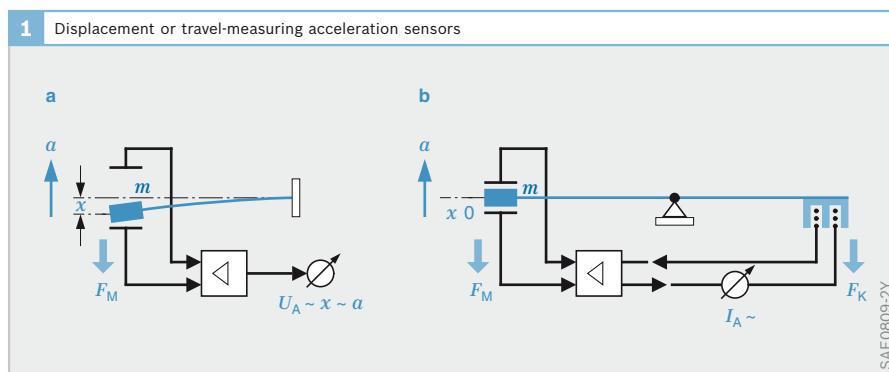
$$(3) \quad S = x/a = m/c$$

This indicates that a large mass together with low spring stiffness result in high measurement sensitivity.

Table 1	
g	Acceleration due to gravity
	$1 \text{ g} \approx 9.81 \text{ m/s}^2$
HHC	Hill hold control
	ABS Antilock brake system
ESP	Electronic stability program

Fig. 1

- Schematic:
- a Excursion-measuring
 - b Closed-loop position controlled
 - a Measured acceleration
 - x System excursion
 - F_M Measuring force (inertial force on the mass m)
 - F_K Compensating force
 - I_A Output current
 - U_A Output voltage



Variable acceleration

In the dynamic case, a damping force and an inertia force are to be taken into consideration in addition to the spring force. The essential damping force is proportional to the velocity \dot{x} and is described using the damping coefficient p . The inertia force is proportional to the acceleration \ddot{x} . The resulting equation (4) describes a resonant system:

$$(4) \quad F = m \cdot a = c \cdot x + p \cdot \dot{x} + m \cdot \ddot{x}$$

Assuming negligible damping ($p \approx 0$), this system has a resonant frequency of:

$$(5) \quad \omega_0 = \sqrt{\frac{c}{m}}$$

This means that in accordance with equation (3), the measuring sensitivity S is directly linked to the resonant frequency ω_0 in the following manner:

$$(6) \quad \omega_0^2 \cdot S = 1$$

In other words, sensitivity can be expected to be reduced by factor 1/4 when the resonant frequency is doubled. It is only clearly below their resonant frequency that such spring-mass systems display adequate proportionality between measured variable and deflection.

The resulting excursion is converted into an electrical signal by a suitable measuring technique (cf. Table 2).

Position control

Travel measuring systems allow the application of the compensation principle, in which the system deflection due to acceleration is controlled out by an equivalent return force (Fig. 1b). The sensor element is now a component part of a closed control loop. The return force or the variable generating it (e.g. current or voltage) acts as a measure of the acceleration. The system operates through closed-loop control very close to the zero point of the excursion and achieves a high degree of linearity.

Closed loop position control produces a greater measuring range, which is only limited by the return force and a higher cutoff frequency than similar, non-motion-controlled systems (Fig. 1a).

Damping

To prevent an excessive resonance, which would cause interference, pure deflection systems demand defined damping that is independent of the temperature. If the damping coefficient p is related to the other parameters in equation (4), this results in Lehr's damping factor D :

$$(7) \quad D = \frac{p}{2 \cdot c} \cdot \omega_0 = \frac{p}{2 \cdot \sqrt{c \cdot m}}$$

2 Measuring principles

Electrical pick-off	Principle	Technical implementation	Examples of application
Piezoresistive	Stress measurement via the change in resistance as a result of elongation caused by acceleration on the spring-mass system	Silicon volume micromechanics, thin-layer systems	Laboratory applications, previously also crash sensors
Piezoelectric	Charge shifts in the crystal as a result of the force applied (spring-mass system)	Ceramic material, PZT, quartz, PVDF	Knock sensors, previously also crash sensors
Capacitive	Excursion of the spring-mass system is determined through capacitance measurement of a plate capacitor	Silicon volume micromechanics, surface micromechanics	Used throughout the motor vehicle
Thermal	Lateral excursion of a heated gas zone and detection of the asymmetry with respect to the heated zone	Volume micromechanics	Rollover sensing

Table 2

This dimensionless variable permits a simple description and comparison of differing resonant systems. Natural vibration and increase in resonance sharpness are largely determined by this damping factor. Values of $D = 0.5$ to 0.7 are preferred in practice (Fig. 2).

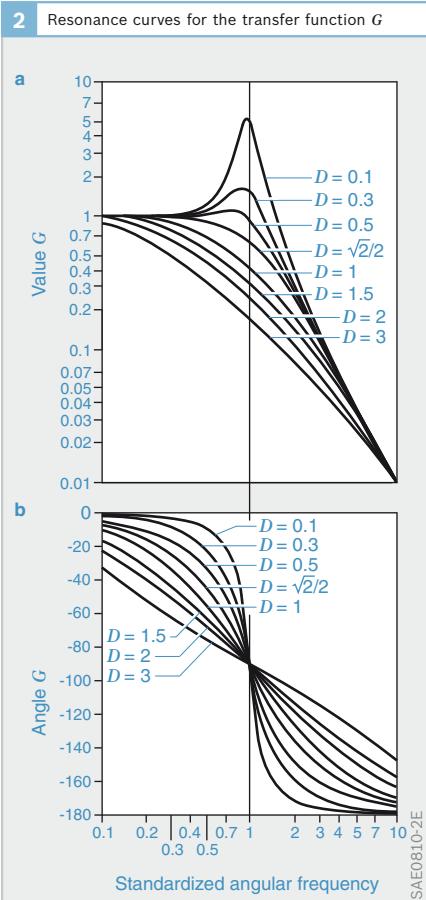
In the case of position-controlled systems, damping is implemented and adjusted in the closed control loop.

Physical implementation

The majority of the acceleration sensors currently in use in the motor vehicle are manufactured in silicon surface micromechanics.

Because of the capacitive evaluation, the measurement signal is determined only by the geometry and stable material parameters and is barely influenced by variables such as temperature. The very small measuring capacities of these sensors, however, requires that evaluation electronics are provided in the immediate vicinity.

The advantage of surface micromechanics is the opportunity it provides to control by position or to deflect the systems by applying electrostatic forces (to the measuring electrodes or to an additionally fitted electrode set). The latter, being a genuine electro-mechanical self test, is an effective means of checking the entire signal path.



Systems for measuring mechanical stress

Applications

Piezoelectric sensors are only used in the motor vehicle in acceleration and yaw-rate sensors.

Longitudinal effect principle of action

Under the action of a force F , piezoelectric materials generate a charge Q on their surfaces to which electrodes are attached (Fig. 3). This charge is proportional to the mechanical stress generated by the force F .

Fig. 2

- a Amplitude resonance curve
- b Phase resonance curve of the complex transfer function $G(i\cdot\Omega) = [x(i\cdot\Omega)]/[a(i\cdot\Omega)]$

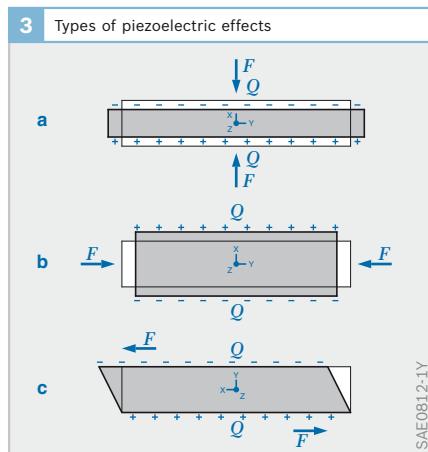


Fig. 3

- a Longitudinal effect
- b Transverse effect
- c Tangential effect
- F Force
- Q Charge

Materials for piezoelectric components can be broadly classified into monocrystals, such as quartz, and ceramics.

Piezoelectrically active ceramics are produced by sintering finely ground ferroelectric materials. These ceramics then consist of very many small crystals randomly oriented in space. These are then polarized in production by the application of a high electric field strength. This largely aligns the randomly oriented polarizations of the microcrystals.

Depolarization, with the consequent loss of the piezoelectric properties, can be caused in piezoceramics by temperatures above the Curie temperature, by intensive mechanical stresses or by polarization with opposite field strengths.

The charges generated are discharged through the external resistance of the measuring circuit or through the piezo sensor's internal resistance. Such sensors cannot measure statically, and are only used where dynamic measurement is needed. The typical limit frequencies of this high pass response, however, lie above 1 Hz, depending on the application.

Transverse effect and shear effect

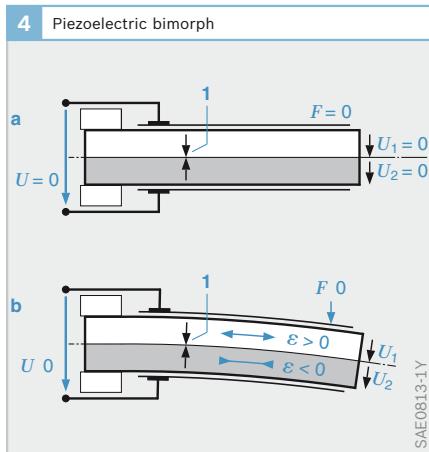
In addition to the longitudinal piezoelectric effect (Fig. 3a), there are also transverse and shear effects (Figs. 3b and 3c).

Fig. 4

- a Non-active state
- b When bent, stretched on top ($\varepsilon > 0$) compressed below ($\varepsilon < 0$)
- 1 Direction of polarization
- F Measuring force
- U Total voltage
- U_1, U_2 Part voltages

Fig. 5

- a Voltage pick-off
- b Charge pick-off
- 1 Supply lead
- 2 Piezoelectric specimen with capacitance C_p
- C_M Measuring capacity
- F Measuring force
- Q Charge
- U Voltage



In practice, these effects occur together. The relationship between the charge generated and the mechanical stress acting can thus be formulated as a tensor equation.

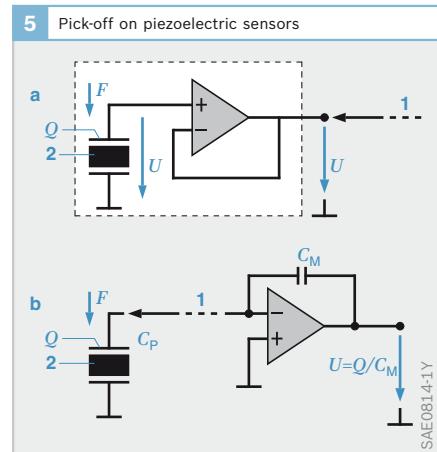
The transverse effect is used for example in a bimorph. These are composed of two oppositely polarized piezoceramics used for the measurement of bending stresses (Fig. 4). When the two-layer ceramic composite bends, one half is stretched ($\varepsilon > 0$) and the other compressed ($\varepsilon < 0$). The opposed-polarity of the ceramic plates means that the resulting part voltages U_1 and U_2 now add to form a total voltage U which can be picked off on the two outside metal layers.

Electrical signal evaluation

Piezoelectric sensor elements may be modeled as a voltage supply with a capacitive internal resistance. The internal resistance is infinitely high for ideal elements. Signal processing must be rated with corresponding care.

There are two options for matching the impedance to subsequent systems:

- **Electrometer amplifier** (Fig. 5a): in this case, the voltage applied across the electrodes is detected and amplified by an amplifier with extremely high impedance input. Parasitic capacitances



however contribute to the transmission function.

- ▶ Charge amplifier (Fig. 5b): here the charge is stored on a second capacitor. Parasitic capacitances have no influence.

In both cases, resistors parallel to the sensor element must be taken into consideration for the transfer function.

Thermal acceleration sensors

Thermal acceleration sensors generate a “bubble of heated gas” over a heater element. The tight heated gas space has a lighter density than the surrounding, cooler gas. In the event of a lateral acceleration, the low density area of gas is displaced within the surrounding, cooler gas. The ensuing asymmetry is detected by thermocouples or resistors as a bridge circuit. The bridge voltage represents the acceleration signal.

Packaging

A packaging which is tailor made to suit each individual application plays a decisive role for the quality of the sensor. Acceleration sensors detect the measured variable without a movable connection to the outside world, it is no problem to encapsulate them hermetically. A rigid mechanical coupling to the body to be measured must, however, be achieved, as additional elastic or loose elements could lead to falsification of the measurement. On the other hand, this fixed coupling must not lead to thermal expansion which may occur at the structure being transferred to the sensor so that measurements are falsified.

Examples of application

- ▶ Piezoelectric acceleration sensors (bimorphous spring elements, longitudinal elements such as knock sensors)
- ▶ Micromechanical acceleration sensors, made in bulk micromechanics and surface micromechanics

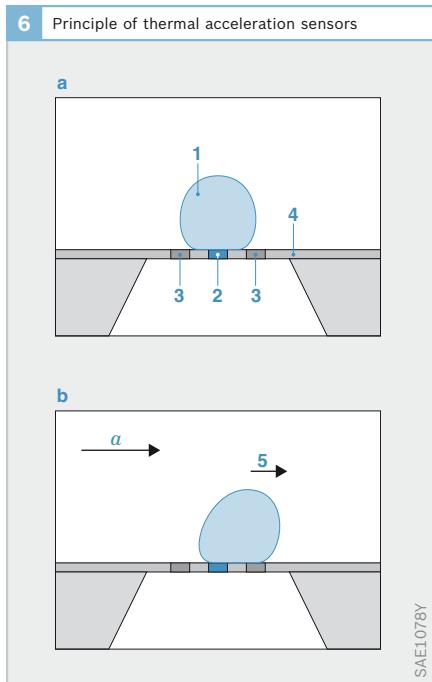


Fig. 6

- | | |
|---|-----------------------|
| a | Acceleration $a = 0$ |
| b | Acceleration $a > 0$ |
| 1 | Heated gas zone |
| 2 | Heater element |
| 3 | Temperature sensor |
| 4 | Substrate coating |
| 5 | Displaced heated zone |

Pressure sensors

Measured variables

Pressure measurement takes place either directly, by way of diaphragm deformation, or using a force sensor for the following applications in the motor vehicle (examples):

- ▶ Intake-manifold or boost pressure (1 to 5 bar) for gasoline injection
- ▶ Brake pressure (10 bar) on electropneumatic brakes
- ▶ Air-spring pressure (16 bar) on pneumatic-suspension vehicles
- ▶ Tire pressure (5 bar absolute) for tire-pressure monitoring
- ▶ Hydraulic reservoir pressure (approximately 200 bar) for ABS and power-assisted steering
- ▶ Shock-absorber pressure (approximately 200 bar) for chassis and suspension control
- ▶ Coolant pressure (35 bar) for air-conditioning systems
- ▶ Modulation pressure (35 bar) for automatic transmissions
- ▶ Brake pressure in master cylinder and wheel-brake cylinder (200 bar), and automatic yaw-moment compensation on the electronically-controlled brake
- ▶ Overpressure/underpressure of the tank atmosphere (0.5 bar)

- ▶ Combustion-chamber pressure (100 bar, dynamic) for detection of misfiring and knock detection
- ▶ Element pressure on the diesel fuel-injection pump (1,000 bar, dynamic) for electronic diesel control
- ▶ Fuel pressure on the diesel common rail (up to 2,000 bar)
- ▶ Fuel pressure on the gasoline direct injection system (up to 200 bar)

Measuring principles

Pressure as a measured variable is a non-directional force acting in all directions which occurs in gases and liquids. It is propagated in liquids, and also very well in gel-like substances and soft sealing compounds. There are static and dynamic measuring sensors for the measurement of these pressures.

The dynamic pressure sensors include for instance all microphones which, since they are insensitive to static pressures, are used to measure pressure pulsations in gaseous and/or liquid media. Since up to now, practically only static pressure sensors have been used in automotive engineering, these will be dealt with in more detail here.

Direct pressure measurement

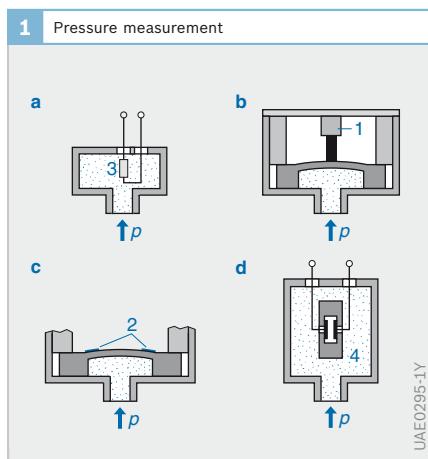
Given that all known resistors are more or less pressure-dependent (volumetric effect), when very high pressures ($>10^4$ bar) are to be measured it would suffice to simply subject an electrical resistor to the pressure medium (Fig. 1a). On the other hand, they are at the same time more or less temperature-dependent, a characteristic which is usually very difficult to suppress. Furthermore, the sealed lead-out of their connections from the pressure medium presents difficulties.

Diaphragm-type sensors

The most common method used for the measurement of pressure (also in automotive applications) uses a thin diaphragm as a mechanical intermediate stage which is

Fig. 1

- a Direct measurement, pressure-dependent resistor (3)
- b Measurement using a force sensor (1)
- c Measurement through diaphragm deformation via strain gage (2)
- d Capacitive measurement using the deformation of a diaphragm cell (4)



UAE0295-1Y

exposed on one side to the pressure to be measured and which deflects to a greater or lesser degree as a function of the pressure. Within a very wide range, its diameter and thickness can be adapted to the particular pressure range. Low-pressure measuring ranges lead to relatively large diaphragms which can easily deform in the range 1 to 0.1 mm. Higher pressures though demand thicker, small-diameter diaphragms which generally only deform by a few μm .

In case (capacitive) pick-offs for spacing or distance measurements are also required, voltage-measuring methods dominate in the medium-pressure to high-pressure ranges. Here, practically only strain gage techniques are used.

Capacitive pick-off

In contrast to their application in inertia sensors (see acceleration/yaw-rate sensors), capacitive pressure sensors are still only rarely encountered even though they could possibly provide similar advantages (particularly with respect to their accuracy). This is more than likely the result of one important difference compared to the other sensors referred to above: pressure sensors need a direct contact with the monitored medium. The dielectric characteristics of which practically always affect the calibration of such capacitive pressure sensors. This means that the calibration would then not only be dependent upon the medium in question, but would also be impossible without it (that is, in the "dry" state). Clear separation of the sensor from the monitored medium has up to now only been achieved at the cost of considerable technical outlay.

Strain gage pick-off

The elongations of the diaphragm occurring on the deflection of a diaphragm-type sensor are detected using strain-gage technology (strain-gage strips or strain-gage resistors). Strain-gage resistors are affixed to the diaphragm (e.g. vapor-deposited). Their electrical resistance

varies under the influence of mechanical stresses. The resistors are connected together to form a Wheatstone bridge. Its voltage is a measure for the pressure.

Table 1 presents a systematic overview of the proven pressure-measurement techniques which to a great extent have already been used in automotive applications. The list is arranged according to the type of diaphragm material and the applied strain gage technology. The combinations have been marked to indicate either those which are already actually in production and which will be described as actual examples in the "sensor designs" chapter (x), or those which are already being closely considered for production (marked in blue):

1 Strain gage pick-off and diaphragm material			
Strain gage pick-off	Diaphragm material		
	Ceramics	Metal (steel)	Silicon
Foils ¹⁾ (glued)			
Thick-film			
Metal thin-film		x	
Silicon thin-film		x	
Diffusion resistors			x

Table 1

¹⁾ Low suitability for large-scale production, (x) current examples
■ Suitability considered

With regard to the particular measuring effect's magnitude and type, the strain-gage techniques listed above have widely varying characteristics. The gage factor (K) defines the magnitude of the measuring effect of strain-gage resistors. It gives the relative change in such a resistor's resistance R referred to the relative change in its length l (equation 1):

$$(1) \quad K = \frac{\Delta R/R}{\Delta l/l} = 1 + 2 \cdot \nu + \frac{d\rho/\rho}{\varepsilon}$$

Here, the symbol ε (elongation) is often used for the relationship $\Delta l/l$, and is used in multiples of 10^{-6} (ppm) as "micron" or "micro strain".

ν is the material's Poisson ratio, ρ representing its electric conductivity. ν characterizes the reduction in cross-section of the material on elongation and, in the idealized case of a volume being kept constant, amounts to $\nu = 0.5$ (in reality $\nu = 0.3$ to 0.4).

Whereas the conductivity term in equation 1 is of hardly any importance in the case of metallic resistors, with regard to Si resistors it plays a dominant role.

One refers to a longitudinal K factor when the resistor is expanded in the current direction and to a transverse K factor when it is expanded transverse to the current direction (Fig. 2). Table 2 provides an overview of typical values for the most important gage factors.

"Creep" (slight mechanical give under the effects of long-term unidirectional loading) is a highly-feared phenomenon which, when it occurs at all, is only encountered on glued foil strain gage. The other strain gage techniques all apply non-glued techniques and are not affected by this phenomena.

To be precise, a diaphragm's deformation depends upon the difference in the pres-

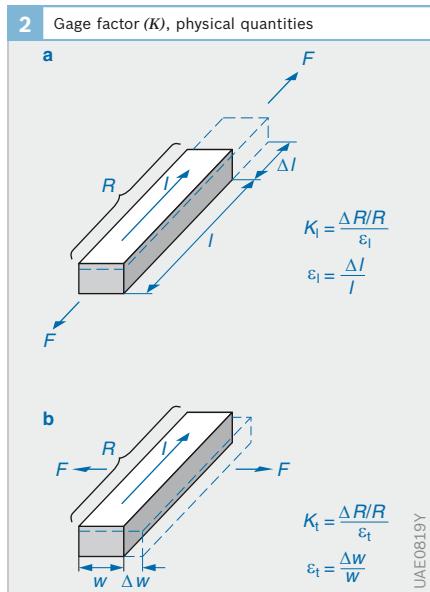


Table 2

Table 3

Fig. 2

- a Longitudinal
- b Transverse
- F Force
- I Current
- R Resistance
- l Length
- w Width
- ε Elongation
- K Gage factor

sure applied to its top and bottom sides. This means that there are four different basic types of pressure sensor (Table 3):

- Absolute pressure
- Reference pressure
- Barometric pressure and
- Differential pressure

Transfer to a force sensor

Instead of directly using the force taken up by their diaphragm, a number of sensors transfer it to a force sensor whose measurement range can remain constant due to the fact that the purely mechanical diaphragm has already performed the adaptation to the pressure-measuring range. Perfect linkage from measuring diaphragm to force sensor (for instance by a tappet) must be ensured though.

Examples of application

- Thick-film pressure sensors
- Micromechanical pressure sensors
- Si combustion-chamber pressure sensors
- Metal-diaphragm high-pressure sensors

2 Gage factors (K) for different materials

Material	Gage factors (K)	
	longitudinal	transverse
Foil strain gage	1.6 to 2.0	≈ 0
Thick-film	12 to 15	12 to 15
Metal thin film	1.4 to 2.0	-0.5 to 0
Si thin film	25 to 40	-25 to -40
Si-monocrystalline	100 to 150	-100 to -150

3 Basic sensor types for pressure measurement

Pressure on diaphragm bottom side p_u	Pressure on upper side of diaphragm p_o	Ambient pressure	Vacuum
Measured pressure	Differential pressure	Reference pressure	Absolute pressure
Measured pressure	Differential pressure	Reference pressure	Absolute pressure
Ambient pressure	Reference pressure	-	Barometric pressure
Vacuum	Absolute pressure	Barometric pressure	-

Force and torque sensors

Measured variables

The following list underlines the wide variety of applications for force and torque sensors in automotive engineering:

- ▶ In the commercial-vehicle sector, coupling force between the tractor vehicle and its trailer or semitrailer for the closed-loop controlled application of the brakes, whereby neither push nor pull forces are active at the drawbar
- ▶ Damping force for use in electronic chassis and suspension control
- ▶ Axle load for electronically controlled braking-force distribution on commercial vehicles
- ▶ Pedal force on electronically-controlled brake systems
- ▶ Braking force on electrically actuated, electronically-controlled brake systems
- ▶ Drive and brake torque
- ▶ Steering and steering servo torque
- ▶ Finger protection on power windows and electrically operated sliding sun-roofs
- ▶ Wheel forces
- ▶ Weight of vehicle occupants (for occupant-protection systems)

In many cases, initial developments failed to lead to the expected results since generally the costs involved in achieving the stipulated accuracy were excessive for the systems in which the sensors were to be installed. Contrary to expectations, it proved impossible to force down the costs for the production of good torque sensors below those for pressure and acceleration sensors. In fact the torque sensors cost more. Matters are aggravated, and this applies particularly to torque sensors, when the measured variable has to be transferred using non-contact methods from a rotating shaft (e.g. steering spindle or drive shaft) to a sensor mounted on the chassis. Collector rings are not acceptable for the motor vehicle.

Since any form of measurement of only part of the force or torque is very problematical and can easily lead to false results, force and torque sensors must be directly connected into the power flux (in other words, the complete measured variable must pass through them). Sensors that measure force are extensive sensors, in other words, their size depends directly on the measuring range.

Although, in line with automotive-industry demands, there are compact force and torque sensors on the market, these only measure accurately enough when the forces are introduced to the sensor in a precisely defined manner, a stipulation that normally can only be complied with under laboratory conditions. The tolerances and misalignment normally encountered in practice generally dictate the connection of homogenization elements which in turn then lead to the sensors becoming too large.

If force and torque-transfer components must be cut in order for previously testable sensors to be fitted, this generally results in an interface problem. This can only be solved by close cooperation between the sensor supplier and the various suppliers of the parts (or vehicle manufacturer) which must be cut in order to install the sensors. Up to now, this problem has not been encountered with other sensor types, at least not with this severity and with such wide-ranging implications.

Even if the force and torque-transfer components do not have to be cut, and mechanical elements are used as measuring springs which only need modifying for installation of the sensor elements, very precise alignment is still necessary.

Even though hardly any series production force and travel sensors are currently available from Bosch for the motor vehicle, there follows a brief overview of the most important measuring principles and proposed developments that have made it to the shortlist.

Measuring principles

Basically speaking, when considering force and torque measurement, a difference must be made between static and dynamic measuring principles, and between measuring principles based on displacement and mechanical strain. Past requirements were predominantly for static sensors, and in the case of force sensors, generally non-resilient strain-measuring principles were preferred. "Soft", resilient sensor systems are only accepted for torque sensors, particularly in the current example of steering-wheel torque sensing, which can also be implemented using angle-measurement pick-offs. This is possible, particularly because this characteristic proved to be tolerable in earlier hydraulic systems which were not equipped with sensors. The use of magnetically active coil systems previously dominated both sectors, force and torque sensing, in automotive and industrial uses.

More recently, however, displacement measuring magnetostatic sensors (Hall sensors) are being used here, too, to allow the application of micro-structured elements manufactured in series production (e.g. measurement of the passenger's weight) and, despite the familiar problems with dirt and the complexity of their installation, optoelectronic pick-offs are also

being considered (e.g. for electronic power-assisted steering systems), which simultaneously permit torque measurement and high-resolution angle of rotation measurement.

Strain measuring force sensors

Magnetoelastic principle

Under the influence of a magnetic field, ferromagnetic materials change their length in the direction of the field (magnetostriiction effect). The length can either increase (positive magnetostriction) or reduce (negative magnetostriction) here depending on the material, for the same field direction. The inversion of this effect, the change in the magnetic characteristics under the effect of tensile and compressive stresses, or elongation and compression, is known as the magnetoelastic effect. This effect is expressed in an anisotropic (directional) behavior of relative magnetic permeability μ_r (relationship between magnetic induction B and magnetic field strength H). If this still has the same value in all directions (isotropic) in a force-free situation, it takes on a somewhat different value (μ_{rl}) in the direction of the force under the influence of a force F introduced than that transverse to it (μ_{rq}) (Fig. 1). The effect is not only observed on crystalline or polycrystalline materials, but also in amorphous substances.

In fact, the permeability change in the direction of force is a true reflection of the sign of the force. Even though practically all ferromagnetic materials demonstrate this effect, it can be optimized by using a specific alloy composition. Unfortunately though, the materials which have good linearity, low hysteresis, and low temperature sensitivity, are not identical with the ones which have a high measuring effect. Whereas the maximum measuring effects observed up to the present are about 30 % (referred to the fundamental isotropic values), and can be utilized without electronics being necessary, the effect displayed by materials which have been optimized from

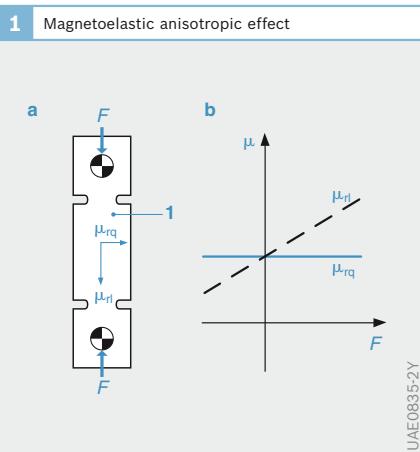


Fig. 1

- a Magnetoelastic measurement structure
- b Measuring effect
- F Force
- μ_r Relative magnetic permeability
- μ_{rq} Transverse to the direction of force
- μ_{rl} In the direction of force

the point of view of measurement techniques is only in an area of a few percent and needs electronic amplification.

The advantages of the magnetoelastic effect are to be found in its broad temperature range and the fact that technical applications are possible up to temperatures as high as approximately 300 °C. On the other hand, it represents a marked volume effect. This means that the coils used for detection not only register a local variation of permeability (as with a strain gage, for instance) caused by the application of force, but more or less the effects throughout the whole of the coil's cross section. This makes the sensor somewhat less sensitive to the possibility of force being applied asymmetrically.

Since the changes in permeability as a function of applied force are practically always registered with the help of alternating fields, the penetration depth of these fields, which is highly frequency-dependent, must be observed:

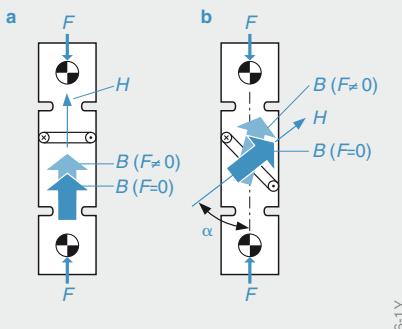
Only those mechanical forces can contribute to the measuring effect which are present in the measuring field's effective penetration depth. In order to be able to put the measuring effect to maximum use, the magnetically active air gap should be

kept as small as possible. Often, this means that the magnetically active measuring circuit is closed with ferromagnetic material, even when this is not included in the power flux.

Figure 2 shows the two most important possibilities of evaluating the magnetoelastic effect: If a coil is arranged on the measurement structure so that its direction of field coincides with the direction of applied force, the change in inductance L can be picked-off and applied directly. Independent of the magnitude of the applied force, the excitation field strength H and the induction B always have the same direction (Fig. 2a).

If the field strength H of the supply coil is not axially parallel to the applied force, the effect of the latter not only changes the magnetic induction B , but also its direction (due to the anisotropy of the permeability, Fig. 2b). Assuming that with no force applied the directions of H and B are superposed one upon the other in the normal manner, these assume increasingly different directions when force is applied and increased. This can be exploited particularly advantageously in the variation of the magnetic coupling of two coils affixed at 90° to one another (Kreuzduktor) (Fig. 3).

2 Influence of the magnetoelastic effect



3 Magnetoelastic tensile-force/compressive-force sensor for measuring brake torque (design specimen)



Fig. 2

- a With direction of force parallel to the direction of field
- b For different directions of field strength H and force F
- B Induction
- α Enclosed angle

Strain gage principle (piezoresistive)

Strain-gage measuring resistors (strain-gage strips) represent the most widespread and probably the most reliable and precise method for measuring force and torque (Fig. 4). Their principle is based on the fact that in the zone of the elastic-member material to which Hooke's Law applies there is a proportional relationship between the mechanical strain σ in the member, caused by the introduction of force, and the resulting elongation ε . In this case, in accordance with Hooke's law:

$$(1) \quad \varepsilon = \Delta l/l = \sigma/E$$

whereby the proportionality constant E is the modulus of *elasticity*. Since it does not directly measure the strain resulting from the applied force, but rather the - locally - resulting elongation, the strain-gage method can be regarded as an indirect measuring method. For instance, if the modulus of *elasticity* decreases by 3 % above 100 K, a figure which is normal for metals, then the force indicated by the strain-gage method is 3 % too high.

Strain-gage resistors in the form of film resistors - by contrast with cemented-on strain gages - are so closely bonded to the

surface of the selected elastic member that they follow its surface elongation perfectly. The change in resistance resulting from the resistor's elongation is defined by the particular gage factor K of the resistor in question (refer to "pressure sensors"):

$$(2) \quad \Delta R/R = K \cdot \varepsilon$$

The K factor generally does not exceed a value of 2 for metal-film resistors, but is generally somewhat below this in practice. Strain gages are designed such that they (in conjunction with a specific elongating material (substrate) and their thermal elongation) have no inherent temperature sensitivity where possible ($T_{CR} \approx 0$). Any residual temperature sensitivity is usually eliminated by depositing the resistors on the elastic member in the form of a half or full bridge. Since temperature effects result in same-direction changes on the strain gage this results in no output signal.

The auxiliary bridge resistors can be (but need not be) located within the elastic member's elongation zone. They can also be fitted as purely compensation resistors (Fig. 4c). It must be noted that often the K factor itself also has a temperature coefficient (TC_K). Usually, this decreases along

4 Strain-gage force sensors

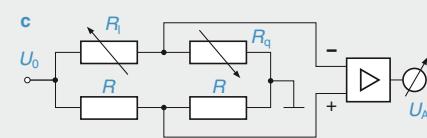
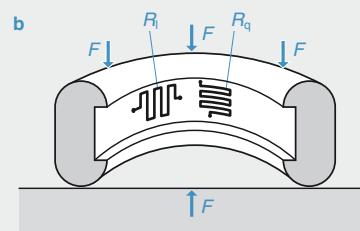
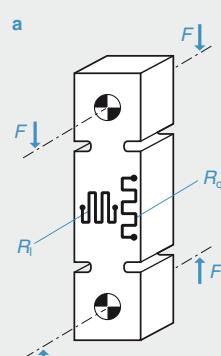


Fig. 4

- a Rod-shaped
- b Toroidal-shaped
- c Electronic evaluation
- F Force
- $R_{l,q}$ Metal-film resistors, lengthwise, crosswise
- R Auxiliary bridge resistors
- U_0 Supply voltage
- U_A Output voltage

with increasing temperature, which means that in favorable cases it can compensate for the signal increase caused by the modulus of *elasticity*. Apart from this, signal reductions caused by the K factor are usually compensated for by a bridge supply voltage U_0 which increases accordingly.

In addition to the longitudinal K factor K_l described above, there is also the transverse K factor K_t , which must then be applied if the direction of strain is transverse to the current flow. It has an inverse sign (negative) and is no greater than 0.5 in value.

Disadvantages/limitations:

Notwithstanding their high levels of accuracy and reliability, since the elongations and therefore also the changes in resistance (at least in the case of metal film resistors) are only in the percentage range of the original state, the strain-gage sensors only generate output signals in the mV range so that in general local amplification is required. A further disadvantage of small strain-gage sensors is the fact that they measure precisely the strain

(and only this strain) at the point at which it is applied. This, of course, no longer holds true in cases in which the strain-gage structure is distributed across the elastic member's complete surface. This necessitates extremely precise and reproducible application of the force to be measured if measuring errors due to uneven force introduction are to be avoided.

Application:

As a rule, in order to carry out force measurements, very small strain-gage resistors must be attached to larger force-carrying parts or elastic members. The traditional method of applying the strain-gage resistors to the elastic member by means of a foil backing (as applied in such devices as high-precision scales), is not inexpensive enough for large-batch "low-cost" production. First attempts are therefore being made to apply low-cost, large batch film/layer techniques by depositing the strain-gage resistors on small metal wafers which are then pressed into, or welded onto the elastic element.

Orthogonally printed resistors

Practically all electrical film resistors change their resistance not only under the influence of laterally applied deformation strain but also when compression is applied vertical to the film plane (orthogonal). Here, the so-called "conductive plastic", commonly used in potentiometers, features very high sensitivity. "Cermet" and "carbon layers" are also highly sensitive (Fig. 5). Up to a certain limit, the resistance of the above materials decreases along with increasing compression. The values that can be achieved without permanent resistance change are similar to those which apply for lateral elongation. In both cases, the limit is a function of substrate strength and not of the resistance material. Of course, sensors of this type are almost only suitable for loading by compression but not by lateral strain.

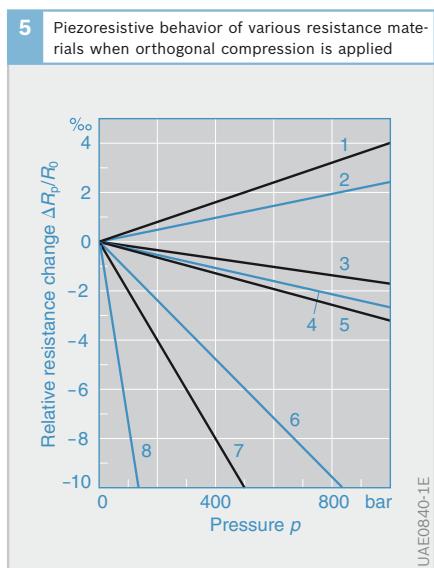


Fig. 5

- 1 84.5 Ag 15.5 Mn
- 2 Manganin
- 3 Cu
- 4 Au
- 5 Ag
- 6 Carbon film/layer
- 7 Cermet
- 8 Conductive plastic

Travel measuring force sensors

The question of whether an integrated travel-measuring weight sensor still has a little additional elasticity or not is not important when it is built into seats for motor vehicle occupants, which are anyway suspended.

This principle is used in the iBolt™ force sensor (see sensor types). Four of these sensors are built into the frame of the front passenger seat for measuring the weight of the passenger. This makes it possible to modify the triggering of the airbag.

Examples of application for force sensors

- ▶ Magnetoelastic load measuring bolts (agricultural tractor with plow force control)
- ▶ iBolt™ force sensor for sensing passenger weight

Torque sensors

A fundamental distinction is made in torque measurement, too, between methods using angle or strain measurement. In contrast to strain-measurement methods (strain-gage resistors, magnetoelastic), angle-measurement methods (e.g. eddy

current) require a certain length l of the torsion shaft via which the torsion angle (approximately 0.4 to 4°) can be picked-off. The mechanical stress σ proportional to the torque is aligned at an angle of less than 45° to the shaft axis (Fig. 6).

The principles described below are all suited for the non-contacting (proximity) transfer of measured values, even from rotating shafts. In the case of steering-torque measurement, it is also desirable for the employed measuring system to measure the steering angle very accurately (through a full rotation of 360°). This is in the sense of modular integration and is to be implemented with only slight modification.

Strain-measuring sensors

Even though magnetoelastic sensors which enclose the shaft are available on the market, these involve very high costs. Since it is often impossible to optimize the shaft material with regard to its magnetoelastic properties, a search is being made for methods to coat the measuring shaft with a magnetoelastic layer. Such a coating, which has to exhibit good measuring qualities, has still not been found.

This is why the strain gage principle has become predominant (Fig. 7): a strain-gage bridge measures the mechanical stress. The bridge is powered using transformer principles (by a rectifier and control-electronics circuitry on the shaft, independent of the air gap). Further local electronic components on the shaft permit the measurement signal to be amplified and converted to an AC form which is independent of the air gap (for instance, frequency-analog) which can then also be outputted using transformer principles.

When large-batch quantities are involved, the electronic circuitry on the shaft can be integrated in a single chip without problems. The strain-gage resistor can be inexpensively formed on a prefabricated round steel plate (for instance, using thin-film techniques) which is then welded to the shaft. The two ring transformers re-

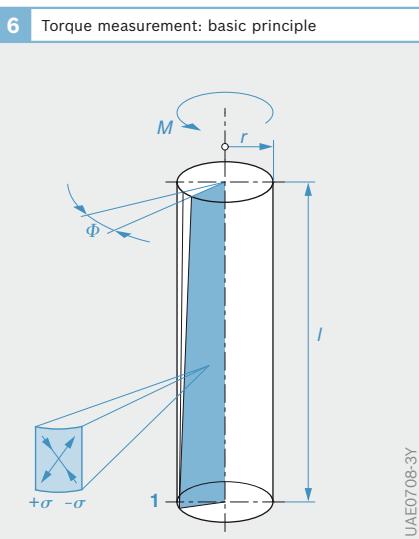


Fig. 10

- | | |
|----------|------------------|
| 1 | Torsion rod |
| Φ | Torsion angle |
| σ | Torsional stress |
| M | Torque |
| r | Radius |
| l | Rod length |

quired for power and signal transmission can be manufactured at an extremely low cost from a soft-magnetic molding compound. High precision levels can be achieved with such a configuration in spite of reasonable manufacturing costs.

Angle-measuring (torsion-measuring) sensors
 Sensors for measuring angular difference
 It is a relatively easy matter to determine the torsion angle when two mutually-independent incremental speed sensors, or an absolute-measuring (analog or digital) non-contacting angular-movement pick-off, are provided at each end of the section of the shaft subject to torsion ($L \approx 5$ to 10 cm) (Fig. 8). The difference in their outputs $\varphi_2 - \varphi_1$ is a measure for the torsion angle:

$$(3) \quad M = \text{const} \cdot L \cdot (\varphi_2 - \varphi_1)$$

where L = length of the section subject to torsion

Up to now, since adequate accuracy demands extremely precise bearings, together with the necessity to provide correspondingly accurate angular or incremental subdivision around the complete periphery, this method has been regarded as too complicated. Nevertheless work is forcing ahead with solving this problem (magnetically or optically), because such a system would incorporate two distinct advantages:

- ▶ Possibility of simultaneous measurement of the angle of rotation with the same system
- ▶ Possibility of measurement without far-reaching modifications to the torsion shaft being necessary, so that essentially the sensor could be in the form of a plug-in sensor providing an efficient interface for a supplier component

The angle may be measured with high precision using the vernier principle, for example: this is done by marking the circumference of the shaft with a sinusoidal track, of which an integer number of N cycles fits in one circumference.

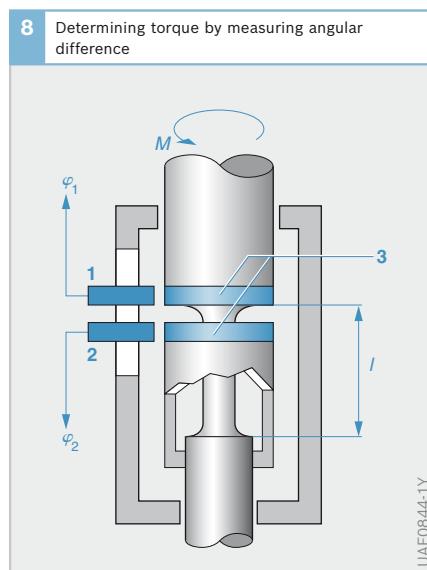
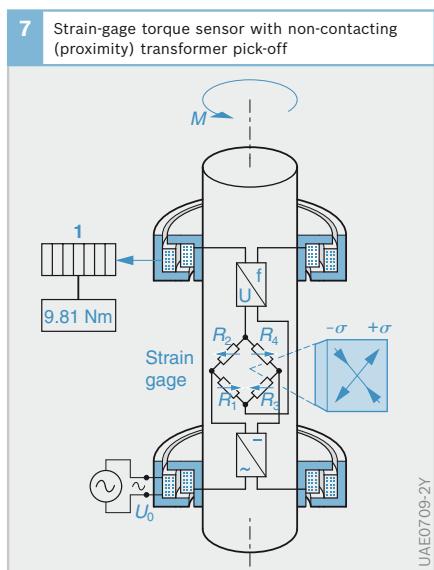


Fig. 7

1 Torque indicator
 σ Torsional stress
 M Torque
 U_0 Supply voltage
 R_1 to R_4 Strain-gage measuring resistors

Fig. 8

1, 2 Angle/speed sensors
 3 Angle markings
 l Torsion-measurement section
 M Torque to be measured
 $\varphi_{1,2}$ Angle signals

By contrast with a pure incremental track, which would permit only a resolution of an N -th of the circumference, this sinusoidal signal structure, however, theoretically permits as fine a resolution per circumference as is required by the application of the arcsin function within a single sine cycle. This fine resolution is, however, only useful if it is always known in which of the N similar cycles one is located.

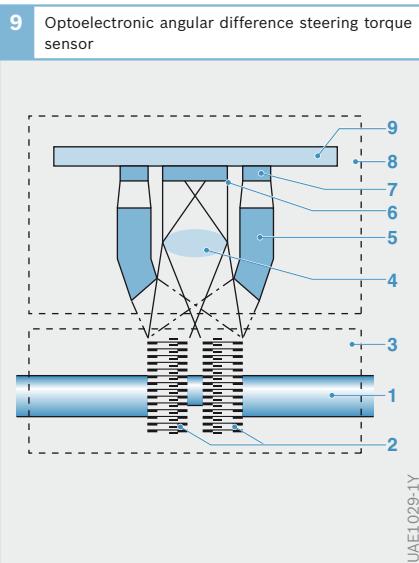
The unambiguity can easily be created if a second track with a little less spatial frequency in which only $N-1$ sine cycles fit on the circumference is applied around the circumference. Although the two signals do not supply an unambiguous signal for an angle-of-rotation measurement, it is, however, possible to draw firm conclusions from the phase difference $\Delta\varphi$ between the two signals, in which sine cycle one is currently located. The pitch difference between the two signals varies over the entire circumference by precisely 360° . Figure 10 shows an example with $N = 10$. If the phase difference between the two signals is, for instance, in the range between 36° and 72° , then one is clearly lo-

cated in the second sine cycle. If the value is between 216° and 252° , one is located in the 6th cycle. The precise measuring angle φ is then formed by adding another angle, in the first case of 360° and in the second case of 216° to the measured, but not unambiguous fine signal φ .

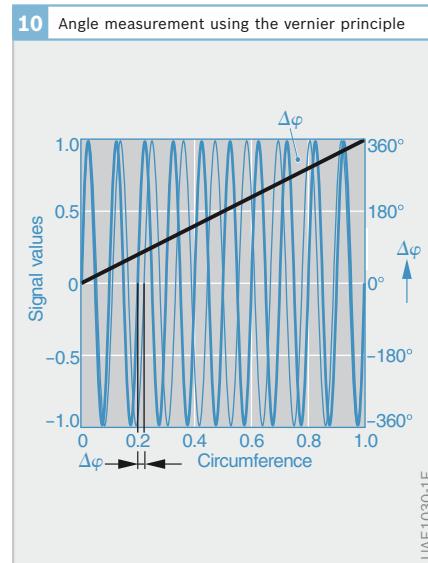
In practice however, the arcsin function is not used to resolve the individual sine cycles finely, as it would be necessary to guarantee a constant and normalized signal amplitude for this. It is rather the case that a second, cosine-wave track, phase-shifted against the first by precisely 90° , is applied next to the sine-wave marking. Because of the spatial vicinity, it may be assumed that both tracks are sensed with the same amplitude u , so that the angle φ within the sine cycle can be determined by the arctan function from the two individual signals u_1 and u_2 independently of u :

$$(4) \quad \varphi = \arctan(u_1/u_2)$$

where $u_1 = u \cdot \sin \varphi$ and
 $u_2 = u \cdot \cos \varphi$

**Fig. 9**

- Optical pick-off of the angle marking tracks
- 1 Steering shaft with torsion rod
 - 2 Code disks with bar code
 - 3 Steering gear housing
 - 4 Lens
 - 5 Optical waveguides
 - 6 Optical ASIC
 - 7 LED
 - 8 Sensor module
 - 9 PCB



Thus a total of four tracks are required for the determination of a precise and unambiguous angle of rotation φ . The steering angle and TAS (Torque Angle Sensor) steering-angle sensor illustrated in Figure 9 thus requires a total of eight tracks, which are read optoelectronically, to measure two angles φ_1 and φ_2 . Its angular precision over 360° is 1° , the angular resolution is 0.0055° , the resolution of the difference angle $\Delta\varphi$ is 0.0044° with a measuring range of $+ 9^\circ$.

Eddy-current sensors

Two slotted sleeves of electrically conducting aluminum are attached to each end of a sufficiently long section of the measuring shaft. These are inserted one inside the other (Fig. 11). Each sleeve is provided with two rows of slots so that when the shaft is twisted, it becomes increasingly visible through one row of slots, and is hidden more and more by the other row. This leads to increasing, or decreasing, damping of the two high-frequency coils (approx. 1 MHz) situated above each row of slots so that coil inductance also varies accordingly. The slotted sleeves

must be precisely manufactured and assembled in order to achieve the stipulated accuracy. The associated electronics are appropriately accommodated very near to the coils.

Although this sensor principle was developed at Bosch for motor vehicle applications, it was only used in electric power tools (torque sensing in industrial power screwdrivers). A license-holder in Japan has further developed this, very cost-effective, sensor principle for motor vehicle applications to product maturity.

Examples of application for torque sensors

- ▶ Strain gage steering torque sensor (developed up to A prototype for electrical power-assisted steering)
- ▶ Angle-measuring eddy-current torque sensors (electric power tools, third-party product in Japan)
- ▶ Angle-measuring optoelectronic steering angle sensor (electrical power-assisted steering)

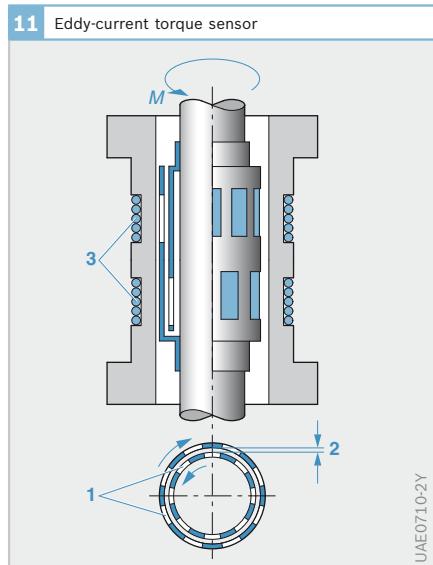


Fig. 11

1	Slotted sleeves
2	Air gap
3	High-frequency coils
M	Torque

Flowmeters

Measured variables

The purpose of flow measurement in the motor vehicle is the detection of the intake air flow rate. This air flow rate must be known precisely so that the engine-management system - both in diesel and in gasoline engines - can set a defined air-fuel mixture. This value can be determined by a flowmeter. The sensors which are used for measuring air flow rate or gas flows in general are also referred to as "anemometers".

As such, the often-used term "air quantity" is incorrect because it does not stipulate whether volume or mass is concerned. Since the chemical processes involved in fuel combustion are clearly based on mass relationships, the object of the measurement is the mass of intake air. On gasoline engines, the air-mass flow rate is the most important load parameter. In diesel engines, the exhaust-gas recirculation rate is regulated using the air-mass flow.

Depending upon engine power, the average (over time) maximum air-mass flow rate to be measured is between 400 and 1,200 kg/h. Due to the low air requirements at engine idle in modern gasoline engines, the ratio of minimum to maximum

flow is 1:50 to 1:100. Because of the higher idle air demand in diesel engines, these ratios must be assumed to be 1:20 to 1:40. The severe exhaust gas and fuel-consumption requirements dictate accuracies of 2 to 3 % of the measured value. Referred to the measuring range, this can easily correspond to a measuring accuracy of $2 \cdot 10^{-4}$, which is unusually high for a motor vehicle.

The air though, is not drawn in continuously by the engine, but rather in time with the opening of the intake valves. Particularly with the throttle valve wide open (WOT) in gasoline engines, this leads to considerable pulsation of the air-mass flow, also at the measuring point which is always in the intake tract between air filter and throttle valve, or between air filter and turbocharger (Fig. 1). Intake-manifold resonance leads to the pulsation in the manifold sometimes being so pronounced that brief return flows can occur. This applies in particular to 4-cylinder engines in which there is no overlap of the air-intake phases. An accurate flowmeter must be capable of registering these return flows with the correct direction.

On a 4-cylinder engine, the pulsations are generated at twice the crankshaft speed. This means that they can easily be in the range 50 to 100 Hz. With a flowmeter featuring a linear characteristic curve and a narrower frequency bandwidth than above, it would suffice for it to follow the mean value of these rapidly fluctuating air flows. The mean value is in any case positive, so that the meter need not necessarily detect the correct sign.

Practically all of the air-mass meters actually in use feature a characteristic curve which is far from linear, so that the measurement signal must be linearized electronically before it can be evaluated. Averaging before linearization can lead to significant dynamic errors (mean-value errors). This fact means that such air-mass meters must be sufficiently rapid in following the pulsations which mostly have a

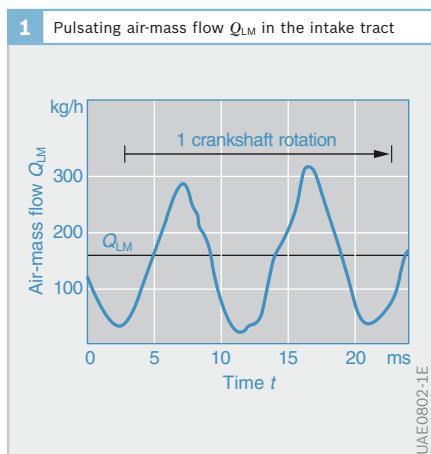


Fig. 1

At full load with
engine speed
 $n = 3,000$ rpm,
intake-manifold
pressure
 $p_s = 0.96$ bar,
mean air-flow rate
 $Q_{LMm} = 157.3$ kg/h

pronounced non-sinusoidal characteristic, and therefore also have a considerable harmonic content. This necessitates a bandwidth of about 1,000 Hz. Apart from this considerable bandwidth, the air-mass meters must also have a high switch-on time constant in order for them to be able to measure correctly during the engine starting phase.

As is the case in all flowmeters, those used in the motor vehicle are calibrated for flow through a pipe with a symmetrical flow profile. This means that a flow with a velocity vector v at practically any point of the flow cross-section of area A depends only on the radius to the central axis, and not on the angle at the circumference. The flow profile (laminar or turbulent, Fig. 2) is related to the Reynolds number R_e :

$$R_e = v \cdot D / \eta$$

where

D typical cross-section dimension and
 η kinematic ductility of the medium.

The flow profiles illustrated by way of example only establish themselves after the flow has been uniformly formed for some time. In the motor vehicle, a profile lying

between the laminar and turbulent profiles illustrated will be established for the case of laminar flow – caused by the short inflow. The laminar or turbulent characteristic of the flow profile depends on whether R_e is below or above a value of approximately 1,200.

A further important influencing factor is the degree of turbulence in the flow, which is determined by the aerodynamic design of the inflow to the flowmeter. If the laminar-turbulent transition is within the measuring range, marked irregularity of the characteristic curve can be expected at this point. As far as motor vehicle applications go, a purely turbulent flow (square profile: $v = const_r$) cannot be assumed at the smallest air masses (gasoline engine idling). A careful design of the inflow, to achieve square profiles even in the event of laminar flow, is then necessary.

Assuming a homogeneous density ρ , and assuming a square profile, the flow is simple to calculate as follows:

$$Q_V = v \cdot A$$

Volume flow rate

$$Q_M = \rho \cdot v \cdot A$$

Mass flow rate

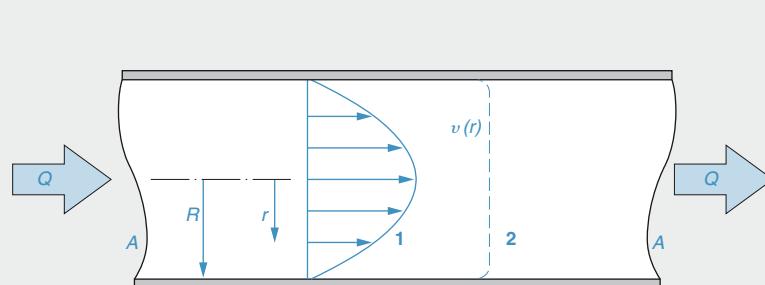


Fig. 2

- 1 Laminar flow profile
- 2 Turbulent flow profile
- A Cross-sectional area of the tube
- Q Flow
- R Tube radius
- r Distance from the tube center
- $v(r)$ Flow profile

Whereas in measurement techniques, long, straight, advance and overshoot sections of constant cross-section are stipulated in order to guarantee a symmetrical profile, such conditions cannot be complied with in a vehicle's cramped under-hood installation space. If pronounced asymmetries occur, the flowmeter must be calibrated as a function of the actual installation conditions.

Measuring principles

Up to now, of the practically unlimited variety of flowmeters on the market, only those which operate according to the impact-pressure principle have come to the forefront for air-quantity measurement in the vehicle. This principle still depends upon mechanically moving parts, and in principle correction measures are still needed to compensate for density fluctuations.

Today, true air-mass meters applying thermal methods (hot-wire or hot-film air flowmeters) are used which can follow sudden flow changes without mechanically moving parts.

Variable orifice plates (sensor plates)

The calculation of the pressure drop across fixed orifice plates is based on two physical laws:

Continuity equation:

$$\rho \cdot v_1 \cdot A_1 = \rho \cdot v_2 \cdot A_2 = \text{const}$$

Bernoulli's equation:

$$p_1 + \frac{1}{2} \cdot \rho \cdot v_1^2 = p_2 + \frac{1}{2} \cdot \rho \cdot v_2^2 = \text{const}$$

These laws are to be applied for two different measuring cross-sections A_1 and A_2 (Fig. 3). Assuming constant density $\rho = \rho_1 = \rho_2$, this results in the pressure drop:

$$\Delta p = Q_v^2 \cdot \rho \cdot \left(\frac{1}{A_2^2} - \frac{1}{A_1^2} \right)$$

This pressure drop can be measured either directly with a differential-pressure sensor, or by means of the force acting against a so-called sensor plate. It should be noted here that impact-pressure gages of this type measure a flow value that corresponds neither to the volume flow rate nor to the mass flow rate, but to a geometrical mean value of the two:

$$Q_{\text{St}} = \text{const} \cdot \sqrt{\rho} \cdot v = \text{const} \cdot \sqrt{Q_v \cdot Q_M}$$

Fig. 3

- a Ring orifice
- b Sensor plate
- 1 Orifice plate
- A_s Disk cross-section
- $A_{1,2}$ Measuring cross-section
- $p_{1,2}$ Measured pressures
- Δp Pressure drop
- Q_{LM} Air-mass flow

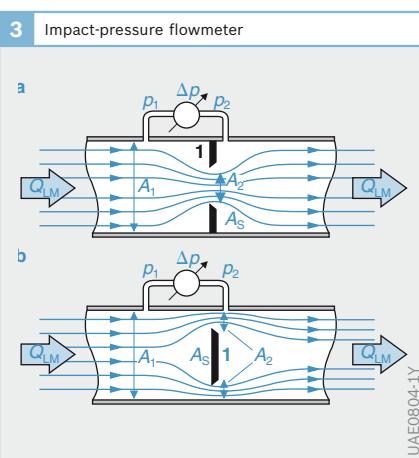
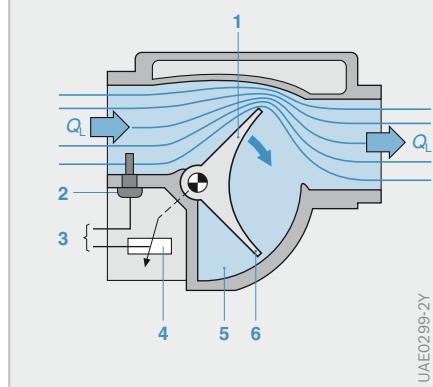


Fig. 4

- 1 Sensor plate
- 2 Air-temperature sensor
- 3 To control unit
- 4 Potentiometer
- 5 Damping chamber
- 6 Compensation flap
- Q_L Intake-air flow

4 Impact-pressure air-mass meter



Because of the square relationship to the flow rate, fixed orifice plates permit only a variation in measured variable of 1:10, as otherwise the accuracy of the differential pressure recording is no longer adequate and significant pressure losses would result which would lead to power losses in the engine. When larger ranges are to be covered, several orifice plates must be used, or such versions which automatically adapt themselves to the measuring range by opening up a larger flow cross-section A_2 in line with the impact pressure.

With such variable, moving sensor plates it is an easy matter to increase the variation to 1:100. Here, the increasing air flow causes the sensor plate to be deflected (usually against a constant counterforce) into an area whose cross-section is specifically shaped so that the resulting deflection/angle relationship complies with the desired characteristic. In other words, linear for K-Jetronic and non-linear for L-Jetronic. The sensor plate's (Fig. 4) setting is then a measure for the air flow which is in relationship to the impact pressure defined above.

The limit frequency for such sensors is typically approx. 10 Hz. Such sensor plates though are unable to follow the high pulsation frequencies which often occur. From the point of view of the pulsation, they can be regarded as fixed orifice plates with a square-law characteristic curve. Under certain load conditions this leads to considerable mean-value errors which can only be compensated for roughly by the use of suitable software.

If the density ρ of the intake air changes because of temperature fluctuations or the altitude, the measurement signal here will only change by $\sqrt{\rho}$. An air-temperature sensor and a barometric pressure sensor must also be used to determine the density fluctuation fully.

Hot-wire/hot-film anemometers

When current I_H flows through a thin wire with electrical resistance R , its temperature increases. If at the same time a medium with density ρ flows across it at velocity v , a balance is set up between the electrical power input P_{el} and the power P_V drawn off by the air flow, whereby

$$P_{el} = I_H^2 \cdot R = P_V = c_1 \cdot \lambda \cdot \Delta \vartheta$$

Here, the power drawn off by the air flow is proportional to the temperature difference $\Delta \vartheta$ established and the coefficient of thermal conductivity λ . The following applies in close approximation:

$$\lambda = \sqrt{\rho \cdot v} + c_2 = \sqrt{Q_{LM}} + c_2$$

Although λ is primarily a function of the mass flow Q_{LM} , with the static medium at standstill ($v = 0$), a certain heat loss takes place (convection) represented by the additive constant c_2 . This results in the familiar interrelationship between the heating current I_H and the mass flow Q_{LM} .

$$I_H = c_1 \cdot \sqrt{(\sqrt{Q_{LM}} + c_2)} \cdot \sqrt{\frac{\Delta \vartheta}{R}}$$

With the supply of a constant heat $I_H^2 \cdot R$, a temperature rise $\Delta \vartheta$ decreasing reciprocally with the root of the air-mass flow Q_{LM} would establish itself. If, on the other hand, the heating current I_H is controlled such that a constant temperature increase (for instance, $\Delta \vartheta = 100$ K) is maintained even when the flow rate increases, this will lead to a heating current which increases approximately with the fourth root of the mass flow, and at the same time serves as a measure for the mass flow.

The advantage of such a control loop lies in the fact that the electrical heating resistor always remains at the same temperature so that its heat content need not be changed by means of time-wasting heat transfer. In fact, with a 70 µm platinum wire for instance, it is possible to achieve

time constants in the 1 ms range for changes in air-flow rate. In cases where closed-loop control is not used the time constants would be 40 to 100 times higher.

If the heater temperature were to be maintained constant simply by keeping its (temperature-dependent) resistance constant, with constant mass flow and higher medium temperature, this would result in a current drop and therefore a false measurement. In practice, this error is avoided by using a bridge circuit containing a second high-ohm compensation resistor R_K of the same type (e.g. platinum). Here, the heating resistor is kept at a constant over-temperature $\Delta\vartheta$ compared to the medium (Fig. 5) by a control loop. In case of a sudden jump in the medium temperature, the sensor reacts with a long time constant since in this case the heat content of the hot wire must be changed. The heating current generates a control voltage U_M proportional to the air-mass flow on a precision resistor (measuring resistor R_M).

The heating resistors in the first air-mass meters (hot-wire air-mass meters) used for automotive applications were of very fine platinum wire. This wire was stretched in trapezoidal form across the flow cross-section so that it was able to

average out any asymmetries in the flow profile (Fig. 6). Adequate service lives only became possible when the platinum wire was stabilized by alloy additives so that its resistance no longer changed due to deposits and cracks on its surface. This meant though that the deposits on the hot wire had to be burned-off automatically (approximately 1,000 °C) following every operating phase.

Notwithstanding a number of functional advantages, this sensor concept was far too costly. A thick-film version (HFM2 hot-film air-mass meter) was able to combine all the resistors concerned with the measurement on a single ceramic substrate. Due to the ceramic substrate's considerable thermal capacity, it was difficult not to exceed the maximum permissible switching constants. Furthermore, a complicated saw cut had to be made to reduce the undesirable heat coupling between heating and compensation resistors. On the other hand though, this version permitted the burn-off process to be dispensed with since the special flow conditions no longer led to unwanted deposits.

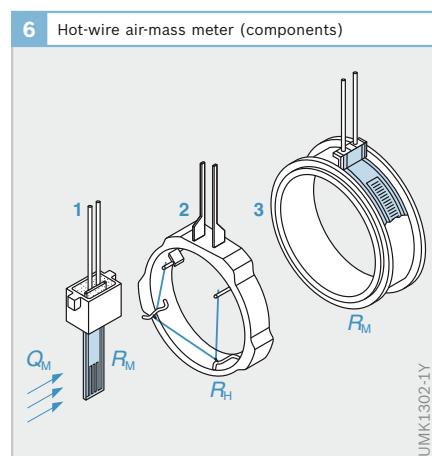
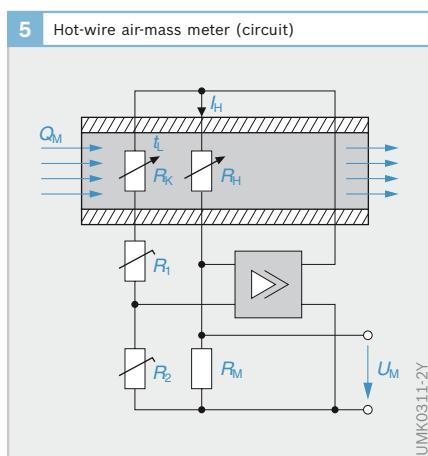
In contrast to both its predecessor types, a further silicon-based micromechanical version (HFM5 and HFM6 hot-

Fig. 5

- Q_{LM} Air-mass flow
- U_M Measurement voltage
- R_H Hot-wire resistor
- R_K Compensation resistor
- R_M Measuring resistor
- $R_{1,2}$ Calibration resistors

Fig. 6

- 1 Temperature compensation resistor R_K
- 2 Sensor ring with hot wire R_H
- 3 Precision measuring resistor (R_M)
- Q_M Air-mass flow



film air-mass meters) fulfilled practically all expectations. In particular, this version is able to measure in both flow directions with the correct sign (Fig. 7). This means that the brief return flows that occur as a result of pulsation no longer lead to measuring errors (Fig. 8).

To this end, in addition to the heater control circuit used in the previous versions, a temperature sensor is located on each side of the heating resistor, in other words upstream and downstream. This principle is similar to the Thomas process often encountered in the literature. When there is no flow ($Q_{LM} = 0$), each of these sensors indicates the same temperature. When flow starts though, since the upstream sensor is cooled by the medium, the higher the flow the higher the temperature difference between the two resistors. The output signal derived from the temperature difference has a similar characteristic curve to the anemometers used up to now, whereby its sign is a clear indication of the flow direction.

Due to its small size, the micromechanical flowmeter is only a partial-flowmeter. In other words, it is no longer in any way able to average-out any non-homogeneity in the flow velocity as a function of the flow cross-sectional area. Rather,

this flowmeter must ensure that the partial flow it measures represents the same fraction of the total flow throughout the whole measuring range. This precondition can be achieved by means of a flow rectifier (grill) built into the measuring pipe.

Examples of application

- HFM5 and HFM6 hot-film air-mass meters for gasoline and diesel engine applications

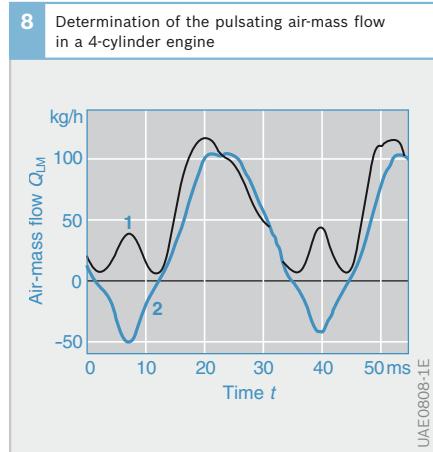
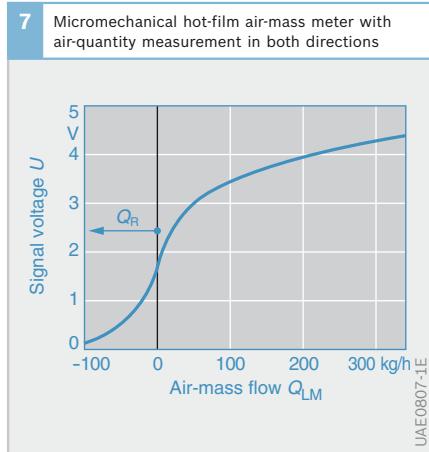


Fig. 7
 Q_R Return flow

Fig. 8
At full load and
engine speed
 $n = 900$ rpm
1 Hot-wire
air-mass meter
2 Hot-film
air-mass meter

Gas sensors and concentration sensors

Measured variables

The concentration of a given material or medium defines the mass or volume percent of a given material in another given material or in a mixture or combination of other materials. With a concentration sensor (also known as a concentration probe), the important thing is that in the ideal case it is sensitive to only one medium, while at the same time practically "ignoring" all other mediums. Of course, in practice, every concentration sensor has its own cross sensitivity to other mediums even though, as is often the case, "temperature" and "pressure" are maintained constant.

In the vehicle, the following parameters must be measured:

- ▶ Oxygen content in the exhaust gas (closed-loop combustion control, catalytic-converter monitoring)
- ▶ Carbon-monoxide and nitrogen-oxide content, as well as air humidity inside the vehicle (air quality, misting of vehicle windows)
- ▶ Humidity in the compressed-air brake system (air-drier monitoring)
- ▶ Dampness of the outside air (black ice warning)
- ▶ Concentration of soot in diesel-engine exhaust gas. A still unsolved problem. In contrast to the above-mentioned gas concentrations, this is a particle concentration. The difficulties inherent in the measuring assignment are further aggravated by the possibility of the sensor being blocked by particles so that it no longer functions.

The introduction of the fuel cell as an automotive drive means that further gas sensors will have to be developed, for instance for the detection of hydrogen.

Measuring principles

Measured mediums occur in gaseous, liquid, or solid state, so that in the course of time countless measuring methods have been developed. For automotive applications, until now only the gas-analysis area, and in particular the measurement of gaseous humidity, has been of any interest. Table 1 presents an overview of the processes applied in general measurement techniques.

Gas measurement in general

Gas sensors are usually in direct unprotected contact with the monitored medium (in other words with foreign matter) so that the danger of irreversible damage exists. This form of damage is referred to as sensor "contamination". For instance, the lead that may be contained in fuel or the exhaust gas can make the electrolytic oxygen concentration sensors (Lambda oxygen sensors) unusable.

Moisture measurement

In addition to the outstanding significance of the Lambda oxygen sensor in dealing with exhaust gases, moisture measurement also plays an important role.

In the broader sense, moisture indicates the moisture content of gaseous, liquid, or solid substances. In the narrower sense, we are dealing here with the gaseous-water (water vapor) content in gaseous media – above all in the air.

1 Gas-analysis processes (without particular attention being paid to the moisture-measurement process)
(X) = for automotive applications

Physical process	Physical-chemical process	Chemical process
Thermal conductivity	Heat of reaction	Selective absorption
Magnetic processes	Heat of absorption	Selective absorption with prior chemical conversion
Radiation absorption	Characteristic color reaction	
Gas chromatography	Electrolytic conductivity	X
Radioactive processes	Electrochemical processes	X

When a damp gas is cooled in an isobaric process, it reaches its saturation point at a specific temperature (known as the dew point τ).

A number of important definitions are given below in connection with humidity measurement (refer also to Figure 1):

m_w Mass of water

m_s Mass of water in the saturation state

m_{tr} Mass of dry gas

M_w Mole mass of water

M_{tr} Mean mole mass of the dry gas

p Total pressure of the gas mixture

p_w Partial pressure of the water vapor

p_s Saturated vapor pressure (vapor pressure of the water at mixture temperature)

Absolute humidity:

$$\chi = \frac{m_w}{m_{tr}} = \frac{M_w}{M_{tr}} \cdot \frac{p_w}{p - p_w} \text{ (in \%)}$$

$$f_a = \frac{m_w}{V_{tr}} \quad \text{(volume-related)}$$

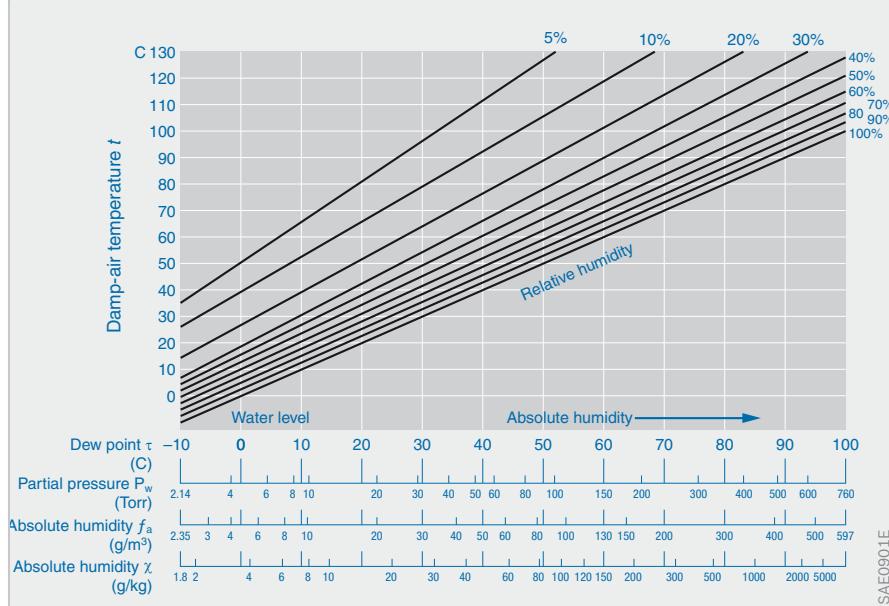
Relative humidity:

$$\Phi = \frac{p_w}{p_s} \text{ (in \%)}$$

For low-cost applications (for instance in the vehicle), resistive and capacitive sensors are used almost exclusively. They are provided with hygroscopic layers which can store water as a function of the relative humidity (and release it again), and thus trigger a usually drastic change in a resistor's value or in the value of a planar capacitor.

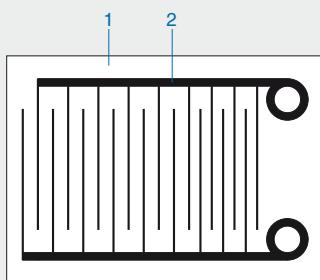
On capacitive humidity sensors, a hygroscopic insulating layer (e.g. Al_2O_3 , or a polymer plastic), which can possibly also

1 Temperature/humidity diagram for air

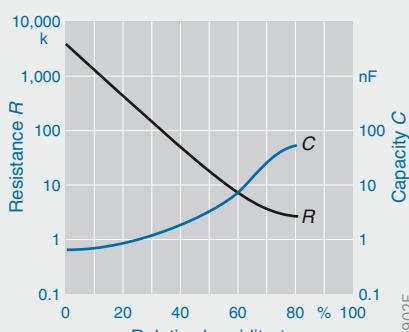


2 Humidity-measurement procedures. X = technologically important

Processes	Serial no.	Measuring device	Measuring method
Saturation processes	1	X Dew point hygrometer	Direct methods (measurement of absolute humidity)
	2	X LiCl dew point hygrometer	
Evaporation processes	3	X Psychrometer	
	4	Volume hygrometer	
	5	X Electrolysis hygrometer	
Absorption processes	6	Condensate volume hygrometer	
	7	X Infrared hygrometer	
	8	Microwave hygrometer	
	9	Electrical-discharge hygrometer	
Energetic processes	10	Diffusion hygrometer	
	11	X Electrical conductive-foil hygrometer	Indirect methods (measurement of relative humidity)
	12	X Capacitor hygrometer	
	13	X Hair hygrometer	
Hygroscopic processes	14	Bimetal strip hygrometer	
	15	Color hygrometer	
	16	Quartz hygrometer	
	17	Gravimetric hygrometer	

Table 2**2** Capacitive sensor wafer with comb electrodes

SAE0903Y

3 Resistive and capacitive humidity sensors (typical characteristic curve)

SAE0902E

be the mounting wafer, serves as the dielectric of a capacitor. Either one of the electrodes is permeable to water, or the electrodes have a comb-like structure (Fig. 2). With increasing humidity, the dielectric absorbs more water and the sensor's capacitance increases considerably (relative dielectric constant of water $\epsilon_{rW} \approx 81$, Fig. 3).

In the case of the resistive sensor, an insulation substrate is coated with a layer of hygroscopic salt (LiCl) held in a paste binder and located between an electrode pair. The layer's conductivity changes drastically along with the relative humidity (Fig. 3). Unfortunately this resistance change is also highly dependent upon temperature so that normally compensation is needed. The dew point can also be defined, and with it the absolute humidity, by an additional measurement of the air temperature. The typical time constants of these sensors are about 30 s.

Table 2 gives an overview of the numerous measurement processes that have been developed in the course of time for humidity measurement.

The piezoelectric effect

In 1880 Pierre Curie and his brother Jacques discovered a phenomenon that is still very little known today, but is present in the everyday lives of millions of people: the piezoelectric effect. For example, it keeps the hands of a quartz clock operating in time.

Certain crystals (e.g. quartz and tourmaline) are piezoelectric: Electric charges are induced on the crystal surface by exerting a compression or elongation force along certain crystal axes. This electrical polarization arises by shifting positive and negative ions in the crystal relative to each other by exerting force (see Fig., Item b). The shifted centers of charge gravity within the crystal balance each other out automatically, but an electric field forms between the end faces of the crystal. Compressing and elongating the crystal create inverse field directions.

On the other hand, if an electrical voltage is applied to the end faces of the crystal, the effect reverses (inverse piezoelectric effect): The positive ions in the electric field migrate toward the negative electrode, and negative ions toward the positive electrode. The crystal then contracts or expands depending on the direction of the electric field strength (see Fig., Item c).

The following applies to piezoelectric field strength E_p :

$$E_p = \delta \Delta x / x$$

$\Delta x / x$: relative compression or elongation

δ : piezoelectric coefficient, numeric values
 10^9 V/cm to 10^{11} V/cm

The change in length Δx when a voltage U is applied results from the following:

$$U / \delta = \Delta x \quad (\text{Example of quartz: deformation of approximately } 10^{-9} \text{ cm at } U = 10 \text{ V})$$

The piezoelectric effect is not only used in quartz clocks and piezo-in-line injectors, it has many other industrial applications, either as a direct or inverse effect:

Piezoelectric sensors are used for knock control in gasoline engines. For example, they detect high-frequency engine vibrations which is a feature of combustion knock. Conversion of mechanical vibration to an electric voltage is also used in the crystal audio pickup of a record player or crystal microphones. In the piezoelectric igniter (e.g. in a cigarette lighter), mechanical pressure produces the voltage needed to generate a spark.

On the other hand, if an alternating voltage is applied to a piezoelectrical crystal, it vibrates mechanically at the same frequency as the alternating voltage. Oscillating crystals of this type are used as stabilizers in electrical oscillating circuits or as piezoelectric acoustic sources to generate ultrasound, for instance.

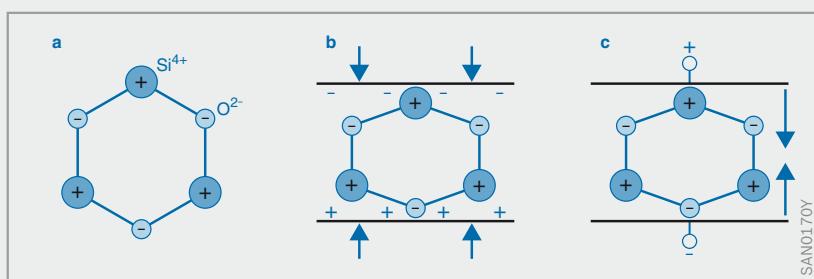
When used in clocks, the oscillating quartz is excited by an alternating voltage whose frequency is the same as the quartz's natural frequency. This is how an extremely time-constant resonant frequency is generated. In a calibrated quartz, it deviates by only approx. 1/1,000 second per year.

Principle of the piezoelectric effect
 (illustrated on the example of a unit cell)

a Quartz crystal SiO_2

b Piezoelectric effect:
 When the crystal is compressed, negative O^{2-} ions shift upward, positive Si^{4+} ions shift downward: electric charges are induced at the crystal surface.

c Inverse piezoelectric effect:
 When an electrical voltage is applied, O^{2-} ions shift upward, Si^{4+} ions shift downward: the crystal contracts.



Temperature sensors

Measured variables

Temperature is defined as a nondirectional quantity which characterizes the energy state of a given medium, and which can be a function of time and location:

$$T = T(x, y, z, t) \quad (1)$$

where: x, y, z are the spatial coordinates, t is time, and T is measured according to the Celsius or Kelvin scale.

Generally speaking, with monitored media which are in gaseous or liquid form, measurements can be taken at any point. In the case of solid bodies, measurement is usually restricted to the body's surface. With the most commonly used temperature sensors, in order for it to assume the medium's temperature as precisely as possible, the sensor must be directly in contact with the monitored medium (direct-contact thermometer). In special cases though, proximity or non-contacting temperature sensors are in use which measure the medium's temperature by means of its (infrared) thermal radiation (radiation thermometer = Pyrometer, thermal camera).

Generally speaking, a temperature sensor should reflect this dependency correctly, in other words, it should reflect as accurately as possible the local distribution of the temperature and its change as a function of time.

In special cases, possibly for functional reasons, this stipulation can be relaxed somewhat. The call for high local resolution and for high-speed response both demand that the sensor should be as small as possible, since it should not falsify the temperature readings by itself absorbing heat. In other words it should have a low thermal capacity.

In order to ensure that the temperature assumed by the sensor remains independent of the usually very different temperature of its mounting, it should be thermally well insulated from it. Since this also falsifies the measurement, the heat developed by the majority of sensors in the active state should be kept to a minimum (e.g. $< 1 \text{ mW}$).

The temperature sensor's dynamic response is specified by the time constant τ . This defines the time taken by the sensor to reach 63 %, 90 %, or 99 % of its final reading when subjected to a jump in temperature. This time depends not only upon the sensor's thermal capacity, but also to a great extent upon the heat-transfer coefficient between the sensor and the monitored medium. The higher it is, the faster the sensor reaches its final reading. Naturally, this figure is far higher for liquid mediums than for gaseous mediums. It must also be noted that the heat-transfer coefficient is highly influenced by any flow velocity v in the medium. It increases with approximately \sqrt{v} . In other words, the time constant of a temperature sensor should always be specified with reference to a specific flow velocity of a defined medium.

1 Temperature measuring points in the motor vehicle	
Measuring point	Temperature range °C
Intake/charge air	-40 to 170
External environment	-40 to 60
Interior	-20 to 80
Exhaust air/heating system	-20 to 60
Evaporator (air-conditioning system)	-10 to 50
Cooling water	-40 to 130
Engine oil	-40 to 170
Battery	-40 to 100
Fuel	-40 to 120
Tire air	-40 to 120
Exhaust gas	100 to 1,000
Brake caliper	-40 to 2,000

Table 1

Temperature measurement in the vehicle makes use almost exclusively of the temperature-dependence of electrical resistance material with positive (PTC) or negative (NTC) temperature coefficient. The direct-contact thermometers apply this phenomena. For the most part, the conversion of the temperature change to an analog voltage takes place by adding a second resistor to form a voltage divider (which also has a linearizing effect). Recently though, non-contacting (pyrometric) temperature sensing has been considered for occupant-protection (passenger-position monitoring for airbag triggering), and for comfort and convenience (air conditioning in accordance with skin-temperature measurement, prevention of windscreen mist-up). It needed the advent of microsystem technology for this (pyrometric) method to become feasible from the costs viewpoint. Table 1 presents a listing of the temperatures which have to be measured in the vehicle.

It is not only the highly differing measuring ranges that demand a variety of different sensor concepts and technologies, but the different accuracies and dynamic response requirements also lead to very different sensor shapes. At many locations, temperature is also measured in order that it can be compensated in cases where temperature variations trigger faults or act as an undesirable influencing variable.

Measuring principles for direct-contact sensors

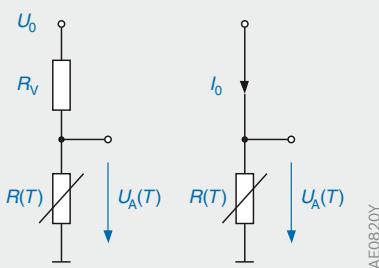
The fact that practically all physical processes are temperature-dependent means that there are almost just as many methods for making temperature measurements. The preferable methods though are those in which the temperature effect is very distinctive and dominant and as far as possible features a linear characteristic curve. Furthermore, the measuring elements should be suitable for inexpensive mass-production, whereby they should be adequately reproducible and non-aging. Taking these considerations into account, the following sensor techniques have come to the forefront, some of which are also applied in automotive technology:

Resistive sensors

In the form of 2-pole elements, temperature-dependent electrical resistors are particularly suitable for temperature measurement, no matter whether in wire-wound, sintered ceramic, foil, thin-film, thick-film, or monocrystalline form. Normally, in order to generate a voltage-analog signal they are combined with a fixed resistor R_V to form a voltage divider, or load-independent current is applied (Fig. 1). The voltage-divider circuit changes the original sensor characteristic $R(T)$ to a slightly different characteristic $U(T)$:

$$U(T) = U_0 \cdot \frac{R(T)}{R(T) + R_V} \quad (2)$$

1 Methods for resistance/voltage conversion



On the other hand, the application of a supply current I_0 permits the resistor curve to be reproduced exactly:

$$U(T) = I_0 \cdot R(T) \quad (3)$$

The measurement sensitivity is reduced more or less depending upon the construction of the voltage-divider circuit. Notwithstanding this fact, in the case of resistance characteristics which display a

Fig. 1

I_0	Current supply
U_0	Supply voltage
R_V	Temperature-independent series resistor
$R(T)$	Temperature-dependent measuring resistor
$U_A(T)$	Output voltage

slightly progressive curve it does have a linearization influence (an effect which is usually very welcome). Very often, in this connection, the auxiliary resistor is so dimensioned that it equals the measuring resistor at a given reference temperature T_0 (e.g. 20 °C):

$$R_V \approx R(T_0) \quad (4)$$

If the precision is inadequate, a resistance sensor can be calibrated to the setpoint value both with regard to its resistance (referred to a reference temperature), and its temperature coefficient (TC), by means of an adjustable parallel resistor R_P and a serial resistor R_S (Fig. 2). Of course, when fixed resistors are added the TC is reduced and the characteristic changes somewhat.

Sintered-ceramic NTC resistors

As a result of their pronounced measuring effects and inexpensive manufacture, the most common semiconductor resistors in use are based on heavy-metal oxides and

oxidized mixed crystals. These are sintered in bead form or disc form (Fig. 3) and have a polycrystalline structure. Because of their very steep temperature curve, they are also known as NTC resistors or thermistors. To a good approximation, and by applying the exponential law, their characteristic curve can be defined as follows:

$$R(T) = R_0 \cdot e^{B \cdot \left(\frac{1}{T} - \frac{1}{T_0} \right)} \quad (5)$$

Where $R_0 = R(T_0)$,
 $B = 2,000$ to 5,000 K = const,
 T absolute temperature

Here, the characteristic-curve gradient (TC) or the percentage resistance change as a function of temperature, is highly dependent upon the working point. This means that it can only be defined at given points:

$$\text{TC} = -B/T^2 \quad (6)$$

Fig. 2
a Resistance sensor
b Characteristic curves
1 Auxiliary contacts
2 Bridge
 R_{Ni} Nickel-film resistor
 $R_{tot}(T)$ Total resistance referred to temperature T
 R_P Adjustable parallel resistor
 R_S Adjustable series resistor

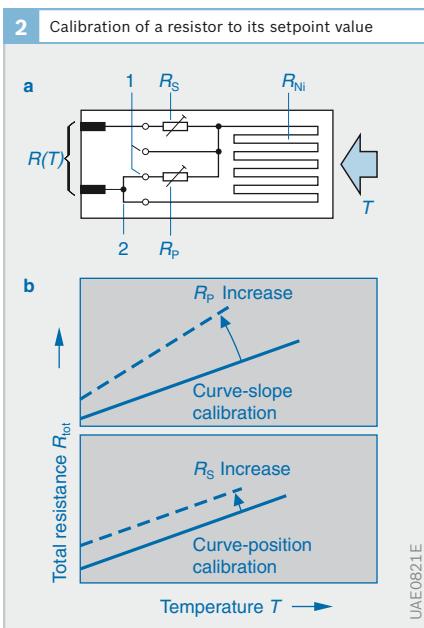
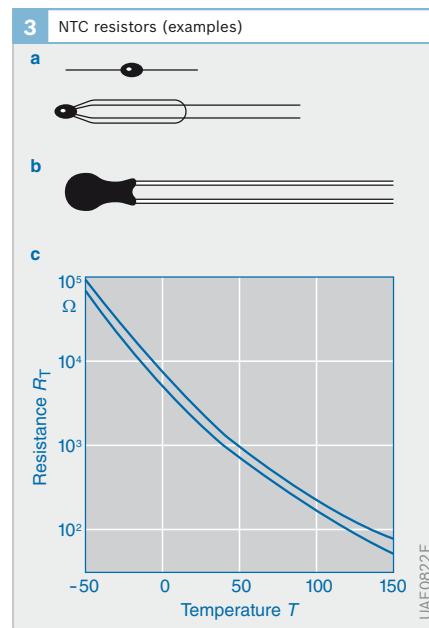


Fig. 3
a Pearl form
b Disk form
c Characteristic curve with limits of variation



It reduces considerably along with increasing temperature, and the resistance value itself frequently varies by 4 to 5 powers of ten, e.g. typically from a few 100 kΩ to a few 10 Ω. This pronounced temperature-dependence means that applications are restricted to a “window” of about 200 K. This though can be selected in the range from -40 to approximately 850 °C. Tighter tolerances of up to ± 0.5 K at a given reference point are complied with by using a selection process, or possibly even by grinding in oil, a method which of course has an affect upon costs. The aging stability of these sensors has been vastly improved compared to earlier versions which means that it is quite possible for the very close tolerances to apply throughout the sensor's useful life.

PTC thin-film/thick-film metallic resistors
 The thin-film metallic resistors placed with two additional temperature-neutral calibration resistors on a common substrate wafers have particularly high accuracy values. They can be tightly tolerated in respect of their characteristic curve and can be manufactured to be stable over the long term and are additionally trimmed by laser cuts. The use of layer technology makes it possible to adapt the substrate (ceramic, glass, plastic foil) and the upper layers (plastic molding or paint, sealed foil, glass and ceramic coatings) to the respective application, and thus provide protection against the monitored medium. Although metallic layers are less sensitive to thermal variations than the ceramic-oxide semiconductor sensors, both linearity and reproducibility are better. The following applies for the computational definition of these sensors:

$$R(T) = R_0 (1 + \alpha \cdot \Delta T + \beta \cdot \Delta T^2 + \dots) \quad (7)$$

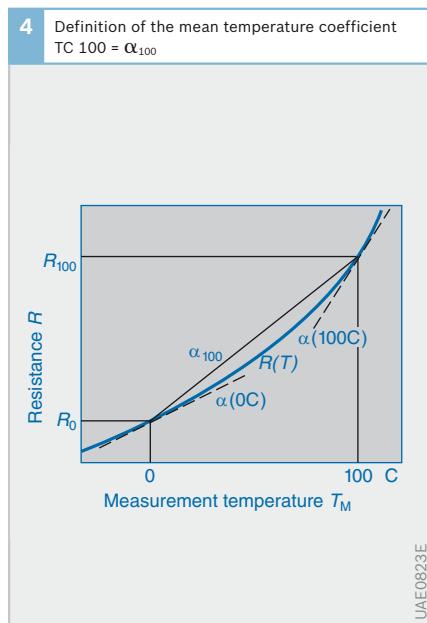
where $\Delta T = T - T_0$ and
 $T_0 = 20$ °C (reference temperature),
 α linear temperature coefficient (TC),
 β square temperature coefficient.

Even though the β coefficient is generally very small for metals, it cannot be ignored completely. This is why the measurement sensitivity of such sensors is usually characterized by means of a mean TC, the “TC 100”. The TC 100 corresponds to the mean curve gradient between 0 °C and 100 °C (Table 2 and Figure 4).

$$\text{Where } \text{TC 100} = \frac{R(100^\circ\text{C}) - R(0^\circ\text{C})}{R(0^\circ\text{C}) \cdot 100\text{K}} \quad (8)$$

2 Temperature coefficient TC 100			
Sensor material	TC 100 $10^{-3}/\text{K}$	Characteristic curve	Measuring range
Nickel (Ni)	5.1	slightly progressive	-60 to 320
Copper (Cu)	4.1	slightly progressive	-50 to 200
Platinum (Pt)	3.5	slightly degressive	-220 to 850

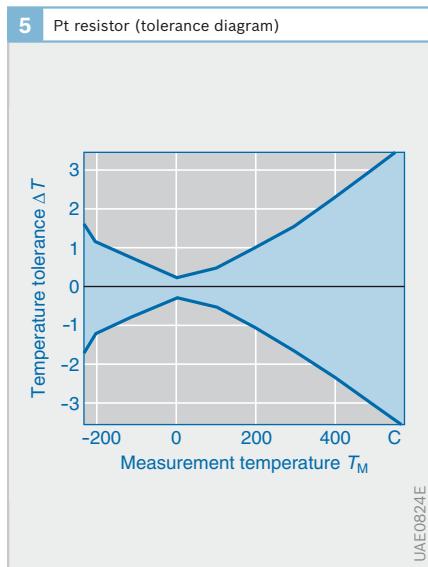
Table 2



Although platinum (Pt) resistors have the lowest TC, they are not only the most precise resistive temperature sensors, but can also boast the best aging stability. They are available on the market under the designation "PT 100" or "PT 1000" (100 Ω or 1,000 Ω nominal resistance at a reference temperature of 20 °C) in a variety of different tolerance classes (up to 0.1 °C) (Fig. 5). At temperatures up to about 1,000 °C, thick-film platinum sensors are suitable whose platinum layer has been stabilized by special additives.

Thick-film resistors (PTC/NTC)

Thick-film pastes with both high resistance (low surface-area requirement) and positive and negative temperature coefficients are generally employed as temperature sensors for compensation purposes. They have a non-linear characteristic curve (which though is not as "bent" as the curve of the solid NTC resistors), and are suitable for laser-trimming. The measuring effect is improved by forming voltage-divider circuits from NTC and PTC materials.

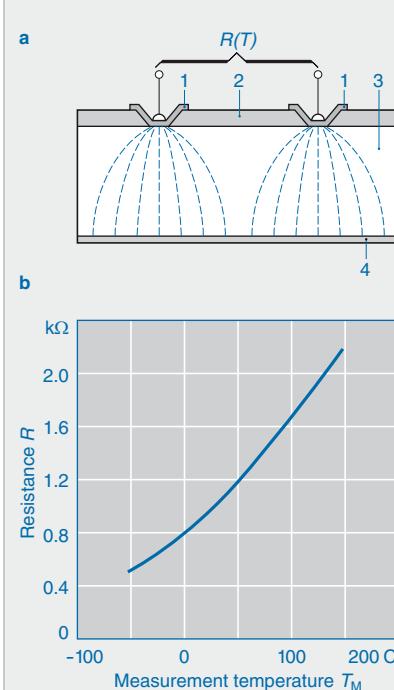
**Fig. 6**

- a Design
- b Characteristic curve
- 1 Contacts
- 2 Passivation (Nitride, oxide)
- 3 Si substrate
- 4 Counter-electrode without connection

$R(T)$ Temperature-dependent resistor

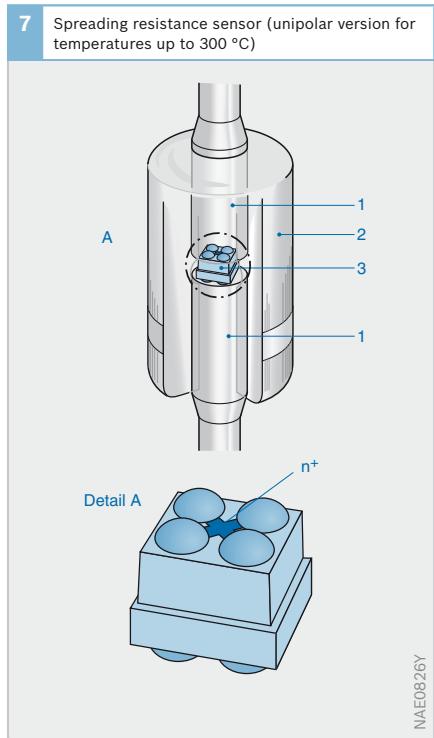
Monocrystalline silicon semiconductor resistors (PTC)

Basically speaking, with temperature sensors of monocrystalline semiconductor materials such as silicon (Si), it is possible to incorporate further active and passive circuit elements on the sensor chip (initial signal conditioning at the measuring point). Due to the closer tolerances involved, their production uses the "spreading resistance" principle (Fig. 6a). The current flows through the measuring resistor and through a surface-point contact before arriving at the silicon bulk material. It then proceeds, widely distributed, to a counter-

6 Spreading-resistance principle (dual-hole version)

electrode covering the base of the sensor chip. In addition to the high reproducibility material constants, the high current density behind the contact point (high accuracy thanks to photolithographic production) mainly defines the sensor's resistance. In order to make the sensor highly independent of polarity, these sensors are usually series-connected in pairs in opposite alignments (dual-hole version, Fig. 6). The base electrode can be in the form of a temperature contact (no electrical function).

Measurement sensitivity is practically double that of a Pt resistor ($\text{TC} = 7.73 \cdot 10^{-3}/\text{K}$). The progressive bend of the temperature curve is more pronounced than on a metallic sensor. The measuring range is limited to approx. +150 °C by the material's intrinsic conductivity (Fig. 6b). There are special versions (Fig. 7) available for uses up to 300 °C.



Thermocouples

Thermocouples are used in particular for measuring ranges $\geq 1,000$ °C. They rely on the Seebeck effect, according to which there is a voltage across the ends of a metallic conductor when these are at different temperatures T_1 and T_2 . This "thermoelectric voltage" U_{th} depends (independently of the development of this) exclusively on the temperature difference ΔT at the ends of the conductors (Fig. 8). It is expressed by the equation:

$$U_{\text{th}} = c (T_2 - T_1) = c \Delta T, \quad (9)$$

where the proportionality constant is material-specific and termed the Seebeck coefficient.

Since the instrument leads used to measure this voltage across the metallic conductor must themselves be equipped with terminals (for instance made of copper),

8 Seebeck effect

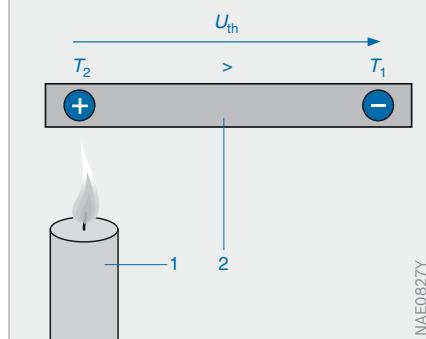


Fig. 8

- 1 Heat source
- 2 Metallic conductor
- + High,
- low thermal velocity of the electrons
- T_2 High temperature
- T_1 Low temperature
- U_{th} Thermoelectric voltage

3 Thermoelectric voltage U_{th} of a number of metals

Material	Thermoelectric voltage U_{th} mV/100 °C
Constantan	-3.40
Nickel	-1.90
Paladium	-0.28
Platinum	0.00
Copper	+0.75
Manganite	+0.60
Iron	+1.88
Silicon	+44.80

Fig. 7

- 1 Metal wire
- 2 Glass
- 3 Si crystal

Table 3

these are also subject to the same temperature difference, so that unfortunately only the difference between the metallic conductor and the connecting cables is measured. Thermoelectric voltages are always listed based on Platinum as the reference material (Table 3).

In order that the generated voltages are as high as possible, a number of material pairs have established themselves (Fig. 9, e.g. iron/constantan etc.). It is important that the “limbs” of such a thermocouple are joined at the end to which the heat is applied in such a manner that the joint is electrically conductive (by means of twisting, welding, soldering, etc., Fig. 10).

Thermocouples themselves are usually short, and the extensions up to the point where the signal is captured can be made with compensating cables which use the same material pair as in the thermocouple itself. It is important that both free ends

of the thermoelement configuration are at the same (reference) temperature, otherwise the temperature difference at the free ends will also be included in the measurement. Thermocouples, therefore, always measure only the temperature difference to a given reference point. If the measuring point's absolute temperature is to be measured, other devices (such as resistive sensors) must be used to also measure the temperature at the reference point.

The characteristic curve of the thermoelectric voltage against temperature is usually not as linear as that given in equation (9). The sensor signals are usually small, and ICs are used for their amplification and for their linearization. In order to increase the measurement voltage, it is common practice to connect a number of identical thermocouples in series. These have their “hot” junctions at the temperature to be measured, and their “cold” junctions at the reference temperature (Fig. 11, thermopile).

Fig. 9

- 1 Copper/Constantan
- 2 Iron/Constantan
- 3 Nickel-chromium/Nickel
- 4 Platinum rhodium/Platinum

Fig. 10

- A/B Material pair (thermocouple legs)
- 1 Measuring point (electrically conductive junction)
 - 2 Connection head
 - 3 Compensating cable
 - 4 Reference point
 - 5 Connection cable (Cu)
- T_M Measurement temperature
 T_R Reference temperature
 U_{th} Thermoelectric voltage

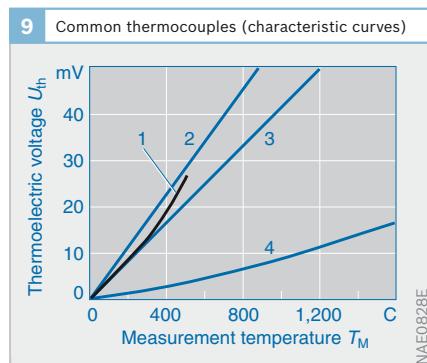
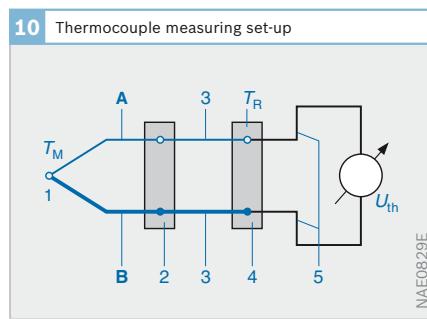
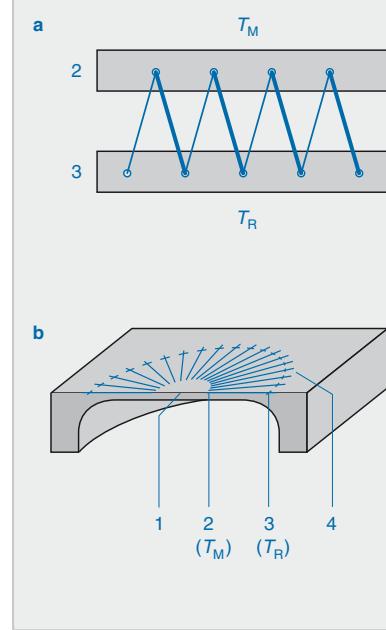


Fig. 11

- a Principle of the thermopile
 - b Example of application
- 1 Sensitive surface
 - 2 “Hot” junctions at the measurement temperature T_M
 - 3 “Cold” junctions at the reference temperature T_R
 - 4 Thermopile



11 Thermocouples connected in series



Although thermocouples are robust (for instance, high-level EMC due to low internal resistance), they are not particularly accurate as a measuring device. Their deviation can easily be in the 5° to 15 °C range, and they are not outstanding regarding their resistance to aging, which means that individual calibration does not result in a permanent improvement of their accuracy.

Of course, thermocouples can be manufactured using both thin-film and thick-film techniques. Metallic films stacked one on top of the other provide for excellent thermal contact, and extremely small thermocouples can be produced by applying microsystem technology. Thermocouples are particularly suitable for use in thermopiles comprised for instance from 50 to 100 individual thermocouples. They are used in non-contacting radiation thermometers (pyrometers).

Semiconductor depletion layers

Presuming a constant current, the forward voltage of semiconductor depletion layers (Fig. 12) such as those in diodes and in the basis-emitter path of a transistor, demonstrate very good linearity as a function of temperature:

$$U_F(T) = \frac{k \cdot T}{q} \cdot \ln\left(\frac{I_F}{I_{sat}} + 1\right) \quad (10)$$

Where:

$I_{sat} = I_{sat}(T)$ and $I_F = \text{constant}$,
 $q = 1.6 \cdot 10^{-19} \text{ C}$ (elementary charge),
 $k = 1.88 \cdot 10^{-23} \text{ JK}^{-1}$ (Boltzmann's constant),
 T absolute temperature.

It is advantageous here when the sensor directly outputs a voltage which is a function of the temperature. The two-pole sensor is of course dependent upon polarity. Whereas, for all sensors, the forward voltage decreases by almost exactly 2 mV/°C, the absolute voltage at each depletion layer differs considerably from sensor to

sensor, and may necessitate additional calibration elements in order for a precise measurement to be made. Above all, it is the temperature-dependent saturation current I_{sat} which is responsible for the negative temperature coefficient (NTC). This current increases dramatically with rising temperature. Here, the silicon's intrinsic conductivity limits this sensor's application to temperature ranges <150 °C.

Sometimes, emitter-coupled transistor pairs are used in a similar manner for temperature measurement. With this form of temperature measurement, the ratios of the collector currents to each other represent a very good reproducible measure for the temperature. Usually, an integrated supplementary circuit is used for the "on-chip" conversion to an analog output voltage.

Zener diodes operated in the reverse direction can also be used as highly practical temperature sensors. Their voltage changes are highly dependent upon the Zener voltage. Here, the option exists of various levels of voltage reduction at Zener voltages <4.7 V, and voltage increases at Zener voltages >4.7 V.

Such sensors are often used for temperature compensation on the chip itself.

12 Semiconductor depletion layers

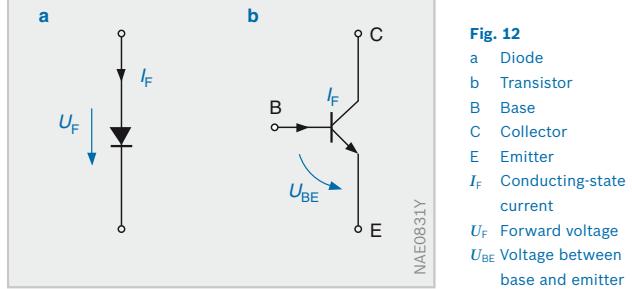


Fig. 12

- a Diode
- b Transistor
- B Base
- C Collector
- E Emitter
- I_F Conducting-state current
- U_F Forward voltage
- U_{BE} Voltage between base and emitter

Measuring principles for non-contacting temperature measurement

The radiation emitted by a body is used for the non-contact measurement (pyrometry) of its temperature. This radiation is for the most part in the infrared (IR) range (wavelength: 5 to 20 μm). Strictly speaking, the product of the radiated power and the emission coefficient of the body is measured. The latter is a function of the material, but for materials which are technically of interest (including glass) it is usually around 1, although for reflective and IR-permeable materials (e.g. air, silicon) it is far less than 1.

The measuring point is reproduced on a radiation-sensitive element, which heats up slightly with respect to its environment (typically 0.01 to 0.001 $^{\circ}\text{C}$). The element's temperature is a measure for the tempera-

ture of the body being measured. A given temperature difference at the object often corresponds to only 1/1,000th of this difference at the measuring point. Nevertheless, the object's temperature can be determined with an accuracy of 0.5 $^{\circ}\text{C}$.

Bolometer

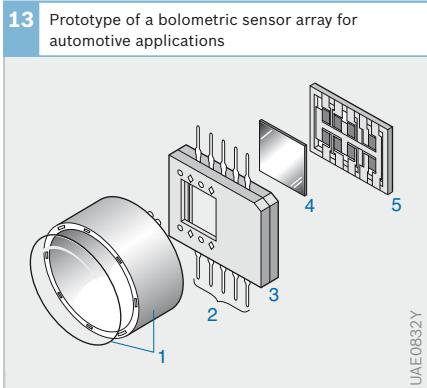
The Bolometer is a highly sensitive resistance temperature sensor for measuring minute temperature increases (Fig. 13). A further sensor is needed for measuring the temperature of the sensor housing. If this device is to operate efficiently across a wide temperature range, it is necessary though for both these sensors to feature an extremely high degree of synchronism. The Bolometer housing is therefore usually thermostatically controlled (and well-insulated to the housing) so that the primary detecting element (sensor) always operates at the same temperature.

Fig. 13

- 1 Lens housing with lens
- 2 Connections
- 3 Infrared detector housing
- 4 Infrared window
- 5 Detector

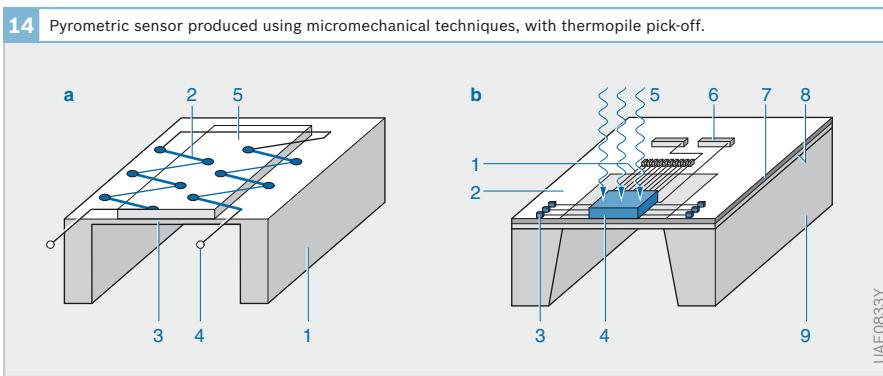
Fig. 14

- a Principle of the measuring cell
- 1 Silicon chip
- 2 Thermocouples connected in series (e.g. Al/poly-Si)
- 3 SiN diaphragm
- 4 Thermopile terminals
- 5 Absorber layer
- b Sensor type
- 1 Thermocouple
- 2 "Cold" junction
- 3 Diaphragm
- 4 Absorber
- 5 Heat radiation
- 6 Electrical connection
- 7 Si_3N_4 layer
- 8 SiO_2 layer
- 9 Heat sink



Thermopile sensor

When a very extensive temperature range is concerned, it is more practical for the temperature difference generated by the radiation from the object to be measured using thermocouples. In order to increase the measuring effect, a number of thermocouples are connected in series (thermopile). Such a thermopile sensor (Fig. 14) is inexpensive to manufacture micromechanically. All its "hot" junctions are located on a thermally well insulated thin diaphragm,



and all its "cold" junctions are in contact with the thicker chip rim (heat sink). Typically, the sensor's response time is approx. 20 ms. Using such a so-called "single-pixel sensor", it is an easy matter to determine the windshield's surface temperature so that measures can be taken to prevent misting should the temperature fall below the dew point.

Single-point sensors, imaging sensors
 If a number of pixels are combined on a single chip (e.g. 4 x 4) to form an array, this provides the basis for a rough form of image analysis (Fig. 16). However, there should not be too much insensitive surface area between the pixels and the pixels themselves must be thermally well insulated against each other. As all the pixels can optionally respond electrically, the chip has a large number of terminals. On a TO5 housing, for example, the ASIC for signal preamplification and series connection of the signal must be located directly adjacent to the sensor chip. Usually, in the case of thermopile sensors, this ASIC also includes a reference-temperature sensor which measures the pixel's absolute temperature. This permits object temperatures to be measured with an accuracy of approx. ± 0.5 K.

In order to reproduce a scene thermally on the sensor array, the array requires an IR optical imaging unit. The very inexpen-

sive curved mirror is usually ruled out due to it needing too much room. Glass lenses are impermeable for IR light, and plastic lenses can only be used for operating temperatures of up to approx. 85 °C max. On the other hand, Si lenses are highly suitable for thermal radiation and up to diameters of approx. 4 mm micromechanical techniques can be used to inexpensively manufacture them in the form of a Fresnel or refraction lens. Fitted in the cover of a TO5 housing, these then also serve to protect the sensor against direct damage (Fig. 15). Even though filling the housing with an inert gas improves the crosstalk between the individual pixels, it also negatively affects their response time.

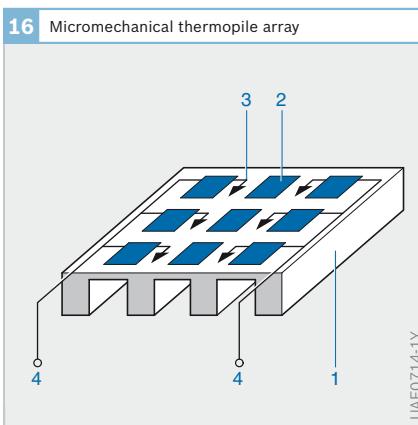


Fig. 16
 1 Silicon chip
 2 Pixel
 3, 4 Pixel connections
 UAE0714-1Y

15 Thermal image analysis

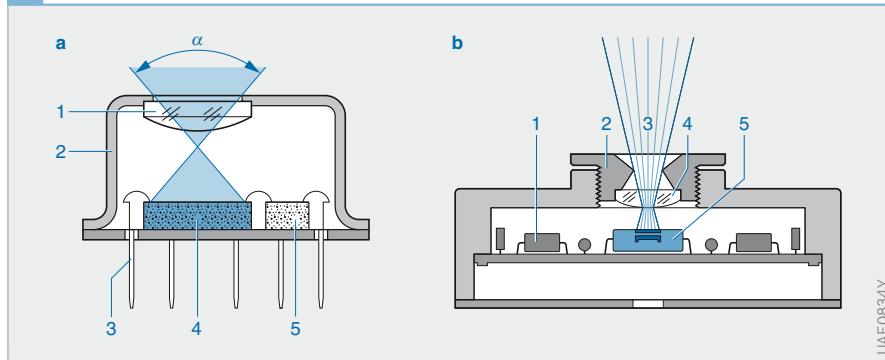


Fig. 15
 a IR imaging sensor
 b Simple IR camera
 1 Electronics
 2 Lens system
 3 Camera's field of view
 4 Si IR lens
 5 Sensor array
 α Viewing angle
 UAE0834Y

Imaging sensors (video)

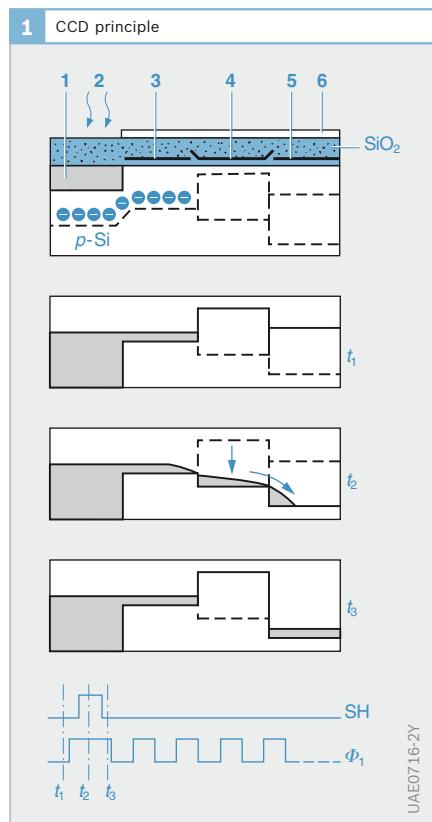
In particular, imaging sensors are beginning to gain a hold in the motor vehicle using visible light or infrared light. They can be used for passenger-compartment monitoring, but are principally aimed at observation outside the vehicle.

All of these sensors have one objective in view, and that is the simulation of the superior capabilities of the human eye and its mental recognition capabilities (of course, only to a very modest degree at first). These were introduced in large numbers to industrial measurements some time ago - in particular on robot handling equipment. The costs of imaging sensors, and the associated very high-performance processors needed for the interpretation of a scene, are already of interest for applications in the automotive sector.

In contrast to the human eye, common imaging sensors are sensitive in the near IR range (wavelength approximately $1 \mu\text{m}$). With appropriate non-visible IR illumination, therefore, all imaginable applications in the vehicle become feasible, including nighttime operation. For applications in the visible light range, the imaging sensors generally have an IR filter in front of them to prevent false recordings due to color.

In future, imaging sensors will be able to play a highly variegated role for the observation of the vehicle's interior (seating position, forward shift in case of a crash, size of seat occupants etc.), and of the vehicle's surroundings (lane guidance, collision prevention, parking and reversing aids, road-sign recognition). A night vision device, that takes the IR image recorded for the road and displays it on a screen in the dashboard, thus offering enhanced driving safety in poor visibility (darkness, fog, etc.) is already in series production.

Imaging sensors are a special case of multisensor structures comprising light-sensitive elements (photodiodes), arranged as pixels in a row or matrix form and receiving the light through a conventional imaging optical system. In the case of the silicon imaging sensors principally available (CCD, Charge-Coupled Devices, Figs. 1 and 2), free charge carriers are generated proportionally to the intensity and exposure time by the light arriving through a transparent electrode; they are then collected in a "potential pot" ($\text{Si}-\text{SiO}_2$ boundary layer). Further electrodes move these charges into an opaque zone which are further transported in analog shift registers (bucket chain principle) line-by-line into an output register which is read serially at a high clock rate.

**Fig. 1**

- 1 Photodiode
- 2 Light
- 3 Storage electrode
- 4 Shift gate
- 5 Transfer electrode
- 6 Optical masking

While CCD sensors can only be used in motor vehicles to a very limited degree because of their limited brightness/darkness dynamics (50 dB), their read time and their temperature range ($< 50^{\circ}\text{C}$), newer, smart, imaging sensors based on CMOS technology appear to be fully suitable for use in the motor vehicle. The logarithmic brightness/signal curve possible here, corresponds to the human eye and has a dynamic response of 120 dB. This makes aperture control, for instance, unnecessary and supplies a constant contrast resolution across the entire brightness range. These sensors permit a digitally controlled, random access to the individual pixels in a matrix sensor (Fig. 3) with simultaneous higher sensitivity (higher read rate).

The first steps in preprocessing the signals on the imaging-sensor chip have already been implemented (intelligent imaging sensor). A resolution of $1,280 \times 1,024$ pixels with typical individual dimensions of $6 \mu\text{m} \times 6 \mu\text{m}$ is already possible in sensors currently on the market.

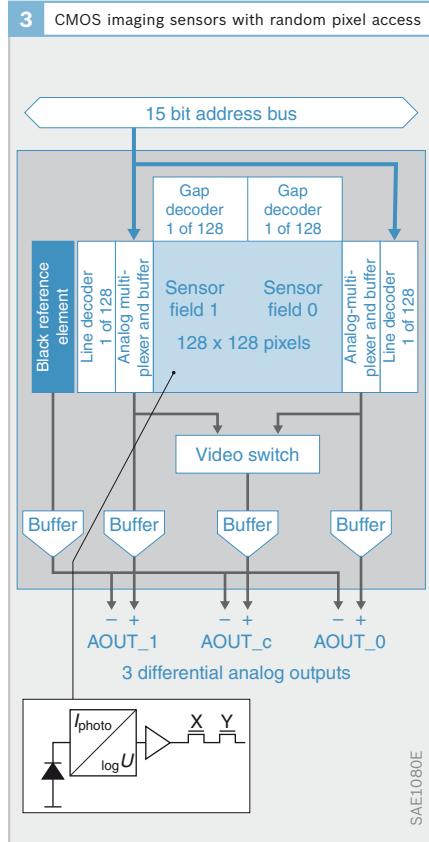


Fig. 3
Pixel processors at each pixel in the active pixel matrix ensure logarithmic signal compression.
Reading is random access and non-destructive to the signal, transmission is loss-free.

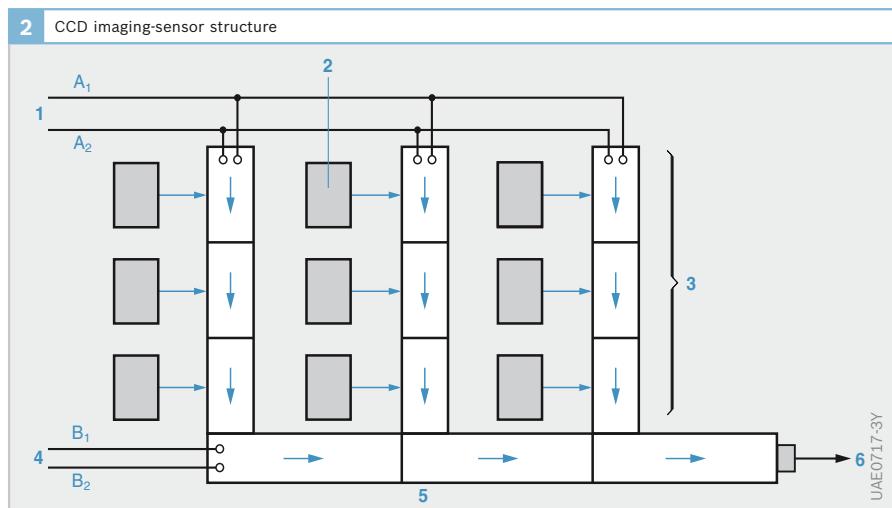


Fig. 2
1 Column clock pulse A_1/A_2
2 Photosensors
3 CCD array
4 Line clock pulse B_1/B_2
5 Output register
6 Video output

Sensor types

Engine-speed sensors

Application

Engine-speed sensors are used in engine-management systems for

- ▶ Measuring the engine speed and
- ▶ Determining the crankshaft position (position of the pistons)

The engine speed is calculated from the interval between the speed sensor's signals.

Inductive speed sensors

Design and operating principle

The sensor is mounted directly opposite a ferromagnetic pulse wheel (Fig. 1, Item 7) from which it is separated by a narrow air gap. It has a soft-iron core (pole pin, 4), which is enclosed by a winding (5). The pole pin is also connected to a permanent magnet (1), and a magnetic field extends through the pole pin and into the pulse wheel. The level of the magnetic flux through the coil depends on whether the sensor is opposite a pulse-wheel tooth or space. Whereas the magnet's leakage flux is concentrated by a tooth, and leads to an increase in the magnetic flux through the coil, it is attenuated by a space. When the pulse wheel rotates, these changes in magnetic flux induce a sinusoidal output voltage in the coil which is proportional to the rate of change of the flux and thus the

engine speed (Fig. 2). The amplitude of the alternating voltage increases strongly along with increasing pulse-wheel speed (several mV to >100 V). At least about 30 rpm are needed to generate an adequate amplitude.

The number of teeth on the pulse wheel depends on the particular application. In Motronic systems, a 60-pitch pulse wheel is normally used, although 2 teeth are omitted (Fig. 1, Item 7) so that the pulse wheel has $60 - 2 = 58$ teeth. The space where the missing teeth would be situated is allocated to a defined crankshaft position and serves as a reference mark for synchronizing the control unit.

The geometries of the pulse-wheel teeth and the pole pin must be matched to each other. An evaluation circuit in the control unit converts the sinusoidal voltage, which is characterized by strongly varying amplitudes, into a constant-amplitude square-wave voltage for evaluation in the control unit microcontroller.

Active speed sensors

Active speed sensors operate according to the magnetostatic principle. The amplitude of the output signal is not dependent on the rotational speed. This makes it possible for very low speeds to be sensed (quasistatic speed sensing).

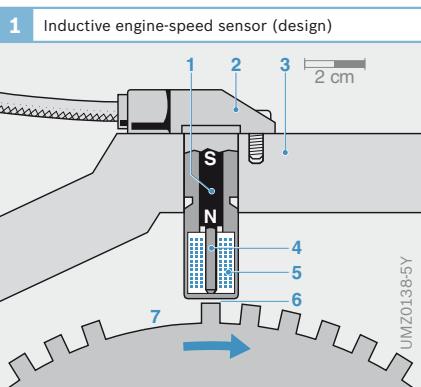
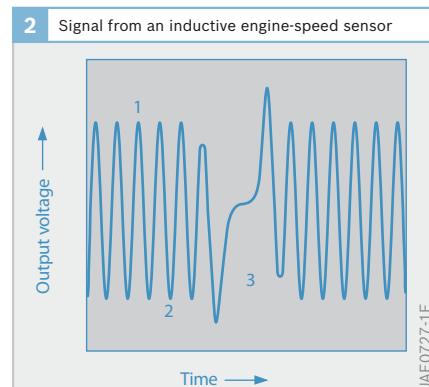


Fig. 1

- 1 Permanent magnet
- 2 Sensor housing
- 3 Crankcase
- 4 Pole pin
- 5 Winding
- 6 Air gap
- 7 Pulse wheel with reference mark



UAE07/7-1E

Fig. 2

- 1 Tooth
- 2 Tooth space
- 3 Reference mark

Differential Hall sensor

A voltage U_H proportional to the magnetic field (Hall voltage) can be picked off horizontally to the current direction at a current-carrying plate which is permeated vertically by a magnetic induction B . In a differential Hall sensor, the magnetic field is generated by a permanent magnet (Fig. 3, Item 1). Two Hall sensor elements (2 and 3) are situated between the magnet and the pulse wheel (4). The magnetic flux by which these are permeated depends on whether the speed sensor is opposite a tooth or a space. Determining the difference between the signals from two sensors reduces the magnetic interference signals and achieves an improved signal/noise ratio. The edges of the sensor signal can be processed without digitization directly in the control unit.

Multipole wheels are used instead of the ferromagnetic pulse wheel. Here, a magnetizable plastic is attached to a nonmagnetic metallic carrier and alternately magnetized. These north and south poles adopt the function formerly performed by the teeth of the pulse wheel.

AMR sensors

The electrical resistance of magnetoresistive material (AMR, Anisotropic Magneto Resistive) is anisotropic, i.e. it depends on the direction of the magnetic field to which it is exposed. This property is utilized in an AMR sensor. The sensor is located between a magnet and a pulse wheel. The field lines change direction when the pulse wheel rotates (Fig. 4). This generates a sinusoidal voltage, which is amplified in an evaluation circuit in the sensor and converted into a square-wave signal.

GMR sensors

The use of GMR technology (Giant Magneto Resistive) is a further development of active speed sensors. Because of the higher sensitivity by comparison with the AMR sensors, larger air gaps are possible, which makes applications in difficult circumstances possible. The higher sensitivity also results in less background noise from the signal edges.

All two-wire current interfaces previously used for Hall speed sensors may also be used on GMR sensors.

Fig. 3

- a Configuration
- b Signal of Hall sensor
 - high amplitude with small air gap
 - low amplitude with large air gap
- c Output signal

- 1 Magnet
- 2 Hall sensor 1
- 3 Hall sensor 2
- 4 Pulse wheel

4 Principle of speed sensing with an AMR sensor

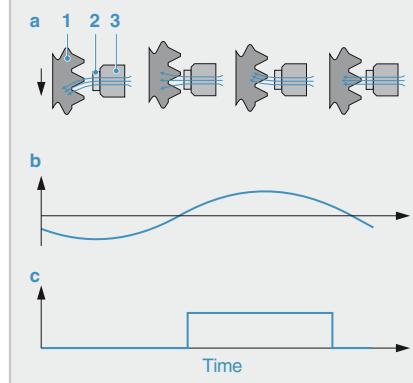
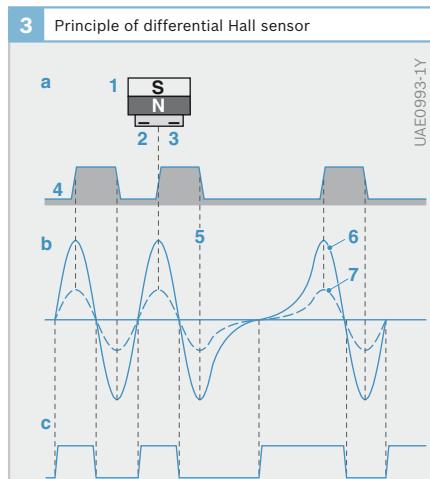


Fig. 4

- a Configuration at various times
- b Signal from AMR sensor
- c Output signal

- 1 Pulse wheel
- 2 Sensor element
- 3 Magnet



Hall phase sensors

Application

The engine's camshaft rotates at half the crankshaft speed. Taking a given piston on its way to TDC, the camshaft's rotational position is an indication as to whether the

piston is in the compression or exhaust stroke. The phase sensor on the camshaft provides the control unit with this information. This is required, for example, for ignition systems with single-spark ignition coils and for sequential fuel injection (SEFI).

Design and operating principle

Hall rod sensors

Hall rod sensors (Fig. 1a) utilize the Hall effect: a rotor of ferromagnetic material (7, pulse wheel with teeth, segments or aperture plate) rotates along with the cam-shaft. The Hall IC (6) is located between the pulse wheel and a permanent magnet (5), which generates a magnetic field strength perpendicular to the Hall element.

If one of the pulse-wheel teeth (Z) now passes the current-carrying sensor element (semiconductor wafer), it changes the magnetic-field strength perpendicular to the Hall element. This results in a voltage signal (Hall voltage) which is independent of the relative speed between sensor and pulse wheel. The evaluation electronics integrated in the sensor's Hall IC condition the signal and output it in the form of a square-wave signal (Fig. 1b).

Fig. 1

- a Positioning of sensor and single-track pulse wheel
- b Output signal characteristic U_A
- 1 Electrical connection (plug)
- 2 Sensor housing
- 3 Crankcase
- 4 Sealing ring
- 5 Permanent magnet
- 6 Hall IC
- 7 Pulse wheel with tooth/segment (Z) and space (L)
- a Air gap
- φ Angle of rotation

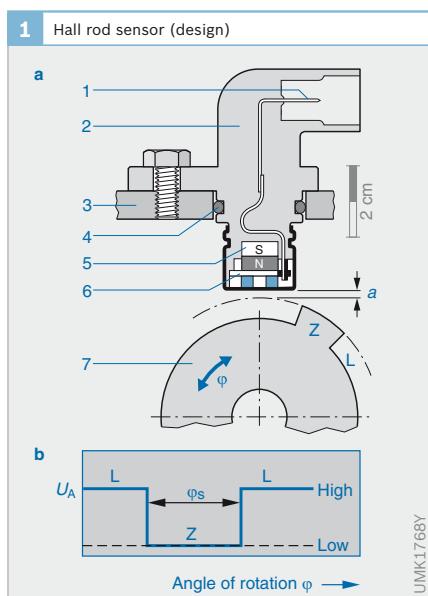
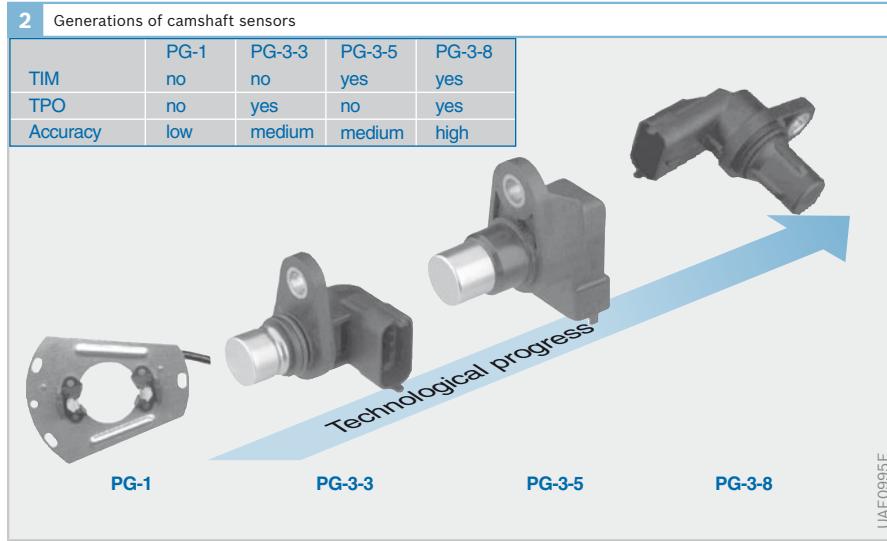


Fig. 2

- TIM = Twist Intensive Mounting (i.e. the sensor may be twisted around its axis any number of times, without losing accuracy. Important for minimizing type diversity).
- TPO = True Power On (i.e. the sensor detects directly on switching on whether it is located over a tooth or a space. Important for short synchronization times between crankshaft and camshaft signals).

2 Generations of camshaft sensors

	PG-1	PG-3-3	PG-3-5	PG-3-8
TIM	no	no	yes	yes
TPO	no	yes	no	yes
Accuracy	low	medium	medium	high



Speed sensors for transmission control

Application

Transmission speed sensors determine the shaft speeds in AT, ASG, DKG and CVT transmissions. These are the turbine and output shaft speeds on AT transmissions with a hydrodynamic torque converter, the speed of the primary and secondary pulleys in CVT transmissions and the speeds of the two input shafts and the driven shaft of double-clutch transmissions. The speed of the power take-up element is also detected for high dynamic demands on power take-up control.

Direction of rotation detection can also be necessary in high-end transmissions in order to optimize clutch management and for rollback prevention.

Both stand-alone sensors and sensors built into electronics modules which project into the transmission from the outside or are mounted internally are used.

Requirements

Transmission speed sensors are exposed to very severe operating stresses because of

- ▶ Extreme ambient temperatures between -40 and +150 °C
- ▶ An aggressive operating environment caused by the transmission oil, also known as ATF (contains special additives for transmissions and has a low condensate content)
- ▶ High mechanical stress with vibrational accelerations up to 30 g
- ▶ Abraded metallic materials and a build-up of particles in the transmission

The package of the electronics used in the sensor is hence subject to high demands as a consequence of these severe stresses. A service life of more than 15 years in transmission oil has been achieved by using a suitable oil-resistant package.

Because of the very compact design of the transmissions, the mechanical customer interface cannot generally be covered by standard geometries. This means that custom sensor types are needed for each transmission model. Module-integrated types differ in respect of insertion length, direction of recording and mounting flange (Fig. 1). A further variation found in stand-alone sensors is the location of the mounting socket and the design of the connector.

Hall ASICs (Application Specific Integrated Circuit) with varying complexities of the evaluation algorithms are used to cover the complete range of the functional requirements (Fig. 2).

If a ferromagnetic trigger wheel or a trigger area (toothed, stamped or embossed) is available on the rotating transmission component, the magnetic field required to operate the Hall sensor is generated by a back-bias magnet.

1 Sensor types

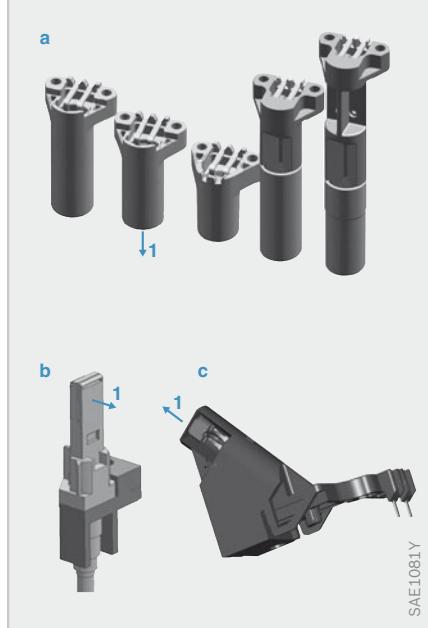


Fig. 1

a Bottom read

b Side read

c Slant read

1 Direction of detection

It is mounted in the sensor immediately behind the ASIC.

Compact transmissions increasingly require speed measurement over greater distances (magnetic air gap) by rotating, nonmagnetic components or through a housing wall. Multipole wheels (magnetized rings) are used for these applications, the back-bias magnet is omitted from the sensor.

Design

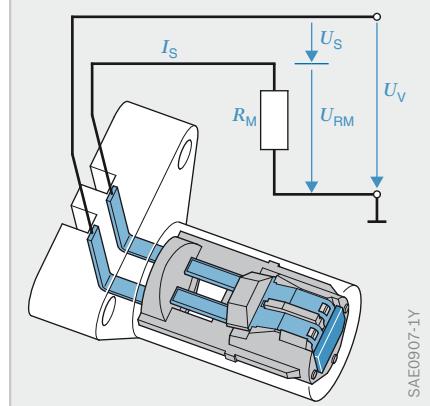
The Hall ASICs used in the transmission speed sensors are mounted in a retainer with or without a back-bias magnet – depending on the magnetic interface – with the electrical contact being made by a welding process, then inserted in a housing, potted with epoxy resin or – in the case of types mounted on the outside of the transmission – have an oil-tight enclosure sprayed onto them (Fig. 3).

The sensor has a two-wire interface, which combines optimum diagnostic options with a minimum number of electrical connections. The two terminals both supply the Hall IC and transmit the signal.

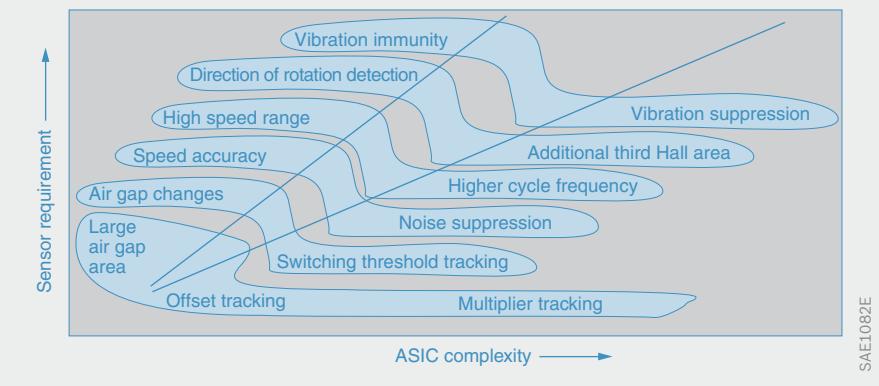
Operating principle

Transmission speed sensors use the differential Hall effect. The difference between the Hall voltages from two Hall plates on the ASIC is determined. The majority of the similar interferences can be compensated for in this way. The difference signal is first amplified on some ASIC types, then converted into a digital signal by trigger algorithms of varying complexities. This constitutes the controlled variable for modulating the output current by means of a power source. A digital signal with two current levels (typically 7 mA for the

3 Hall sensor with two-wire current interface



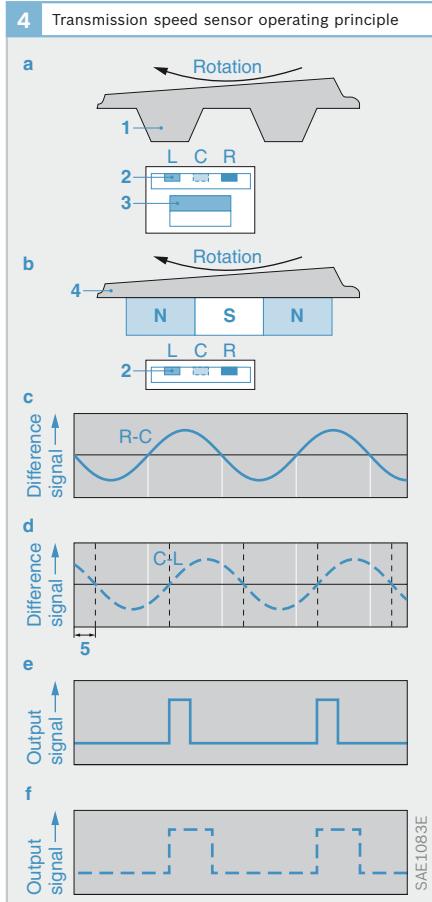
2 Requirements complexity



low level and 14 mA for the high level) is obtained, the modulation frequency of which corresponds to the frequency of tooth change on the trigger wheel and thus represents the speed of rotation. The sensor signal is evaluated in the electronic control unit by means of a measuring shunt R_M , which converts the sensor current I_S into the signal voltage U_{RM} .

On principle, it does not matter for the operation of a differential Hall ASICS whether the sensor is operated with a steel trigger wheel or on a multipole wheel (Figs. 4a and 4b).

Some transmission controls include functions that require detection of a standstill. For this application, the sensor must have as great an immunity as possible against vibration-induced variations in the air gap and the torsional vibration of the trigger wheel. This property of the sensor, known as immunity against vibration, can only be implemented to a very limited degree in differential Hall sensors with only two Hall plates, by the use of adaptive trigger thresholds, for example. Two phase-shifted differential signals become available with the use of a third Hall plate. This makes both the detection of the direction of rotation (Figs. 4c to f) and additional function algorithms for increasing immunity against vibration possible.



The typical values for “value” and “high feature” sensors differ in the air gap range achievable (distance between the sensitive area on the sensor and the trigger wheel), the signal frequency range and the additional functionalities implemented (Table 1).

The complexity of transmission type, mounting space restrictions including all the derived marginal design conditions and the functional requirements leads to application-specific solutions in most applications. These are characterized by a combination of ASIC, package design and sensor mechanical and magnetic interface suitable for the system requirements.

1 Typical characteristics		
Type	Value	High feature
Maximum air gap at trigger wheel	2.5 mm	3.5 mm
at pole wheel	5 mm	7 mm
Signal frequency	0 to 8 kHz	0 to 12 kHz
Direction of rotation detection	No	Yes
Trigger wheel vibration	-	± 1.5°

1 Trigger wheel
2 Hall plates L and R (C optional for direction of rotation detection)
3 Permanent magnet (back-bias)
4 Multipole wheel
5 Phase shift dependent on the direction of rotation

Table 1

Wheel-speed sensors

Application

Wheel-speed sensors are used to measure the rotational speed of the vehicle wheels (wheel speed). The speed signals are transmitted via cables to the ABS, TCS or ESP control unit of the vehicle which controls the braking force individually at each wheel. This control loop prevents the wheels from locking up (with ABS) or from spinning (with TCS or ESP) so that the vehicle's stability and steerability are maintained.

Navigation systems also use the wheel-speed signals to calculate the distance traveled (e.g. in tunnels or if satellite signals are unavailable).

Design and operating principle

The signals for the wheel-speed sensor are generated by a steel pulse generator that is fixed to the wheel hub (for passive sensors) or by a multipole magnetic pulse generator (for active sensors). This pulse generator has the same rotational speed as the wheel and moves past the sensitive area of the sensor head without touching it. The sensor "reads" without direct contact via an air gap of up to 2 mm (Fig. 2).

The air gap (with strict tolerances) ensures interference-free signal acquisition. Possible interference caused for instance by oscillation patterns in the vicinity of the brakes, vibrations, temperature, moisture, installation conditions at the wheel, etc. is therefore eliminated.

Since 1998 active wheel-speed sensors have been used almost exclusively with new developments instead of passive (inductive) wheel-speed sensors.

Passive (inductive) wheel-speed sensor

A passive (inductive) speed sensor consists of a permanent magnet (Fig. 2, 1) with a soft-magnetic pole pin (3) connected to it, which is inserted into a coil (2) with several thousand windings. This setup generates a constant magnetic field.

The pole pin is installed directly above the pulse wheel (4), a gear wheel attached to the wheel hub. As the pulse wheel turns, the continuously alternating sequence of teeth and spaces induces corresponding fluctuations in the constant magnetic field. This changes the magnetic flux through the pole pin and therefore also through the coil winding. These fluctuations in the magnetic field induce an alternating voltage in the coil suitable for monitoring at the ends of its winding.

The frequency and amplitude of this alternating voltage are proportional to wheel speed (Fig. 3) and when the wheel is not rotating, the induced voltage is zero.

Tooth shape, air gap, rate of voltage rise, and the control unit input sensitivity define the smallest still measurable vehicle speed and thus, for ABS applications, the minimum response sensitivity and switching speed.

1 Passive (inductive) wheel-speed sensors

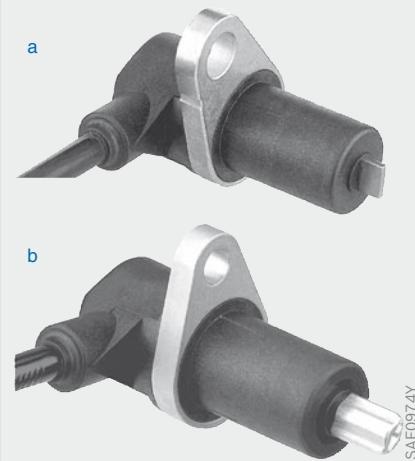
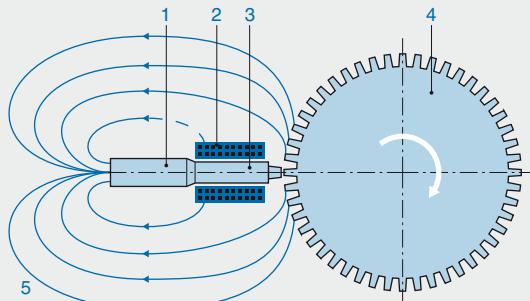


Fig. 1

- a Chisel-type pole pin (flat pole pin)
- b Rhombus-type pole pin (lozenge-shaped pole pin)

2 Figure illustrating the principle of the passive wheel-speed sensor



SAE0975Y

Fig. 2

- 1 Permanent magnet
- 2 Solenoid coil
- 3 Pole pin
- 4 Steel pulse wheel
- 5 Magnetic field lines

Various pole-pin configurations and installation options are available to adapt the system to the different installation conditions encountered with various wheels. The most common variants are the chisel-type pole pin (Fig. 1a, also called a flat pole pin) and the rhombus-type pole pin (Fig. 1b, also called a lozenge-shaped pole pin). Both pole-pin designs necessitate precise alignment to the pulse wheel during installation.

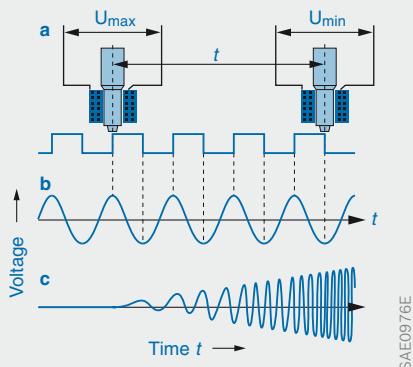
Active wheel-speed sensor

Sensor elements

Active wheel-speed sensors are used almost exclusively in today's modern brake systems (Fig. 4). These sensors usually consist of a hermetic, plastic-cast silicon IC that sits in the sensor head.

In addition to magnetoresistive ICs (the electrical resistance changes as the magnetic field changes) Bosch now uses Hall sensor elements almost exclusively. These sensors react to the smallest changes in the magnetic field and therefore allow greater air gaps compared to passive wheel-speed sensors.

3 Signal output voltage of passive wheel-speed sensor



4 Active wheel-speed sensors



SAE0977Y

Fig. 3

- a Passive wheel-speed sensor with pulse wheel
- b Sensor signal at constant wheel speed
- c Sensor signal at increasing wheel speed

Pulse wheels

A multipole ring is used as a pulse wheel for active wheel-speed sensors. The multipole ring consists of alternately magnetized plastic elements that are arranged in the shape of a ring on a nonmagnetic metal carrier (Fig. 6 and Fig. 7a). These north and south poles adopt the function formerly performed by the teeth of the pulse wheel. The sensor IC is exposed to the continuously changing fields generated by these magnets. The magnetic flux through the IC therefore changes continuously as the multipole ring turns.

Fig. 7

- a Hall IC with multipole pulse generator
 - b Hall IC with steel pulse wheel and magnet in sensor
- 1 Sensor element
2 Multipole ring
3 Magnet
4 Steel pulse wheel

Fig. 5

- 1 Wheel hub
2 Ball bearing
3 Multipole ring
4 Wheel-speed sensor

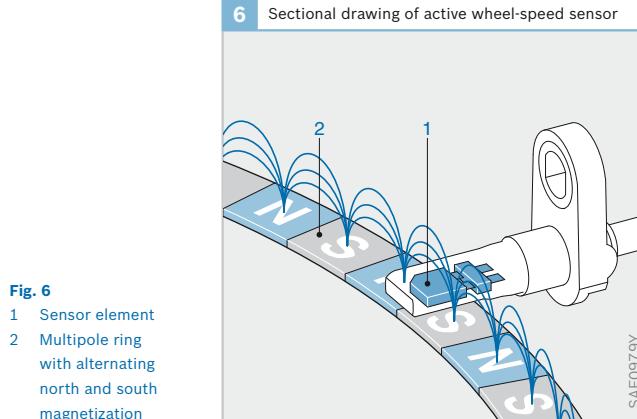
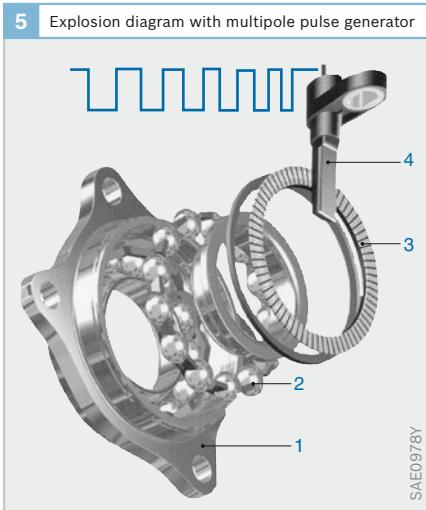


Fig. 6

- 1 Sensor element
2 Multipole ring with alternating north and south magnetization

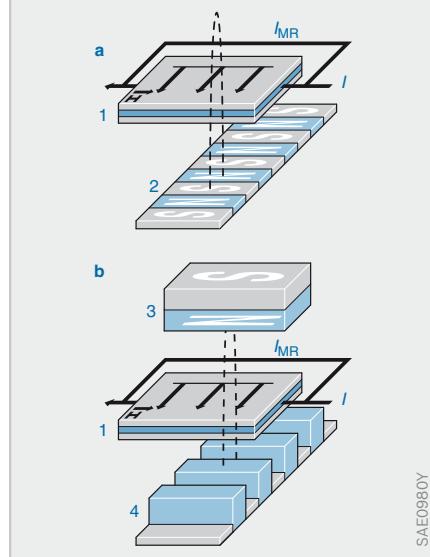
A steel pulse wheel can also be used instead of the multipole ring. In this case a magnet is mounted on the Hall IC that generates a constant magnetic field (Fig. 7b). As the pulse wheel turns, the continuously alternating sequence of teeth and spaces induces corresponding fluctuations in the constant magnetic field. The measuring principle, signal processing and IC are otherwise identical to the sensor without a magnet.

Characteristics

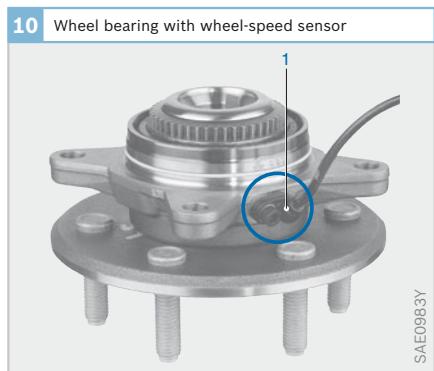
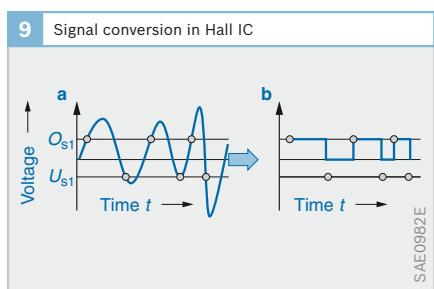
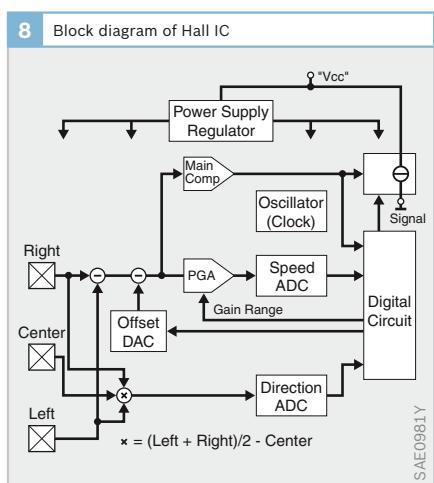
A typical feature of the active wheel-speed sensor is the integration of a Hall measuring element, signal amplifier and signal conditioning in an IC (Fig. 8). The wheel-speed data is transferred as an impressed current in the form of square-wave pulses (Fig. 9). The frequency of the pulses is proportional to the wheel speed and the speed can be detected until the wheel is practically stationary (0.1 km/h).

The supply voltage is between 4.5 and 20 volts. The square-wave output signal level is 7 mA (low) and 14 mA (high).

7 Diagram illustrating principle for measuring speed



This type of data transmission using digital signals is less sensitive to interference than the signals from passive inductive-type sensors. The sensor is connected to the control unit by a two-conductor wire.



Compact dimensions combine with low weight to make the active wheel-speed sensor suitable for installation on and even within the vehicle's wheel-bearing assemblies (Fig. 10). Various standard sensor head shapes are suitable for this.

Digital signal conditioning makes it possible to transfer coded additional information using a pulse-width-modulated output signal (Fig. 11):

► Direction of wheel rotation recognition:

This is especially significant for the "hill hold control" feature, which relies on selective braking to prevent the vehicle from rolling backwards during hill starts. The direction of rotation recognition is also used in vehicle navigation systems.

► Standstill recognition:

This information can also be evaluated by the "hill hold control" function. The information is also used for self-diagnosis.

► Signal quality of the sensor:

Information about the signal quality of the sensor can be relayed in the signal. If a fault occurs the driver can be advised that service is required.

Fig. 9

- a Raw signal
- b Output signal

O_{S1} Upper switching threshold

U_{S1} Lower switching threshold

11 Coded information transfer with pulse-width-modulated signals

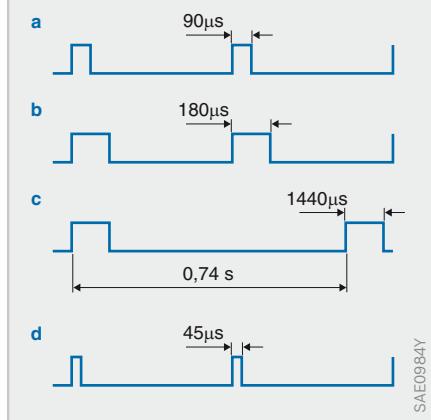


Fig. 10

- 1 Wheel-speed sensor

Fig. 11

- a Speed signal when reversing
- b Speed signal when driving forwards
- c Signal when vehicle is stationary
- d Signal quality of sensor, self-diagnosis

Micromechanical yaw-rate sensors

Applications

In vehicles with electronic stability program (ESP), the rotation of the vehicle about its vertical axis is registered by micromechanical yaw-rate (or yaw-speed) sensors (also known as gyroscopes) and applied for vehicle-dynamics control. This takes place during normal cornering, for instance, but also when the vehicle breaks away or goes into a skid. These sensors are reasonably priced as well as being very compact. They have forced out the conventional high-precision mechanical sensors.

A further area of application is rollover detection in airbag electronic control units for triggering restraints (e.g. sidebags/window airbags, rollover bar) in the event of the vehicle rolling over. MM2 yaw-rate sensors have been specially developed for this area. These are particularly suited for the detection of rotary motion around the longitudinal axis because of their optimized sensing direction. They permit a very compact design and space-saving installation in airbag control units, installed along the longitudinal axis of the vehicle. Packaging the sensor element and the evaluation electronics together in a standard IC housing contributes to the cost reductions in this area.

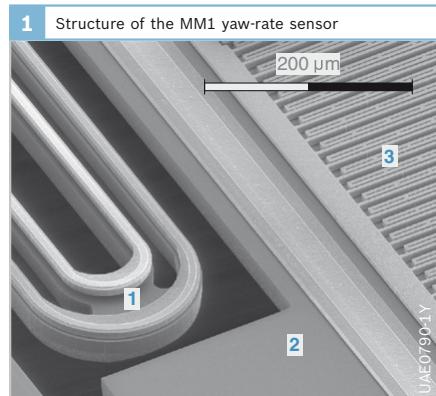


Fig. 1

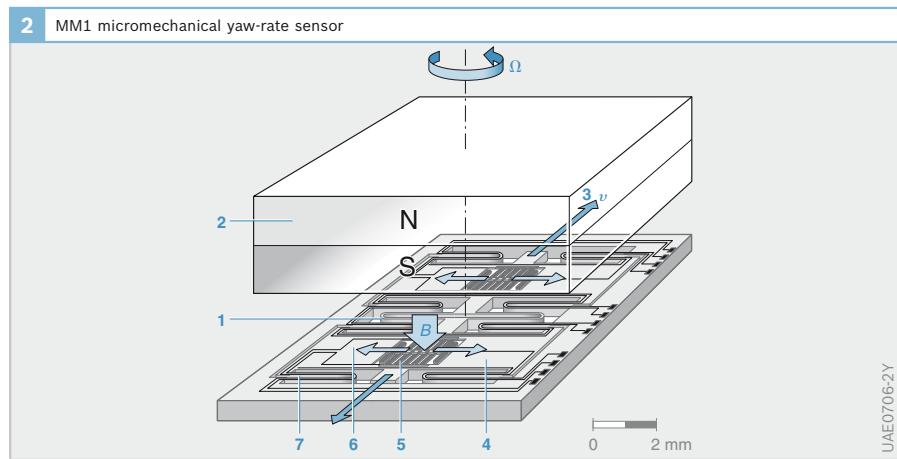
- 1 Retaining/guide spring
- 2 Part of the oscillating element
- 3 Coriolis acceleration sensor

Fig. 2

- 1 Frequency-determining coupling spring
- 2 Permanent magnet
- 3 Direction of oscillation
- 4 Oscillating element
- 5 Coriolis acceleration sensor
- 6 Direction of Coriolis acceleration
- 7 Retaining/guide spring
- Ω Yaw rate
- v Oscillation velocity
- B Permanent-magnet field

MM1 micromechanical yaw-rate sensor

A mixed form of technology is applied in order to achieve the high accuracies needed for vehicle-dynamics systems. That is, two somewhat thicker oscillating elements (mass plates) which have been machined from a wafer using bulk micro-mechanics oscillate in counter-phase to their resonant frequency which is defined by their mass and their coupling-spring stiffness (>2 kHz). On each of these oscillating elements, there is a miniature, surface-type micromechanical capacitive acceleration sensor. When the sensor chip



UAE0706-2Y

rotates about its vertical axis at yaw rate Ω , these register the Coriolis acceleration in the wafer plane vertical to the direction of oscillation (Figs. 1 and 2). These accelerations are proportional to the product of yaw rate and the oscillation velocity which is maintained electronically at a constant value.

To drive the sensor, all that is required is a simple, current-carrying conductor track on each oscillating element. In the permanent-magnet field B vertical to the chip surface, this oscillating element is subjected to an electrodynamic (Lorentz) force. Using a further, simple printed conductor (which saves on chip surface), the same magnetic field is used to directly measure the oscillation velocity by inductive means. The different physical construction of drive system and sensor system serves to avoid undesirable coupling between the two sections. In order to suppress unwanted external acceleration effects (common-mode signal), the opposing sensor signals are subtracted from each other. The external acceleration effects can be measured by applying summation. The high-precision micromechanical construction helps to suppress the effects of high oscillatory acceleration which is several factors of 10 higher than the low-level Coriolis acceleration (cross sensitivity far

below 40 dB). Here, the drive and measurement systems are rigorously decoupled from each other.

MM2 micromechanical yaw-rate sensor

If the silicon yaw-rate sensor is manufactured completely in surface micromechanics (SMM) and, at the same time, the magnetic drive and control system is replaced by an electrostatic system, the decoupling of drive and measuring system does not need to be so thorough. Comb-like structures (Figs. 3 and 4) electrostatically force a centrally mounted rotary oscillator to oscillate. The amplitude of these oscillations is held constant by means of a similar capacitive pick-off. Coriolis forces result at

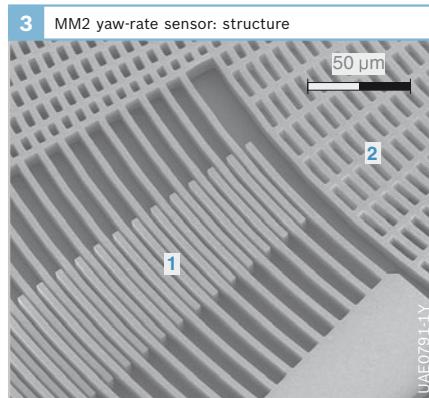


Fig. 3

- 1 Comb-like structure
- 2 Rotary oscillator

4 MM2 surface-micromechanical yaw-rate sensor

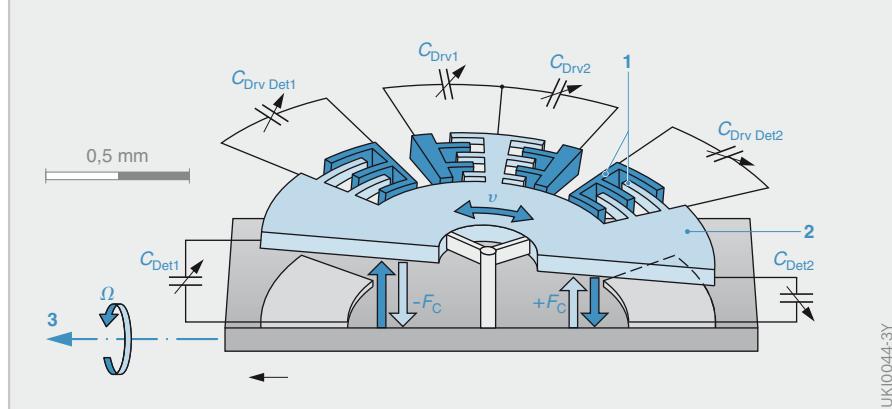


Fig. 4

- 1 Comb-like structure
- 2 Rotary oscillator
- 3 Measuring axis
- C_{Drv} Drive electrodes
- C_{Det} Capacitive pick-off
- F_C Coriolis force
- v Oscillation velocity
- $\Omega = \Delta C_{Det}$, measured yaw rate

the same time in an out-of-plane tilting movement, the amplitude of which is proportional to the yaw rate Ω , and which is detected capacitively by the electrodes underneath the oscillator. To avoid excessive damping of this movement, the sensor must be operated in a vacuum. Although the chip's small size and the somewhat simpler production process result in considerable cost reductions, this miniaturization is at the expense of reductions in the measuring effect, which in any case is not very pronounced, and therefore of the achievable precision. It also places more severe demands on the electronics. The system's high flexural stability, and mounting in the axis of gravity, serve to mechanically suppress the effects of unwanted external acceleration from the side.

DRS MM3x sensor cluster

A new generation of micromechanical elements are used in the DRS MM3x sensor cluster. These can measure and digitally process angular velocities and accelerations. They create a modular concept for hardware and software based on circuit board technology with many new safety features which lead to a versatile and reliable solution for many applications in the vehicle.

Application

The ESP system, the connection to further chassis comfort systems and the development of advanced driving stability systems called for inertia signals with high requirements, in particular in respect of signal quality and robustness, additional measuring axes and high reliability. This is why Bosch has developed a third generation, the versatile yet inexpensive DRS MM3.x sensor cluster, to meet the requirements of functions such as hill hold control (HHC), automated parking brake (APB),

navigation (Travel Pilot), adaptive cruise control (ACC), roll over mitigation (ROM), electronic active steering (EAS), active suspension control (ASC), Steer-by-Wire.

DRS MM3.7k is the base variant of the MM3 generation for ESP applications. It comprises a yaw-rate sensor and an integrated lateral acceleration module.

Operating principle

The new micromechanical measuring element for measuring yaw rate belongs to the familiar group of vibrating gyroscopes operating according to the Coriolis principle (CVG = Coriolis Vibrating Gyros). It consists of an inverse tuning fork with two linear vibration modes perpendicular to one another, the drive circuit and the evaluator circuit. Drive and evaluation take place electrostatically in a comb-like structure. The Coriolis acceleration is measured electrostatically using electrodes that extend into each other. The measuring element consists of two masses connected by a coupling spring. The resonant frequency the same for the two vibration modes. It is typically around 15 kHz and is thus outside the usual interference spectrum in the vehicle and is hence resistant to interfering vibration. The evaluation circuit (ASIC) and the micromechanical measuring element are contained in a prefabricated housing with 20 terminals (Premold 20).

The acceleration module is comparable in form with the yaw-rate sensor module and consists of a micromechanical measuring element, an electronic evaluating circuit and a housing with 12 terminals (Premold12). The spring-mass structure is deflected in its sensitive axis by external accelerations and is evaluated by means of a differential capacitor in a comb-like form.

Piezoelectric "tuning-fork" yaw-rate sensor

Application

In order that it can use the digital road map stored on the CD-ROM to calculate the distance driven, the computer in the vehicle's navigation system needs information on the vehicle's movements (composite navigation).

When the vehicle is cornering, the navigation system's yaw-rate sensor records the vehicle's rotation about its vertical axis, thus permitting determination of the direction of travel. The vehicle's position can hence be calculated in conjunction with the determination of the distance traveled from the tachometer or wheel sensor signals by "composite positioning".

Initially in navigation systems, the vehicle's direction was determined from the earth's magnetic field by a magnetic field sensor. However, these could easily be influenced by the magnetic interference in the vehicle. As oscillation gyroscopes are based on the measurement of mass forces, they are insensitive to magnetic interference.

Design

The piezoelectric "tuning fork" yaw-rate sensor comprises a steel element shaped like a tuning fork. This incorporates four piezo elements (two above, two below, Fig. 1) and the sensor electronics. The tuning forks are about 15 mm long.

Operating principle

When voltage is applied, the bottom piezo elements start to oscillate and excite the upper section of the "tuning fork", together with its upper piezo elements, which then starts counter-phase oscillation. The frequency is approximately 2 kHz.

Straight-ahead driving

With the vehicle being driven in a straight line there are no Coriolis forces acting on the tuning fork, and since the upper piezo elements always oscillate in counter-phase and are only sensitive perpendicularly to the direction of oscillation, they do not generate a voltage.

Cornering

The rotational movement around the vehicle's vertical axis as it travels around a curve causes the upper portion of the tuning fork to leave the oscillatory plane so that an alternating voltage is generated in the upper piezo elements which is transferred to the navigation computer by an electronic circuit in the sensor housing. The voltage-signal amplitude is a function of both the yaw rate and the oscillation velocity. Its sign depends on the direction (left or right) taken by the curve.

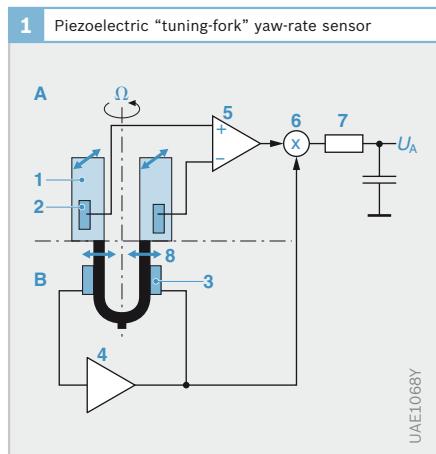


Fig. 1

- A Sensing section of the oscillating element
- B Stimulating section of the oscillating element
- 1 Oscillating element
- 2 Acceleration sensor
- 3 Actuator (piezoelectric element for vibration excitation)
- 4 Regulator for constant vibration excitation
- 5 Charge amplifier
- 6 Multiplication (demodulation)
- 7 Low pass
- 8 Vibration excitation
- U_A Output voltage (proportional to the yaw rate)
- Ω Yaw rate

Micromechanical pressure sensors

Application

Pressure is a non-directional force acting in all directions which occurs in gases and liquids. Micromechanical pressure sensors detect the pressure of various media in the motor vehicle, e.g.:

- ▶ Intake-manifold pressure, e.g. for load sensing in engine-management systems
- ▶ Boost pressure for boost-pressure control
- ▶ Ambient pressure for taking into account air density, e.g. in boost-pressure control
- ▶ Oil pressure to take engine load into account in the service display
- ▶ Fuel pressure for monitoring the level of fuel-filter contamination

Fig. 1

- a Sectional drawing
b Bridge circuit

- 1 Diaphragm
2 Silicon chip
3 Reference vacuum
4 Glass (Pyrex)
 p Measured pressure
 U_0 Supply voltage
 U_M Measurement voltage
 R_1 Strain-gage resistor (compressed)
 R_2 Strain-gage resistor (elongated)

Micromechanical pressure sensors determine the absolute pressure of liquids and gases by measuring the pressure differential in relation to a reference vacuum.

Version with the reference vacuum on the component side

Design

The measuring cell of a micromechanical pressure sensor consists of a silicon chip (Fig. 1a, Item 2), in which a thin diaphragm is micromechanically etched (1). Four strain-gage resistors (R_1, R_2), are diffused into the diaphragm. The electrical resistance in these resistors changes in response to mechanical elongation. The measuring cell is surrounded and sealed on its component side by a cap (Fig. 2, Item 6), which encloses the reference vacuum underneath.

A temperature sensor (Fig. 3, Item 1), the signals of which can be evaluated independently, can also be integrated in the pressure-sensor housing.

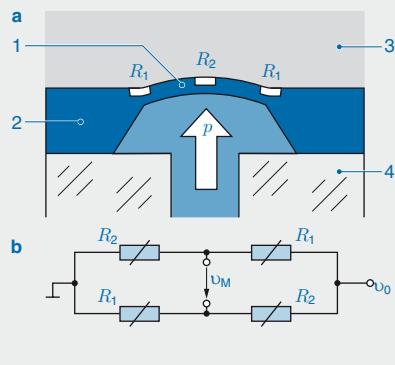
Operating principle

The diaphragm of the sensor cell is deflected to varying degrees depending

on the external pressure acting on it (the center of the diaphragm is deflected by 10 to 1,000 μm). The four strain-gage resistors on the diaphragm change their electrical resistance as a function of the mechanical stress resulting from the applied pressure (piezoresistive effect).

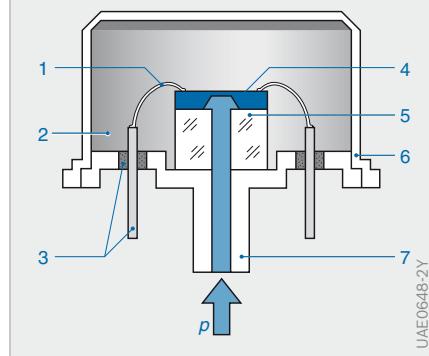
The four measuring shunts are arranged on the silicon chip so that when the diaphragm is deformed, the resistance of two of them increases and that of the other two decreases. These measuring shunts are arranged in a Wheatstone bridge circuit (Fig. 1b) and a change in their resistance values leads to a change in the ratio

1 Pressure-sensor measuring element with reference vacuum on the components side



UAE001/-2Y

2 Measuring element of pressure sensor with cap and reference vacuum on the components side (design)



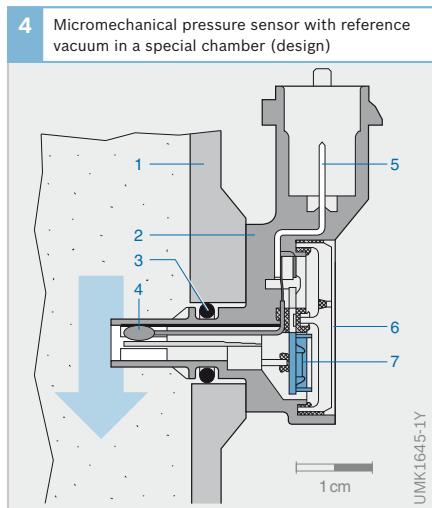
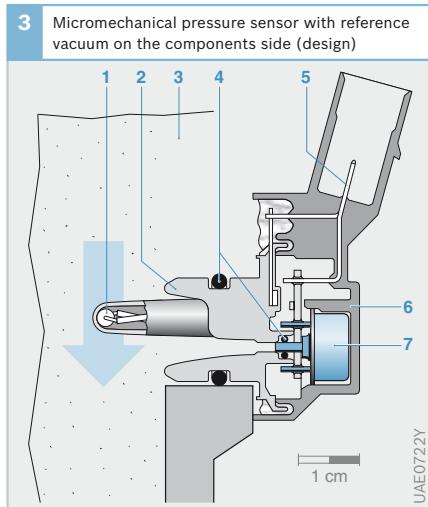
UAE0648-2Y

Fig. 2

- 1, 3 Electrical terminals with glass-enclosed lead-in
2 Reference vacuum
4 Measuring element (chip) with evaluation electronics
5 Glass base
6 Cap
7 Supply for measured pressure p

of the voltages across them. This leads to a change in the measurement voltage U_M . This as yet unamplified voltage is therefore a measure of the pressure applied to the diaphragm.

The measurement voltage is higher with a bridge circuit than would be the case if an individual resistor were used. The Wheatstone bridge circuit thus permits a higher sensor sensitivity.



The signal-conditioning electronic circuitry is integrated on the chip. Its function is to amplify the bridge voltage, compensate for temperature influences, and linearize the pressure curve. The output voltage is in the range of 0 to 5 V and is supplied via electrical connections to the engine control unit (Fig. 3, Item 5). The control unit uses this output voltage to calculate the pressure.

Version with reference vacuum in a special chamber

Design

A pressure sensor with the reference vacuum in a special chamber (Fig. 4) for use as an intake-manifold or boost-pressure sensor is easier to install than a sensor with the reference vacuum on the component side: A silicon chip (Fig. 5, Item 6) with etched diaphragm and four strain-gage resistors in a bridge circuit is located - like the pressure sensor with cap and reference vacuum on the component side - as a measuring cell on a glass base (3). In contrast to the sensor with the reference vacuum on the component side, there is no passage in the glass base through which the measured pressure can be applied to the sensor element. Instead, pressure is

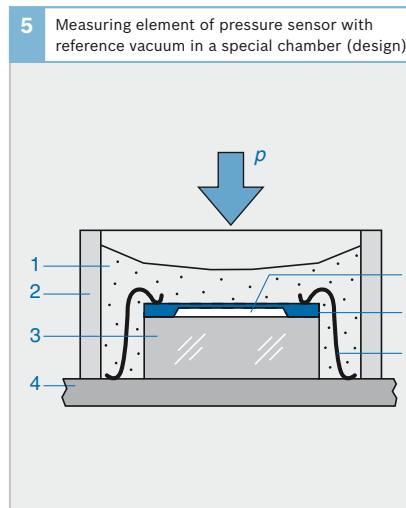


Fig. 3

- 1 Temperature sensor (NTC)
- 2 Housing base
- 3 Manifold wall
- 4 Sealing rings
- 5 Electrical connection (plug)
- 6 Housing cover
- 7 Measuring cell

Fig. 4

- 1 Manifold wall
- 2 Housing
- 3 Sealing ring
- 4 Temperature sensor (NTC)
- 5 Electrical connection (plug)
- 6 Housing cover
- 7 Measuring cell

Fig. 5

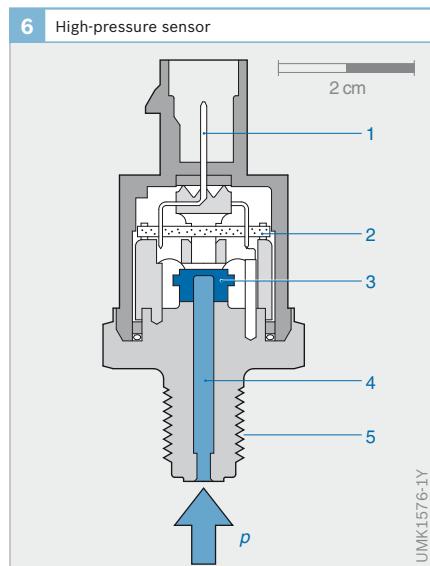
- 1 Protective gel
- 2 Gel frame
- 3 Glass base
- 4 Ceramic hybrid
- 5 Chamber with reference vacuum
- 6 Measuring cell (chip) with evaluation electronics
- 7 Bonded connection
- p Measured pressure

applied to the silicon chip from the side on which the evaluation electronics is situated. This means that a special gel must be used on this side of the sensor to protect it against environmental influences. The reference vacuum (5) is enclosed in the cavity (special chamber) between the silicon chip and the glass base. The complete measuring element is mounted on a ceramic hybrid (4), which incorporates the soldering surfaces for electrical contacting inside the sensor.

A temperature sensor can also be incorporated in the pressure-sensor housing. It protrudes into the air current, and can therefore respond to temperature changes with a minimum of delay (Fig. 4, Item 4).

Operating principle

The operating principle, and with it the signal conditioning and signal amplification together with the characteristic curve, correspond to that used in the pressure sensor with cap and reference vacuum on the component side. The only difference is that the measuring element's diaphragm is deformed in the opposite direction and therefore the strain-gage resistors are "bent" in the other direction.



High-pressure sensors

Application

High-pressure sensors are used in a motor vehicle to measure fuel pressure and brake-fluid pressure, e.g.:

- Rail-pressure sensor for gasoline direct injection (pressure up to 200 bar)
- Rail-pressure sensor for common-rail diesel-injection system (pressure up to 2,000 bar)
- Brake-fluid pressure sensor in the hydraulic modulator of the electronic stability program (pressure up to 350 bar)

Design and operating principle

High-pressure sensors operate according to the same principle as micromechanical pressure sensors. The core of the sensor is a steel diaphragm, onto which strain-gage resistors have been vapor-deposited in the form of a bridge circuit (Fig. 6, Item 3). The sensor's measuring range is dependent on the thickness of the diaphragm (thicker diaphragms for higher pressures, thinner diaphragms for lower pressures). As soon as the pressure to be measured is applied to one side of the diaphragm via the pressure connection (4), the strain-gage resistors change their resistance values as a result of the deflection of the diaphragm. The output voltage generated by the bridge circuit is proportional to the applied pressure. This voltage is transmitted via connecting lines (bonding wires) to an evaluation circuit (2) in the sensor. The evaluation circuit amplifies the bridge signal to 0 to 5 V and transmits it to the control unit, which calculates the pressure with the aid of a characteristic curve.

Temperature sensors

Application

Engine-temperature sensor

This is installed in the coolant circuit (Fig. 1). The engine management uses its signal when calculating the engine temperature (measuring range -40 to +130 °C).

Air-temperature sensor

This sensor is installed in the air-intake tract. Together with the signal from the boost-pressure sensor, its signal is applied in calculating the intake-air mass. Apart from this, setpoint values for the various control loops (e.g. EGR, boost-pressure control) can be adapted to the air temperature (measuring range -40 to +120 °C).

Engine-oil temperature sensor

The signal from this sensor is used in calculating the service interval (measuring range -40 to +170 °C).

Fuel-temperature sensor

Is incorporated in the low-pressure stage of the diesel fuel circuit. The fuel temperature is used in calculating precisely the injected fuel quantity (measuring range -40 to +120 °C).

Exhaust-gas temperature sensor

This sensor is mounted on the exhaust-gas system at points which are particularly critical regarding temperature. It is applied in the closed-loop control of the systems used for exhaust-gas treatment. A platinum measuring shunt is usually used (measuring range -40 to +1,000 °C).

Design and operating principle

Depending upon the particular application, a wide variety of temperature sensor types are available. A temperature-dependent semiconductor measuring shunt is fitted inside a housing. This resistor is usually of the NTC (Negative Temperature Coefficient, Fig. 2) type. Less often a PTC (Positive Temperature Coefficient) type is used. With NTC, there is a sharp drop in resistance when the temperature rises, and with PTC there is a sharp increase.

The measuring shunt is part of a voltage-divider circuit to which 5 V is applied. The voltage measured across the measuring shunt is therefore temperature-dependent. It is supplied through an analog-digital (A/D) converter and is a measure of the temperature at the sensor. A characteristic curve is stored in the control unit which allocates a specific temperature to every resistance or output voltage value.

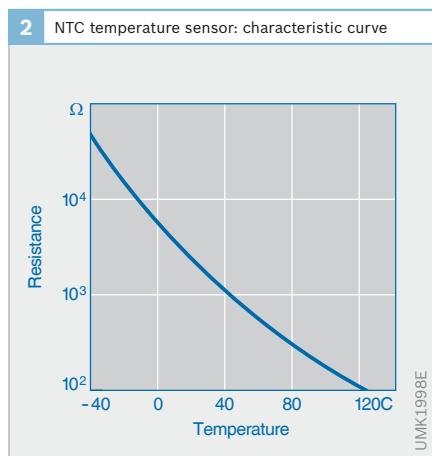
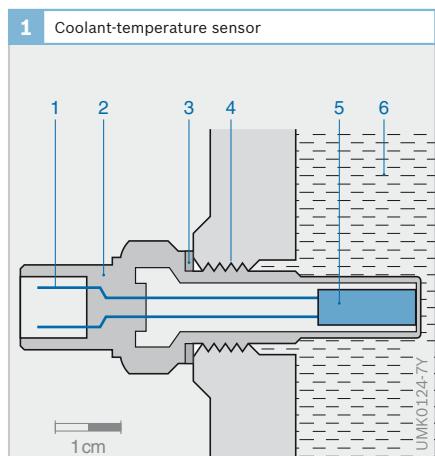


Fig. 1

- 1 Electrical connection
- 2 Housing
- 3 Sealing ring
- 4 Thread
- 5 Measuring shunt
- 6 Coolant

Accelerator-pedal sensors

In Motronic systems with electronic throttle control (ETC), the accelerator-pedal sensor records the travel or the angular position of the accelerator pedal. For this purpose, potentiometers are used in addition to proximity-type sensors. The accelerator-pedal sensor is incorporated with the accelerator pedal in the accelerator-pedal module. These ready-to-install units make adjustments on the vehicle a thing of the past.

Fig. 1

- 1 Potentiometer (reference potentiometer)
- 2 Potentiometer (half voltage)

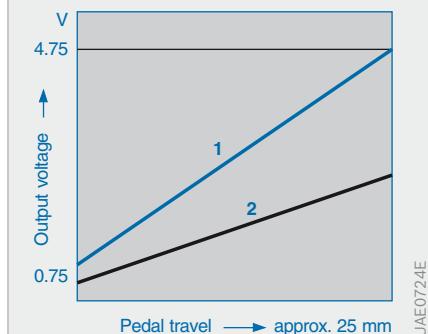
Fig. 2

- 1 Rotor disk (permanent-magnetic)
- 2 Pole shoe
- 3 Conductive element
- 4 Air gap
- 5 Hall sensor
- 6 Shaft (soft magnetic)

Fig. 3

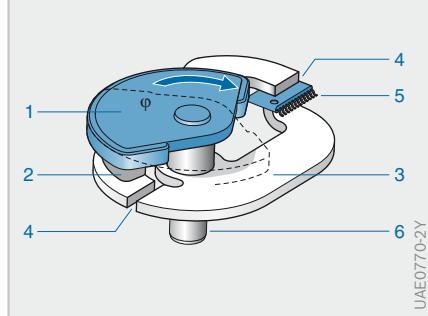
- a Principle of operation
- b Characteristic curve
- 1 Hall IC positioned in the midpoint of the circular path
- 2 Hall IC located outside the midpoint (linearization)
- 3 Magnet

1 Characteristic curve of an accelerator-pedal sensor



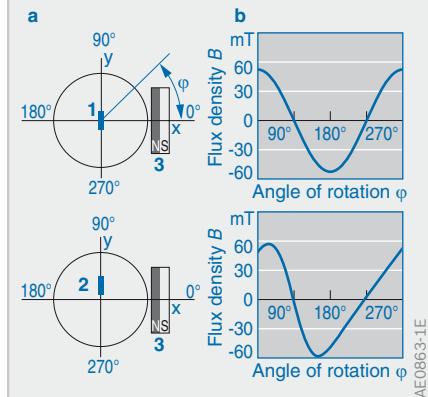
UAE0724E

2 ARS1 Hall angular-position sensor



UAE0770-2Y

3 Principle of ARS2 Hall angular-position sensor



UAE0863-1E

Hall angular-position sensors

Hall sensors are used to measure the movement of the accelerator pedal on a non-contact basis. In the case of the Type ARS1 Hall angular-position sensor, the magnetic flux of a roughly semicircular, permanent-magnetic disk is fed back via a pole shoe, two further conductive elements and the similarly ferromagnetic shaft to the magnet (Fig. 2). Depending upon the angular setting, the flux is led to a greater or lesser degree through the two conductive elements, in the magnetic path of which a Hall sensor is also situated. Using this principle, it is possible to achieve a practically linear characteristic curve in the measuring range of 90°.

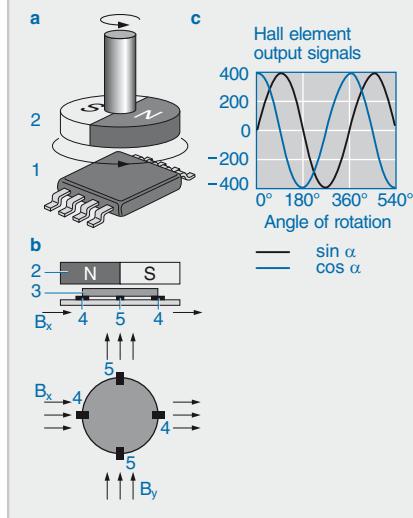
The Type ARS2 is a simplified version which does without soft magnetic conductive elements (Fig. 3). In this version, the magnet moves around the Hall sensor in a circular arc. Only a relatively small section of the resulting sinusoidal characteristic curve features good linearity. If the Hall sensor is located slightly outside the center of the circular arc, the characteristic curve increasingly deviates from the sinusoidal, and now features a short measuring range of almost 90°, and a longer measurement range of more than 180° with good linearity. A great disadvantage though is the low level of shielding against external fields, as well as the remaining dependence on the geometric tolerances of the magnetic circuit, and the intensity fluctuations of the magnetic flux in the permanent magnet as a function of temperature and age.

In the case of the FPM2.3 type Hall angular-position sensor, it is not the field strength but rather the direction of the magnetic field which is used to generate the output signal. The field lines are recorded by four radially arranged measuring elements lying in one plane in the x- and y-directions (Fig. 4). The output signals are derived in the ASIC from the raw data (cos and sin signals) using the arctan function. The sensor is positioned between two magnets to generate a homogenous

magnetic field. The sensor is therefore insensitive to component tolerances and is temperature-resistant.

As with the accelerator-pedal module with a potentiometric sensor, these proximity-type systems also contain two sensors in order to receive two redundant voltage signals.

4 Measuring principle of FPM2.3

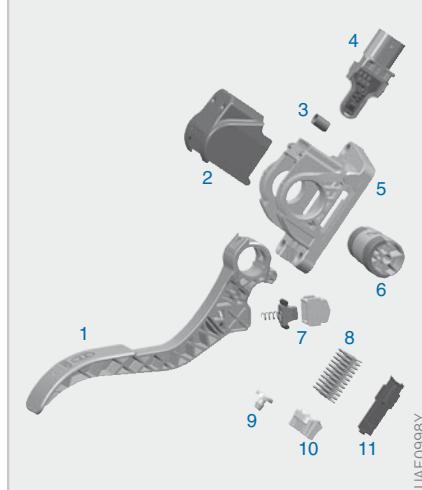


UAE0997Y

Fig. 4

- a Design
- b Principle
- c Measurement signals
- 1 Integrated circuit (IC) with Hall elements
- 2 Magnet (opposing magnet not shown here)
- 3 Conductive element
- 4 Hall elements (for recording x-components of B)
- 5 Hall elements (for recording y-components of B)
- B_x Homogenous magnetic field (x-components)
- B_y Homogenous magnetic field (y-components)

5 Exploded view of FPM2.3 accelerator-pedal module



UAE0998Y

Fig. 5

- 1 Pedal
- 2 Cover
- 3 Spacer sleeve
- 4 Sensor block with housing and plug
- 5 Bearing block
- 6 Shaft with two magnets and hysteresis elements (round magnets not visible)
- 7 Kickdown (optional)
- 8 Two springs
- 9 Stop damper
- 10 Thrust member
- 11 Floor cover

Steering-angle sensors

Application

The electronic stability program (ESP) applies the brakes selectively to the individual wheels in order to keep the vehicle on the desired track selected by the driver. Here, a control unit compares the steering angle and the applied brake pressure with the vehicle's actual rotary motion (around its vertical axis) and its road speed. If necessary, the brakes are applied at individual wheels. These measures serve to keep the float angle (deviation between the vehicle axis and the actual vehicle movement) down to a minimum and, until the physical limits are reached, prevent the vehicle breaking away.

Basically speaking, practically all types of angular-position sensors are suitable for registering the steering angle. Safety considerations, though, dictate that only those types are used which can be easily checked for plausibility, or which in the ideal case automatically check themselves. Potentiometer principles are used, as well as optical code-registration and magnetic principles. Whereas a passenger-car steering wheel turns through $\pm 720^\circ$ (a total of 4 complete turns), conventional angular-position sensors can only measure maximum 360° . This means that with the majority of the sensors actually used for this purpose it is necessary to continually register and store the data on the steering wheel's actual setting.

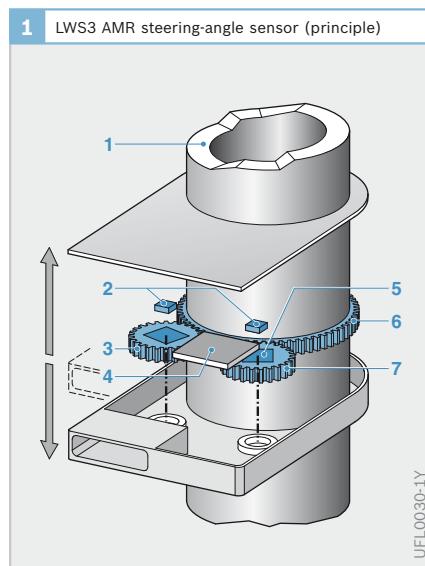
Design and operating principle

Steering-angle sensor with AMR element

The LWS3 steering-angle sensor uses AMR (Anisotropic Magnetoresistive Sensors) for its operation. The AMR's electrical resistance changes according to the direction of an external magnetic field. In the LWS3, the information on angle across a range of four complete rotations is provided by measuring the angles of two gear wheels which are rotated by a third gear wheel on the steering shaft. The first two gear

wheels differ by one tooth which means that a definite pair of angular variables is associated with every possible steering-wheel position.

By applying a mathematical algorithm (a computing process which follows a defined step-by-step procedure) referred to here as a modified vernier principle, it is possible to use the above AMR method for calculating the steering angle in a mi-



croprocessor. Here, even the measuring inaccuracy of the two AMR sensors can be compensated for. In addition, a self-check can also be implemented so that a highly plausible measured value can be sent to the control unit.

Figure 1 shows the schematic representation of the LWS3 steering-angle sensor. The two gear wheels, with magnets inserted, can be seen. The sensors are located above them together with the evaluation electronics.

A different design is the LWS4 steering-angle sensor which unambiguously measures a steering angle of 180° . It is mounted on the shaft end of the steering axle (Fig. 3)

Steering-angle sensor with GMR element

The LWS5 is the first steering-angle sensor based on the GMR effect (Giant Magneto Resistive). Its mechanical structure and operating principle have been taken from the LWS3. The LWS3 and LWS5 are mechanically and electrically compatible.

The GMR layers are processed on the planarized surface of the evaluation circuit (vertical integration). The GMR resistor bridge and the evaluation module are connected by means of plated-through holes.

These short connections enhance the sensor's robustness against external interference.

The two GMR elements measure the direction of the field lines of the two magnets. The steering angle is calculated from this in the microprocessor. The sensor element and microprocessor communicate via a digital interface (SPI interface). The calculated steering angle value is placed on the CAN by the microprocessor.

Because of the greater sensitivity, by comparison with the AMR effect, the LWS5 can work with weaker magnets and greater air gaps. This brings considerable cost benefits in materials and design. The 360° angular measuring range of a single GMR element (a 180° measuring range is typical for AMR), allows the use of smaller gear wheels on the LWS5. It thus requires a significantly smaller installation space than the LWS3. In addition, it has a significant level of scalability, reflected in the measuring range ($\pm 90^\circ$ to $\pm 780^\circ$) and in the degree of redundancy. This means that the sensor can meet the specific requirements of the various vehicle manufacturers as closely as possible.

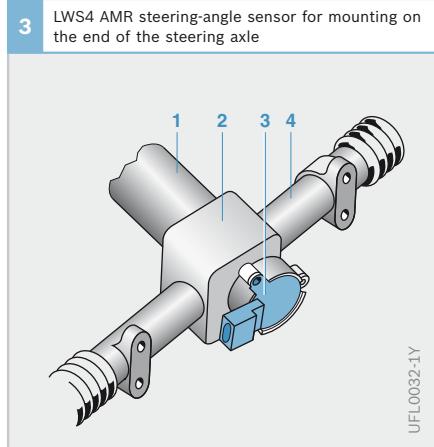


Fig. 3

- 1 Steering column
- 2 Steering gear
- 3 Steering-angle sensor
- 4 Steering rack

Position sensors for transmission control

Application

The position sensor detects the positions of an actuator (e.g. selector lever shaft, selector slide or park lock cylinder) in AT, ASG, DKG and CVT transmissions. The sensors may be fitted stand-alone or built into electronics modules.

Requirements

Sensors built into the transmission must be designed to cope with the challenging operating conditions prevailing there:

- ▶ Ambient temperatures between -40 and +150 °C,
- ▶ Transmission oil, which is an aggressive ambient medium,
- ▶ High mechanical stress up to 30 g
- ▶ Abraded metallic materials and a build-up of particles in the transmission.

Resistance to the media and temperature in these applications is provided by an oil-resistant enclosure for the electronics and the use of high temperature PCBs and AVT for the electronics. The electronic components of ASIC, capacitors, resistors and interfaces (bonds, soldering points) must

be designed to be sufficiently robust as to be able to withstand the severe mechanical stresses over the service life of the vehicle.

Surface-mounted sensors, on the other hand, are made in the standard packages familiar from engine compartment applications. They must meet the corresponding requirements such as being splash-proof and handling operating temperatures from -40 to 130 °C.

Because of the complex requirements (Table 1) arising from the various transmission topologies, and the requirements on mounting space and function (safety concept, required precision, etc.) different physical measuring principles (Hall, AMR, GMR, eddy-current principles) and shapes (linear and rotary detection) may be used. These are explained in greater detail below using the example of the Hall effect and the eddy current principle.

Linear position determination on the basis of Hall switches

Design

Four digital Hall switches are mounted on a printed-circuit board in such a way that they detect the magnetic orientation of a linearly displaceable, multipolar permanent magnet (Fig. 1). The magnet carriage is coupled with the linearly actuated selector slide (hydraulic slide in the transmission control plate) or the parking lock cylinder.

In addition to the Hall switch, the printed-circuit board carries resistors to represent diagnosis functions and EMC capacitors.

The sensor electronics are protected against the influence of the transmission oil by being potted in a dense, oil-resistant epoxy resin.

1 Requirements on position detection

Installation location	Inside the transmission On the transmission
Design	Linear Rotational
Measuring principle	Hall AMR GMR Eddy current
Signal acquisition	Digital (Gray code) Analog
Output signal	Digital (2-/4-bit) PWM Analog
System security	Redundancy High availability P/N recognition
Shifting type	M and E shifting

Table 1

Operating principle

In the case of automatic transmission with manual shifting, also known as M-shifting, the position sensor detects the positions of the selector slide, P, R, N, D, 4, 3, 2 and the intermediate ranges and passes these to the transmission control in the form of a 4-bit code (Fig. 2). For safety reasons, the position is encoded in one-step encoding, i.e. two-bit changes are always required before a new position is detected. Single-bit changes caused by malfunctions can be detected as false by the control unit by plausibility considerations. The sequence of these bit changes corresponds to a Gray code.

In some applications, the sensor transmits the controlled variable for starter release, the P/N signal, directly via a processor-independent hardware route. This increases the availability in operating states (e.g. when the vehicle's electrical system voltage is low) in which the electronic control unit would not yet be operational.

In the case of automatic transmission with electronic shifting, also known as E-shifting, the position sensor detects only the positions of the park lock cylinder P_{in} and P_{out} and the intermediate range and passes these to the transmission control unit in the form of a 2-bit code.

Rotational position determination on the basis of eddy currents

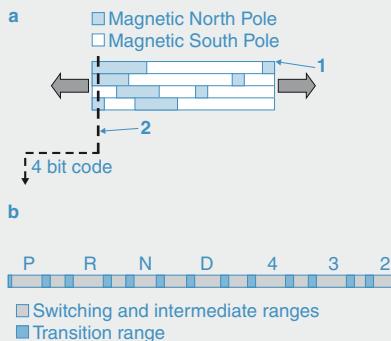
Design

A position sensor with rotational design is shown in Figure 3. In this example, a specially shaped rotor with a return coil is mounted to rotate on the selector lever shaft outside the transmission. It represents the interface to the transmission actuator. Redundant transmit and receive coils are mounted on the fixed sensor board with their associated evaluating ASICs. In this solution, the actuator (shaft) passes through the circuit board.

Fig. 1
a Front view
b Rear view

- 1 Oil-resistant package
- 2 Electrical connection by terminal plate
- 3 Printed-circuit board potted Hall elements
- 4 Carriage with permanent magnet
- 5 Locating pin

2 Coding on the magnetic carriage



UTS0363Y

SAE1084E

- Fig. 2**
- a Magnetic encoding
 - b Position ranges
 - 1 Moving carriage
 - 2 Fixed position of the Hall elements

Operating principle

The sensor uses the eddy-current principle. The transmitting coil induces in the return coil an eddy current opposite to its cause, the magnetic field of which induces a voltage in the receiver coils. The geometries of the return coil and the receiver coil are tuned to one another, so that continuously changing positions of the rotor can be detected.

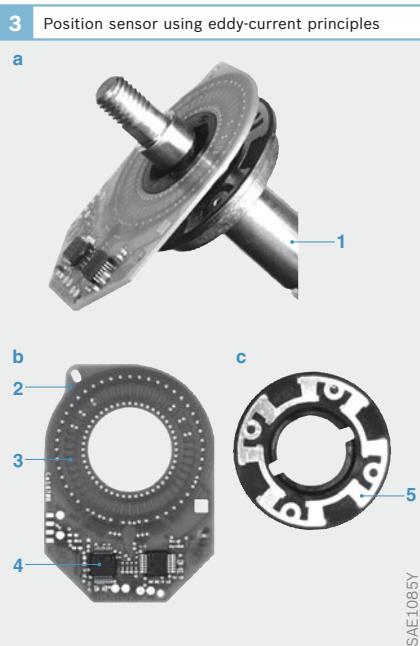
Similarly to the example of linear position detection, the rotational position sensor detects the P, R, N, D, 4, 3 and 2 positions. In this case, however, on the selector lever shaft projection out of the transmission. The positions of the shaft correspond to the positions of the selector slide inside the transmission.

One benefit of analog signal conditioning is that the individual shift and intermediate ranges are assigned to the angle positions detected by the sensor in the software, and so can easily be adapted to transmission design variants.

For safety reasons, the sensor makes two independent, opposing output signals available, with the upper and lower 5 % of the signal stroke being used as a diagnostic range. In the event of a fault being detected inside the sensor, a voltage level in the upper diagnostic range is output. Faults that can occur on the transmission path to the control unit (e.g. open circuit or short-circuit) also lead to voltage levels in the diagnostics range. In addition to this, the electronic control unit tests the signal for plausibility by adding the two signals together. Sensor or transmission errors can be detected by the control unit in this way, and an appropriate emergency running program can be selected, depending on the fault.

Table 2 compares the two examples of position sensor and illustrates the complexity of the topic.

The complex requirements on the actuator's mechanical interface, the mounting space, the environmental influences and the safety concept frequently lead to application-specific sensor system solutions.

**Fig. 3**

- a Components (without housing)
 - b Printed-circuit board
 - c Rotor
- 1 Selector lever shaft
 - 2 Sensor PCB
 - 3 Redundant transmitter and receiver coil
 - 4 Redundant electronics
 - 5 Rotor with return coil

Table 2

2 Typical characteristics

Position sensor	Hall switch	Eddy current
Measuring principle	Hall effect (digital)	Eddy current (analog)
Sensitivity to external magnetic fields	Yes	No
Sensitivity to metallic surroundings	No	Yes
Flexible position scaling	No	Yes (software)
Operating voltage range	4 to 12 V	4.5 to 5.5 V
Sensitivity to the air gap	High	Low

Axle sensors

Applications

The automatic headlamp range control (ALWR) adjusts the vehicle's headlamp range automatically. With the headlamps on low beam, ALWR compensates for the vehicle tilt so that the driver has adequate vision without dazzling oncoming traffic. Whereas, the static ALWR compensates for the vehicle tilt resulting from the vehicle loading, the dynamic version also takes into account the front-end up and down pitching movement of the vehicle caused by braking and acceleration. Here, the axle sensors precisely register the bodywork's tilt angle.

Design and operating principle

Axle sensors (angular-position sensors) are used to measure the vehicle tilt. These are attached to the bodywork at the front and rear of the vehicle. Spring compression is transmitted to the sensors by a pivot lever which is connected through a connecting rod to the particular axle or to the wheel suspension. The vehicle's tilt is then calculated from the difference between the voltages of the front and rear axle sensors.

The axle sensors function according to the Hall effect. A Hall IC is built into the stator (Fig. 1, Item 5) where it is located in

a homogenous magnetic field. The magnetic field generates a Hall voltage in the Hall IC proportional to the magnetic-field strength. When the ring magnet (6) is rotated by the shaft (2), the magnetic field through the Hall IC changes accordingly.

When the springs are compressed due to loading and/or acceleration/braking, the connecting rod (Fig. 2, Item 4) transfers the spring compression movement to the pivot lever of the axle sensor so that it is converted into a voltage signal which is proportional to the angle of rotation.

The control unit registers the axle-sensor signals and from them generates the difference between front and rear axle. Using this voltage difference, and taking into account the vehicle's speed, the control unit now calculates the setpoint value for the servomotor settings. With the vehicle driving at constant speed, the dynamic ALWR remains in the status which features a high level of damping whereby the stepping motors are adapted slowly to the vehicle tilt. This prevents bumps, unevenness, or holes in the road causing continuous correction of the headlamp settings. When the vehicle accelerates, or the brakes are applied, the system automatically switches to the dynamic mode, and adjusts the headlamp range within a few milliseconds. The system then switches back to the slower mode.

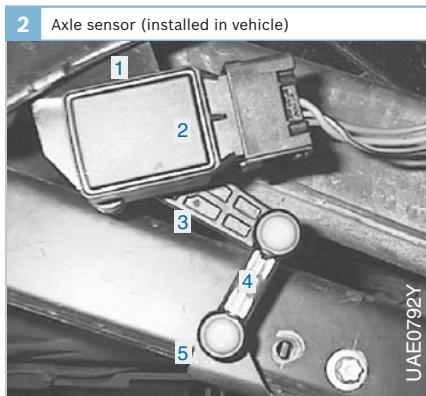
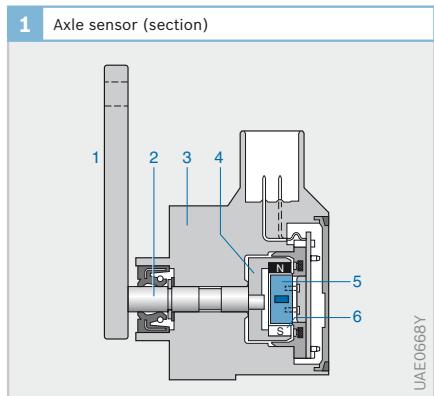


Fig. 1

- 1 Pivot lever
- 2 Shaft
- 3 Housing
- 4 Ring-magnet mount
- 5 Stator with Hall IC
- 6 Ring magnet

Fig. 2

- 1 Bodywork mounting
- 2 Axle sensor with plug-in connection
- 3 Pivot lever
- 4 Connecting rod
- 5 Vehicle axle

Hot-film air-mass meters

Application

To provide precise pilot control of the air-fuel ratio, it is essential for the supplied air mass to be exactly determined in the respective operating status. The hot-film air-mass meter measures some of the actually induced air-mass flow for this purpose. It takes into account the pulsations and reverse flows caused by the opening and closing of the engine's intake and exhaust valves. Intake-air temperature or air-pressure changes have no effect upon measuring accuracy.

HFM5 design

The housing of the HFM5 hot-film air-mass meter (Fig. 1, Item 5) projects into a measuring tube (Fig. 3, Item 6) which, depending upon the engine's air-mass requirements (370 to 970 kg/h) can have a variety of diameters. The measuring tube normally

contains a flow rectifier, which ensures that the flow in the measuring tube is uniform. The flow rectifier is either a combination of a plastic mesh with straightening action and a wire mesh, or is a wire mesh on its own (Fig. 3, Item 8). The measuring tube is installed in the intake tract downstream from the air filter. Plug-in versions are also available which are installed inside the air filter.

The most important component parts in the sensor are the measuring cell (Fig. 1, Item 4) in the air intake (8) and the integrated evaluation electronics (3).

The sensor measuring cell consists of a semiconductor substrate. The sensitive surface is formed by a diaphragm which has been manufactured in micromechanical processes. This diaphragm incorporates temperature-sensitive resistors. The elements of the evaluation electronics (hybrid circuit) are installed on a ceramic substrate. This principle permits very compact design. The evaluation electronics are connected to the control unit by means of electrical connections (1).

The partial-flow measuring passage (6) is shaped so that the air flows past the measuring cell smoothly (without swirl effects) and back into the measuring tube via the air outlet (7). The length and location of the inlet and outlet of the partial-flow measuring passage have been chosen to provide good sensor performance even in the event of sharply pulsating flows.

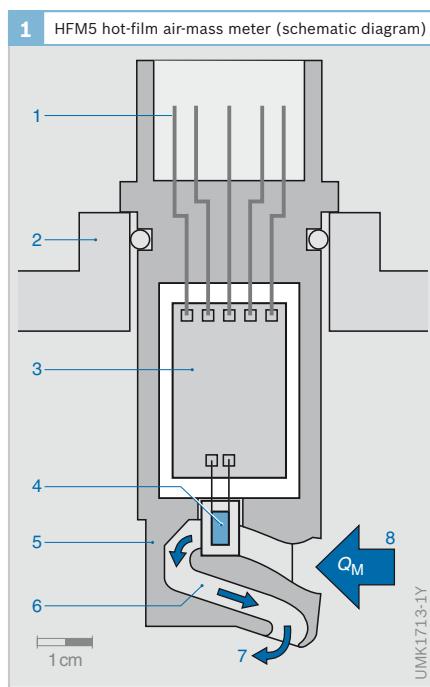
Operating principle

The HFM5 hot-film air-mass meter is a thermal sensor which operates according to the following principle: A centrally situated heating resistor on the measuring cell (Fig. 3, Item 3) heats a sensor diaphragm (5) and maintains it at a constant temperature. The temperature drops sharply on each side of this controlled heating zone (4).

The temperature distribution on the diaphragm is registered by two temperature-dependent resistors which are mounted upstream and downstream of the heating

Fig. 1

- 1 Electric connections (plug)
- 2 Measuring-tube or air-filter housing wall
- 3 Evaluation electronics (hybrid circuit)
- 4 Sensor measuring cell
- 5 Sensor housing
- 6 Partial-flow measuring passage
- 7 Outlet, partial air flow Q_M
- 8 Intake, partial air flow Q_M



UMK1713-1Y

resistor so as to be symmetrical to it (measuring points M_1 , M_2). Without the flow of incoming air, the temperature profile (1) is the same on each side of the heating zone ($T_1 = T_2$).

As soon as air flows over the measuring cell, the uniform temperature profile at the diaphragm changes (2). On the inlet side, the temperature characteristic is steeper since the incoming air flowing past this area cools it off. On the opposite side, the temperature characteristic only changes slightly, because the incoming air flowing past has been heated by the heater element. The change in temperature distribution leads to a temperature difference (ΔT) between the measuring points M_1 and M_2 .

The heat dissipated to the air, and therefore the temperature characteristic at the measuring cell is dependent on the air mass flowing past. The temperature difference is (irrespective of the absolute temperature of the air flow past) a measure of the air-flow mass. It is also direction-dependent so that the air-mass meter can record both the amount and the direction of an air-mass flow.

Due to its very thin micromechanical diaphragm, the sensor has a highly dynamic response (< 15 ms), a point which is of particular importance when the incoming air is pulsating heavily.

The evaluation electronics integrated in the sensor convert the resistance differential at the measuring points M_1 and M_2 into an analog voltage signal of between 0 and 5 V. Using the sensor curve (Fig. 2) stored in the engine control unit, the measured voltage is converted into a value representing the air-mass flow (kg/h).

The shape of the characteristic curve is such that the diagnosis facility incorporated in the control unit can detect such malfunctions as an open-circuit line. An additional temperature sensor for evaluation functions can be integrated in the HFM5 hot-film air-mass meter. It is not required for measuring the air mass.

Incorrect air-mass readings will be registered if the sensor diaphragm is contaminated with dust, dirty water or oil. For the purpose of increasing the robustness of the HFM5, a protective device has been

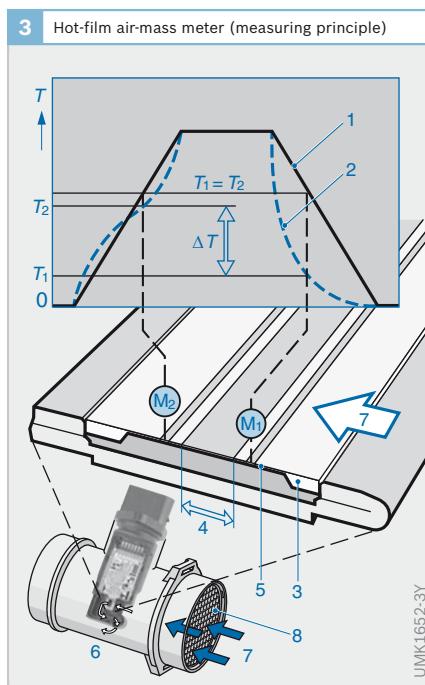
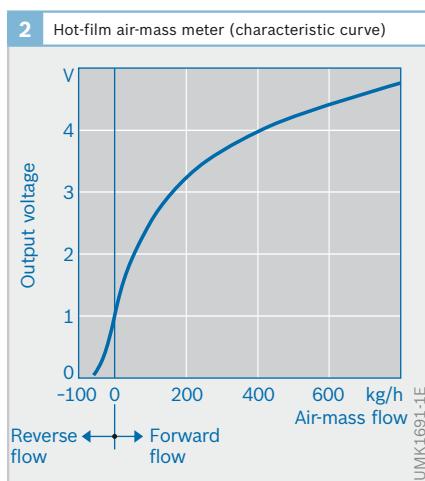


Fig. 3

- 1 Temperature profile without air flow
- 2 Temperature profile with air flow
- 3 Measuring cell
- 4 Heating zone
- 5 Sensor diaphragm
- 6 Measuring tube with air-mass meter
- 7 Intake-air flow
- 8 Wire mesh
- M_1 , M_2 Measuring points
- T_1 , T_2 Temperature values at measuring points
- M_1 and M_2
- ΔT Temperature difference

developed which, in conjunction with a deflector mesh, keeps dirty water and dust away from the sensor element (HFM5-CI; with C-shaped bypass and inner tube (1), which together with the deflector mesh protects the sensor).

HFM6 hot-film air-mass meter

The HFM6 uses the same sensor element as the HFM5 and has the same basic design. It differs in two crucial points:

- ▶ The integrated evaluation electronics operate digitally in order to obtain greater measuring accuracy
- ▶ The design of the partial-flow measuring passage is altered to provide protection against contamination directly upstream of the sensor element (similar to the deflector mesh in the HFM5-CI)

Digital electronics

A voltage signal is generated with a bridge circuit from the resistance values at the measuring points M₁ and M₂ (Fig. 3); this voltage signal serves as the measure of the air mass. The signal is converted into digital form for further processing.

The HFM6 also takes into account the temperature of the intake air when determining the air mass. This increases significantly the accuracy of the air-mass measurement.

The intake-air temperature is measured by a temperature-dependent resistor, which is integrated in the closed control loop for monitoring the heating-zone temperature. The voltage drop at this resistor passes through an analog-digital converter to give a digital signal representing the intake-air temperature. The signals for air mass and intake-air temperature are used to address a program map in which the correction values for the air-mass signal are stored.

Improved protection against contamination

The partial-flow measuring passage is divided into two sections in order to provide better protection against contamination (Fig. 4). The passage which passes the sensor element has a sharp edge (1), around which air must flow. Heavy particulates and dirty-water droplets are unable to follow this diversion and are separated from the partial flow. These contaminants exit the sensor through a second passage (5). In this way, significantly fewer dirt particles and droplets reach the sensor element (3) with the result that contamination is reduced and the service life of the air-mass meter is significantly prolonged even when operated with contaminated air.

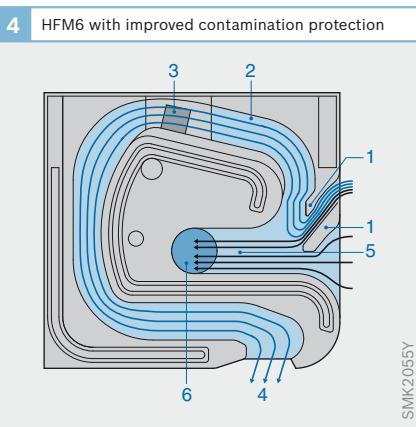


Figure 4

- 1 Diverting edge
- 2 Partial-flow measuring passage (first duct)
- 3 Sensor element
- 4 Air outlet
- 5 Second passage
- 6 Particulate and water outlet

Piezoelectric knock sensors

Application

Regarding their principle of functioning, knock sensors are basically vibration sensors and are suitable for detecting structure-borne acoustic oscillations. These occur as "knock", for instance, in gasoline engines when uncontrolled combustion takes place. They are converted by the knock sensor into electrical signals (Fig. 1) and transmitted to the Motronic control unit, which counteracts the engine knock by adjusting the ignition angle.

Design and operating principle

Due to its inertia, a mass (Fig. 2, Item 2) excited by a given oscillation or vibration exerts a compressive force on a toroidal piezoceramic element (1) at the same frequency as the excitation oscillation. These forces effect a charge transfer within the ceramic element. An electrical voltage is generated between the top and bottom of the ceramic element which is picked off via contact washers (5) and processed in the Motronic control unit.

2 Knock sensor (design and mounting)

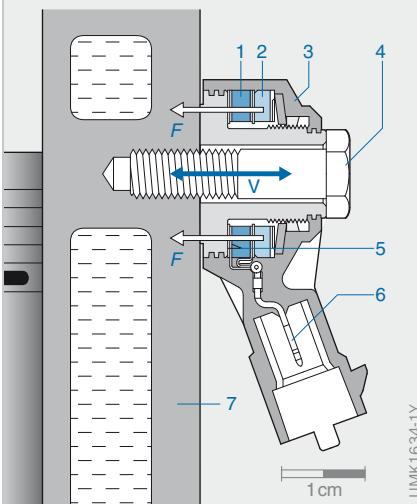


Fig. 2

- 1 Piezoceramic element
- 2 Seismic mass with compressive forces F
- 3 Housing
- 4 Bolt
- 5 Contact washer
- 6 Electric connection
- 7 Engine block
- V Vibration

Mounting

In four-cylinder engines, one knock sensor is sufficient to record the knock signals for all the cylinders. Engines with more cylinders require two or more knock sensors. The knock-sensor installation point on the engine is selected so that knock can be reliably detected from each cylinder. The sensor is usually bolted to the side of the engine block. It must be possible for the generated signals (structure-borne noise vibrations) to be introduced without resonance into the knock sensor from the measuring point on the engine block. A fixed bolted connection satisfying the following requirements is required for this purpose:

- The mounting bolts must be tightened to a defined torque
- The contact surface and the bore in the engine block must comply with prespecified quality requirements
- No washers of any type may be used

1 Knock-sensor signals (oscilloscope display)

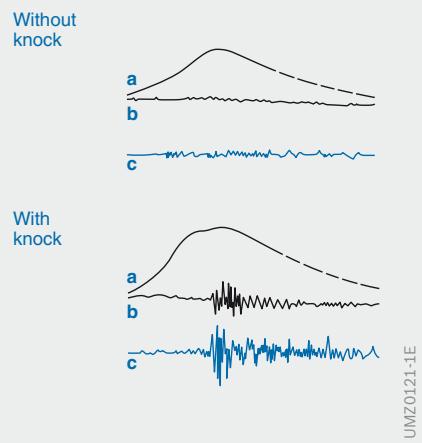


Fig. 1

- a Cylinder-pressure characteristic curve
- b Filtered pressure signal
- c Knock-sensor signal

SMM acceleration sensors

Application

Surface micromechanical (SMM) acceleration sensors have a variety of acceleration detection roles in the vehicle. These include the detection of acceleration values for a frontal or lateral impact in order to trigger the seat-belt pretensioners, the airbag and the rollover bar in vehicle occupant-protection systems. Surface micromechanical (SMM) acceleration sensors are used in the field of active vehicle safety in ABS, ESP and HHC (Hill Hold Control, low-speed traction control) systems. Further possible applications of acceleration sensors are found in chassis and suspension control (active suspension) and vehicle alarm systems, where a change in the vehicle tilt will trigger an alarm.

Design and operating principle

The surface micromechanical sensor elements are designed for different measuring ranges, depending on the application. These measuring ranges lie between 1 g and 400 g ($1\text{ g} \approx 9.81\text{ m/s}^2$). The design of an acceleration sensor for side or frontal impact sensing is described below. The SMM sensor element is fitted together with the evaluation electronics (ASIC) in a first

packaging stage (module). Figure 1 shows a module in a SO16 housing. This module on a small PCB, with further circuit components, is pressed into the second packaging stage, a plastic housing. This housing is closed and sealed with a welded plastic lid. This sensor is intended for installation in the side or fender of the vehicle after calibration and testing and sends its acceleration signal via plug and cable to the central airbag control unit.

The functional layers of the sensor element representing a spring-mass system are applied to the surface of the silicon wafer by an additive method (Fig. 3, surface micromechanics).

The mass with its comb-like electrodes (Figs. 2 and 3, Item 1) is connected via spring elements (2) to the anchor points (Fig. 3, Item 4) in the core of the sensor. There are fixed comb-like electrodes (3, 6) on the chip on each side of these movable electrodes. A useful capacitance in the range of 300 fF to 1 pF results from the connection in parallel of the individual capacitances formed by the fixed and moving electrode fingers. The two rows of electrode fingers connected in parallel result in two useful capacitances ($C_1 - C_M$ and $C_2 - C_M$), which change in opposite directions as the central mass is deflected.

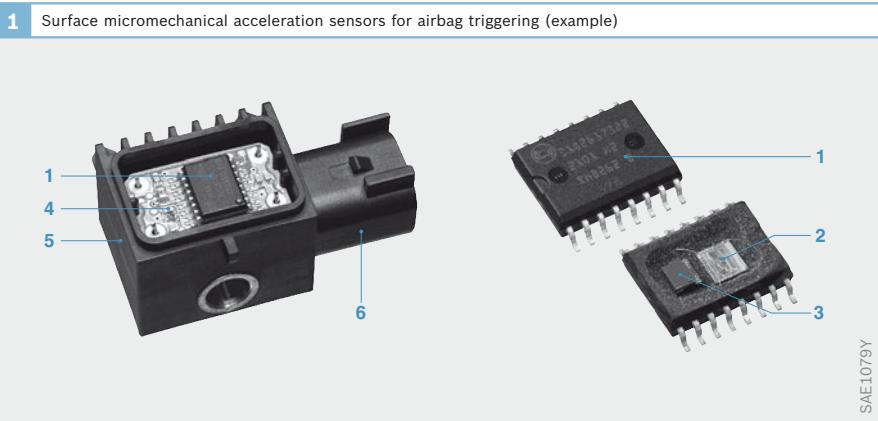


Fig. 1

- 1 First packaging stage (module) in SO16 housing
- 2 Evaluator chip (ASIC)
- 3 SMM sensor element
- 4 Populated PCB
- 5 Second packaging stage (device)
- 6 Plug

The acceleration applied to the spring-mass system deflects it by an amount linearly related to the acceleration applied over the spring return force. An electrical output signal that is linearly dependent on the acceleration is obtained by evaluation of this differential capacitor.

The acceleration signal obtained in the first stage of the evaluation circuit is further conditioned in the ASIC, i.e. it is amplified, filtered and prepared for the output interface.

Analog voltages, pulse-width-modulated signals, SPI protocols or current loop interfaces are usual as the output interface. Tolerances arising from the manufacture of the sensor element, the evaluation circuit and package influences on sensitivity and zero point are eliminated by means of a comparison of programmable memory cells at the end of the production line. A self-test function tests the full mechanical and electrical signal path. In the course of this self-diagnosis, an electrostatic force deflects the sensor structure, i.e. simulating an acceleration in the vehicle, and the response of the measurement signal is compared with a setpoint value.

Up to three acceleration sensors (for ESP and HHC) and one or two yaw-rate sensors (for ESP) are incorporated in the ESP sensor cluster. This clustering of sensor modules reduces the number of individual components and signal lines by comparison with separate sensor devices. It also means that fewer mountings and less installation space are needed in the vehicle.

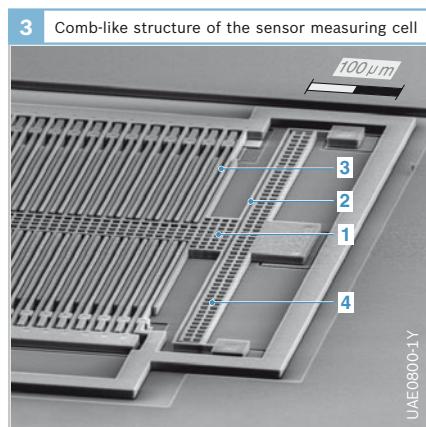


Fig. 3
1 Spring-mounted seismic mass with electrode
2 Spring
3 Fixed electrodes
4 Anchor areas

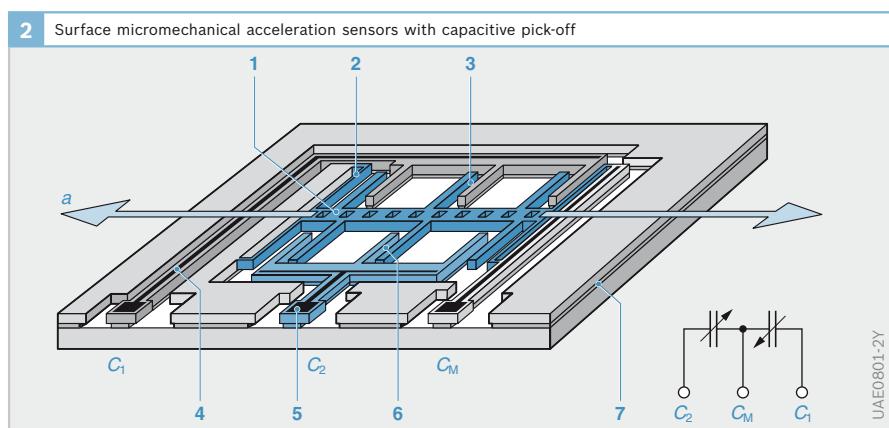


Fig. 2
1 Spring-mounted seismic mass with electrodes
2 Spring
3 Fixed electrodes with capacity C_1
4 Printed Al conductor track
5 Bond pad
6 Fixed electrodes with capacity C_2
7 Silicon oxide
a Acceleration in the sensing direction
 C_M Measuring capacity

Micromechanical bulk silicon acceleration sensors

Application

Micromechanical bulk silicon acceleration sensors detect the acceleration signals needed for ABS, ESP and shock absorber adjustment.

This type of sensor is today predominantly used in the low acceleration ranges with high demands on signal to noise ratio ($< 2 g_n$).

Design and operating principle

Anisotropic and selective etching techniques are used to form the required spring-mass system for the sensor from the solid wafer (bulk or volume micromechanics). Capacitive pick-offs have proven especially effective for the high-precision measurement of this seismic-mass deflection. They require a wafer-thick silicon or glass plate (Fig. 1, Item 4) with counter-electrodes, on each side of the spring-held mass (2). The wafers and their counter-electrodes here also provide overload protection.

This configuration corresponds to a series circuit of two capacitors C_{1-M} and C_{2-M} . Alternating voltages are applied across terminals C_1 and C_2 , and their superimpositions picked off between the capacitors at C_M .

In the idle condition, the capacitances C_{1-M} and C_{2-M} are ideally equal. This means that the difference ΔC is equal to zero. If an acceleration a is acting in the measuring direction, the Si center plate (the seismic mass) is caused to deflect. This causes a change in the distance to the upper and lower plates, and with it a capacitance change in the capacitors C_{1-M} and C_{2-M} and thus a difference ΔC not equal to zero. This leads to a change in the electrical signal at C_M which is amplified and filtered in the evaluation electronics.

The air space between the plates allows effective damping with low temperature sensitivity.

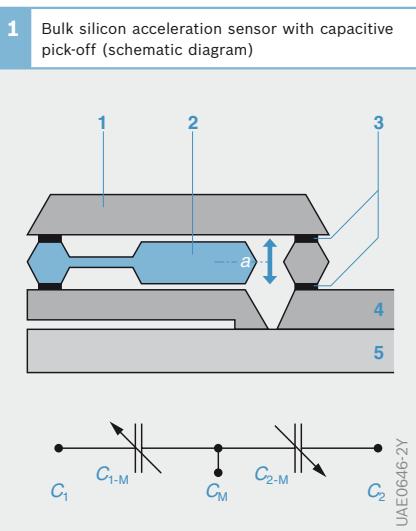


Fig. 1

- 1 Upper Si plate
- 2 Center Si plate (spring-held or seismic mass)
- 3 Si-oxide
- 4 Bottom Si plate
- 5 Glass substrate
- a Acceleration in the sensing direction
- C_M Measuring capacity

Piezoelectric acceleration sensors

Application

Piezoelectric spring elements are used as acceleration sensors in passenger-restraint systems for triggering the seat-belt pretensioners, the airbags, and the rollover bar.

Design and operating principle

The core of this acceleration sensor is a spring element (bending beam). A bonded structure comprising two piezoelectric layers of opposing polarity (bimorphous spring element). An acceleration acting on this element generates a mechanical tensile stress in one layer and a compressive stress in the second layer (Fig. 1).

The metallized surfaces on the top and underneath of the spring element act as electrodes from which the resultant voltage is picked off. This structure is packaged together with the electronics required in a hermetically sealed housing (Fig. 2).

The electronic circuit consists of an impedance converter and a tunable amplifier with predefined filter characteristics. The design means that no static signals

2 Piezoelectric acceleration sensor (two channel sensor for PCB mounting)

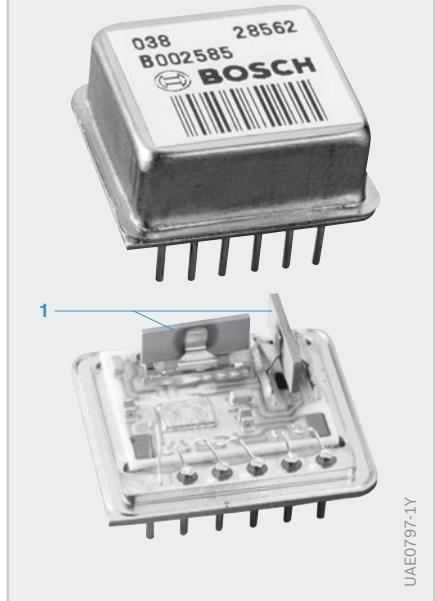


Fig. 2
1 Spring element

can be measured (lower cutoff frequency is typically 1 to 10 Hz).

Piezo-spring elements do not require any additional seismic mass. Their own mass is adequate for an easily evaluated signal.

1 Spring element in a piezoelectric acceleration sensor

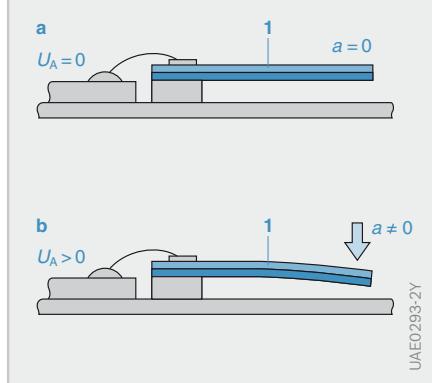


Fig. 1
a Not under acceleration
b Under acceleration a
1 Piezoceramic bimorphous spring element
 U_A Measurement voltage

iBolt™ force sensor

Applications

NHTSA FMVSS-208 (National Highway Traffic Safety Administration; Federal Motor Vehicle Safety Standards and Regulations 208) came into force in the USA in 2004. These regulations were introduced to prevent or reduce injuries to children traveling in the front passenger seat caused by the airbag triggering. Classifying the front passenger by means of a weight measurement makes it possible to switch the airbag off selectively if a small child is occupying the seat.

The BOSCH iBolt™ (intelligent bolt) sensor was developed to perform this weight classification reliably and robustly. This is done by fitting four iBolts™ in the seat frame of the front-passenger seat (a sensor in each corner of the seat, Fig. 1). A control unit also built into the seat analyses the four analog, electrical weight sig-

nals and transmits the result of the classification to the airbag control unit.

Design and operating principle

The iBolt™ sensor's operating principle is based on the measurement of the deflection of a bending beam caused by the passenger's weight. The amount of the deflection is detected by measurement of the magnetic field strength in a special Hall sensor/magnet configuration (Fig. 2a).

The iBolt™ is designed such that it is the z component of the passenger's weight that causes the deflection of the bending beam. The vehicle's co-ordinates system here defines the x axis as the direction of travel, the z and y axes being perpendicular and horizontal to this. The configuration of the magnet and the Hall IC in the sensor is selected such that the static magnetic field which penetrates the Hall IC results in an electrical signal linear to the deflection of the bending beam. The special design of the iBolt™ sensor prevents a horizontal deflection of the Hall IC in relation to the magnet, to minimize the influence of transverse forces and moments. In addition, the maximum stress in the bending beam is limited by a mechanical overload stop (Fig. 2b). This protects the iBolt™, especially in the event of overloads in a crash.

Calibration

The linear Hall sensor uses the spinning current measuring principle which allows calibration of the sensitivity, the offset and the temperature sensitivity of measurement. The calibration data is stored in an EEPROM incorporated in the substrate for the Hall sensor.

Linearity of the output signal

A linear output signal has been achieved by means of a special design feature. The force generated by the weight of the front passenger is passed from the upper seat structure through the sleeve into the bending beam (Fig. 2a). The force is then



Fig. 1

- 1 Seat base frame
- 2 BOSCH iBolt™ sensor

passed from the bending beam into the lower seat structure. The bending beam was designed as a double bending beam as this has an S-shaped deformation line. The two vertical connection points of the double bending beam remain vertical over the entire deflection range. This guarantees a linear and parallel movement of the Hall IC in relation to the magnet, resulting in a linear output signal (Fig. 3).

Symmetrical measuring range

Tests of the system in vehicle seats have shown that both positive and negative forces can act on the sensors. This has a number of causes: on the one hand, negative forces can act on an individual tension as a result of initial tensions as a consequence of tolerances in the assembly of the seat and of its installation in the vehicle. On the other hand, negative forces arise on individual sensors as a function of the distribution of forces on the sensors

generated by the seat occupant's position, which also depends on the position of the seat back. For this reason, the measuring range of the iBolt™ sensor has been designed such that the forces can be detected in both the positive and negative z directions. This permits an unambiguous determination of the front passenger's weight.

Thanks to its symmetrical measuring range, the iBolt™ sensor detects compression and tensile forces with the same sensitivity and the same tolerances. This means that the same sensors can be used in both vertical mounting orientations for all four points of connection between the upper and lower seat structures.

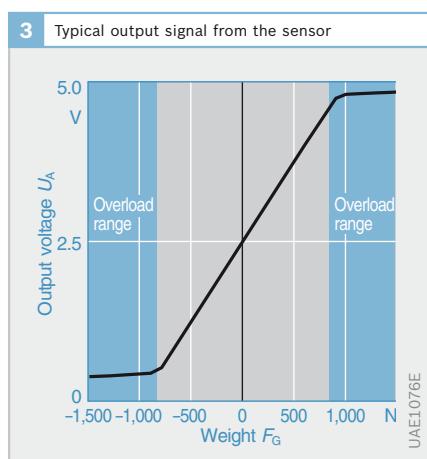
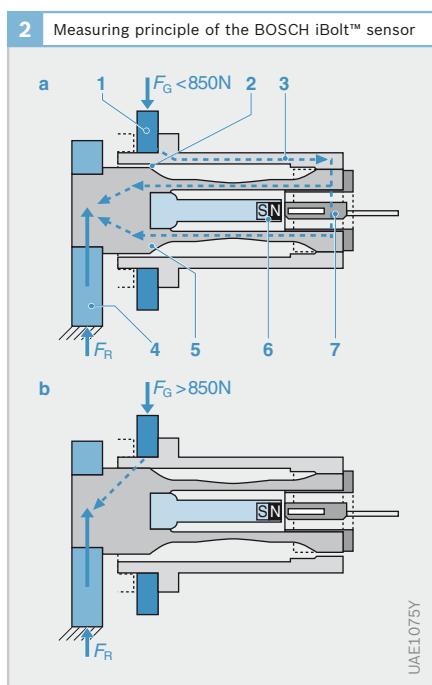


Fig. 2

- a Ratios for weight $F_G < 850$ N (within the measuring range)
- b Ratios for weight $F_G > 850$ N (outside the measuring range)

- 1 Oscillating crank
- 2 Air gap
- 3 Sleeve
- 4 Seat rail
- 5 Double bending beam
- 6 Magnet
- 7 Hall IC

Fig. 3

- Output signal as a function of the force applied

Torque sensor

Application

Electromechanical power-assisted steering is increasingly being introduced in small, compact and medium class vehicles. The essential advantages are simple installation and commissioning in the vehicle, energy saving and the suitability of these systems in the vehicle's network of control units for assistance systems enhancing comfort and safety.

Design and operating principle

In order to sense the driver command in electromechanical power-assisted steering, it is necessary to measure the torque applied by the driver. In the sensor currently in series production use for this, a torsion rod is fitted in the steering shaft. The rod is twisted by a defined amount in linear proportion to the torque applied by the driver when the driver turns the steering wheel (Fig. 1). This twisting can, for its part, be measured by suitable means and converted into electrical signals. The measuring range required for a torque sensor to be used in electromechanical power-assisted steering is usually around ± 8 to ± 10 Nm. The maximum angle of rotation is mechanically limited by special components to protect the torsion rod against overload or destruction.

A magnetoresistive sensor is mounted on one side of the torsion rod to allow measurement of the twisting and hence the torque applied. This scans the field of a magnetic multipole wheel mounted on the other side. The number of poles selected for this wheel is such that the sensor issues an unambiguous signal within its maximum measuring range and thus makes a clear statement of the torque applied at any time.

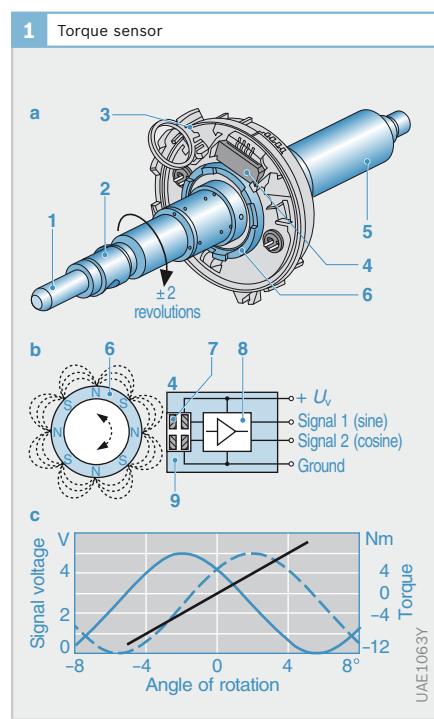
Fig. 1

- a Sensor module
 - b Measuring principle
 - c Electrical output signal
-
- 1 Torsion rod (twisting area internal)
 - 2 Input shaft (from the steering wheel)
 - 3 Housing for the coil spring making the electrical connections
 - 4 Sensor module with magnetoresistive sensor chip and signal amplification
 - 5 Steering pinion/balancer shaft
 - 6 Magnetic multipole wheel

Here the magnetoresistive sensor uses supplies two signals over the measuring range which describe a sine and a cosine signal plotted against the angle of rotation of the torsion rod. The angle of rotation and hence of the torque is calculated in an electronic control unit using an arc-tangent function.

As there is always a fixed assignment of the two signals over the defined measuring range, a deviation from this allows faults in the sensor to be detected and the necessary substitution measures to be introduced.

A coil spring with the necessary number of contacts is used to provide the electrical contacts for the sensor over the rotational range of approximately ± 2 revolutions of the steering wheel. The power is supplied and the measured values are transmitted via this coil spring.



Rain/light sensor

Application of the rain sensor

The rain sensor detects fine raindrops on the windshield, so that the windshield wipers can be triggered automatically. The sensor controls the speed of the wiper system according to the amount of rain measured on the windshield (intermittent, speed 1 and speed 2). Together with the electronically controlled wiper drive, intermittent operation can be varied infinitely. For instance, when masses of rain water flood the windshield when overtaking, the wiper is automatically switched to the highest wiper speed.

Various activation strategies are available:

- ▶ Permanently active
- ▶ Activated via steering column switch
- ▶ Activation after steering column switch movement (transients)

This assistance function frees the driver from the other control operations used on conventional wiper systems. Nevertheless, manual control is retained as an additional intervention.

The rain sensor can also be used for other functions such as closing the windows and the sliding sunroof.

Application of the light sensor

A light sensor is also incorporated in the rain sensor, it may be used to control the low beam. This can detect the various situations (e.g. twilight, driving into and out of a tunnel, driving under long bridges) and switch the low beam on or off accordingly. The light sensor can also be used to control all the lighting functions on the vehicle. These functions include adjusting the brightness of the instrument cluster lighting, "coming home leaving home" or selectively switching on the rear lamps.

Design and operating principle

A rain sensor comprises an optical transmission and reception path (Fig. 1). An LED (6) emits light at a determined angle into the windshield (2) which is reflected on the outer boundary layer (glass/play) (total reflection) and then is evaluated on an aligned receiver (photodiode, 4). If there is moisture on the sensitive outer surface, a part of the light is coupled out as a function of the droplet size and number, which attenuates the return signal.

The integrated light sensor generally comprises two or three photodiodes (3, 5) which receive and evaluate the light coming from various directions. Depending on the function of the light sensor, diodes are used that reflect the receptiveness of the human eye (silicon eyes) or have their maximum sensitivity rather in the near infrared range.

Complex filter functions and trigger strategies are stored in the rain/light sensor. The status information derived from this (e.g. wiper speed 1, intermittent wiper, intermittent timing, lights on, lights off) are made available to other control units via a data bus (e.g. CAN, LIN).

Fig. 1 Rain/light sensor for windshields

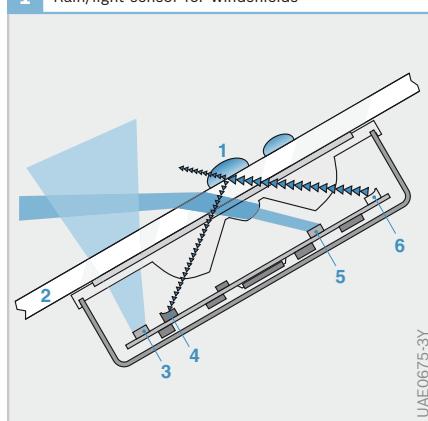


Fig. 1

- | | |
|---|----------------------------------|
| 1 | Raindrops |
| 2 | Windshield |
| 3 | Ambient-light sensor |
| 4 | Photodiode |
| 5 | Light sensor set up for distance |
| 6 | LED |
- UAE0675-3Y

Two-step Lambda oxygen sensors

Application

These sensors are used in gasoline engines equipped with two-step Lambda oxygen sensor. They extend into the exhaust pipe between the engine's exhaust manifold and the catalytic converter, and record the exhaust-gas flow leaving each cylinder. Thanks to the Lambda oxygen sensor being heated, it can be installed further away from the engine so that long periods of full-load engine operation are unproblematical. The LSF4 sensor is also suitable for operation with exhaust-gas control systems featuring a number of sensors (for instance with OBDII).

Two-step Lambda oxygen sensors compare the residual-oxygen content in the exhaust gas with the oxygen content in the reference atmosphere (surrounding air inside the sensor) and indicate whether a rich ($\lambda < 1$) or lean air-fuel mixture ($\lambda > 1$) is present in the exhaust gas. The sudden jump in the characteristic curve of these sensors permits air-fuel control of $\lambda = 1$ (Fig. 1).

Operating principle

The two-step Lambda oxygen sensors operate in accordance with the principle of the galvanic oxygen-concentration cell with solid-state electrolyte (Nernst principle). The ceramic element is conductive for oxygen ions from a temperature of approximately 350 °C (safe, reliable operation at >350 °C). Due to the abrupt change in the residual-oxygen content on the exhaust-gas side in the range of $\lambda = 1$ (e.g. $9 \cdot 10^{-15}$ % vol. for $\lambda = 0.99$ and 0.2 % vol. for $\lambda = 1.01$), the different oxygen content on both sides of the sensor generates an electrical voltage between the two boundary layers. This means that the oxygen content in the exhaust gas can be used as a measure of the air-fuel ratio. The integrated heater ensures that the sensor

functions even at extremely low exhaust-gas temperatures.

The sensor's output voltage U_S is a function of the oxygen content in the exhaust gas. In the case of a rich air-fuel mixture ($\lambda < 1$) it reaches 800 to 1,000 mV. For a lean mixture ($\lambda > 1$), only approximately 100 mV are generated. The transition from the rich to the lean area at U_{reg} is at about 450 to 500 mV.

The temperature of the ceramic element influences its ability to conduct the oxygen ions, and thus the shape of the output-voltage curve as a function of the excess-air factor λ (the values in Figure 1 are therefore temperature-dependent). In addition, the response time for a voltage change when the mixture composition changes is also strongly dependent on temperature.

Whereas response times at ceramic temperatures below 350 °C are in the seconds range, at optimum temperatures of around 600 °C the sensor responds in less than 50 ms. When the engine is started therefore, the Lambda control is switched off until the minimum operating temperature of about 350 °C is reached. During this period, the engine is open-loop-controlled.

1 Voltage curve of a two-step Lambda oxygen sensor for different operating temperatures

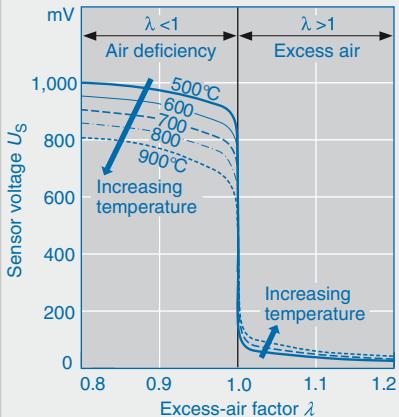


Fig. 1

- a Rich mixture (air deficiency)
- b Lean mixture (excess air)

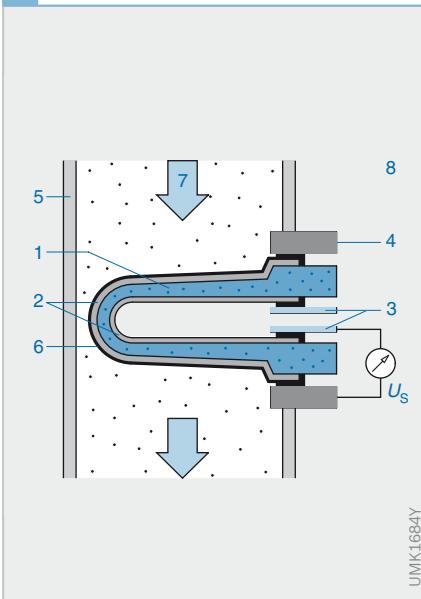
Design

LSH25 finger-type sensor

Sensor ceramic element with protective tube
The solid electrolyte is a ceramic element which is impermeable to gas. It is a mixed oxide comprising the elements zirconium and yttrium in the form of a tube closed at one end (finger, Fig. 3). The surfaces have been provided on both sides with electrodes made from a microporous, thin noble-metal layer.

The ceramic body protrudes into the exhaust pipe, and the platinum electrode on its outside surface acts as a catalytic converter in miniature. Exhaust gas which reaches this electrode is processed catalytically and brought to a stoichiometrical balance ($\lambda = 1$). In addition, the side that is exposed to the exhaust gas has a porous, ceramic multiple layer (spinel layer) to protect it against contamination and

3 Configuration of a finger-type Lambda oxygen sensor in the exhaust pipe

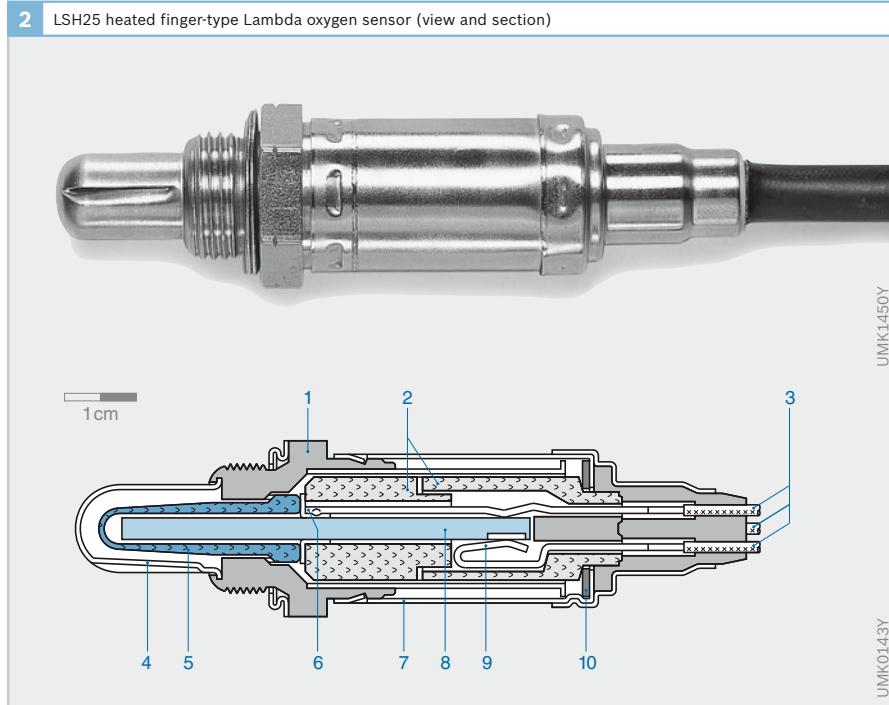


UMK1684Y

Fig. 3

- 1 Sensor ceramic element
- 2 Electrodes
- 3 Contacts
- 4 Housing contact
- 5 Exhaust pipe
- 6 Ceramic protective layer (porous)
- 7 Exhaust gas
- 8 Outside air
- U_s Sensor voltage

2 LSH25 heated finger-type Lambda oxygen sensor (view and section)



UMK1450Y

UMK0143Y

Fig. 2

- 1 Sensor housing
- 2 Ceramic support tube
- 3 Connecting cable
- 4 Protective tube with slots
- 5 Active sensor ceramic
- 6 Contact element
- 7 Protective sleeve
- 8 Heater element
- 9 Clamp-type connections for the heater element
- 10 Disk spring

erosive damage. The ceramic element is also protected against mechanical impact and thermal shocks by a metal tube. A number of slots in the protective tube are specially shaped so that on the one hand they are particularly effective against extensive thermal and chemical stresses, while on the other hand preventing a sharp drop in sensor-ceramic temperature when the exhaust gas is "cool".

The sensor's "open" inner chamber is connected to the surrounding air, which acts as a reference gas (Fig. 3).

Sensor element with heater element and electrical connection

A ceramic support tube (Fig. 2, Item 2) and a disk spring (10) serve to locate, fix, and seal-off the active, finger-shaped sensor ceramic in the sensor housing. A contact element (6) between the support tube and the active sensor ceramic element provides the contact between the inner electrode and the connection cable.

The outer electrode is connected to the sensor housing by the metal sealing ring.

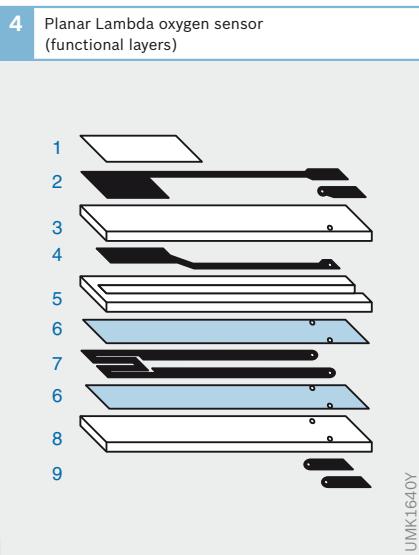
The sensor's complete internal structure is located and held in place by a protective metal sleeve (7) which at the same time acts as the support for the disk spring. The protective sleeve also guards against contamination of the sensor's interior. The connecting cable is crimped to the contact element which protrudes from the sensor, and is protected against moisture and mechanical damage by a temperature-resistant cap.

The tube-type (finger) Lambda oxygen sensor is also equipped with an electrical heater element. This ensures that the ceramic-element temperature remains sufficiently high, even at low engine load and thus low exhaust-gas temperature.

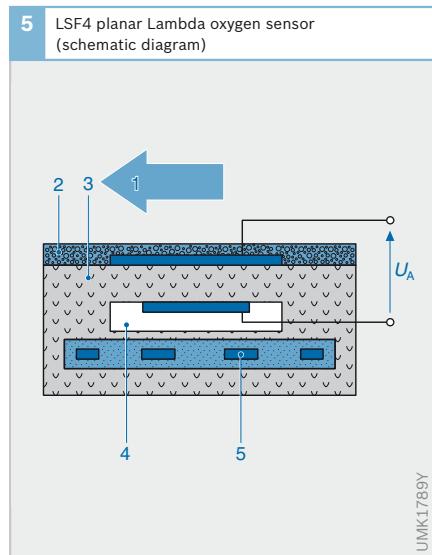
This external heating is so quick that the sensor reaches operating temperature 20 to 30 s after the engine has started and therefore Lambda control can come into operation. Finally, sensor heating provides for an optimal ceramic-element operating temperature above the operating limit of 350 °C and thus ensures low and stable exhaust-gas emissions.

Fig. 4

- 1 Porous protective layer
- 2 Outer electrode
- 3 Sensor foil
- 4 Inner electrode
- 5 Reference-air-channel foil
- 6 Insulation layer
- 7 Heater
- 8 Heater foil
- 9 Connection contacts

**Fig. 5**

- 1 Exhaust gas
- 2 Porous ceramic protective layer
- 3 Measuring cell with microporous noble-metal layer
- 4 Reference-air-channel
- 5 Heater
- U_A Output voltage



LSF4 planar Lambda oxygen sensor

Regarding their function, planar Lambda oxygen sensors correspond to the heated finger sensors with their voltage-jump curve at $\lambda = 1$. On the planar sensor though, the solid-state electrolyte is comprised of a number of individual laminated foils stacked one on top of the other (Fig. 4). The sensor is protected against thermal and mechanical influences by a double-walled protective tube.

The planar ceramic element (measuring cell and heater are integrated) is shaped like a long stretched-out wafer with rectangular cross-section.

The surfaces of the measuring cell are provided with a microporous noble-metal layer, which also has a porous ceramic layer on the exhaust-gas side to protect it against the erosive effects of the exhaust-residues. The heater is a wave-shaped element containing noble metal.

It is integrated and insulated in the ceramic wafer and ensures that the sensor heats up quickly even in the event of low power input.

The reference-air channel inside the LSF4 Lambda oxygen sensor - operating as a reference-gas sensor (Figs. 5 and 6) - has access to the ambient air. This enables it to compare the residual oxygen content of the exhaust gas with the oxygen content of the reference gas. In other words, with the air inside the sensor. This means that the sensor voltage also features a characteristic jump in the case of the planar sensor, in the area of stoichiometric air-fuel mixture ($\lambda = 1$) (Fig. 1).

6 LSF4 planar Lambda oxygen sensor (view and section)

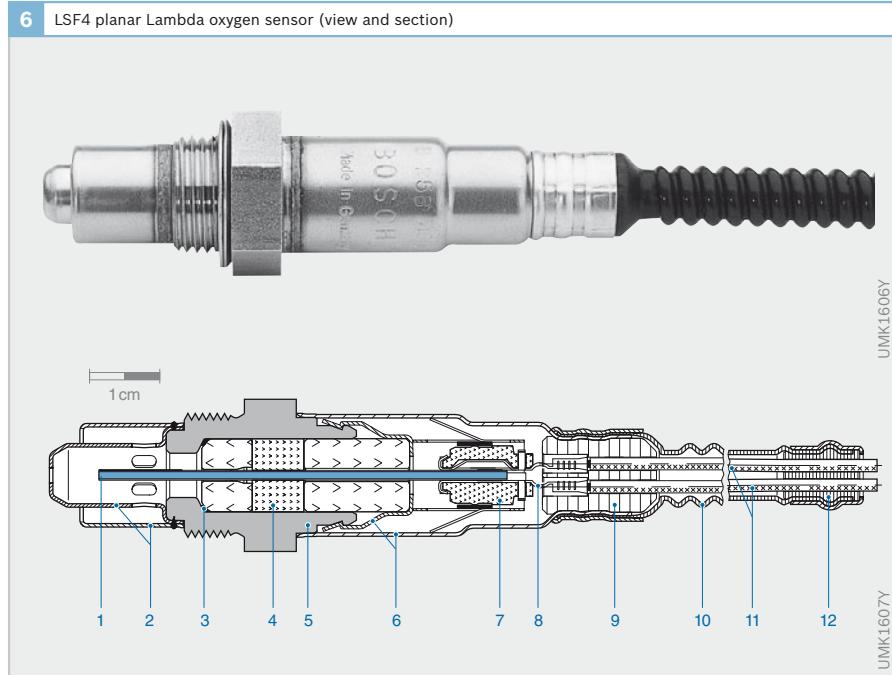


Fig. 6

- 1 Planar measuring cell
- 2 Double protective tube
- 3 Sealing ring
- 4 Seal packing
- 5 Sensor housing
- 6 Protective sleeve
- 7 Contact holder
- 8 Contact clip
- 9 PTFE grommet
- 10 PTFE shaped sleeve
- 11 Five connection cables
- 12 Seal

LSU4 planar wide-band Lambda oxygen sensor

Application

As its name implies, the wide-band Lambda oxygen sensor is used across a very extensive range to determine the oxygen concentration in the exhaust gas. The figures provided by the sensor are an indication of the air-fuel ratio in the combustion chamber. The excess-air factor λ is used when defining the air-fuel ratio.

The sensor projects into the exhaust pipe and registers the exhaust-gas mass flow from all cylinders. Wide-band Lambda oxygen sensors make precise measurements not only at the stoichiometric point $\lambda = 1$, but also in the lean range ($\lambda > 1$) and in the rich range ($\lambda < 1$). In combination with control electronics, these sensors generate an unmistakable, continuous electrical signal in the range from $0.7 < \lambda < \infty$ (air with 21 % O₂) (Fig. 3). This means that the wide-band Lambda oxygen sensor can be used not only in engine-management systems with two-step control ($\lambda=1$), but also in control concepts with rich and lean air-fuel mixtures. This type of Lambda oxygen sensor is therefore also suitable for Lambda control with lean-burn concepts on gasoline engines, as well as for diesel engines, gas-

powered engines and gas-powered central heaters and water heaters (hence the German designation LSU: Lambda-Sonde-Universal = universal Lambda oxygen sensor).

In a number of systems, several Lambda oxygen sensors are installed for even greater accuracy. Here, for instance, they are fitted upstream and downstream of the catalytic converter as well as in the individual exhaust tracts (cylinder banks).

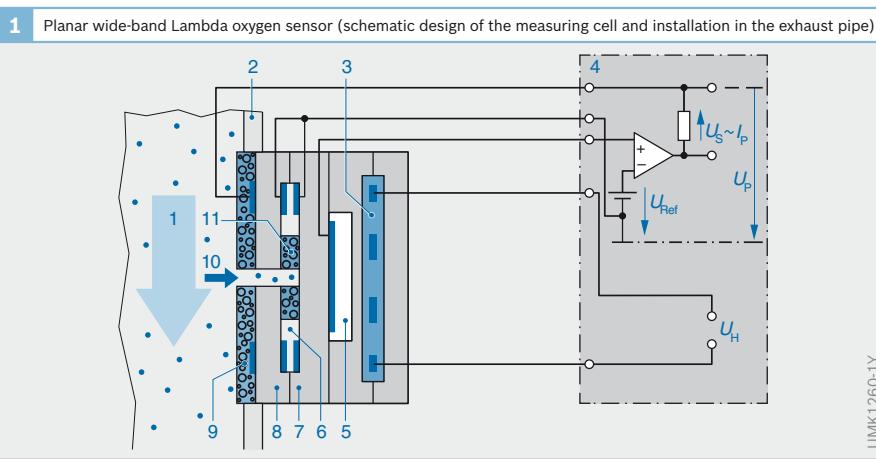
Design

The LSU4 wide-band Lambda oxygen sensor (Fig. 2) is a planar dual-cell current-limit sensor. It features a measuring cell (Fig. 1) made of zirconium-dioxide ceramic (ZrO₂), and is a combination of a Nernst concentration cell (sensor cell which functions in the same way as a two-step Lambda oxygen sensor) and an oxygen pump cell for transporting the oxygen ions. The oxygen-pump cell (Fig. 1, Item 8) is so arranged with respect to the Nernst concentration cell (7) that there is a 10 to 50 µm diffusion gap (6) between them. Here, there are two porous platinum electrodes: one pump electrode and one Nernst measuring electrode. The diffusion gap is connected to the exhaust gas by way of a gas-access passage (10). A porous diffusion barrier (11) serves to limit the flow of oxygen molecules from the exhaust gas.

Fig. 1

- 1 Exhaust gas
- 2 Exhaust pipe
- 3 Heater
- 4 Control electronics
- 5 Reference cell with reference-air channel
- 6 Diffusion gap
- 7 Nernst concentration cell with Nernst measuring electrode (on the diffusion-gap side) and reference electrodes (on the reference-cell side)
- 8 Oxygen-pump cell with pump electrode
- 9 Porous protective layer
- 10 Gas-access passage
- 11 Porous diffusion barrier

- I_p Pump current
- U_p Pump voltage
- U_h Heating voltage
- U_{ref} Reference voltage (450 mV, corresponds to $\lambda = 1$)
- U_s Sensor voltage



On the one side, the Nernst concentration cell is connected to the atmosphere by a reference-air channel (5), on the other it is connected to the exhaust gas in the diffusion gap.

The sensor requires control-electronics circuitry to generate the sensor signal and to regulate the sensor temperature.

An integral heater (3) heats up the sensor quickly so that it soon reaches its operating temperature of 650 to 900 °C needed for generating a usable signal. This function decisively reduces the effects that the exhaust-gas temperature has on the sensor signal.

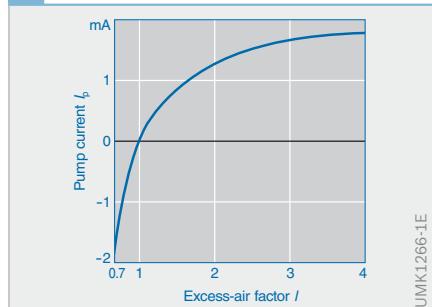
Operating principle

The exhaust gas enters the actual measuring chamber (diffusion gap) of the Nernst concentration cell through the pump cell's small gas-access passage. In order that the excess-air factor λ can be adjusted in the diffusion gap, the Nernst concentration cell compares the gas in the diffusion gap with that in the reference-air channel.

The complete process proceeds as follows: by applying the pump voltage U_P across the pump cell's platinum electrodes, oxygen from the exhaust gas can be pumped through the diffusion barrier and into or out of the diffusion gap. With the help of the Nernst concentration cell, an electronic circuit in the control unit con-

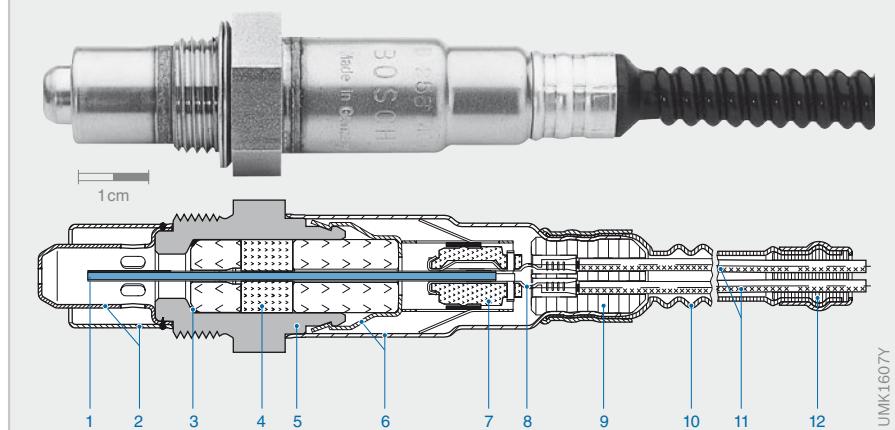
trols the voltage (U_P) across the pump cell in order that the composition of the gas in the diffusion gap remains constant at $\lambda = 1$. If the exhaust gas is lean, the pump cell pumps the oxygen to the outside (positive pump current). On the other hand, if the exhaust gas is rich, the oxygen (due to the decomposition of CO₂ and H₂O at the exhaust-gas electrode) is pumped from the surrounding exhaust gas and into the diffusion gap (negative pump current). At $\lambda = 1$, no oxygen needs to be transported, and the pump current is zero. The pump current is proportional to the exhaust-gas oxygen concentration and is thus a nonlinear measure for the excess-air factor λ (Fig. 3).

3 Pump current I_P of a wide-band Lambda oxygen sensor as a function of the excess-air factor λ of the exhaust gas



- Fig. 2**
- 1 Measuring cell (combination of Nernst concentration cell and oxygen-pump cell)
 - 2 Double protective tube
 - 3 Sealing ring
 - 4 Seal packing
 - 5 Sensor housing
 - 6 Protective sleeve
 - 7 Contact holder
 - 8 Contact clip
 - 9 PTFE grommet
 - 10 PTFE shaped sleeve
 - 11 Five connecting cables
 - 12 Seal

2 LSU4 planar wide-band Lambda oxygen sensor (view and section)

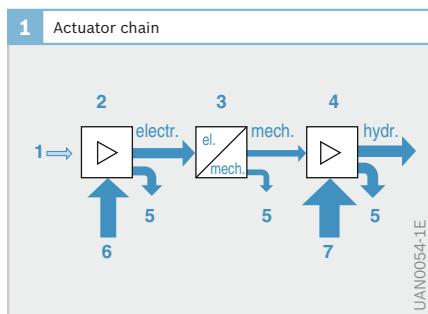


Actuators

Actuators (final-control elements) form the interface between the electronic signal processor (data processing) and the actual process (mechanical motion). They convert the low-power signals conveying the positioning information into operating signals of an energy level adequate for process control. Signal transducers are combined with amplifier elements to exploit the physical transformation principles governing the interrelationships between various forms of energy (electrical - mechanical - fluid - thermal).

Fig. 1

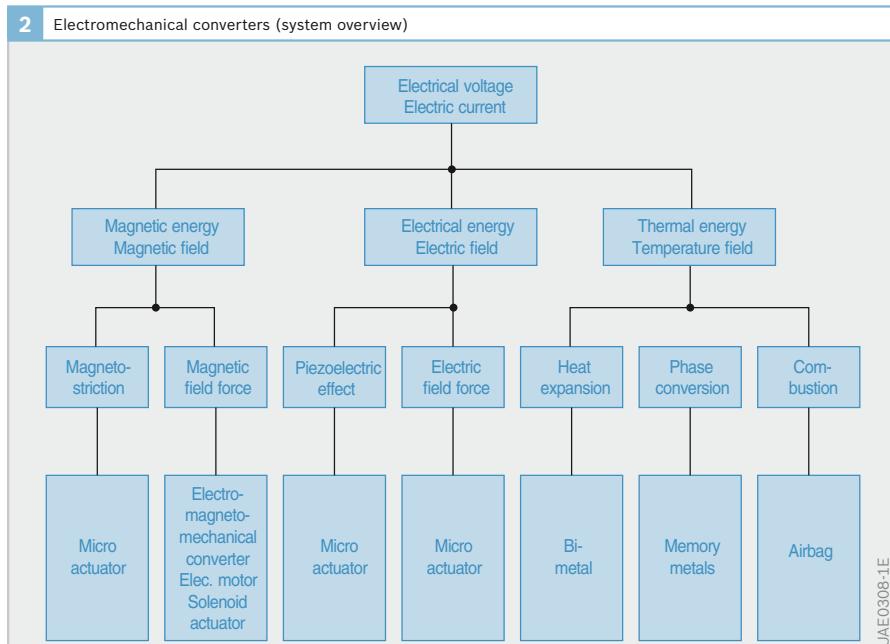
1	Information
2	Actuator
3	Converter
4	Actuator
5	Losses
6	External electrical energy
7	External hydraulic energy



Electromechanical actuators

This type of energy conversion represents one option for classifying electromechanical actuators. The energy emanating from the source is transformed into magnetic or electrical field energy, or converted to thermal energy. The individual force-generation principle is determined by these forms of energy, and is based on either field forces or certain specific material characteristics. Magnetostrictive materials make it possible to design actuators for applications in the micropositioning range. This category also includes piezoelectric actuators, which are built according to a multilayer design similar to ceramic capacitors, and are actuators for high-speed fuel injectors. Thermal actuators depend exclusively on the exploitation of characteristics of specific materials.

Actuators in a motor vehicle are mostly electromagneto-mechanical converters and, by extension, electrical servomotors, translational, and rotational solenoid actuators.



An exception is the pyrotechnic airbag system. The solenoid actuators can themselves be the servo element, or they can assume a control function by governing a downstream amplifier (e.g. mechanical-hydraulic).

Force generation in the magnetic field

The distinction between the electrodynamic and the electromagnetic actuator principles stems from the manner in which forces are generated in the magnetic field. Common to both principles is the magnetic circuit with soft-magnetic material and the coil for excitation of the magnetic field. A major difference lies in the force which can be extracted from the unit under technically feasible conditions. Under identical conditions, the force produced by application of the electromagnetic principle is greater by a factor of 40. The electrical time constant for this type of actuator is comparable to the mechanical time constants. Both force-generation principles are applied in linear and rotary drive mechanisms.

Electrodynamic principle

The electrodynamic principle is based on the force exerted on moving charges or charged conductors (Fig. 3a, current I) within a magnetic field (Lorentz force).

A field coil or a permanent magnet generates a DC magnetic field (magnetic flux density B). The electrical energy destined for conversion is applied (current I) to the moving armature coil (immersion coil). A high degree of actuator precision is achieved by designing the armature coil with low mass and low inductance. The two energy accumulators (one on the fixed and one on the moving component) produce two active force directions via current-direction reversal in the armature and field coils.

The secondary field produced by the armature current flows in an open magnetic circuit, thereby diminishing the effects of saturation. Approximately speaking, the force (torque) exerted by an electrodynamic actuator over its setting range is proportional to current and independent of travel.

Electromagnetic principle

The electromagnetic principle exploits the mutual attraction displayed by soft iron materials in a magnetic field. The electromagnetic actuator is equipped with only one coil, which generates both the field energy and the energy to be transformed. In accordance with the operating principles, the field coil is equipped with an iron core to provide higher inductance.

3 Electrodynamic and electromagnetic converters

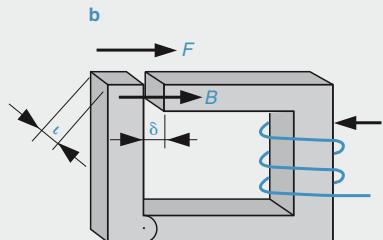
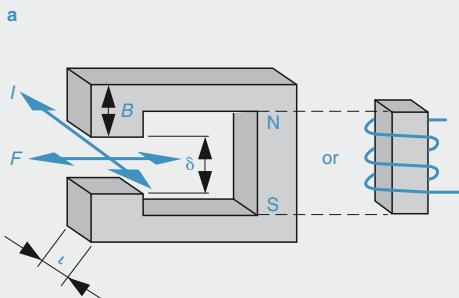


Fig. 3

a Electrodynamic principle:
Force on moving charge
 $F = B \cdot I \cdot l \cdot \sin \alpha$

b Electromagnetic principle:
Force between pole faces
 $F = B^2 \cdot A / 2\mu_0$

However, as the force is proportional to the square of the magnetic flux density B , the unit is operative in only a single force-transfer direction. The electromagnetic actuator thus requires a return element (such as a mechanical spring or a magnetic return mechanism).

Dynamic response

The dynamic response of an electromechanical actuator, i.e. the activation and deactivation operations, is defined by the equation of mechanical motion, the differential equation of electrical circuits and Maxwell's equations of dynamics. The current- and position-dependent force follows from Maxwell's equations.

The most basic electrical circuit consists of an inductance with an ohmic resistor. One means of enhancing the dynamic response is through over-excitation at the instant of activation, while deactivation can be accelerated by a Zener diode, for example. In each case, increasing the dynamic response of the electric circuit involves additional expenditure and increased losses in the actuator's triggering electronics.

Field diffusion is a delay effect which is difficult to influence in actuators with high dynamic response. Rapid switching operations are accompanied by high-frequency field fluctuations in the soft-magnetic material of the actuator's magnetic circuit. These fluctuations, in turn, induce eddy currents, which counteract their cause (build-up and decay of the magnetic field). The resultant delay in the build-up or reduction of forces can only be reduced by selecting appropriate materials with low electric conductivity and permeability.

Design

Design selection is essentially determined by operating conditions (i.e. installation space, required force/travel characteristic curve, and dynamic response).

Fig. 4

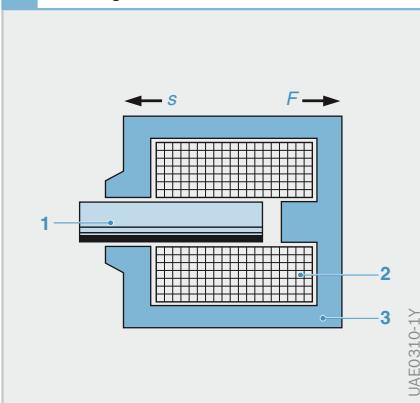
- 1 Armature
- 2 Coil
- 3 Magnetic yoke

Electromagnetic actuators

A typical form for translational electromagnetic actuators is the switching solenoid (Fig. 4) with a force/travel characteristic curve which falls as a function of the square of positioning travel. The precise shape of the curve (Fig. 5) is determined by the type of working air gap (e.g. conical or solenoid armature).

Rotational electromagnetic actuators are characterized by a defined pole arrangement in stator and rotor. When current is applied to one of the coils, the rotor and stator poles respond with mutual attraction, and in doing so generate a torque.

4 Switching solenoid



5 Switching solenoid (characteristic curves)

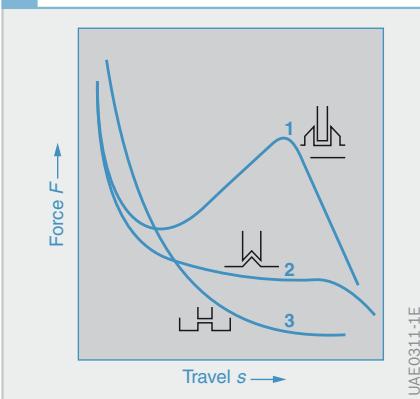
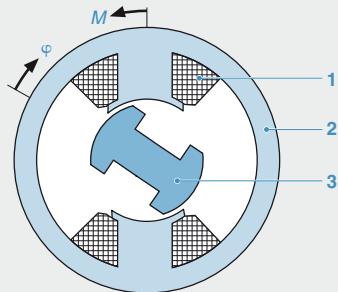


Fig. 5

- 1 Solenoid plunger
- 2 Conical armature
- 3 Cylindrical armature

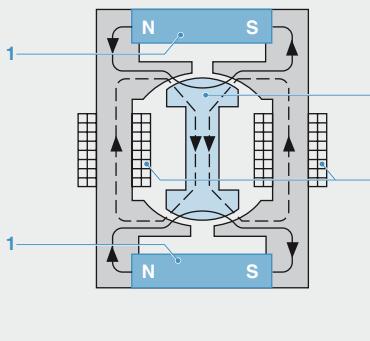
6 Electromagnetic single-winding rotary actuator

The single-winding rotary actuator (Fig. 6) incorporates a pair of poles in each of the two main sections, as well as a coil in the stator. Its maximum adjustment range is approx. 45°.

The torque motor (Fig. 7) is a bidirectional electromagnetic rotary actuator featuring a stable operating point and without counterforces. The rotor is maintained in a stable position by the excitation field of the permanent magnet in the stator. The magnetic field generated by one or two stator windings produces torque and provides unilateral compensation for the excitation field. This type of layout is suitable for applications in which substantial torque is required over small control angles. The relationship between the applied current and the torque motor's force is roughly linear (Fig. 8). The torque-motor principle is also employed in translational actuators.

Fig. 6

- 1 Coil
- 2 Stator
- 3 Armature

7 Torque motor

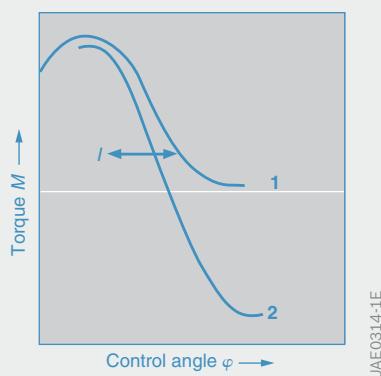
Electrodynamic actuators

In a pot magnet (immersion-coil actuator, Fig. 9) a cylindrical immersion coil (armature coil) is set in motion in a working air gap. The adjustment range is limited by the axial length of the armature coil and by the air gap.

The short-stroke linear motor (Fig. 10) is an actuator with a virtually round disk coil.

Fig. 7

- 1 Magnets
- 2 Armature
- 3 Control coils

8 Electromagnetic rotary actuators (characteristics)

A distinction is made between single-winding and dual-winding rotary actuators (Fig. 11). Both types include a permanent magnet within the rotor and one or two stator windings. The rotor magnet is magnetized at both ends to produce magnetic flux density in the rotary magnet's working air gap, which combines with the armature current to produce a torque. Originating from the position illustrated, the adjustment range is less than $\pm 45^\circ$. The setting range of the single-winding rotary actuator also varies according to both the torque requirement and the angle range in which the necessary flux density can be provided.

Fig. 8

- 1 Single-winding rotary actuator
- 2 Torque motor

The dual-winding rotary actuator can be described as a combination of two single-winding rotary actuators with a 90° peripheral offset and designed to produce opposed torque flows. A stable operating point is achieved at the zero transition point on the resulting torque curve without auxiliary counterforces.

Application

Electromechanical actuators are direct-action control elements, and without an intermediate ratio-conversion mechanism, they convert the energy of the electrical control signal into a mechanical positioning factor/work. Typical applications include positioning of flaps, sleeves, and valves. The described actuators are final-control elements without internal return mechanisms, i.e. without a stable operating point. They are only capable of carrying out positioning operations from a stable initial position (operating point) when a counterforce is applied (e.g. return spring and electrical control).

A solenoid armature provides a stable static operating point when its force/travel curve is superimposed on the characteristic response of a return spring. A variation of the coil current in the solenoid shifts the operating point. Simple positioning is achieved by controlling the current. However, particular attention must be paid here to the nonlinearity of the force-current characteristic and the positioning system's sensitivity to interference factors (e.g. mechanical friction, pneumatic, and hydraulic forces). The temperature sensitivity of the coil resistance results in positioning errors, making corrective current control necessary. A high-precision positioning system with good dynamic response must incorporate a position sensor and a controller.

Fig. 9

- 1 Immersion coil
- 2 Permanent magnet
- 3 Magnetic yoke

Figure 10

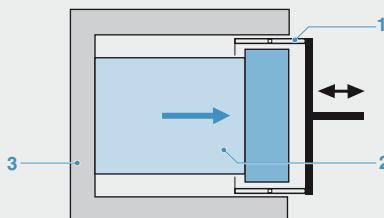
- 1 Coil
- 2 Permanent magnet
- 3 Magnetic yoke

Fig. 11

- a Single-winding rotary actuator
- b Dual-winding rotary actuator

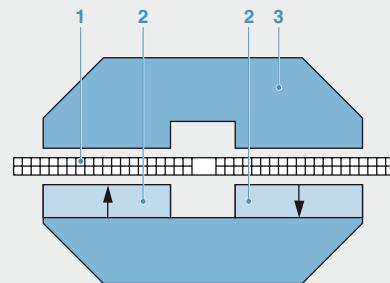
- 1 Coil 1
- 2 Coil 2
- 3 Stator
- 4 Permanent magnet
- 5 Shaft

9 Electrodynamic immersion-coil actuator



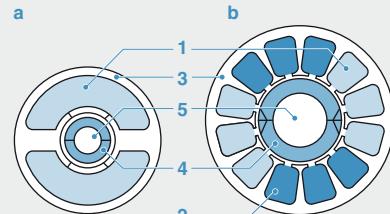
UAE0315-1Y

10 Electrodynamic short-stroke linear motor

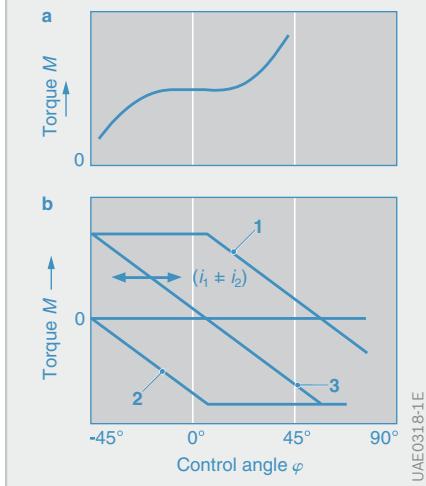


UAE0316-1Y

11 Electrodynamic rotary actuator



UAE0317-1Y

12 Electrodynamic rotary actuator (characteristic curves)

Fluid-mechanical actuators

Hydraulic and pneumatic actuators utilize similar principles for the conversion and regulation of energy. Table 1 shows the differences in characteristics and applications.

In most applications, fluid-mechanical actuator drives are in the form of hydrostatic energy converters. These operate according to the displacement principle, converting the pressure energy of the fluid medium into mechanical work and vice versa (Fig. 14).

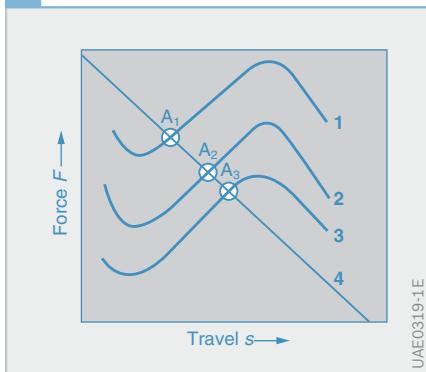
In contrast, hydrodynamic transformers operate by converting flow energy (kinetic energy of the moving fluid) into mechanical work (example: hydrodynamic clutch).

Losses during energy conversion stem from leakage and friction. Fluid-thermal losses are caused by flow resistance, in which throttle action transforms the hydraulic energy into heat. A portion of this heat is dissipated into the environment, and some of it is absorbed and carried away by the fluid medium:

$$Q_{\text{heat}} = Q_1 \cdot p_1 - Q_2 \cdot p_2$$

With incompressible fluids:

$$Q_{\text{heat}} = Q_1 \cdot (p_1 - p_2)$$

13 Operating points (A) of a linear solenoid**Fig. 12**

- a Single-winding rotary actuator
- b Dual-winding rotary actuator

1 Coil 1

2 Coil 2

3 Coils 1+2

1 Comparison of hydraulic and pneumatic actuators

	Hydraulic actuators	Pneumatic actuators
Medium	<ul style="list-style-type: none"> - Fluid, mainly oil - Supplied from tank, oil sump - Virtually incompressible - Self-lubricating - Viscosity highly temperature-dependent 	<ul style="list-style-type: none"> - Gas, mainly air - Supplied from ambient air - Compressible - Additional lubrication necessary - Viscosity changes practically irrelevant
Pressure section	<ul style="list-style-type: none"> - To approx. 30 MPa (200 MPa for diesel-fuel injectors) 	<ul style="list-style-type: none"> - To approx. 1 MPa or greater (approx. 0.05 MPa for vacuum actuators)
Line connections	<ul style="list-style-type: none"> - Supply and return connection (possible leakage connection) 	<ul style="list-style-type: none"> - Pressure connection only, return directly to environment
Applications	<ul style="list-style-type: none"> - Positioning applications with high load rigidity, demanding requirements for synchronization and positioning precision in closed-loop control system 	<ul style="list-style-type: none"> - Actuators with low power requirement, positioning by mechanical stops, in open control loop

Fig. 13

- 1, 2, 3 Characteristic curves for various currents
- 4 Characteristic curve of a return spring

Table 1

The flow develops into turbulence at restrictions. The flow rate of the fluid is then largely independent of viscosity. On the other hand, viscosity does play a role in the case of laminar flow in narrow pipes and apertures.

Fluid-mechanical amplifiers control the conversion of energy from fluid to mechanical state. The regulating mechanism must be designed for control with only a very small proportion of the energy flow required for the ultimate positioning operation.

Switching valves open and close the orifice governing the flow to and from a fluid-mechanical energy converter (Fig. 15). Provided that the control-element opens sufficiently, the throttling losses remain negligible. Pulse-width-modulated opening and closing can be applied to achieve quasi-continuous control of the fluid-mechanical energy conversion process with virtually no losses. In practice, however, pressure fluctuations and mechanical contact between the valve elements result in undesirable vibration and noise.

14 Fluid-mechanical actuators

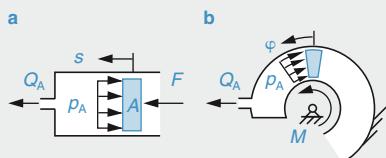


Fig. 14

a Linear actuator
 $F = p_A \cdot A$
 $s = Q_A / A$

b Rotary actuator
 $M = (p_A \cdot V_{th}) / 2\pi$
 $\varphi = (Q_A / V_{th}) \cdot 2\pi$

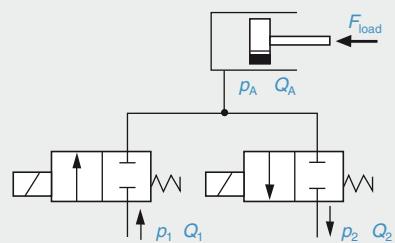
Electrical machines

Electrical machines are used to convert electrical and mechanical energy. An electric motor converts electrical energy into mechanical energy, and an alternator converts energy in the opposite direction. Electrical machines (Fig. 16) consist of a stationary component (the stator) and a rotating component (the rotor or armature). There are special designs which depart from this configuration, such as linear machines which produce linear motion.

Permanent magnets or several coils (windings) are used to produce magnetic fields in the stator and rotor. This causes torque to develop between the two machine components. Electrical machines have iron stators and rotors in order to control the magnetic fields. As the magnetic fluxes change over time, stators and rotors must consist of stacks of individual laminations which are insulated with respect to one another (to minimize eddy-current losses).

The spatial arrangement of the coils and the type of current used (direct current, alternating current, or three-phase current) permit a number of different electrical machine designs. They differ from one another in the way they operate, and therefore have different applications.

15 Energy regulation with 2/2-way switching valves



UAH0104-1E

Direct-current machines

The stator of a direct-current machine contains salient poles which are magnetized by the direct-current excitation windings. In the rotor (here also called the armature), the coils are distributed among slots in the laminated iron core and connected to a commutator. Carbon brushes in the stator housing wipe against the commutator as it rotates, thereby transferring direct current to the armature coils. The rotary motion of the commutator causes a reversal in the direction of current flow in the coils. The different rotational speed vs. torque characteristics result from the method selected for connecting the excitation winding and armature.

Series connection (series characteristics, Fig. 17)

- ▶ Clear relationship between speed and load
- ▶ High starting torque
- ▶ Unacceptably high speed possible at the moment of load reduction, requiring therefore a rigid connection with the load
- ▶ Torque reversal (change in direction of rotation) caused by a change in the current direction in the armature or excitation winding
- ▶ Application, e.g. as vehicle drive motor or starter for internal-combustion engines

Parallel connection (shunt characteristic, Fig. 18)

- ▶ Speed varies only slightly with load
- ▶ Change in direction of rotation caused by a change in the current direction in the armature or excitation winding
- ▶ Application, e.g. as drive motor for machine tools or direct-current generators

16 Layout of an electrical machine

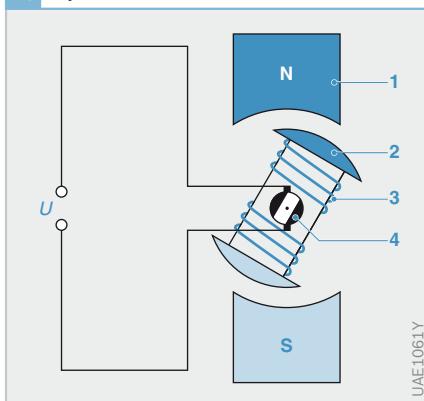
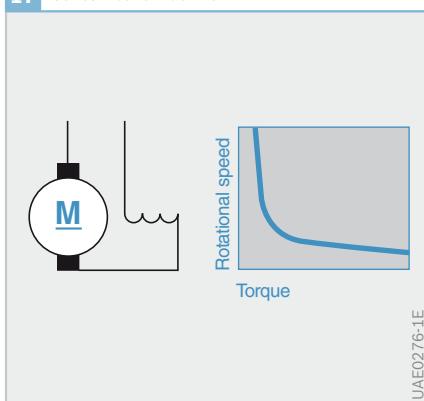


Fig. 16

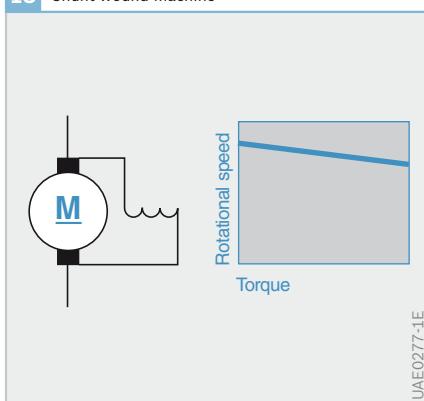
- 1 Stator
- 2 Rotor
- 3 Rotor coil
- 4 Commutator

Stator field generated by permanent magnet or electromagnet

17 Series-wound machine



18 Shunt-wound machine



A shunt characteristic can also be obtained by using a separate power supply for the excitation winding (external excitation) or by using permanent-magnet excitation in the stator.

Applications for permanent-field motors in motor vehicles: starter, windshield wiper, and low-power motors for various drives.

If the motor incorporates both series and shunt excitation windings (compound-wound motor), intermediate levels in the rotational speed/torque characteristic can be obtained.

Application: e.g. large starters.

All direct-current machines are easily capable of speed control over a wide range. If the machine incorporates a static converter which allows adjustment of the armature voltage, the torque and therefore the rotational speed are infinitely variable. The rotational speed can be further increased by reducing the excitation current (field weakening) when the rated armature voltage is reached. A disadvantage of direct-current machines is carbon-brush and commutator wear which makes regular maintenance necessary.

Three-phase machines

A three-phase winding is distributed among the stator slots in a three-phase machine. The three phases of current produce a rotating magnetic field. The speed n_0 (in rpm) of this rotating field is calculated as follows:

$$n_0 = 60 \cdot f / p$$

f = frequency (in Hz)

p = number of pole pairs

Three-phase machines are either synchronous or asynchronous, depending on rotor design.

Fig. 19

a Stator winding, star-connected

b Stator winding, delta-connected

Asynchronous machines

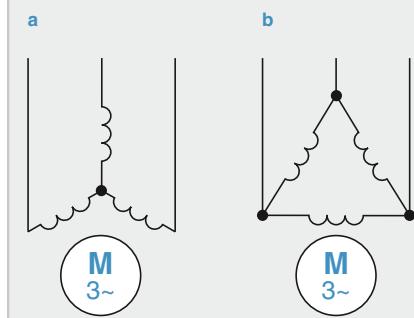
The laminated rotor contains either a three-phase winding, as in the stator, or a rod winding. The three-phase winding is connected to sliprings which are short-circuited either directly or via series resistors (Fig. 19). In the case of the rod winding, the bars are connected to one another by two short-circuiting rings (squirrel-cage rotor). As long as the rotational speed of the rotor deviates from n_0 , the rotating stator field induces current in the rotor windings, thereby generating torque. Deviation of the rotational speed of the rotor n from n_0 is termed slip s :

$$s = (n_0 - n) / n_0$$

Continuous-running operation is only economical in the vicinity of n_0 because losses increase as slip increases (nominal slip $\leq 5\%$). In this range the asynchronous machine has a shunt characteristic. The machine operates as a motor when $n < n_0$, and as an alternator when $n > n_0$. The direction of rotation is changed by reversing two of the phases.

The asynchronous machine is the most frequently used electric motor in the field of drive engineering. With a squirrel-cage rotor it is easy to operate, and requires little maintenance.

19 Three-phase asynchronous motor with squirrel-cage rotor



2 Examples of rotating-field speeds			
No. of poles (2 p)	50 Hz	150 Hz	200 Hz
Rotating-field speed in rpm			
2	3,000	9,000	12,000
4	1,500	4,500	6,000
6	1,000	3,000	4,000
8	750	2,250	3,000
10	600	1,800	2,400
12	500	1,500	2,000

Synchronous machines

In the rotor (here, also called the pole wheel), the poles are magnetized by direct-current coils. The excitation current is usually transferred via two sliprings to the rotor. The pole wheel can be made of solid steel, because the magnetic flux remains constant over time. Constant torque is generated as long as the rotor rotates at a speed of n_0 . At other speeds, the torque fluctuates periodically between a positive and a negative maximum value, and excessively high current is produced. For this reason, a synchronous machine is not self-starting. The synchronous machine also differs from the asynchronous machine in that the reactive power absorption and generation are adjustable.

The synchronous machine is most frequently used as a generator of energy supply in electric power plants. Synchronous motors are used in cases where constant

motor speed based on constant line frequency is desired, or where a reactive power demand exists. The automotive three-phase alternator is a special type of synchronous machine.

The rotational speed of all three-phase machines is determined by the stator frequency. Such machines can operate over a wide range of speeds if used in conjunction with static converters which vary the frequency.

Table 2

EC motors

The “electronically commutated direct-current” or “EC” motor is becoming increasingly popular (Fig. 21). It is essentially a permanent-magnet synchronous machine, and is brushless. The EC motor is equipped with a rotor-position sensor, and is connected to the DC power source through its control and power electronics. The electronic transfer circuit switches the current in the stator winding according to rotor position – the magnets which induce the excitation current are attached to the rotor – to provide the interdependence between rotational speed and torque which is normally associated with a separately excited direct-current machine. The respective magnetic functions of the stator and rotor are the opposite of what they would be in a classical direct-current machine.

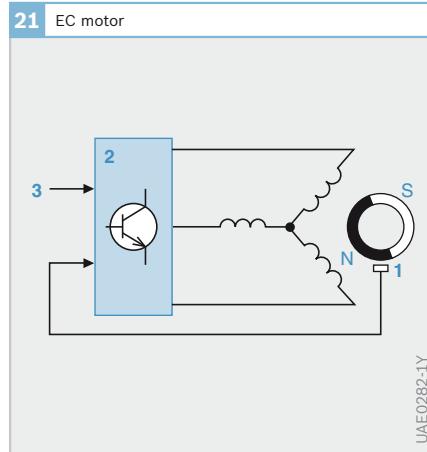
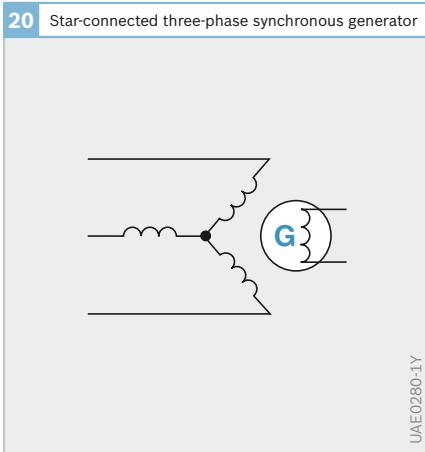


Fig. 20
Slip-ring rotor with excitation winding

Fig. 21
1 Electrical machine with rotor-position sensor
2 Control and power electronics
3 Input

The EC motor's potential applications are a result of the advantages which this drive principle provides: commutator and carbon brushes are replaced by electronics, dispensing with both brush noise and wear. EC motors are maintenance-free (long service life) and can be constructed to meet high degrees of protection. The electronic control feature makes it easy for drive units with EC motors to incorporate auxiliary functions such as infinitely variable speed governing, direction reversal, soft starts, and antilock protection.

The main areas of automotive application are in the HVAC (Heating/Ventilation/Air-Conditioning) sectors, and for pumps and servo units. In the area of production machinery, EC motors are chiefly employed as precision drive units for feed-control in machine tools. Here the decisive advantages are freedom from maintenance, favorable dynamic properties and consistent torque output with minimal ripple.

Single-phase alternating-current machines

Universal motors

The direct-current series-wound motor can be operated on alternating current if a laminated rather than a solid iron stator is used. It is then called a universal motor.

When operated on alternating current, a torque component at twice the frequency of the current is superposed on the constant torque component.

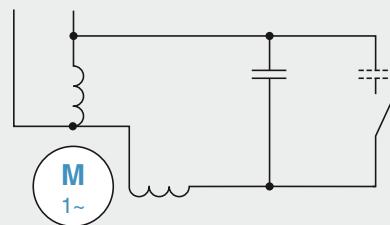
Single-phase asynchronous motors with squirrel-cage rotor

The simplest design of a single-phase asynchronous motor is a three-phase asynchronous machine in which alternating current is supplied to only two stator phases. Although its operation remains largely the same, the power and the maximum torque are reduced. In addition, the single-phase asynchronous machine is not self-starting.

Machines which are intended only for single-phase operation have only a single-phase main winding in the stator, as well as auxiliary starting circuits. The stator also contains an auxiliary winding connected in parallel with the main winding for this purpose. The necessary phase displacement of the auxiliary winding current can be achieved through increased winding resistance (low breakaway torque) or by means of a capacitor connected in series with the auxiliary winding (somewhat greater breakaway torque).

The auxiliary winding is switched off after the motor starts. The direction of rotation of the motor is changed by reversing the two auxiliary or main winding connections. A motor which has a capacitor in series with the auxiliary winding is called a capacitor motor. Capacitor motors with a starting and running capacitor also operate continuously with capacitor and auxiliary winding. Optimum operation for a specific operating point can be achieved by correct selection of capacitor. An additional capacitor is often used in order to increase the starting torque; this capacitor is then disconnected after the motor starts.

22 Two-value capacitor motor



Duty-type ratings for electrical machines

(VDE 0530)

S1: Continuous-running duty

Operation under constant load (rated output) of sufficient duration to reach the thermal steady-state condition.

S2: Short-time-duty type

Operation under constant load is so brief that the thermal steady-state condition is not reached. The rest period is so long that the machine is able to cool down to the temperature of the coolant.

Recommended short-time-duty types: 10, 30, 60 and 90 min.

S3 to S5: Intermittent-periodic duty

Continuous alternating sequence of load and idle periods. The thermal steady-state condition is not reached during the load period or during the cooling period of one duty cycle.

- S3 intermittent-periodic duty without influence of starting on temperature
- S4 intermittent-periodic duty with influence of starting on temperature
- S5 intermittent-periodic duty with influence of starting and braking on temperature

S6: Continuous operation with intermittent loading

Operation with intermittent loading. Continuous alternating sequence of load periods and no-load periods, otherwise as S3.

S7: Uninterrupted duty

Operation with starting and braking.

S8: Uninterrupted duty

Operation with pole-changing. For S3 and S6, the duty cycle time is 10 mins unless otherwise agreed; and recommended values for relative switch-on duration are 15, 25, 40 and 60 %. For S2, S3 and S6, the operating time or the duty cycle time and the switch-on duration are to be specified after the rating; the duty cycle time is only to be specified if it differs from 10 mins. Example: S2 - 60 mins, S3 - 25 %.

Relative switch-on duration

The relative switch-on duration is the ratio of the loading period, including starting and braking, to the cycle time.

Winding temperature

The mean temperature t_2 of the windings of an electrical machine can be determined by measuring the resistance (R_2) and referring it to an initial resistance R_1 at a temperature t_1 :

$$t_2 = \frac{R_2 - R_1}{R_1} (\tau + t_1) + t_1$$

where

$$\tau = \frac{1}{\alpha} - 20 \text{ K}$$

α = temperature coefficient

Hybrid drives

As well as optimizing conventional drive concepts, motor vehicle manufacturers also rely on alternative concepts in order to ensure that existing and future noise and emission restrictions are complied with and reduce fuel consumption. The use of hybrid cars as a means of reducing environmental pollution and increasing driving pleasure and comfort is becoming increasingly popular.

A design which combines electric motor(s) with an internal-combustion engine makes it possible to use smaller engines (down-sizing) and other consumption-reducing measures such as stopping the engine at traffic lights, operating the engine within the optimum efficiency range, and recuperative braking. Hybrid drives allow most of the basic weaknesses of electric or internal-combustion engine drive concepts to be eliminated.

Drive concepts

Definition

Hybrid cars are vehicles whose accelerative force is provided by more than one source. These drive sources obtain the energy that is required from a variety of sources. The term that is currently used to describe these vehicles is HEV (Hybrid Electrical Vehicle), i.e. the combined design of an internal-combustion engine with one or two electric motors and the associated energy accumulators (fuel tank and battery).

Types of hybrid drive

A distinction is made between two basic structures:

- serial
- parallel hybrid drives.

A design which combines both structures is described as a split-power hybrid drive.

Fig. 1

- 1 Internal-combustion engine
- 2 Fuel tank
- 3 Alternator
- 4 Electric motor
- 5 Inverter
- 6 Battery

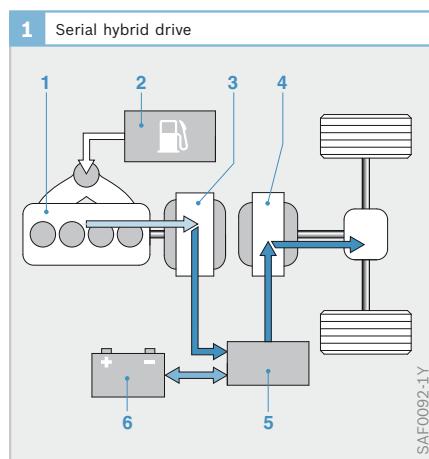
Serial hybrid drive

The series connection of the energy converters is indicative of a serial hybrid concept. The serial arrangement of the hybrid components requires two electric motors and an internal-combustion engine. An electric motor operates as an alternator, and the other as an engine. The internal-combustion engine is not connected to the drive. It charges the battery via the alternator and/or provides the electric motor with the required energy directly.

All of the power that is needed to move the vehicle is therefore transmitted to the drive shaft by the electric motor.

An advantage of this arrangement is that the internal-combustion engine can be operated almost stationary, i.e. within its optimum efficiency range. Because there is no mechanical connection, the electric motor and the internal-combustion engine can be installed in different locations.

One disadvantage of this arrangement lies in the multiple energy conversions and associated loss of efficiency. If you examine the normal mean loss of the individual components, a total loss of about 30 % must be assumed.



It has only been possible to make meaningful use of serial drive concepts since the mid-80's. Until this time, neither electrical machines nor storage media with the required capacity for realizing serial concepts were available.

A serial hybrid system is used in the Toyota Coaster (Bus) Hybrid, for example. Other usage areas for the serial hybrid drive are diesel trains and busses that frequently travel in stop-and-go traffic in urban areas. The serial hybrid drive is not used in passenger cars, however.

Parallel hybrid drive

The term "parallel hybrid drive" is used to describe an arrangement of hybrid components in which the electric motor and the internal-combustion engine can both individually deliver their output to the drive. If necessary, the outputs can be combined by means of mechanically coupling the motor and the engine to the drive. The possibility of power addition allows the machines to be made with relatively small dimensions without adversely affecting the driving performance. The components that are required are designed according to the most frequent area of use.

There are different types of implementation. The electric motor can be directly flanged to the crankshaft (crankshaft-mounted starter-alternator) or coupled to the internal-combustion engine by a belt drive. Both drive sources can be used together or individually to propel the vehicle. Since only one electric motor is incorporated in the system, the vehicle cannot be electrically powered whilst the battery is being charged. The electric motor can only operate as an alternator or for drive purposes.

If a coupling is integrated between the electric motor and the internal-combustion engine (Fig. 2), the engine does not have to be carried along when the vehicle is being propelled using purely electric power. Drag torque losses can therefore be avoided.

The drive-side positioning of the electric motor, i.e. directly on the drive shaft downstream of the transmission, makes it possible to prevent transmission efficiency losses.

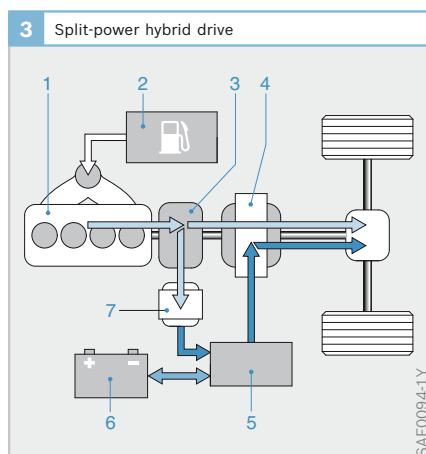
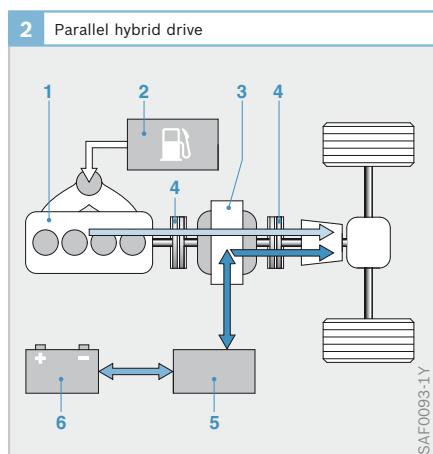


Fig. 2

- 1 Internal-combustion engine
- 2 Fuel tank
- 3 Electric motor/ alternator
- 4 Coupling
- 5 Inverter
- 6 Battery

Fig. 3

- 1 Internal-combustion engine
- 2 Fuel tank
- 3 Planetary gear
- 4 Electric motor
- 5 Inverter
- 6 Battery
- 7 Alternator

Split-power hybrid drive

The split-power hybrid drive arrangement (Fig. 3) is a combination of the two principles that were mentioned previously.

The use of a split-power transmission (planetary gear system) allows some of the internal-combustion engine power to be relayed directly to the drive, i.e. mechanically, whereas the other part is converted into electrical energy via the alternator and optionally stored in the HV battery (high-voltage battery) or directly relayed to the driving electric motor downstream of the transmission.

Both the electric motor and the internal-combustion engine can be used as drives in this combined design. The most consumption-efficient drive method is used for the prevailing operating conditions. Both drives can also provide their power simultaneously for propelling the vehicle. It is also possible to use electricity only.

The use of the planetary gear system in combination with the alternator as a torque support allows the operating point of the internal-combustion engine to be varied within its optimum range, similarly to a combination with a CVT transmission¹⁾. The planetary gear system is one of the main component parts of the split-power hybrid concept. No additional transmission is required for the internal-combustion engine.

This is a form of the strong hybrid that is used in the Toyota Prius, for example.

Use of internal-combustion engine and electric motor in the different hybrid systems

In the serial hybrid, the internal-combustion engine supplies the electric motor with energy. The vehicle is only propelled by the electric motor.

The parallel hybrid with an electric motor fixed to the crankshaft uses the internal-combustion engine as its main energy source. The electric motor is used for assistance during acceleration. Otherwise it is mainly the internal-combustion engine that is used.

In the parallel hybrid that can use electric power only, the electric motor is used more frequently and the internal-combustion engine less frequently than in a parallel hybrid with an electric motor fixed to the crankshaft.

In a split-power hybrid, the output of the internal-combustion engine is apportioned. This process is infinitely variable. For this reason the electric drive continuously operates when the vehicle is being propelled by the internal-combustion engine, without affecting the frequency at which purely electric propulsion is used.

Hybrid drive classification

Figure 4 shows an example of classification according to degree of hybridization. The engine start-stop, recuperative braking, boosting (electric motor support) and electric-only propulsion methods that are shown can be realized using the following hybrid concepts. It remains to be said that both the structural and financial cost and the fuel-saving potential can vary considerably depending on the concept that is used.

The way in which the electrical and internal-combustion engine components are combined and interconnected is another means of distinguishing between types of hybrid. A distinction is also made between the degree of hybridization and the type of energy flow. The operating strategy that is used is another distinguishing feature. There are no fixed definitions or names for the many distinguishing features. They depend on the respective vehicle manufacturer or supplier.

¹⁾ With CVT transmissions (Continuously Variable Transmission) the force is transmitted via pulling media such as wide V-belts or laminated toothed chains. The main components are two axially movable V-pulleys with a pulling medium running between them. A continuous conversion ratio can be achieved.

Micro Hybrid (start/stop system)

The simplest form and the first step in the direction of hybrid vehicles is the micro hybrid, e.g. a starter/alternator system with start/stop function. In this system the starter and the alternator are replaced with an electrical machine. The internal-combustion engine is switched off when the vehicle stops. When the vehicle sets off again it is quietly accelerated to idle speed by the electrical machine, and then the ignition is activated. In the simplest micro hybrid system, the conventional starter is replaced with a starter that has been optimized for start-stop functionality. These systems generally only have a small energy accumulator and comparatively little electrical power.

Mild Hybrid

If this system is combined with the start/stop, booster (internal-combustion engine assisted by the electric motor when starting off and accelerating) and brake energy recovery strategies, it is usually referred to as a mild hybrid concept. Electric propulsion is possible for an extremely short time in combination with cylinder shutoff, but the internal-combustion engine must also be carried along.

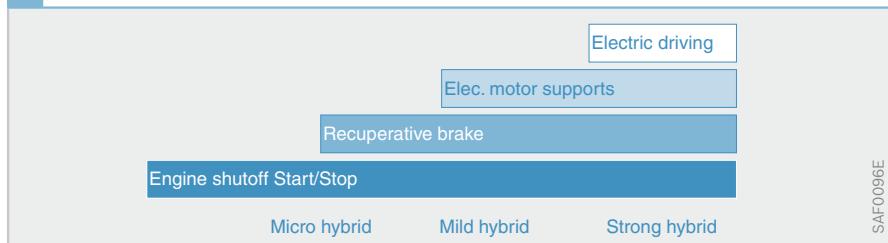
The electric motor can be directly flanged to the crankshaft in a mild hybrid. In mild hybrid systems, additional electrical energy of up to 20 kW is provided by the electric traction accumulator (battery). This is mainly used for starting off and accelerating at low engine rpm.

Strong Hybrid

Strong hybrid systems make it possible to drive in urban traffic using electrical energy only and therefore without producing emissions. The traction accumulator of these systems is more powerful than those of the above-mentioned micro hybrid and mild hybrid variants. It has a typical output of about 25 to 50 kW.

The mild hybrid system is more expensive to implement than other hybrid systems. On the one hand, the vehicle electrical system has to be adapted to the voltages and currents that are required, and on the other hand, additional components such as couplings and transmissions are needed to ensure that the vehicle can be propelled using electric power only. Examples of strong hybrid systems are the parallel hybrid with two couplings (Fig. 2) and the split-power system such as the one used by Toyota (Fig. 3).

4 Hybrid drive classification



Operating strategies for electric hybrid vehicles

The integration of electrical components in the drivetrain has wide-ranging effects on the overall vehicle system. Because of the complexity of the drivetrain in an electric hybrid vehicle (HEV: Hybrid Electric Vehicles), system solutions with mutually coordinated components are of special significance. As well as having a favorable combination of appropriate components, the optimization of the operating strategy is the most important task in order to achieve the goals of reducing consumption and emissions while increasing driving pleasure and driving comfort. HEV functions that reduce consumption and emissions are start/stop operation of the internal-combustion engine, recuperative braking and operating point optimization.

The electric propulsion function is an important constituent of operating point optimization. Booster operation of the elec-

tric drive can increase the torque and therefore provide improved acceleration capability, particularly at low engine rpm. An integrated examination of design and strategy optimization is required in order to utilize the maximum potential. In this case, operating strategy means a driving situation-dependent torque distribution between the two drive sources of internal combustion engine and electric motor.

Figure 5 shows the networking of the individual components in the drivetrain using the example of a vehicle with a parallel hybrid drive. The entire system is coordinated by a supervisory hybrid controller. This involves battery management, engine management, electric drive management, transmission management and brake system management. It becomes clear that the development of these control systems (particularly the supervisory hybrid controller) and the development of the components have to be closely interlinked in order to achieve optimum functionality.

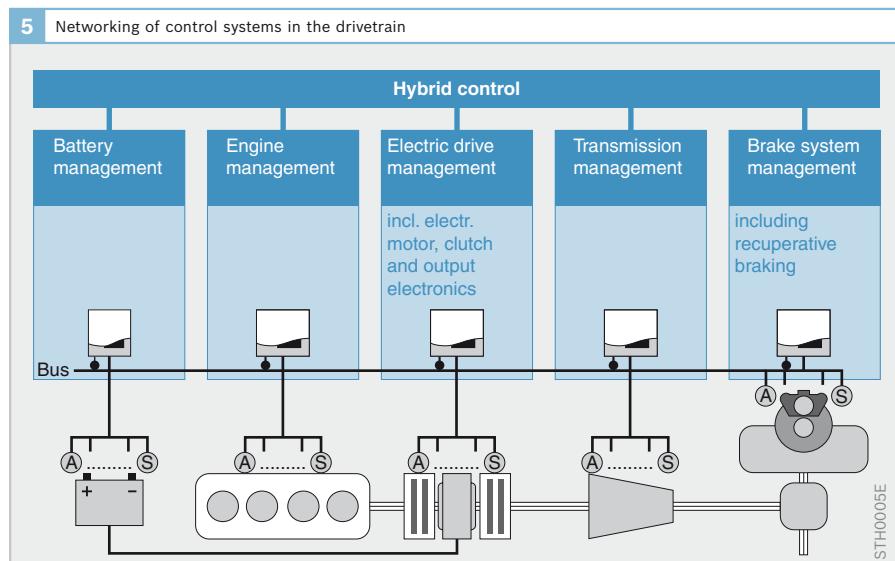


Fig. 5

A Actuator
S Sensor

Influence of operating strategy on the design of the internal-combustion engine

Use of suitable internal-combustion engines

Any internal-combustion engine from vehicles with a conventional drivetrain can generally be used in a hybrid vehicle. Gasoline, natural gas and diesel engines can be combined with an electrical drive, but with differing optimization targets (see sections "Operating strategy for diesel hybrid vehicles/gasoline hybrid vehicles").

The additional possibilities provided by the HEV combination with regard to operating point adjustments, for example, can be pursued for hybrid vehicles and other internal-combustion engine concepts, as well as for conventionally propelled vehicles. Because the operating range has been made smaller, the efficiency optimization that is needed can be restricted to this area and the high costs of providing additional components can be avoided.

If an adequate amount of power is provided by the electrical accumulator, the electric drive can compensate for torque deficits and slower response behavior found in certain internal-combustion engine concepts.

The dynamic requirements that are made of the drivetrain are realized by combining the electric drive and the internal-combustion engine. The torque characteristic, which is particularly favorable at low engine speeds, and the rapid response behavior of the electric drive relieve the strain on the internal-combustion engine during dynamic processes. Internal-combustion engine peak loads can therefore be avoided.

Internal-combustion engine friction optimization

Part of the primary energy reduction in a hybrid vehicle is achieved by reclaiming braking energy (recuperation). This can be carried out during active braking and under overrun conditions, e.g. when driving downhill. In order to make maximum use of this saving potential, the internal-combustion engine must be shut down in these operating ranges. If this is not possible, the engine must be carried along and its drag friction restricts the recuperation potential. In this case, optimizing the friction of the engine represents an important internal-combustion engine requirement.

Emission and fuel consumption optimization

In order to achieve emission and consumption reduction targets, degrees of freedom for internal-combustion engine operation can be exploited that depend on the topology of the drivetrain.

Avoiding operating points where the internal-combustion engine is less efficient or has high emissions is an important strategy. The basic operating strategy must be optimized with regard to the improvement target (e.g. consumption reduction and CO₂ or NO_x reduction). The modified operating conditions can be exploited to optimize the internal-combustion engine concept and the exhaust-gas treatment. The modified requirements result in changes to the functionality and application of the engine management system that are not examined in further detail at this point.

A requirement currently exists for improvements to be made in vehicles with diesel engines with regard to NO_x and particulate emissions. Improvements can be achieved here by shifting the engine operating points in lower emission ranges.

If a gasoline or natural gas engine is used, the operating strategy is designed to save fuel and reduce CO₂ emissions.

Operating strategy for diesel hybrid vehicles

Particularly in part-load operation, vehicles with lean combustion concepts achieve low consumption values but have higher exhaust-gas values than vehicles with stoichiometrically operated internal-combustion engines. The main focus as far as diesel vehicles are concerned is therefore on reducing emissions. The post-combustion reduction of hydrocarbon and carbon monoxide emissions from lean internal-combustion engines using oxidation-type catalytic converters is in an advanced stage of development. Diesel particulate filters with high exhaust cleaning efficiency are already being used to reduce particulate emissions. The primary goal as far as diesel hybrid vehicles are concerned is to reduce nitrogen-oxide emissions (NO_x), whereby the intention is also to reduce CO₂ emissions at the same time.

Figure 6 shows the best ranges in which to operate the internal-combustion engine in order to optimize NO_x emissions.

Figure 6a shows the operating point range for an internal-combustion engine during a driving cycle. In the New European Driving Cycle (NEDC), internal-combustion engines are operated at low part loads and

medium loads in conventional vehicles, and therefore in ranges with high NO_x emissions.

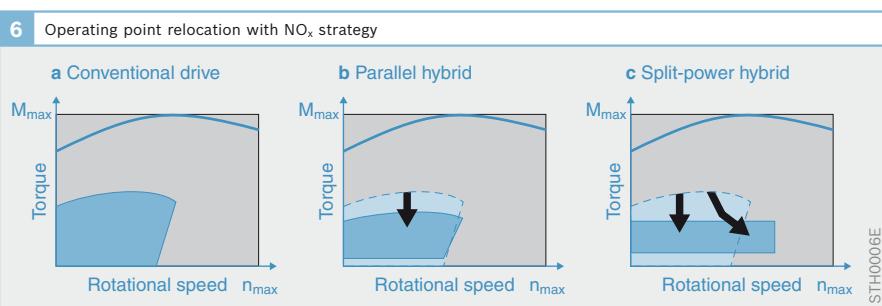
Figure 6b shows a hybrid vehicle with a parallel concept that avoids low internal-combustion engine loads (unfavorable efficiency) by operating with electricity only. Since the characteristic curve ranges of low NO_x emissions are at low loads, the electric motor and the internal-combustion engine should be used in conjunction (boost) in order to avoid engine operation at high loads.

The electrical energy that is required cannot usually be recovered by means of recuperation alone, and the electrical machine is subsequently operated in alternator mode. This means that the electric drive alternates between engine mode and electric mode almost continuously, depending on the cycle. A compromise must be found between two opposing goals: on the one hand, the NO_x and CO₂ emissions must be lowered as much as possible, and on the other hand, the energy throughput from the electric accumulator must be restricted. The energy throughput has a significant influence on the service life of the traction battery.

Figure 6c shows the situation for a split-power hybrid vehicle in which flexible use is made of the internal-combustion engine speed (within certain limits), independently of the vehicle speed (ECVT effect). Engine operating points can be relocated

Fig. 6

- a Operating point range in driving cycle
- b Boost: combined operation of internal-combustion engine and electric motor
- c ECVT effect: operating point relocation to faster engine speeds



to faster speeds, which reduces the internal-combustion engine torques that are needed, leading to additional NO_x reduction.

Because of the serial operating mode on the electrical path (simultaneous alternator operation of the first electric motor and engine operation of the other), the traction battery energy throughput and cycling can be kept lower than with a parallel hybrid.

Operating strategy and internal-combustion engine concepts for gasoline hybrid vehicles

The lowest emission values are achieved with vehicles with stoichiometrically operated gasoline internal-combustion engines due to the three-way catalytic converter technology. The lowest emissions can also be achieved with large-displacement internal-combustion engines in hybrid vehicles by using appropriate warm-up strategies. It may even be possible to reduce the demands that are made on the exhaust-gas treatment system. The main objective with the gasoline hybrid vehicle is therefore to save fuel and increase performance.

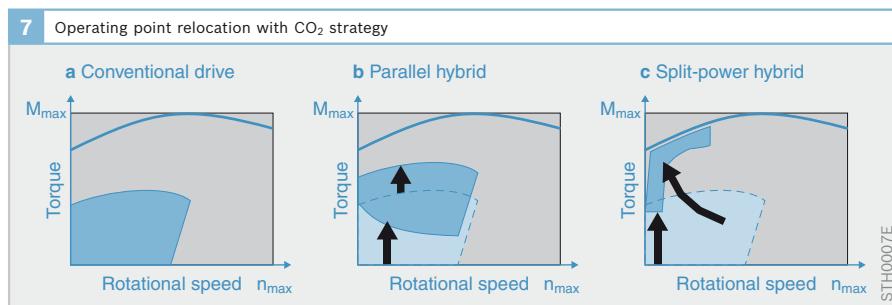
Figure 7 shows a possible internal-combustion engine operating range optimization with regard to minimal CO₂ emissions for the three HEV topologies (i.e. consumption reduction).

In the New European Driving Cycle (NEDC), internal-combustion engines are operated in conventional vehicles with a low part load and therefore at sub-optimal efficiency. In vehicles with a parallel hybrid drive, low internal-combustion engine loads can be avoided by driving with electric propulsion only (Fig. 7b).

Since the electrical energy that is required cannot usually be recovered by means of recuperation alone, the electrical machine is subsequently operated in alternator mode. In comparison to a conventional vehicle, this moves the operation of the internal-combustion engine into the higher load range and therefore to a higher level of efficiency. This means that more electrical energy can be made available than with the previously described NO_x strategy, making it possible to use a greater proportion of electrical propulsion. However, a compromise must be found between CO₂ emissions and energy throughput because of the service life requirements that are made of the traction battery. The energy throughput has a significant influence on the service life of the traction battery.

The operating range of the internal-combustion engine in the split-power hybrid vehicle (Fig. 7c) is more restricted than it is in the parallel hybrid vehicle. It is usually operated speed-dependently when engine loading is such that it allows the entire

- Fig. 7**
- a Operating point range in driving cycle
 - b Avoiding low internal-combustion engine loads by using electrical propulsion and subsequent charging
 - c ECVT effect: operating point relocation to optimum energy range of drivetrain



drivetrain to operate in the most energy-efficient way. Because of the serial operating mode on the electric path (simultaneous alternator and engine operation of both electrical machines), here too the energy throughput and cycling of the traction battery can be kept low in comparison to a parallel hybrid.

Operating point optimization: drive torque distribution

Different forms of hybrid vehicle control and different operating strategy optimizations have significant effects on fuel consumption, emissions, the amount of available torque and the design of the components (e.g. operating range of the electrical machine and the internal-combustion engine, energy throughput and cycling of the electric accumulator), since their operating points are directly dependent on the operating strategy. It has already become clear that the system-encompassing hybrid controller is extremely important. There are many options and degrees of freedom for optimizing operation. The distribution of the requested drive torque to the internal-combustion engine and electrical machine drive sources is extremely important with regard to exploiting the fuel saving potential.

Definition in the finite state machine

However, torque distribution is not needed in all driving conditions. Figure 8 shows the different driving conditions of a hybrid vehicle, which are defined in a finite state machine depending on driver request, the status of the electric accumulator and vehicle speed.

The internal-combustion engine is shut down in the event of all-electric propulsion or recuperation and the maximum available torque is requested from both drive sources during boost operation. All-electric driving mode is restricted to slow vehicle speeds and low acceleration. Recuperation only occurs during vehicle deceleration. Boost operation is mainly used when maximum accelerative force is requested by the driver (kick-down).

Distribution by operating strategy

The expansive scope of hybrid propulsion in which the drive torque distribution must be specified lies between the operating statuses that are defined by the finite state machine. Because of the many degrees of freedom and the large number of interdependent factors, an optimization is required that can be achieved most effectively using model-based procedures.

8 Hybrid vehicle driving conditions

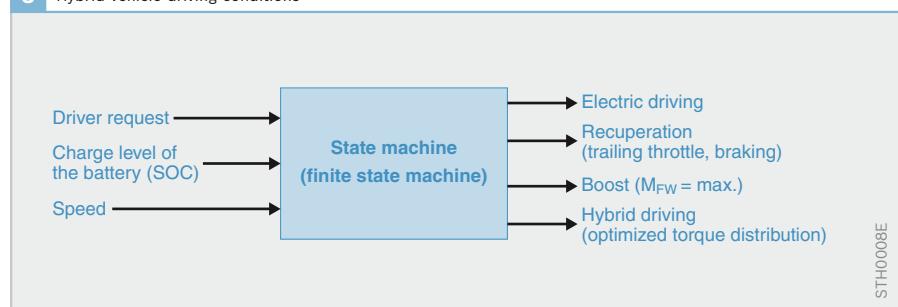


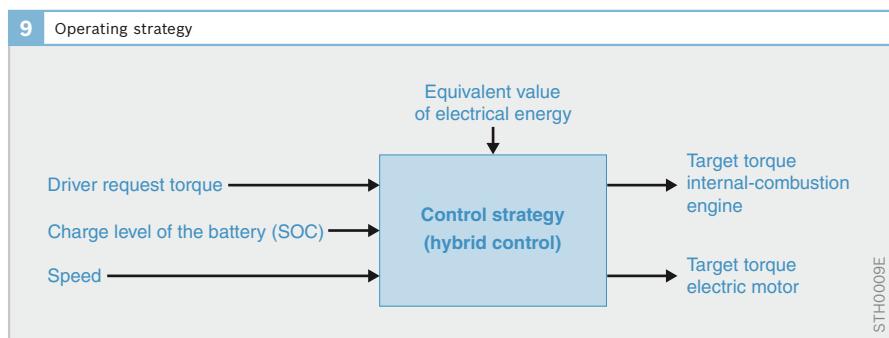
Figure 9 shows the operating strategy dependencies. The hybrid controller distributes the required drive torque between the drive sources of internal-combustion engine and electrical machine on the basis of vehicle speed and the status of the electric accumulator, among other things. The operating strategy also needs an equivalence value of the stored electrical energy that indicates how much fuel has been consumed in generating this electrical energy.

The different types of electrical power generation (recuperation and internal-combustion engine charging) are examined in order to assign an optimization parameter equivalence value (e.g. fuel consumption) to the energy content of the battery. This equivalence value forms the basis for deciding which energy source to use.

Power generation strategy

In a hybrid vehicle, electrical energy can be generated by means of internal-combustion engine charging and recuperation (brake energy recovery). Whereas the latter is acquired without using fuel, the former requires fuel consumption. The efficiency of this charging process depends on the current operating point of the internal-combustion engine during this process.

Since it is not usually possible to generate enough energy from recuperation alone and the battery has limited storage capability, internal-combustion engine charging is unavoidable. In order to minimize the amount of fuel that is used to do this, this type of electrical-energy generation should only be performed when the internal-combustion engine is being used in low-efficiency operating ranges and the additional load can be used to improve efficiency. The optimum utilization of the efficiency improvement during internal-combustion engine charging is an operating strategy issue, since it involves torque distribution between the internal-combustion engine and the electric machine.



Recuperative brake system

In order to make better use of a hybrid drive system, it must be possible to charge the electrical energy accumulator efficiently. An adequate amount of electrical energy must be made available for

- ▶ The frequent internal-combustion engine repeat starts that occur under certain circumstances with the micro hybrid
- ▶ Electric torque support or electric driving mode in mild or strong hybrid systems

The electrical energy for charging the battery can be provided by increasing the load of the internal-combustion engine or by operating the electric motor as an alternator. On the other hand, the vehicle's kinetic energy during deceleration processes can also be used. This energy is converted to heat in conventional vehicles either by the engine-drag torque or by the vehicle service brake by operating the brake pedal.

By using the electric motor as an alternator, hybrid vehicles open up the possibility of reclaiming at least some of the energy and using it for electric consumers or the vehicle's electric drive. This procedure is referred to as *recuperative* or *regenerative* braking.

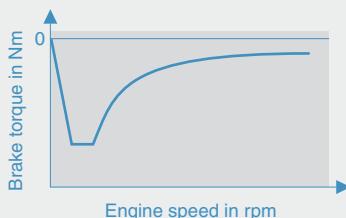
Recuperative braking strategies

Concept

In a strong hybrid the internal-combustion engine is decoupled for recuperative braking and the drag torque is replaced by an equivalent electric motor alternator torque (drag torque simulation). The energy that is released is stored. If the internal-combustion engine cannot be decoupled (as with a mild hybrid), a lower alternator torque can be applied to the drivetrain in addition to the drag torque of the internal-combustion engine (drag torque increase).

However, taking the driving behavior into consideration, no major braking deceleration can be achieved by simulating or increasing the drag torque. The different recuperative brake torques that occur during individual braking procedures and the resulting braking power variations are problematic. These must be adapted to the battery charge level and the thermal load of the electric drive. If significant battery temperature increase occurs after several braking procedures, the recuperative output may need to be reduced in order to prevent the system from overheating.

10 Maximum torque trend with electric machine in alternator mode



Cooperative recuperative brake system

In order to exploit more kinetic energy, the service brake system must be modified at higher braking deceleration levels.

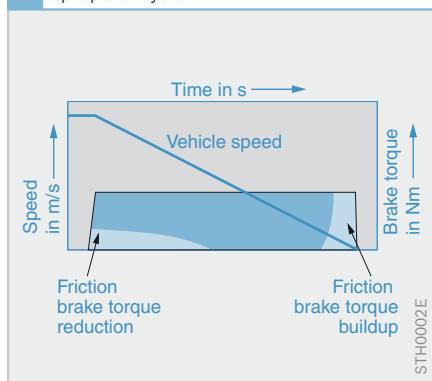
This is done by replacing some or all of the service brake friction torque with recuperated brake torque without changing the vehicle deceleration with the brake pedal position in the same position and using the same amount of pressure. This is achieved in the cooperative recuperative brake system by making the vehicle controller and the brake system interact in such a way that the friction brake torque is always reduced by exactly the amount of alternator brake torque that can be replaced by the electric motor.

Requirements

A cooperative recuperative brake system must be able to perform the following requirements:

- ▶ Determine driver braking deceleration request
- ▶ Maintain the service brake performance capability and operation that is usual in conventional vehicles
- ▶ Determine a suitable brake torque distribution between the friction brake and the recuperative brake, taking stability, comfort and efficiency criteria into consideration

11 Cooperative braking maneuver with a split-power hybrid



- ▶ Determine a suitable brake torque distribution between the vehicle axles
- ▶ Adjust the friction brake torque

Replacing the recuperative brake torque with the friction brake torque requires a suitable data interface between the drivetrain and the brake control unit.

Cooperative braking maneuver

The maximum alternator torque that can be achieved by the electric motor changes continuously over a wide speed range during braking (Fig. 10). This is due to the fact that the output of the electric motor (i.e. torque * speed) is constant within this range. The electric motor only has a constant maximum torque range at slow speeds. If the speed becomes even slower, the recuperative brake torque that can be achieved drops back down to zero.

There must be constant torque at the wheels in order to have constant vehicle braking deceleration. If the electric motor that is being operated as an alternator is utilized up to its limit torque, the friction brake torque must be continuously reduced as the speed drops due to the fact that the alternator torque increases. Figure 11 shows an example of the braking procedure of a split-power hybrid. At the beginning of the maneuver, the alternator torque is increased until it reaches its maximum (if this corresponds to the required total brake torque). Towards the end of the braking maneuver, the alternator torque is reduced and completely replaced with friction braking torque because the electric motor cannot provide any more alternator torque at extremely low rpm (Fig. 10).

When braking from a fast speed to a standstill, the distribution between friction brake torque and recuperative brake torque is continuously adjusted if the pedal is operated with constant pressure.

Fig. 11
Friction brake torque reduction and increase, and replacement with alternator brake torque

Braking-force distribution

Like conventional brake systems, when a recuperative brake system is being designed, the braking-force distribution between the front axle and the rear axle is extremely important for the directional stability of the vehicle. The normal force on the wheels of the front axle increases together with braking deceleration, whereas the normal force on the wheels of the rear axle decreases.

If the electric motor is connected to the wheels of the front axle, additional wheel torque can be transmitted at the front axle as the braking deceleration and therefore the normal force increases. The friction value utilization at the front axle should therefore not increase the friction value utilization of the rear axle in order to obtain directional stability.

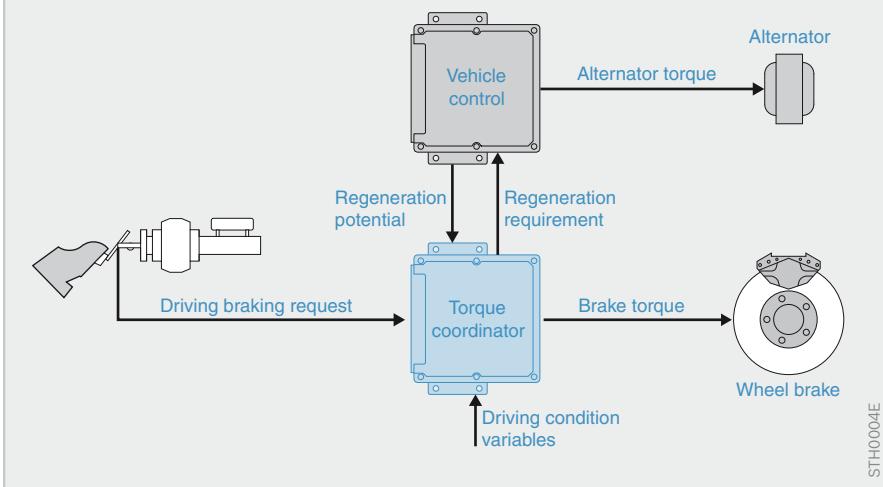
If the vehicle has rear-wheel drive or an electrically powered rear axle (i.e. internal-combustion engine at the front axle and electric motor at the rear axle), the amount of recuperative torque that can be applied decreases as the braking deceleration increases.

The electrical output of the battery has a major influence on the use of recuperation, since it is the restrictive element for the uptake of electrical energy from vehicle movement. As the output from the energy accumulator increases, the maximum possible recuperative braking deceleration increases. However, for reasons of vehicle stability, a large alternator braking torque can only be transmitted to the front axle. High levels of recuperative braking deceleration can therefore only be achieved with a vehicle possessing front-wheel drive or all-wheel drive. In the latter case, the brake torque is distributed to both axles in such a way that it approximately corresponds to the ideal braking-force distribution of the friction brake, depending on the type of differential transmission.

Influences on stability control

Since the recuperative brake system affects braking stability, ABS and ESP control systems must be adapted to the change in automotive physics.

12 System circuit diagram of the cooperative recuperative brake system



If unstable driving conditions are detected or the amount of brake slip is excessive, it is advisable to suppress the recuperative part of the braking and only use the friction brake system for braking deceleration and stabilization corrections. Otherwise instability and drivetrain oscillation could have a detrimental effect on the wheel slip control system.

Compared to partial braking operations, brake stabilization system interventions via the vehicle systems occur so infrequently that suppressing recuperative braking in these situations has no noticeable effect on the average consumption of the vehicle.

Friction brake implementation

There are different types of friction brake in the cooperative recuperative brake system.

These are usually mechatronic brake systems that decouple the brake pedal and the wheel brake and simulate brake pedal characteristics by adding a pedal simulator. The energy storage that is needed for brake boosting can be provided hydraulically or pneumatically.

The block diagram of the cooperative recuperative brake system (Fig. 12) is something that all versions have in common. The vehicle control system continuously monitors all relevant parameters of the hybrid drive and determines the torque that the electric motor can provide for braking. When the brake pedal is operated, the torque coordinator of the brake control unit calculates the distribution of the brake torque to the friction brake and the recuperative brake system.

The proportion of recuperative torque is fed back to the vehicle control system and relayed to the electric motor actuator. The residual brake torque is set by the friction brake while vehicle stability and wheel slip are monitored.

Electrical energy accumulators

Vehicle electrical systems of hybrid systems

Irrespective of the design of the hybrid vehicle (serial, parallel or split-power hybrid), the vehicle electrical system consists of at least two subsystems with different voltage conditions:

- ▶ The traction power supply for supplying the electrical part of the drivetrain; the nominal voltage is significantly above 42 V in this case, whereby the battery voltage can be less than the traction power supply voltage
- ▶ The conventional 12 V vehicle electrical system for supplying the body components (power windows, radio etc.) and the internal-combustion engine (engine control unit, radiator fan, additional water pump etc.)
- ▶ A sub-power supply with different voltage conditions if necessary

Both the traction part and the body part of the vehicle electrical system are equipped with an energy accumulator. A facility for connecting the voltage levels using a DC/DC converter is usually provided.

The provision of a power supply network with different voltage conditions and the way in which it is connected to the other sub-power supplies depends on the design of the respective hybrid vehicle.

In order to ensure that the battery operates reliably with a high voltage condition in a hybrid vehicle, a control unit (battery management system) is needed that monitors the individual battery components and controls charging and discharging, thermal management and various safety functions.

Energy accumulator requirements for traction vehicle electrical systems in hybrid vehicles

Whereas lead-acid systems are used in the 12V vehicle electrical systems of motor vehicles for powering the starter motor, the ignition, the engine management system and the chassis components, the traction vehicle electrical systems of hybrid vehicle require additional energy storage systems. The selection of suitable electrochemical systems depends on the specific requirements:

- ▶ High steady-state voltage of the individual cells
- ▶ Good cold startability
- ▶ Operation at average charge levels without affecting service life, SOC (State Of Charge) charge level approx. 60 %
- ▶ Good charging/discharging performance at average charge level
- ▶ High efficiency in charging and discharging mode (recuperation, acceleration) at all temperatures
- ▶ High cycle strength with as much energy throughput as possible
- ▶ Long traction battery service life (at least the same as the vehicle service life)
- ▶ Thermal management system
- ▶ Easy operation and diagnosis
- ▶ Good safety level during operation and in the event of accidents
- ▶ Preferential use of recyclable and non-toxic materials

State of charge

The energy accumulator is operated at an average charge level in order to achieve good recuperation current absorption and good power output in order to assist acceleration. High charge levels provide high power output, and low charge levels make it possible to feed back large quantities of recuperated electrical energy.

The traction battery of plug-in hybrids is fully charged from the mains power supply at regular intervals. The vehicle is operated with electricity up to a preselected average charge level, after which it is operated like other hybrid vehicles.

Thermal management

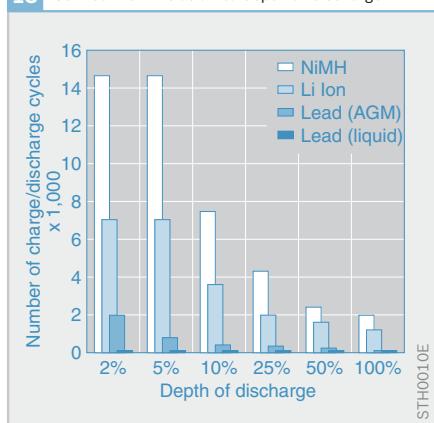
The acceleration and recuperation phases alternate during vehicle operation. Even at high charge/discharge efficiency levels, this leads to heat losses that cause the battery to become warm. In order to remove the heat in a suitable way (air/liquid cooling) the battery system must have a thermal management system.

Charge quantity

During battery system selection it must be taken into consideration that only a limited charge quantity can be used during the lifetime of the battery in cycle operation; this can occur in many flat cycles with small changes to the charge level, or fewer cycles with a large discharge depth (Fig. 13). The charge quantity that can be used tends to be less with low discharge cycles.

It must also be noted that the internal resistance of the battery increases with age, and performance deteriorates.

13 Service life in relation to depth of discharge



Energy accumulator for hybrid systems

Nickel metal hydride batteries (NiMH) are currently used as standard for hybrid vehicles. Nickel cadmium batteries are not used in modern vehicles because of the fact that the cadmium material is toxic and damages the environment.

Lithium battery systems are undergoing further development for use in hybrid vehicles.

Nickel metal hydride systems (NiMH)

Nickel metal hydride batteries are particularly useful for hybrid vehicles because designs with a high power density can be realized, given the materials that are used. The alkaline electrolyte (caustic potash solution, KOH) does not take part in the electrode reaction (compared to lead/acid systems).

Also, operation in a partially charged condition does not adversely affect the service life. High efficiency levels with high charge/discharge currents can be achieved in other charge condition ranges.

The disadvantages of the NiMH system are the high self-discharge rate and extensive power loss at low temperatures.

The relative low steady-state voltage of NiMH cells is also unfavorable. In order to achieve a battery voltage that is comparable with lead systems, approximately 1.6 times as many NiMH cells are required. In order to ensure that the entire battery system is reliable, the quality and uniformity of the individual NiMH cells must be somewhat greater.

Nickel oxide hydroxide is used as the active electrode material (as in NiCd batteries) and (instead of cadmium) a material that stores hydrogen (misch metal). Misch metal is an alloy with high lanthanum, cerium and neodymium content.

The cell is operated with slight hydrogen overpressure and has a pressure relief valve or bursting disk so that it can blow out hydrogen in critical operating condi-

Fig. 13

NiMH: Nickel metal hydride system
Li-Ion: Lithium ion system
AGM: Absorbent Glass Mat
Liquid lead: Lead-acid system/liquid battery

tions. During operation, it must be ensured that excessive overcharging does not occur and that ventilation facilities are provided.

Clouds of caustic alkali can be expected to occur when the overpressure is being released. The fact that hydrogen can form oxyhydrogen gas when it is mixed with air must also be taken into consideration.

Lithium ion systems (Li ions)

Compared to NiMH systems, lithium systems provide even higher energy and power densities at cell voltages of approximately 3.6 V. These systems are therefore mainly used in portable electrical systems (mobile phones, laptops), where they have essentially replaced NiMH batteries.

Efforts are currently being made to subject these cells to further development for use in hybrid vehicles. Special attention is being paid to the use of low-cost and reliable electrode materials (e.g. LiMn_2O_4 , LiFePO_4).

The use of lightweight Li metal and the properties of the other materials involved (graphite as the anode material) make it possible to manufacture extremely thin electrodes (< 0.5 mm), which can be used to create structures with extremely high output (e.g. 3 kW/kg with SOC 60 %, 25 °C

and 10 s pulse duration). Because of the high energy content of the electrode material and the high cell voltage, special measures are required:

- ▶ Use of organic electrolyte with special conductive minerals
- ▶ Safety structures that prevent the cell from exploding if it becomes damaged, for example
- ▶ Monitoring of individual cells in order to avoid overcharging and overheating

The positive electrode consists of special metal oxides (Ni, Mn, Co or mixtures thereof) that can store lithium ions. These ions can move to the counter-electrode and back during the discharging/charging process in a reversible mechanism. The counter-electrode is made from graphite and can also absorb Li ions because of its stratified structure.

Lithium-polymer batteries

A special version of the Li ion battery is the Li polymer battery. It contains electrolytes in non-liquid form. This version is particularly suitable for manufacturing pliable, flexible cells. It is currently being investigated whether Li polymer batteries can be used in hybrid vehicles.

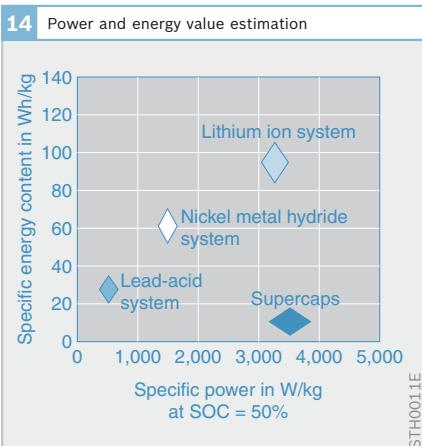


Fig. 14
The comparison only provides an estimate of the basic possibilities of the different systems. Each electrochemical system can be optimized towards either energy or power within certain limits.

► Greenhouse effect

Short-wave solar radiation penetrates the earth's atmosphere and continues to the ground, where it is absorbed. This process promotes warming in the ground, which then radiates long-wave heat or infrared energy. Some of this radiation is reflected in the atmosphere and heats the earth.

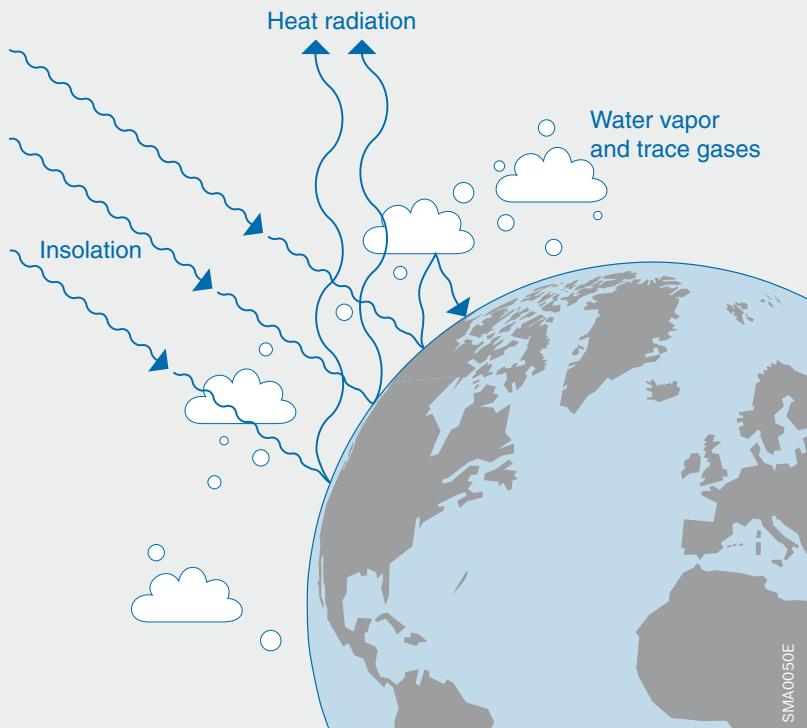
Without this natural "greenhouse effect" the earth would be inhospitable with an average temperature of -18°C . The greenhouse gasses in the atmosphere (water vapor, carbon dioxide, methane, ozone, dinitrogen monoxide, aerosols and cloud particles) provide average temperatures of approximately $+15^{\circ}\text{C}$. Water vapor, in particular, retains substantial amounts of heat.

Since the beginning of the industrial age over 100 years ago, the concentration of carbon dioxide has been increasing rapidly.

The primary cause of this increase has been the burning of coal and petroleum products. In this process, the carbon bound in the fuels is released in the form of carbon dioxide.

The processes that influence the greenhouse effect within the earth's atmosphere are extremely complex. The fact that anthropogenic emissions (i.e. caused by man) are the main cause of the climate change is disputed by another scientific theory. According to this theory, the warming of the earth's atmosphere is being caused by increased solar activity.

There is, however, a large degree of unanimity in calling for reductions in energy consumption in order to lower carbon dioxide emissions and combat the greenhouse effect.



Vehicle electrical systems

The vehicle electrical system of a motor vehicle comprises the alternator as the energy converter, one or more batteries as the energy accumulators and the electrical equipment as consumers. The energy from the battery is supplied to the starter (consumer), which then starts the vehicle engine. During vehicle operation, the ignition and fuel-injection system, the control units, the safety and comfort and convenience electronics, the lighting, and other equipment have to be supplied with power.

Electrical energy supply in the passenger car

When the engine is running, the alternator supplies electricity which, depending on the voltage level in the vehicle electrical system (determined by the alternator speed and the consumers drawing current), is normally enough to power the consumers and charge the battery as well. If the consumer current draw I_V in the vehicle electrical system is greater than the alternator current I_G (e.g. when the engine is idling), the battery is discharged. The vehicle system voltage falls to the voltage level of the battery from which current is drawn. If the consumer current draw I_V is less than the alternator current output I_G , a proportion of the current flows to the

battery and acts as a battery charging current I_B . The vehicle system voltage increases to the setpoint value specified by the voltage regulator.

With careful selection of the battery, alternator, starter and the other electrical system consumers, it must be ensured that the charge balance of the battery is indeed balanced so that:

- ▶ It is always possible for the internal-combustion engine to be started
- ▶ It allows operation of specific electrical consumers for a reasonable period of time when the engine is off

The lowest temperature at which the engine can be started depends on a number of factors, including the battery (capacity, low-temperature test current, state of charge, internal resistance, etc.) and the starter (design, size, and performance). If the engine is started at a temperature of -20 °C, for example, the battery must have a given minimum state of charge p .

Apart from the battery itself, the current output of the alternator and the power output of the consumers have a decisive influence on the charge balance of the battery.

Current output of the alternator

The current output of the alternator is speed-dependent. For an engine idling

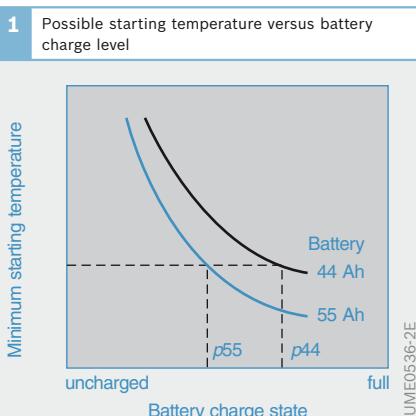
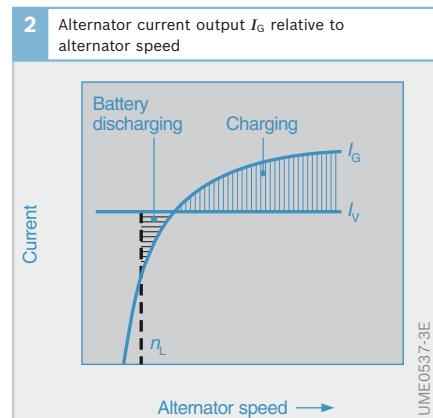


Fig. 1

p Minimum charge level

Fig. 2

I_V Consumer current draw
 n_L Engine idling speed



UIME0537-3E

speed of n_L , the alternator can only supply some of its rated current if it has a conventional turns ratio ranging from 1:2 to 1:3 (crankshaft to alternator). By definition, the rated current is output at an alternator speed of 6,000 rpm.

Power output of the consumers

The electrical consumers have a variety of switch-on durations. A distinction is made between continuous consumers (ignition system, fuel injection, etc.), long-time consumers (lighting, rear-window heating, etc.) and short-term consumers (turn signals, stop lamps, etc.).

The use of a number of consumers is dependent on the time of year (air-conditioning system, seat heating). The on-time of electrical radiator blowers depends on the ambient temperature and vehicle operation. In winter, the vehicle is driven with the lights on most of the time.

The electrical load requirements encountered during vehicle operation are not constant. The first minutes after startup are generally characterized by high demand, followed by a sharp drop in electrical load requirements:

- ▶ A windshield heater requires up to 2 kW for 1 to 3 minutes after the engine is started in order to de-ice the windshield

- ▶ The secondary-air pump, which injects air immediately after the combustion chamber to burn the exhaust gas, runs for about 3 minutes after the engine is started
- ▶ Other electrical consumers such as heaters (rear window, seats, mirrors, etc.), fans and lights are switched on for longer or shorter periods depending on the situation, while the engine-management system is in operation all the time

Charging the battery

Due to the chemical processes that take place in the battery, the battery charge voltage has to be higher at low temperatures and lower at high temperatures. The gassing voltage curve shows the maximum voltage at which the battery does not produce gas. A regulator limits the voltage if the alternator current I_G is greater than the sum of the consumer equipment current draw I_V and the temperature-dependent, maximum permissible battery charging current I_B . Regulators are normally mounted on the alternator. If there is a significant deviation between the temperatures of the voltage regulator and the battery electrolyte, it is better to monitor the voltage regulation temperature directly at the battery. The voltage drop across the alternator/battery charging cable can be accounted for by a regulator which

1 Installed consumers taking account of the switch-on duration (examples)

Electrical consumer	Power input	Average electrical load requirements
Motronic, electric fuel-supply pump	250 W	250 W
Radio	20 W	20 W
Side-marker lamp	8 W	7 W
Low beam (dipped beam)	110 W	90 W
License-plate lamp, tail lamp	30 W	25 W
Indicator lamp, instruments	22 W	20 W
Heated rear window	200 W	60 W
Interior heating, fan	120 W	50 W
Electrical radiator ventilator	120 W	30 W
Windshield wiper	50 W	10 W
Stop lamp	42 W	11 W
Turn signal	42 W	5 W
Fog lamps	110 W	20 W
Rear fog lamp	21 W	2 W
Total		
Installed electrical load requirements	1,145 W	
Average electrical load requirements		600 W

Table 1

measures the actual value of the voltage at the battery.

The arrangement of the alternator, battery and consumers influences the voltage drop on the charging cable and thus the charge voltage. The total current $I_G = I_B + I_V$ flows through the charging cable if all electrical consumers are connected to the battery. In response to the relatively high voltage drop, the charge voltage falls proportionately sharply.

If all consumers are connected on the alternator side, the voltage drop is less and the charging voltage is higher. In the process, consumers that are sensitive to voltage peaks or voltage ripple (electronic circuits) may be damaged or suffer malfunctions. Electrical consumers which feature a high power input and relative insensitivity to overvoltage should therefore be connected close to the alternator, whereas voltage-sensitive consumers with a low power input should be connected close to the battery.

Voltage drops can be minimized by suitable conductor cross-sections and good connections with low contact resistance, even after a long service life.

Design of the vehicle electrical system

Dynamic system characteristic curve

The dynamic system characteristic curve maps the relationship between battery

voltage and battery current during a driving cycle. The envelopes reflect the interrelationships between the battery, alternator, consumers, temperature, engine speed, and engine/alternator speed ratio. A large area in the envelope means that, with this type of vehicle electrical system, significant voltage fluctuations are occurring in the selected driving cycle and that the battery undergoes more powerful cyclization, i.e. its state of charge experiences powerful temporal changes. The system characteristic curve is specific to every different combination and every set of operating conditions, and is therefore a dynamic curve. Measuring systems can be connected to the battery terminals to plot the dynamic system characteristic curve.

Charge-balance calculation

The charge-balance calculation is used as the basis for defining the design of the alternator and battery. By means of a computer program, the battery charge level is calculated from the consumer load and the alternator output power at the end of a specified driving cycle. *Rush-hour driving* (low engine speeds) combined with *winter operation* (low charging-current input to the battery) is regarded as a normal passenger-car driving cycle. The battery must maintain a stable charge balance even under these very unfavorable conditions for

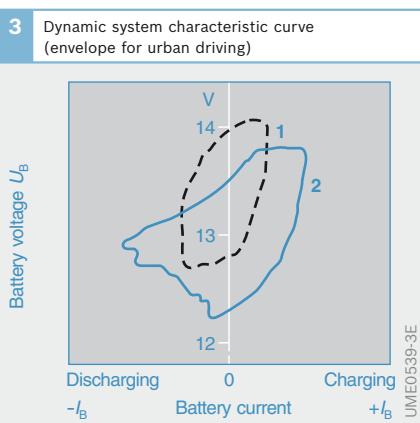
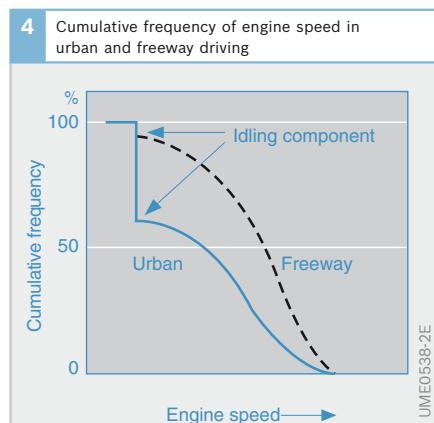


Fig. 3

- 1 With large alternator and small battery
- 2 With small alternator and large battery



the energy balance of the vehicle electrical system.

Driving profile

The driving profile as an input variable for the charge-balance calculation is represented by the cumulative-frequency curve of the engine speed. It describes how often a specific engine speed is reached or exceeded.

In urban driving situations in rush-hour driving, a car's engine is running at idle speed for a large part of the time due to the frequency of stops at traffic lights and consequently high traffic density.

A city bus running on scheduled routes has a high idle percentage time due to interruptions in driving at bus stops. Another factor that has a negative effect on the battery's charge balance is electrical consumers that are operated when the engine is switched off. Long-distance buses generally spend only a small proportion of the time with the engine at idle speed, but on the other hand may have periods when consumers with a high power input are operated with the engine switched off.

Vehicle electrical system simulation

Contrary to the momentary view using a charge-balance calculation, model-based

computer simulations can calculate electrical system power supply status at any time during the trip. They can also include electrical system management systems and assess their effectiveness.

In addition to adjusting current levels in the battery, it is possible to record the vehicle-system voltage characteristic curve and the battery cycle at any time during a trip. Calculations performed using vehicle electrical system simulations are always useful when it comes to comparing electrical-system typologies and assessing the effectiveness of consumers with a high dynamic response or only brief switch-on durations.

Fuel consumption

Since one of the factors that affects the fuel consumption of a vehicle is its mass, the mass of the alternator is also an influence on consumption.

Even the generation of power by the alternator has an effect on fuel consumption: the additional consumption at 100 W of generated electrical output is in the order of 0.17 l per 100 km and depends on the efficiency of the alternator. For this reason, alternators with higher mid-range efficiency levels generally contribute to the engine's fuel efficiency despite being slightly heavier in weight.

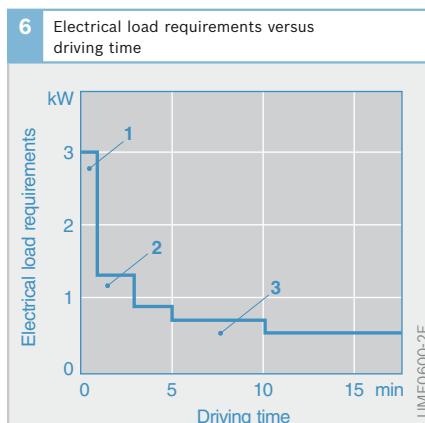
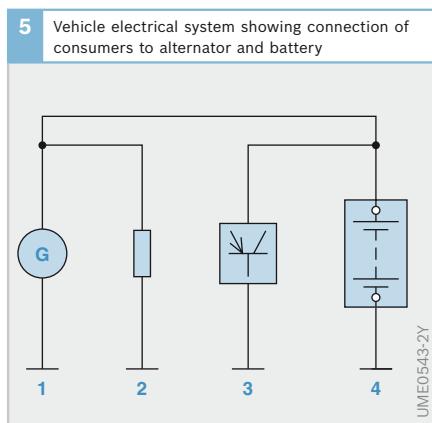


Fig. 5

- 1 Alternator
- 2 Consumer with relatively high power input
- 3 Consumer with low power input
- 4 Battery

Fig. 6

- 1 Windshield heater
- 2 Secondary-air pump
- 3 Heater, fan, engine management, etc.

Electrical energy management

An electrical energy management (EEM) system coordinates the interaction between alternator, voltage transformer, batteries and electrical consumers when the vehicle is in use. When the vehicle is parked, the EEM monitors the batteries, and switches standstill-draw and constant-draw consumers off as soon as the battery charge reaches a critical limit. The EEM regulates the entire electrical energy balance. It compares the power demand from the consumers with the power available within the vehicle electrical system and maintains a constant balance between supply and power output.

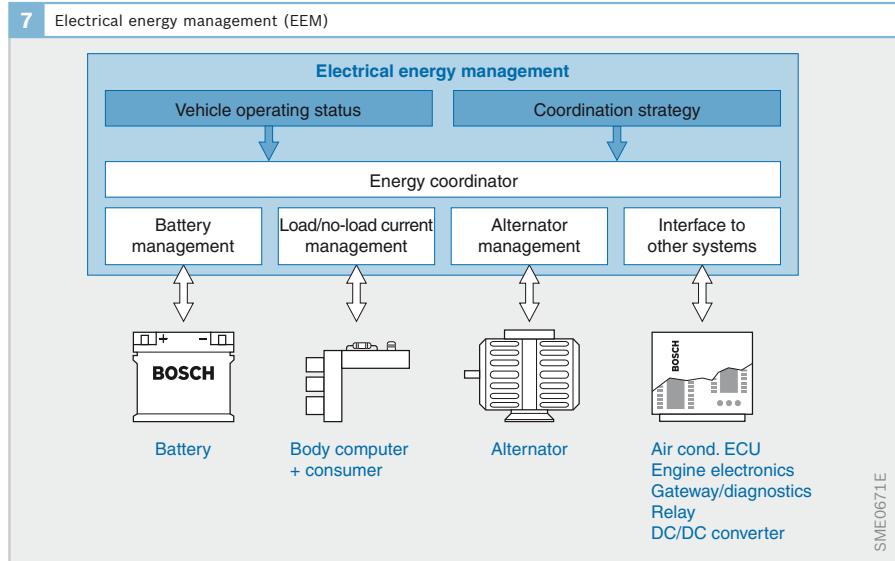
The basis for the EEM is the battery management. The objective of the battery management is to communicate to the EEM information about the current state of the battery and about predicted electrical behavior. Using this information, it is possible to implement operating strategies for increasing vehicle availability and profitability.

The battery management communicates to the EEM the variables relevant to the battery, e.g. the state of charge (SOC), the state of health (SOH) and the state of function (SOF) of the battery. SOF is a prediction of how the battery would react to a pre-defined load profile, e.g. whether a starting operation would be successful with the current battery status.

These values are calculated using complex, model-based algorithms from the measurement of battery current, voltage and temperature.

Using this battery data, the EEM is able to determine the optimal charge voltage and reduce the load on the electrical system (switch off consumers) in response to a degrading state of function and/or increase power generation (e.g. by increasing the idling speed).

If the battery's state of function falls below a specified threshold value despite the measures that have been implemented, the EEM can warn the driver that certain functions (e.g. engine start) will not be available with the current battery status.



Battery status recognition

The control unit that makes battery status recognition (BSR) possible is the electronic battery sensor EBS (sometimes even EEM functions are implemented in this control unit). The sensor with integrated evaluation electronics records the fundamental battery variables of voltage, current and temperature. From these variables, it uses complex software algorithms to calculate the variables that describe the status of the car battery.

The electronic battery sensor comprises a chip, which contains the entire electronics, and a resistor element for current measurement. Together with the terminal clip, both of these form a single assembly unit that can be connected directly to the car battery and fits into the terminal recess of conventional car batteries.

The following tasks of the electrical energy management are made possible by the use of battery status recognition:

- ▶ Assurance of startability (SOF) through compliance with defined limit values of the battery state of function and an increase in vehicle availability
- ▶ Reductions in electrical load requirements and reductions in fuel consumption by means of alternator management with adaptation of alternator voltage
- ▶ Greater flexibility in the design of battery and alternator size by means of superordinate energy management (optimal economic efficiency)
- ▶ Extension of battery life (e.g. through prevention of exhaustive discharge)
- ▶ Battery change indication

For stop/start applications, the following functions can also be fulfilled:

- ▶ Prediction of startability after a defined stationary period (time, temperature, no-load current, power consumption during the stop phase)
- ▶ Assurance of a battery charge reserve in the stop phase (e.g. by switching off consumers)

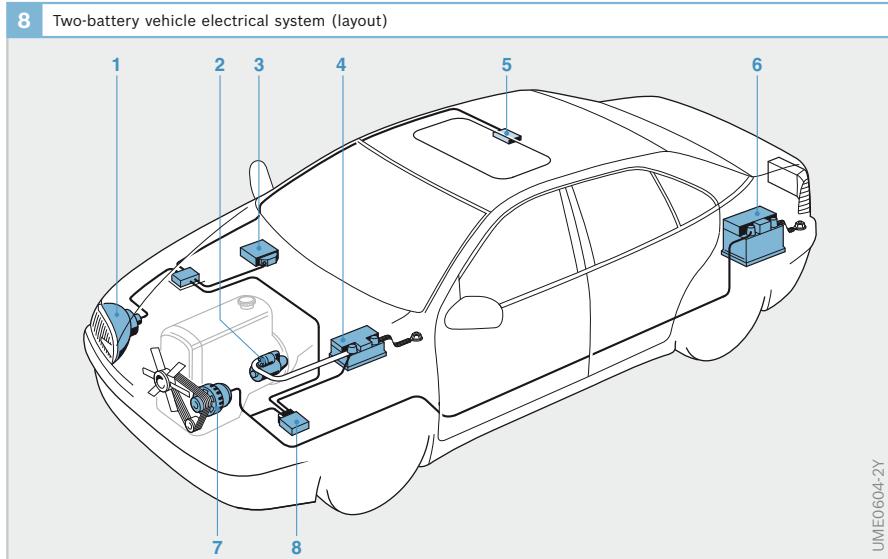


Fig. 8

- | | |
|---|---|
| 1 | Lighting system
(vehicle electrical system) |
| 2 | Starter motor |
| 3 | Engine management
(vehicle electrical system) |
| 4 | Starter battery |
| 5 | Other electrical-system consumers
(e.g. power sunroof) |
| 6 | General-purpose battery |
| 7 | Alternator |
| 8 | Charging/isolating module |

Two-battery vehicle electrical system

In the design of a vehicle battery, which supplies both the starter and the other consumers in the vehicle electrical system, a compromise has to be found between different requirements.

During the engine starting sequence, the battery is subjected to high current loads (300 to 500 A). The associated voltage drop has an adverse effect on certain electrical consumers (e.g. units with microcontroller) and should be as low as possible. On the other hand, only comparatively low currents flow during vehicle operation; for a reliable power supply, the capacity of the battery is the decisive factor. Neither properties – rated output nor capacity – can be optimized simultaneously.

In vehicle electrical systems with two batteries (starter battery and general-purpose battery), the "high power for starting" and "general-purpose electrical supply" functions are separated by the electrical system control unit to make it possible to avoid the voltage drop during the starting process, while ensuring reliable cold starts, even when the charge level of the general-purpose battery is low.

Starter battery

The starter battery must supply a high amount of current for only a limited period of time (during starting). It is therefore de-

signed for a high power density (high power for low weight). Compact dimensions allow installation in the immediate vicinity of the starter motor with short connecting cables. The capacity is reduced.

General-purpose battery

This battery only supplies the vehicle electrical system (excluding the starter). It provides currents for supplying the consumers of the vehicle electrical system (e.g. approx. 20 A for the engine-management system) but has a high cyclic capability, i.e. it can supply and store large amounts of power. Dimensioning is based essentially on the capacity reserves required for activated consumers, the consumers that operate with the engine switched off (e.g. no-load-current consumers, parking lights, hazard warning flashers, immobilizer), and the minimum permissible charge level.

Vehicle electrical system control unit

The vehicle electrical system control unit (EN ECU) in a two-battery vehicle electrical system separates the starter battery and the starter from the rest of the vehicle electrical system provided this can be supplied with sufficient power by the general-purpose battery. It therefore prevents the voltage drop that occurs during starting, affecting the performance of the vehicle electrical system. When the vehicle is parked, it prevents the starter battery from becoming discharged by activated

9 Two-battery vehicle electrical system (schematic)

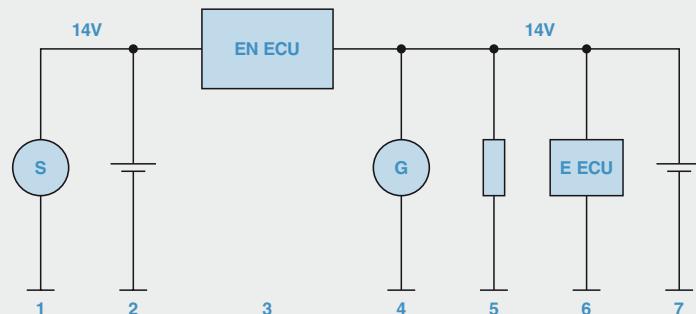


Fig. 9

- 1 Starter
- 2 Starter battery
- 3 Vehicle electrical system control unit
- 4 Alternator
- 5 Consumer
- 6 Engine control unit
- 7 General-purpose battery

consumers with the engine switched off and standstill-draw consumers.

By separation of the starting system from the remainder of the vehicle electrical system, there are theoretically no limits for the voltage level within the starting system. Consequently, the charge voltage can be optimally adapted to the starter battery by a DC/DC converter to minimize the charging time.

If there is no charge in the general-purpose battery, the control unit is capable of provisionally connecting both vehicle electrical systems. This means that the vehicle electrical system can be sustained using the fully-charged starter battery. In another possible configuration, the control unit for the starting operation would connect only the start-related consumers to whichever battery was fully charged.

Vehicle electrical systems for commercial vehicles

Battery changeover 12/24 V

Various heavy commercial vehicles have a combined 12/24 V system, i.e. the supply voltage can be switched between 12 and 24 V. In such cases, the alternator for voltage generation and the electrical components (starter excepted) are designed for nominal 12 V operation. The starter has a nominal voltage of 24 V. It is therefore possible to achieve the power output required to start larger diesel engines, for example.

The system consists of two 12 V batteries that are connected in parallel during vehicle operation and with the engine switched off. The voltage is no different with the parallel connection; the vehicle electrical system is supplied with 12 V. The total

10 Circuit of a starting system with battery changeover relay

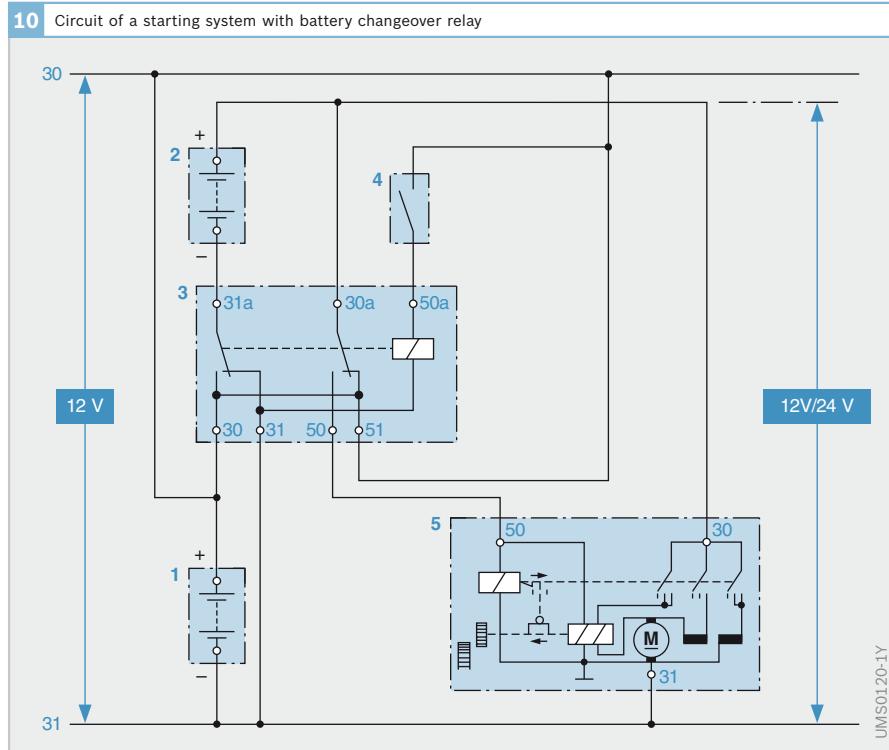


Fig. 10

- 1 12 V battery I
- 2 12 V battery II
- 3 Battery changeover relay
- 4 Ignition switch
- 5 24 V starting motor

capacity of the two batteries is the sum of the individual capacities.

When the ignition switch is turned, a battery changeover relay automatically switches the two batteries in series so that 24 V is applied across the starting-motor terminals during the cranking process. All other consumers are still supplied with 12 V.

As soon as the engine has started, that is when the ignition switch has been released and the starting motor switched off, the battery changeover relay automatically connects the batteries in parallel again. With the engine turning, the 12 V alternator recharges both batteries.

The capacities of two batteries connected in parallel should be equal to achieve uniform current distribution during the charging and discharging process. From the wiring viewpoint, connections should also be as symmetrical as possible, with identical lengths of connection cable and conductor cross-sections.

Components in the vehicle electrical system

At the present time, the components described can also be used in passenger-car electrical systems. Here, however, they are not common and tend to be fitted as optional equipment.

11 Battery master switch



SKE0563-1Y

Battery master switch

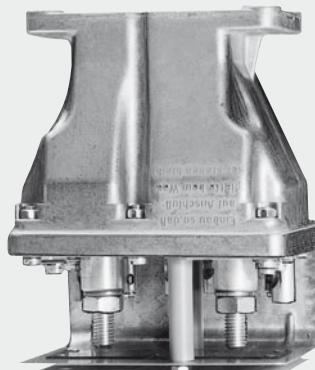
Generally, the vehicle's electrical installation is wired such that when the key is pulled from the ignition switch the electrical lines leading from the switch to, for example, the ignition system, the control units (Motronic, ABS), the wipers, etc. are no longer supplied with current.

On the other hand, the lines leading to the ignition switch, to the starter, and to the light switch remain "live". In other words there is still voltage on these lines, and if they have frayed or worn-through points these can lead to low resistances which can cause leakage currents or short circuits which result in a discharged battery. The consequences are a discharged battery or the possibility of a fire. A battery master switch makes it possible to completely isolate the battery from the vehicle electrical system to eliminate the risk of these dangers.

The single-pole battery master switch is installed in the battery's ground cable (negative terminal) as near to the battery as possible. It should be within convenient reach of the driver.

On installations equipped with three-phase current alternators, due to the danger of voltage peaks (with the attendant destruction of electronic components), it is forbidden to operate the vehicle without the bat-

12 Battery relay



SKE0159-1Y

tery connected. On such installations therefore, the battery master switch may only be actuated with the engine at standstill.

Battery relay

Legislation stipulates that in buses, road tankers etc. a battery relay must be installed as the master switch to separate the vehicle electrical system from the battery. Not only short circuits are avoided (during repairs for instance), but also the decomposition effects due to leakage currents on current-carrying components.

For this type of installation with three-phase current alternator, in order to prevent excessive voltage peaks it is necessary to fit a 2-pole electromagnetic battery main switch. This prevents the alternator being separated from the battery when the engine is running.

Battery-cutoff relay

The battery-cutoff relay (NO contact) separates the starter battery from a second battery used for ancillary equipment. It protects the starter battery against discharge when the three-phase current alternator is not delivering charge current. This relay is provided with a diode for reverse-polarity protection, and a decay diode to suppress the inductive voltage peaks caused by switching.

Battery charging relay

The battery charging relay is needed when an additional 12V battery is to be charged in a 24 V vehicle system voltage. It is provided with resistors across which at 10 A charging current a voltage drop takes place which reduces the charge voltage to 12 V. This of course necessitates the 24 V alternator being able to generate the additional 10 A.

13 Battery-cutoff relay symbol

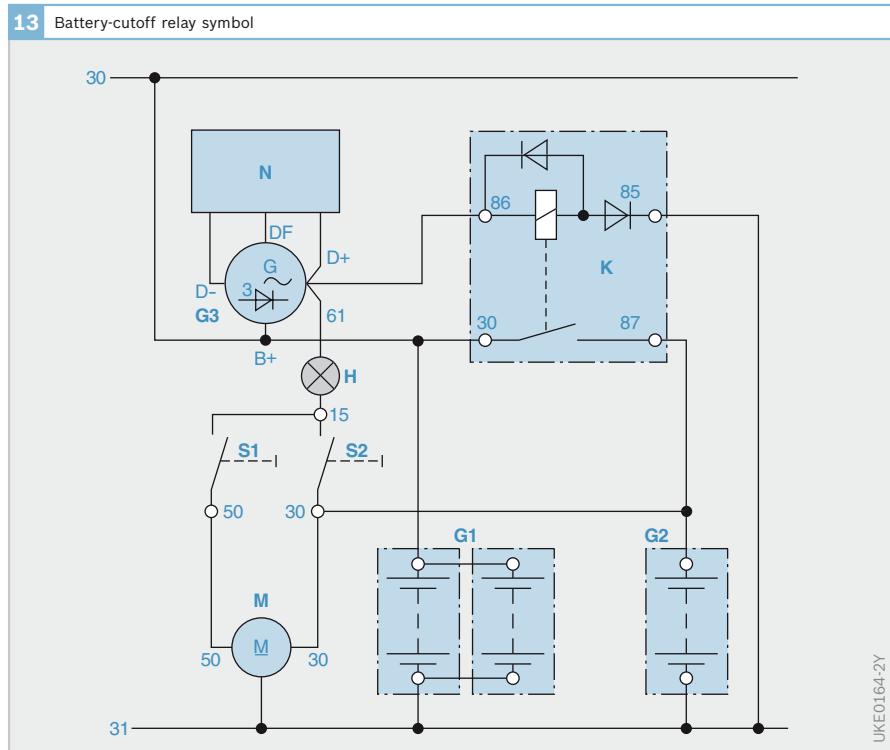


Fig. 13

- G1 Battery for ancillary equipment
 - G2 Starter battery
 - G3 Three-phase current alternator
 - H Charge-indicator lamp
 - K Battery-cutoff relay
 - M Starter
 - N Alternator regulator
 - S1 Ignition switch
 - S2 Driving switch
- UKE0164-2Y

Wiring harnesses

Requirements

The purpose of the wiring harness is to distribute power and signals within a motor vehicle. A wiring harness in the present day, mid-class passenger car with average equipment has approximately 750 different lines, their length totaling around 1,500 meters. In recent years the number of contact points has practically doubled due to the continuous rise in functions in the motor vehicle. A distinction is made between the engine compartment and body wiring harness. The latter is subject to less demanding temperature, vibration, media and tightness requirements.

Wiring harnesses have considerable influence on the costs and quality of a vehicle. The following points must be taken into consideration in wiring harness development:

- ▶ Leak-tightness
- ▶ EMC compatibility
- ▶ Temperatures
- ▶ Damage protection for the lines
- ▶ Line routing
- ▶ Ventilation of the wiring harness

It is therefore necessary to involve wiring harness experts as soon as in the system definition stage. Figure 1 shows a wiring harness that was developed as a special intake-module wiring harness. Thanks to the optimization of routing and security in conjunction with engine and wiring harness development, it was possible to achieve an advancement of quality as well as to yield cost and weight advantages.

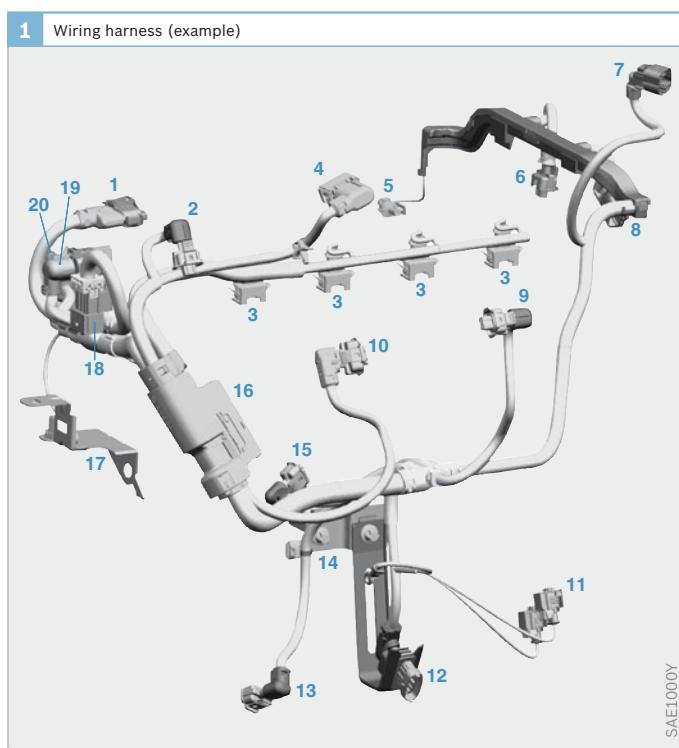
Dimensions and material selection

The most important tasks for the wiring harness developer are:

- ▶ Dimensioning the line cross-sections
- ▶ Material selection
- ▶ Selection of suitable plug-in connections
- ▶ Routing of lines under consideration of ambient temperature, engine vibrations, acceleration and EMC
- ▶ Consideration of the environment in which the wiring harness is routed (topology, assembly stages in vehicle manufacture and equipment on the assembly line)

Fig. 1

- 1 Ignition coil module
- 2 Channel deactivation
- 3 Fuel injectors
- 4 Throttle device DV-E
- 5 Oil-pressure switch
- 6 Engine-temperature sensor
- 7 Intake-air temperature sensor
- 8 Camshaft sensor
- 9 Canister-purge valve
- 10 Intake-manifold pressure sensor
- 11 Charge-current indicator lamp
- 12 Downstream Lambda oxygen sensor
- 13 Speed sensor
- 14 Terminal 50, starter switch
- 15 Knock sensor
- 16 Engine control unit
- 17 Engine ground
- 18 Disconnector plug for engine and transmission wiring harness
- 19 Upstream Lambda oxygen sensor
- 20 EGR valve



SAE1000Y

Line cross-sections

Line cross-sections are defined based on permissible voltage drops. The lower cross-section limit is determined by the line strength. Convention has it that no lines with a cross-section of less than 0.5 mm^2 are used. With additional measures (e.g. supports, protective tubes, tension relief), even a cross-section of 0.35 mm^2 may be permissible.

Materials

Copper is usually used as the conductive material. The insulation materials for the lines are defined by the temperature to which they are exposed. It is necessary to use materials that are suitable for the high temperatures of continuous operation. Here, the ambient temperature must be taken into consideration as much as the heating caused by the flow of current. The materials used are thermoplastics (e.g. PE, PA, PVC), fluoropolymers (e.g. ETFE, FEP) and elastomers (e.g. CSM, SIR).

If the lines are not routed past particularly hot parts (e.g. exhaust pipe) in the engine topology, one of the criteria for the selection of the insulation material and the cable cross-section could be the derating curve of the contact with its associated line. The derating curve represents the relationship between current, the temperature increase that it causes, and the ambient temperature of the plug-in connection. Normally, the heat generated in the contacts can only be carried away along the lines themselves. It should also be noted that the change in temperatures results in a change in the modulus of elasticity of the contact material (metal relaxation). It is possible to influence the relationships described by means of larger line cross-sections and the use of suitable contact types and more noble surfaces (e.g. gold, silver) and thus higher limit temperatures. For highly fluctuating current intensities, it is often useful to measure the contact temperature.

Plug-in connections and contacts

The type of plug-in connections and contacts used depends on various factors:

- ▶ Current intensity
- ▶ Ambient temperatures
- ▶ Vibration load
- ▶ Resistance to media and
- ▶ Installation space

Line routing and EMC measures

Lines should be routed in such a way as to prevent damage and line breaks.

This is achieved by means of fasteners and supports. Vibration loads on contacts and plug-in connections are reduced by fastening the wiring harness as close to the plug as possible and at the same level as the vibration where possible. The line routing must be determined in close cooperation with the engine and vehicle developers.

Where EMC problems arise, it is recommended to route sensitive lines and lines with steep current flanks separately.

Shielded lines are not straightforward to produce and are therefore expensive. They also need to be grounded. The twisting of lines is a more cost-favorable and effective measure.

Line protection

Lines need to be protected against chafing and against making contact with sharp edges and hot surfaces. Adhesive tapes are used for this purpose. The level of protection is determined by the interval and winding density. Corrugated tubing (material savings from corrugation) with the necessary connecting pieces are often used as line protection. However, tape fixing is still an essential means of preventing movement of individual lines inside the corrugated tube. Optimal protection is offered by cable ducts.

Plug-in connections

Design and requirements

The high integration density of electronics in the motor vehicle places high demands on a car's plug-in connections. Not only do they carry high currents (e.g. activation of ignition coils), they also carry analog signal currents with low voltage and current intensity (e.g. signal voltage of the engine temperature sensor). Throughout the service life of the vehicle, the plug-in connections must ensure the reliable transmission of signals between control units and to the sensors whilst maintaining tolerances.

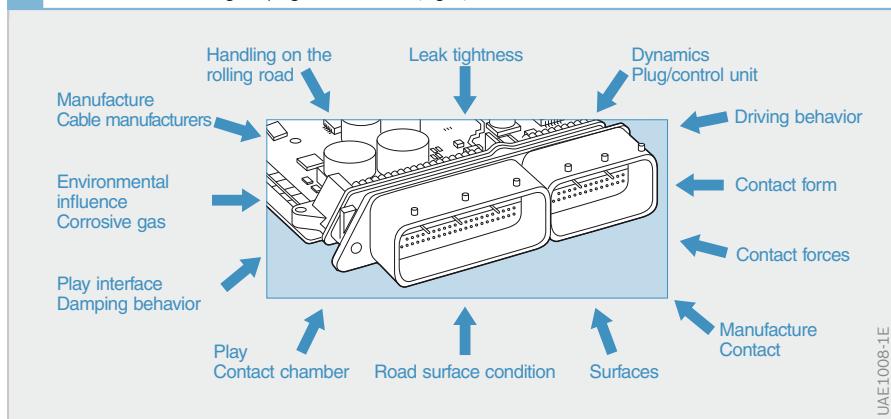
The increasing demands of emission-control legislation and active vehicle safety are forcing the ever more precise transmission of signals through the contacts of the plug-in connections. A large number of parameters must be taken into consideration in the design, arrangement and testing of the plug-in connections (Fig. 1).

The most common cause of failure of a plug-in connection is wearing of the contact caused by vibrations or temperature change. The wear promotes oxidation. This results in an increase in ohmic

resistance – the contact may, for example, be subjected to thermal overload. The contact part may be heated beyond the melting point of the copper alloy. In the case of highly resistant signal contacts, the vehicle controller often detects a plausibility error by comparison with other signals; the controller then enters fault mode. These problem areas in the plug-in connection are diagnosed by the on-board diagnosis (OBD) required by emission-control legislation. However, it is difficult to diagnose the error in the service workshops because this defect is displayed as being a component failure. It is only possible to diagnose the faulty contact indirectly.

For the assembly of the plug-in connection, there are various functional elements on the plug housing intended to ensure that the cables with their crimped contacts can be joined to the plug-in connection reliably and defect-free. Modern plug-in connections have a joining force of < 100 N so that the assembly operative is able to reliably join the connector to the component or control-unit interface. The risk of plug-in connections being connected to the interface incorrectly increases with higher connecting forces. The plug would come loose during vehicle operation.

1 Parameters for the design of plug-in connections (Fig. A)



1 Areas of application for plug-in connections

	No. of poles	Special features	Applications
Low-pin-count	1 to 10	No joining force support	Sensors and actuators (many different requirements)
High-pin-count	10 to 150	Joining force support by slide, lever, modules	Control units (several, similar requirements)
Special connectors	any	e.g. integrated electronics	Special applications (individual, matched requirements)

Design and types

Plug-in connections have different areas of application (Table 1). These are characterized by the number of poles and ambient conditions. There are different classes of plug-in connection: hard engine attachment, soft engine attachment, and body attachment. Another difference is the temperature class of the installation location.

High-pin-count plug-in connections

High-pin-count plug-in connections are used for all control units in the vehicle. They differ in their number of poles and the geometry of the pins. Figure 2 shows a typical design of a high-pin-count plug-in connection. The complete plug-in connection is sealed at the connector strip of the respective control unit by means of a continuous radial seal in the plug housing. This, together with three sealing lips, ensures a reliable seal against the control unit sealing collar.

The contacts are protected against the ingress of humidity along the cable by a seal plate, through which the contacts are inserted and the line crimped to them. A silica-gel mat or silica mat is used for this purpose. Larger contacts and lines may also be sealed using a single-core seal (see “Low-pin-count plug-in connections”).

When the plug is assembled, the contact with the line attached is inserted through the seal plate that is already in the plug. The contact slides home into its position in the contact carrier. The contact latches on its own by a locking spring that engages in an undercut in the plastic housing of the plug.

Table 1

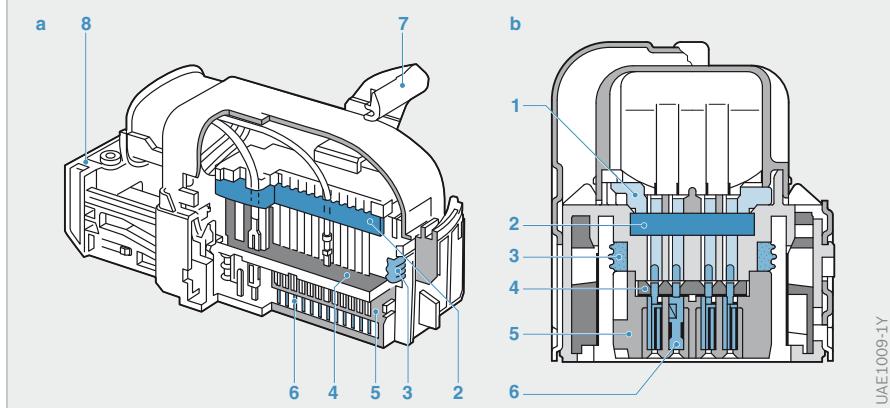
2 High-pin-count plug-in connection (Fig. B)

Fig. 2

- a 3D view
- b Sectional view

- 1 Pressure plate
- 2 Seal plate
- 3 Radial seal
- 4 Slide pin (secondary lock)
- 5 Contact carrier
- 6 Switch contact
- 7 Lever
- 8 Slide mechanism

Once all contacts are in their final position, a slide pin is inserted to provide a second contact safeguard, or secondary lock. This is an additional security measure and increases the retaining force of the contact in the plug-in connection. In addition, the sliding movement is a means of checking that the contacts are in the correct position. The operating force of the plug-in connection is reduced by a lever and a slider mechanism.

Low-pin-count plug-in connections

Low-pin-count plug-in connections are used for actuators (e.g. fuel injectors) and sensors. Their design is similar in principle to that of a high-pin-count plug-in connection (Fig. 3). The operating force of the plug-in connection is not usually supported.

The connection between a low-pin-count plug contact system and the interface is sealed with a radial seal. Inside the plastic housing, however, the lines are sealed with single-core seals secured to the contact.

Contact systems

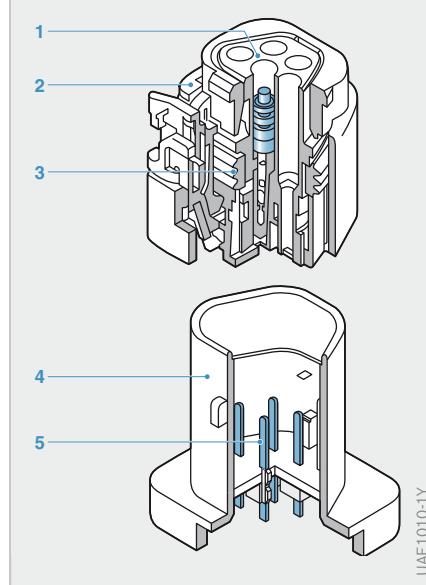
Two-part contact systems are used in the motor vehicle. The inner part (Fig. 4) – the live part – is pressed from a high-quality copper alloy. It is protected by a steel overspring, which at the same time increases the contact forces of the contact by means of an inwardly acting spring element. A catch arm pressed out from the steel overspring engages the contact in the plastic housing part. Contacts are coated with tin, silver or gold, depending on requirements. To improve the wear characteristics of the contact point, not only are different contact coatings used but also different structural shapes. Different decoupling mechanisms are integrated into the contact part to decouple cable vibrations from the contact point (e.g. meandering routing of the supply lead).

Fig. 3

- 1 Contact carrier
- 2 Housing
- 3 Radial seal
- 4 Interface
- 5 Flat blade

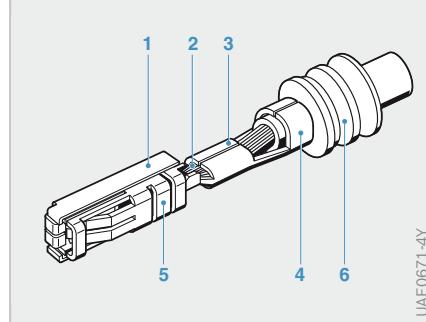
The cables are crimped onto the contact. The crimp geometry on the contact must be adapted to the cable concerned. Pliers or fully automatic, process-monitored crimping presses with contact-specific tools are available for the crimping process.

3 Low-pin-count plug-in connection (Fig. C)



UAE1010-1Y

4 Contact (Fig. D)



UAE0671-4Y

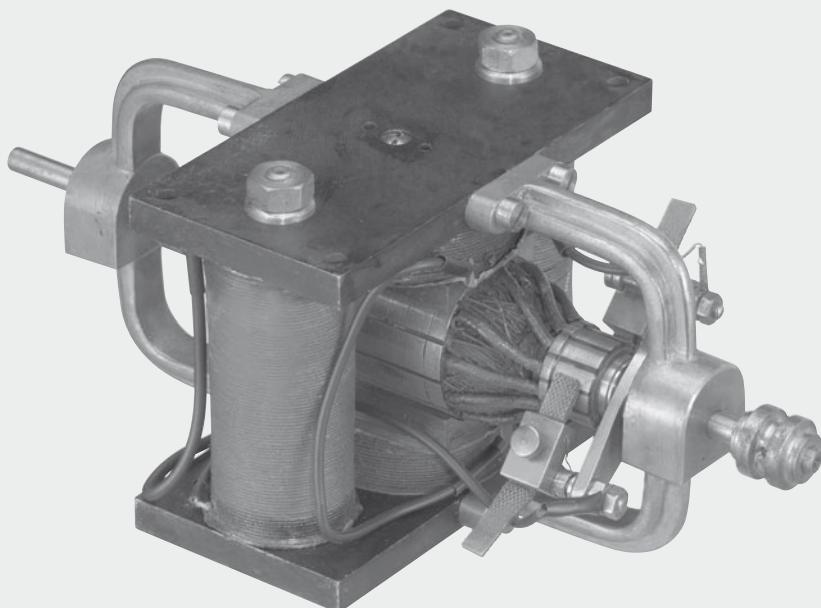
Fig. 4

- 1 Steel overspring
- 2 Single conductor (single core)
- 3 Conductor crimp
- 4 Insulation crimp
- 5 Wave-shaped interior design
- 6 Single-core seal

► History of the alternator

At the turn of the century, the introduction of electrical lighting to motor vehicles to take the place of the previously used horse-and-carriage lighting meant that a suitable source of electrical power had to be available in the vehicle. The battery alone was completely unsuitable since, when discharged, it had to be removed from the vehicle for re-charging. In around 1902, the model for a dynamo (the basis for today's alternator) was created at Robert Bosch. It mainly comprised permanent magnets as stators, an armature with commutator and a contact breaker for ignition (see Fig.). The only difficulty here, though, was the fact that the dynamo's voltage was dependent on the engine's speed which varied considerably.

Endeavors, therefore, concentrated on the development of a DC dynamo with voltage regulation. Finally, electromagnetic control of the field resistor as a function of the machine's output voltage proved to be the answer. Around 1909, using the knowledge available at that time, it thus became possible to build a complete "Lighting and Starting System for Motor Vehicles". This was introduced to the market in 1913 and comprised a dynamo (splashwater-protected, encapsulated 12-V DC dynamo with shunt regulation and a rated output of 100 W), a battery, a voltage-regulator and switch box, a freewheeling starter with pedal-operated switch, and a variety of different lighting components.



UME0664Y

Starter batteries

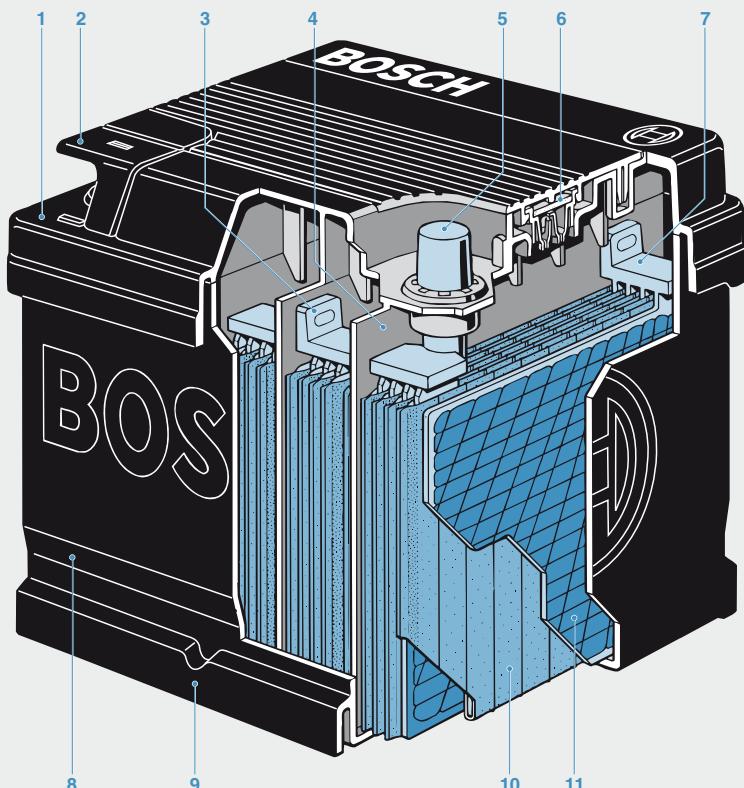
The starter battery is an electrochemical storage facility for the excess electrical energy that is generated by the alternator while the engine is running. This stored energy is needed during vehicle operation in the phases when the energy required by the active consumers is greater than the energy that is generated by the alternator (e.g. at idle speed). The battery also provides the energy that is required by the electrical consumers when the engine is stopped and for starting the engine. The battery can be recharged again when it has discharged. It is therefore a storage battery accumulator, in this case a lead storage battery.

Function and requirements

The starter battery is the electrical energy storage facility in the vehicle electrical system. It has the following tasks:

- ▶ Provide electrical energy for the starter
- ▶ Compensate for the deficit between generation and consumption if the vehicle electrical system is not being supplied with an adequate amount of energy by the alternator (e.g. at idle speed or if the engine is stopped)
- ▶ Damping of voltage peaks from the vehicle system voltage to protect sensitive electronic and electrical components (e.g. bulbs, semiconductors)

1 Design of a lead storage battery (example: maintenance-free battery)



UWE009-3Y

Fig. 1

- 1 One-piece cover
- 2 Terminal-post cover
- 3 Direct cell connector
- 4 Cell partition
- 5 Terminal post
- 6 Sealing plugs underneath the cover plate
- 7 Plate connector
- 8 Battery case
- 9 Bottom rail
- 10 Positive plates inserted into envelope-type separators
- 11 Negative plates

The starter is only switched on for a short time, but it has the highest power input of all electrical consumers (passenger cars with gasoline engines: 0.7 to 2.0 kW; passenger cars with diesel engines: 1.4 to 2.6 kW; busses, commercial vehicles: 2.3 to 9.0 kW). The battery terminal voltage drops during starting because of the high current. It must not drop below a certain level so that the operation of the various control units such as the engine management system is not affected. These will not work if the supply voltage is too low.

A large number of electrical consumers must be supplied with energy when the vehicle is being driven (e.g. the engine management system, the lighting system, the air-conditioning system, the electronic stability program). If the alternator does not generate enough power to supply every electrical consumer that is switched on (e.g. when idling or at low engine speeds), the vehicle system voltage drops to the battery voltage level and electrical energy is drawn from the battery. The battery must therefore be able to provide the vehicle electrical system components with some of their power for a limited period or time or, if the engine and therefore the alternator is switched off, with all of their power. If an adequate amount of power is being generated, the battery can be recharged via the voltage regulator. The procedures that are described are known as the charging cycle and the discharging cycle.

The amount of power that is drawn from a starter battery varies considerably. The power requirements of the vehicle electrical system with the engine and the alternator stationary is approximately 10 to 50 mA (e.g. for the clock, the antitheft alarm system or the remote-controlled central locking system). When the engine is running at idle speed or the vehicle is being driven slowly, 20 to 70 A is required from the battery at certain times. The engine starting procedure requires about 300 A for a period of 0.3 to 3 s, and peak currents of 1,000 A are even possible. The amount of power that is

required and the time that is taken to start the engine are significantly greater at low temperatures (by up to a factor of 2).

The power demand, which results from the electrical load requirements of a given vehicle, is determined in accordance with the vehicle operating conditions. It is decisive not only for the dimensioning of the battery but also of the alternator.

If auxiliary equipment is chosen for the vehicle, this can lead to considerable increases in power demand. Such equipment includes comfort and convenience systems with additional servomotors for power sunroof and power-window drives, seat and steering-column adjustment; as well as seat heating, air-conditioning system, cooler unit or similar. These electrical load requirements are taken into consideration in the vehicle by the manufacturer when the electrical components are dimensioned. This means that a vehicle equipped with such auxiliary equipment is delivered with a larger battery, and in some cases with a more powerful alternator. Other mechanical, cyclical, or climatic loadings are also taken into consideration depending upon the vehicle's application. For example, special vibration-proof batteries with a mat pressed between the plates are frequently used for all-terrain commercial vehicles. AGM batteries, for example, are extremely suitable for special cyclic loading. A corrosion-resistant lead alloy is required for warmer climates.

Caravans and mobile homes are often equipped with such electrical appliances as lighting, refrigerator, heating, radio and TV. Here, it is common to fit extra batteries with a separate electric circuit of their own.

Generally speaking, the lead-acid storage battery is sufficient for meeting these demands, as well as at present still being the most cost-effective energy-storage medium for such assignments. Typical voltages are 12 volts for passenger cars (with a 14-volt vehicle electrical system) and 24 volts for commercial vehicles (two 12-volt batteries connected in series).

Design

Classification

Starter batteries can be subdivided into two types:

► *Closed batteries:*

These are EN 50 342 batteries with freely moving electrolytes in which any gas that builds up can escape through holes in the cover. The vast majority of all starter batteries are closed batteries.

► *Sealed batteries:*

These EN 50 342 batteries only allow gas to escape if the pressure in the battery exceeds a certain value. The battery cannot be topped up with sulfuric acid. The electrolyte is fixed, i.e. it is not free-moving. This can be achieved by absorbing the electrolyte in a glass fiber mat (AGM battery) or by using a gel electrolyte.

The following distinctions also exist:

- Car batteries, the dimensions of which are defined as per the EN 60 095-2 standard and
- Commercial vehicle batteries (mainly in accordance with EN 60 095-4)

With regard to maintenance, starter batteries are categorized into

- Conventional and low-maintenance batteries
- Maintenance-free batteries (in accordance with EN) and
- Completely maintenance-free batteries

The basic design of these batteries is essentially the same, the main difference between them being the alloy from which the structural and discharge materials (grids) of the electrode plates are made.

Components

A 12 V starter battery contains six series-connected individually partitioned cells in a polypropylene case (Fig. 1). Each cell

comprises a plate block consisting of a positive and negative plate set which are composed of lead plates (lead grid and an active material), together with a microporous insulating material (separators) between plates of opposite polarity. The electrolyte is in the form of diluted sulfuric acid which fills the pores in the plates and separators, and the voids in the cells. Terminals, cell connectors, and plate connectors are made of lead. The openings in the partitions for the cell connectors are sealed. A hot-molding process is employed to permanently bond the one-piece cover to the battery case thus providing the battery's top seal.

Each cell of a conventional battery has its own sealing plug that permits initial battery filling, topping-up during maintenance and removes the gasses that occur during charging. Only maintenance-free batteries are installed in new vehicles nowadays because they do not require regular checking of the electrolyte by the driver. They appear to be completely sealed. However, they contain ventilation openings so that the small amounts of gas that occur when the battery is charged by the alternator can escape.

Battery case

The battery case (Fig. 1, Item 8) - the housing of the battery - consists of acid-resistant insulating material (polypropylene) and many versions have bottom rails (9) at the sides for securing the battery in the vehicle.

The battery case is subdivided into cells by partitions. These cells are the basic elements of a battery. They contain the plate blocks (10, 11) with positive and negative plates, as well as the separators inserted between them. The cells are connected in series using cell connectors (3), which provide the connection through openings in the cell walls.

One-piece cover

The cells with the plate blocks are covered and sealed by a shared one-piece cover (Fig. 1, Item 1). The one-piece cover consists of a cap and a basic cover (Fig. 2).

Plate blocks

The plate blocks consist of negative and positive plates (grid plates) connected in parallel, with intermediate separators (Fig. 1, Item 10 and Fig. 3). The capacity of the cells essentially depends on the number of plates and their surface area. The thickness of the plates is selected depending on the battery application and is normally between 1 and 3 mm.

The plates (also known as grid plates) consist of lead grids and active material with which the grids are coated (pasted). The active material of the positive plate contains porous lead dioxide (PbO_2 , orange-brown in color), and that of the negative plate contains pure lead in the form of "sponge lead" (Pb , gray-green in color). In other words, the pure lead is also in an extremely porous form. All negative and positive plates are connected via a plate strap (and/or plate connector), which is located above the set of plates in the battery. This plate strap consists of solid lead. The plates and the plate strap are securely connected together in a melted joint.

Each plate block usually has one negative plate more than it has positive plates.

Active material

The active material is that part of the grid plates that changes chemically when current flows, i.e. during the charging and discharging processes (see DIN 40 729). The material is porous and therefore forms a large effective inner surface area. For example, the negative plates of a battery with 100 Ah have an inner surface area of about $2,000 \text{ m}^2$, and the negative plates have approximately $30,000 \text{ m}^2$. During battery manufacture, the active material is produced by putting lead oxide (PbO) containing 5 to 15 % of finely distributed metallic lead ("gray oxide") into a mixer with water (H_2O), dilute sulfuric acid (H_2SO_4) and possibly other additional materials and short synthetic fibers to form a dough-like paste. This produces basic lead sulfates. Some lead oxide and metallic lead is retained. The paste, which still has a dough-like consistency, is coated onto the lead grids and is left to harden.

The active material of the finished plate is created during subsequent forming, i.e. the electrochemical conversion of this material when the battery is charged for the first time, which all takes place at the battery manufacturer's.

Fig. 2

- a Cap
- b Basic cover

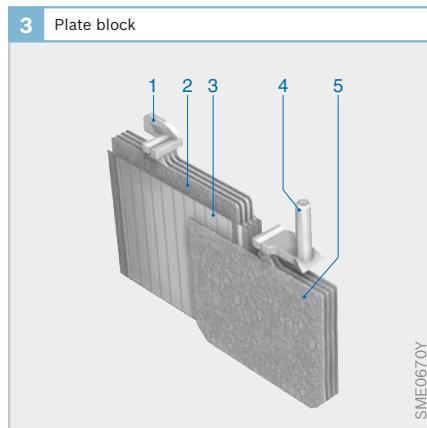
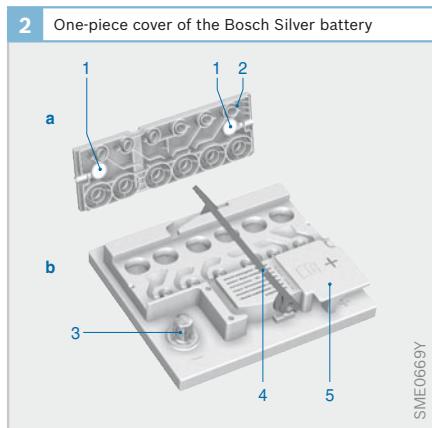


Fig. 3

- 1 Cell connector
- 2 Positive plate
- 3 Separator
- 4 Terminal post
- 5 Negative plate

Separators

Since vehicle batteries have to be as small and light as possible the positive and negative plates are extremely close together, usually 0.8 to 1.5 mm apart. However, they must not come into contact with each other neither when they bend nor when particles crumble away from their surface. Otherwise the battery will be destroyed immediately by the resulting short-circuit. This is why batteries had ribs on the bottom for fixing the plates until recently. Separating walls (*Separators*) were also inserted between the individual plates in a plate block.

The envelope-like separators (Fig. 4) in modern batteries ensure that plates with different polarity are an adequate distance apart and are electrically isolated (galvanically isolated). The bottom ribs are therefore no longer needed. Porous polyethylene foil which is resistant to the effects of oxidation and acid is used as the separator material. This is formed into envelopes which surround (and separate) the positive and negative plates in the plate block. The separators must not produce any significant resistance to the migration of ions in the electrolyte (sulfuric acid). They must also consist of an acid-resistant but permeable (microporous) material so that they can be penetrated by the electrolyte.

The microporous structure is also needed because fine lead fibers that could penetrate the separators would cause short circuiting and therefore have to be held back.

Cell connectors

The individual battery cells are connected in series by the cell connectors (Fig. 1, Item 3 and Fig. 3, Item 1). To decrease the battery's internal resistance and weight, direct cell connectors are used in high-quality batteries. The plate connectors of individual cells are connected by the shortest path, that is directly through the cell partition. This also reduces the danger of short-circuits due to external contact.

Terminal posts and battery-cable terminals
The plate connector of the positive plates in the first cell is connected to the positive terminal post of the battery, and the plate connector of the negative plates in the last cell is connected to the negative terminal post (Fig. 1, Item 5). The two terminal posts are the connecting links between the vehicle electrical system and the battery (Figs. 5 and 6) and consist of a lead alloy. Their conical shape provides a firm seating and a good contact with the battery terminals. The terminal voltage is then available between these two battery terminal posts (i.e. approx. 12 V).

4 Battery plate with envelope-type separator

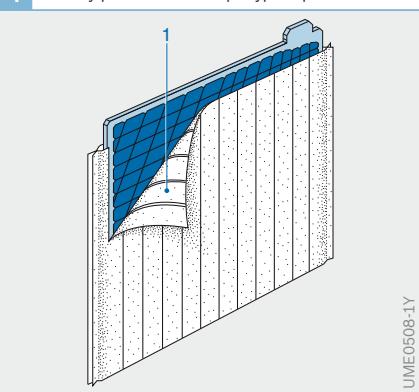
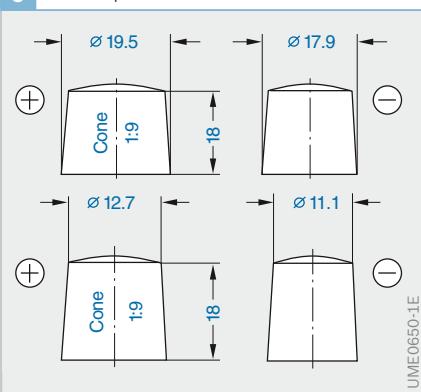


Fig. 4

- 1 Envelope-type separator cut open

5 Terminal posts



The vehicle connecting cables are attached to the terminal posts by special battery terminals. To prevent one pole from being mistaken for the other (polarity reversal), the posts are specially marked and designed differently (the positive post is thicker than the negative post). The battery terminals that are needed to connect the battery in the vehicle come in two versions (Fig. 6):

- Screw-type terminals and
- Solder-type terminals

Designs

All batteries are described in standard lists which contain the electrical specifications and stipulations for the geometrical dimensions of the battery case and the terminal posts (EN 60 095-2 for passenger cars). In addition, in order to ensure full interchangeability irrespective of manufacturer, details of mounting variants are given, as well as of the cell configuration and interconnection. When ordering batteries from the Aftermarket Program, they are listed with these design variants as product features. The design variants are discussed in more detail in the following chapters.

Electrical connections

Depending upon the available space and the equipment layout in the vehicle, a wide variety of batteries with different dimensions and terminal post configurations is required. A wide range of battery dimensions is possible depending on cell arrangement (lengthwise or transverse installation) and interconnection design. An overview of the most common connection designs is given in Figure 7.

Battery covers

Passenger car batteries

Depending on the battery type there are two battery-cover versions for passenger car batteries (Figs. 8 and 9):

- One-piece cover and
- Mono cover

In one-piece covers with a gas passage, the gas that is given off during charging is led out of the battery centrally via a hose. One-piece covers have a sealing plug for each cell that can be removed for topping up with electrolyte and for maintenance purposes. Many maintenance-free batteries no longer have sealing plugs.

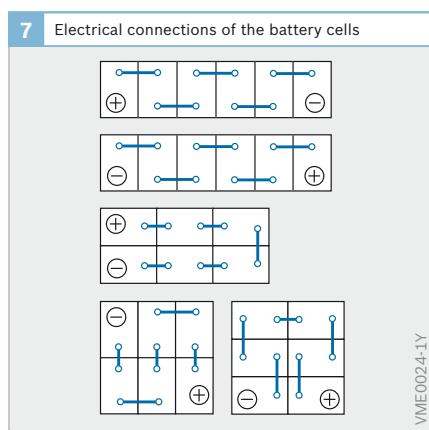
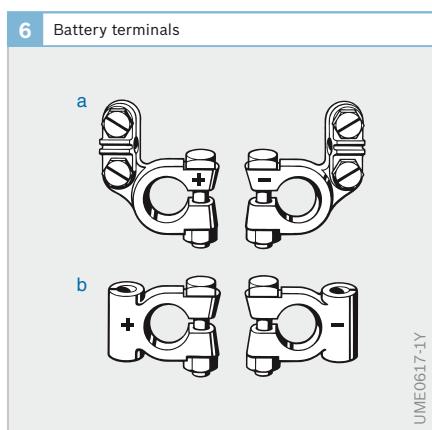


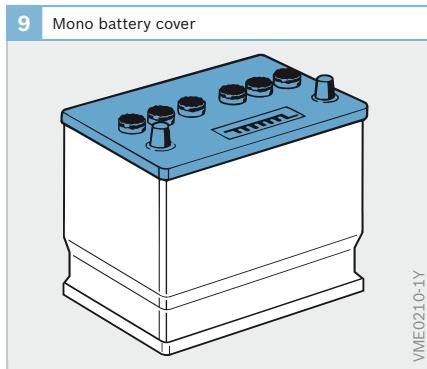
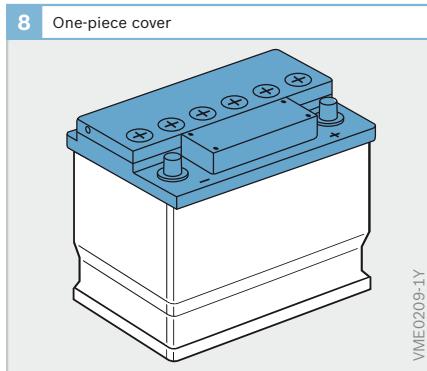
Fig. 6

- a Screw-type terminals
- b Solder-type terminals

The one-piece cover of a maintenance-free Bosch Silver battery has the following characteristics:

- ▶ Continuous, smooth cover surface (no plugs)
- ▶ Handles for easy transportation (Fig. 11)
- ▶ Labyrinth for avoiding loss of water by evaporation
- ▶ Frits (sintered microporous elements made from polyethylene) in the gas passage to prevent backfiring. This means that sparks or flames occurring in the immediate vicinity of the battery cannot reach the interior.
- ▶ Labyrinth and frits to help prevent electrolyte from escaping

The mono cover has neither a gas passage nor a labyrinth. The gas escapes through plugs with ventilation openings.



Commercial vehicle batteries

There are two main types of commercial vehicle battery:

- ▶ Mono cover and
- ▶ Labyrinth cover

The gasses that occur during charging are removed via openings in the plug in mono covers. There are no frits.

The labyrinth cover essentially has the advantages of a passenger car one-piece cover, except for the fact that the labyrinth is lower down because of the external dimensions that are required for a commercial vehicle battery. The advantages are essentially the same:

- ▶ Frits to prevent backfiring
- ▶ Labyrinth to help prevent electrolyte from escaping and
- ▶ Central gas removal

Mounting

The battery must be mounted in the vehicle so that it cannot move of its own accord. It is therefore clamped to a base using one of the following methods:

- ▶ Clamping frame
- ▶ Bracket, each with clamping screw or
- ▶ Secured at the base with a clamping claw or clamping screw (Fig. 10)

The various forms of recess in the base of the battery case for the above forms of mounting are common, and are therefore

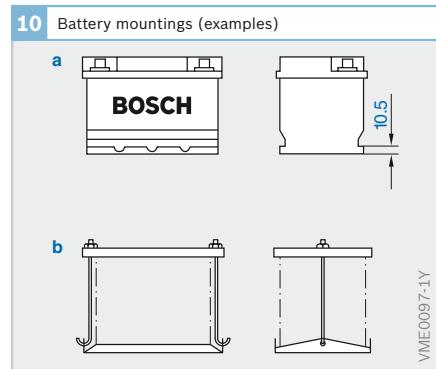


Fig. 10

- a Base mounting
- b Clamping frame mounting

also regulated by standardization.

Reliable mounting is important for safety. Incorrectly mounted batteries can move in the event of minor accidents or even during sudden maneuvers and cause a fire due to short-circuiting.

The form of attachment provided complies with all safety regulations and should therefore not be modified.

Operating principle

Electrochemical processes in the lead cell

Generation of the cell voltage

If a metal electrode (e.g. made from lead) is immersed in an electrolyte (e.g. sulfuric acid), metal ions are released into the electrolyte. These metal ions are positively charged and leave electrons in the electrode.

1 Comparison of different battery versions		EN-water consumption (g/Ah)	Battery case	Cover/plug	Separator	Grid alloy Positive	Grid alloy Negative	Miscellaneous
Conventional starter battery	> 4.0	Bottom ribs	Mono cover and degassing plug	Plate separator	Lead/antimony	Lead/antimony	Extremely deep-cycle resistant; commercial vehicle use: some separators on fiberglass mats	
Low-maintenance	Usually < 4.0	Smooth base		Envelope-type separator	Antimony < 3.5 %	Lead/antimony		
Maintenance-free in acc. with EN	< 4.0, usually < 2.0	Smooth base	One-piece cover with central degassing; plugs	Envelope-type separator	Lead/antimony	Lead/calcium	Hybrid battery; negligible OEM use	
Completely maintenance free	Liquid electrolyte	< 1.0	Smooth base	One-piece cover ¹⁾ with central degassing ^{7); frit²⁾}	Envelope-type separator	Lead/calcium/silver ⁴⁾	Lead/calcium	State of the art ⁵⁾ OEM and IAM
	AGM ³⁾	< 1.0	Reinforced structure, smooth base	One-piece cover with central degassing; valve and frit	Microporous fiberglass mat or gel ⁶⁾ and plate separator	Lead/calcium/silver	Lead/calcium	Extremely deep-cycle resistant; luxury class vehicles and 2-battery vehicle electrical systems

¹⁾ Frequently with labyrinth for optimum electrolyte retention, ²⁾ Frits provide backfiring protection, ³⁾ Absorbent Glass Mat,

⁴⁾ With Bosch batteries, ⁵⁾ Increasingly used for commercial vehicles, ⁶⁾ Not as passenger car starter battery,

⁷⁾ OEM with plugs, aftermarket without plugs

11 Bosch Silver battery and its components

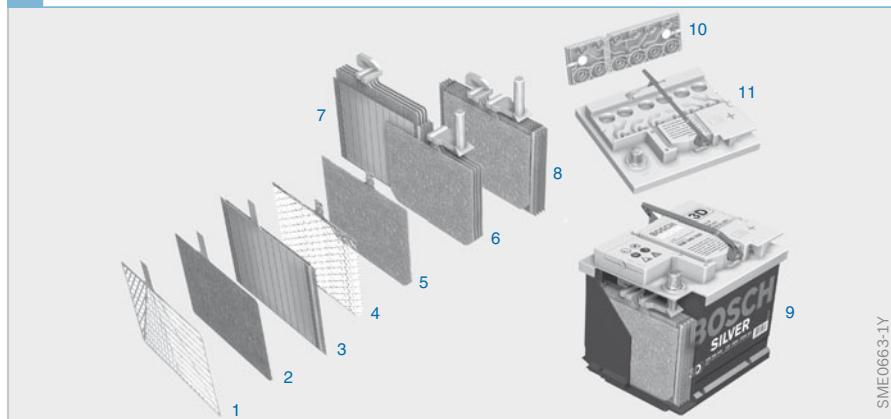


Fig. 11

- 1 Positive grid
- 2 Positive plate
- 3 Positive plate in plate separator (separator)
- 4 Negative grid
- 5 Negative plate
- 6 Negative plate set
- 7 Positive plate set
- 8 Plate block
- 9 Case with bottom rails
- 10 Cap
- 11 Basic cover

The electrode therefore has electrical potential. This counteracts the release of additional ions, creating an equilibrium between the release and depositing of metal ions whilst absorbing electrons.

A PbO_2 electrode behaves in a similar way, except for the fact that a conversion of differently charged lead ions takes place (Pb^{2+} and Pb^{4+}). Different electrodes have different potential, resulting in a cell voltage as a difference between these potentials. In a charged lead cell (Fig. 12) the positive electrode essentially consists of lead oxide (PbO_2), and the negative electrode consists of pure lead (Pb). Dilute sulfuric acid is used as the electrolyte (H_2SO_4 plus H_2O). The sulfuric-acid component means that the pure water becomes conductive so that it can be used as an electrolyte.

The electricity is transported via ion conduction. In the water-based solution the sulfuric acid molecules split into positively charged hydrogen ions (H^+) and negatively charged residual-acid ions ($\text{S}_4\text{O}_4^{2-}$). The splitting is the prerequisite for the electrolyte's conductivity and therefore for the flow of charge or discharge current.

The particle transfers involved in the charge and discharge of the lead cell are described in the two paragraphs below in more detail.

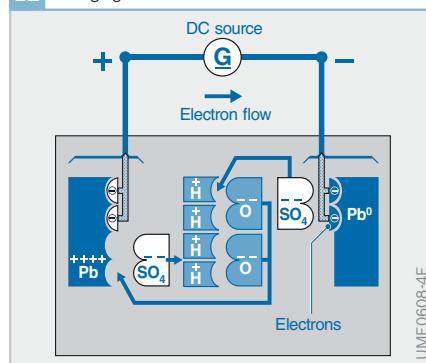
Charging a battery

The battery is charged while the vehicle is being driven, provided that the alternator is supplying enough charging current. A discharged battery can also be recharged using a battery charger.

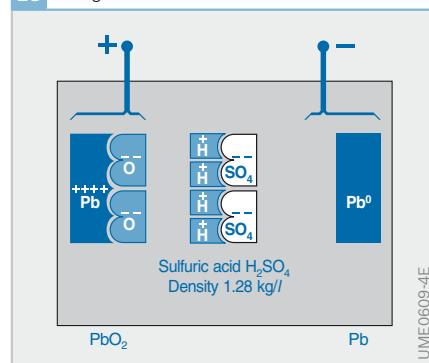
During the charging procedure, the positive electrode of the lead cell is connected to the positive terminal, and the negative electrode is connected to the negative terminal of the DC source (alternator or battery charger). Unlike the discharging process that is described later, the charging process introduces electrical energy so that all cells achieve a higher energy level after charging. Figures 12 to 15 demonstrate in schematic form the processes which take place between the individual particles in the electrode mass and the electrolyte.

The charging voltage source transports a charge from the positive electrode to the negative electrode in the cell. It forces the electrons of the negative electrode, which turns the bivalent positive lead (Pb^{2+}) into "zero value" (metallic) lead (Pb) at this electrode by breaking up the lead sulfate molecules. At the same time, the negatively charged residual-acid ions (SO_4^{2-}) are transferred from the negative electrode to the electrolyte (Fig. 12).

12 Charging



13 Charged cell



At the positive electrode, bivalent positive lead (Pb^{2+}) is turned into tetravalent positive lead (Pb^{4+}) due to the removal of electrons. The lead sulfate (PbSO_4) is electrochemically split by the applied charge voltage. The tetravalent positive lead combines with the oxygen removed from the water (H_2O) to form lead dioxide (PbO_2). At the same time, the sulfate ions released at the positive electrode during this oxidation process (SO_4^{2-} , from the lead sulfate PbSO_4) and hydrogen ions (H^+ , from the water) pass into the electrolyte. The reaction equation of the charging procedure is:



The charging procedure increases the number of hydrogen ions (H^+) and sulfate ions (SO_4^{2-}) in the electrolyte. This means that sulfuric acid (H_2SO_4) is formed, whereby the density ρ of the electrolyte increases (normally $\rho = 1.28 \text{ kg/l}$ for a charged cell). This corresponds to a sulfuric acid content of approx. 37 %. This means that the state of charge of the battery can be determined by measuring the electrolyte density.

Charging is complete (Fig. 13), when

- The lead sulfate (PbSO_4) has been converted into lead dioxide (PbO_2) at the positive electrode and
- The lead sulfate (PbSO_4) has been converted into metallic lead (Pb) at the negative electrode
- The charge voltage and the electrolyte density ρ no longer increase, even if charging continues

Gassing

The charging process causes the supplied electrical energy to be converted into chemical energy and stored. If charging continues after the battery has been fully charged, only electrolytic water decomposition takes place. Oxygen forms at the positive plate (O_2), and hydrogen (H_2) forms at the negative plate. This procedure

is known as gassing. Water may subsequently need to be added.

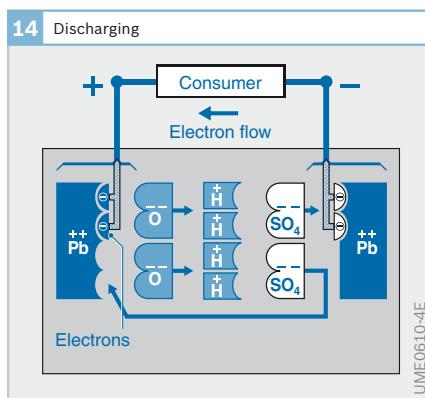
Overcharging can be reduced by restricting the charging time, for example. Overcharging in the vehicle can be prevented by means of charge status-controlled charging, which requires some form of battery status detection, however.

Impact of engine speed

The battery charge largely depends on how the vehicle is driven (e.g. in traffic jams, stop-and-go, or without stopping). The alternator is driven by the engine, and the amount of current that is generated by the alternator increases together with the engine speed. Long waiting times in traffic jams and at traffic lights when the engine is only running at idle speed result in low alternator speeds and a low battery charge current. Short-distance driving has a negative effect on the battery's charge balance. However, if longer distances are driven on a highway or interstate, the engine speed is in the medium to upper range and the battery charge current is accordingly high.

Discharge (current drain from the battery)

Compared to the battery-charging process, the direction of current flow and the electrochemical process are reversed when discharge takes place. If a consumer (e.g. lamp bulb) is connected between the poles



of a battery, due to the potential difference between the poles (6 times the cell voltage) electrons flow from the negative pole through the consumer to the positive pole.

This electron transition causes the tetravalent positive lead (Pb^{4+}) of the positive electrode to be converted into bivalent positive lead (Pb^{2+}), and the binding of the previously tetravalent lead at the oxygen atoms (O) is canceled out (Fig. 14). The oxygen atoms that have been released combine with hydrogen atoms (H^+) that have been taken from the sulfuric acid (H_2SO_4) and form water (H_2O). The density of the electrolyte therefore drops. Depending on the type of battery, this is usually significantly below $\rho = 1.12 \text{ kg/l}$. This cor-

responds to a sulfuric acid content of approx. 17 %.

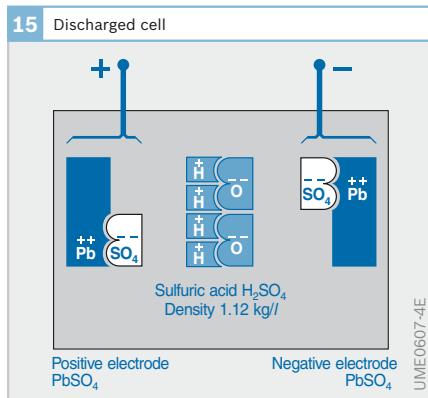
At the negative electrode, bivalent positive lead also forms as a result of electrons moving from the metallic lead to the positive electrode (Pb^{2+}). The bivalent negatively charged sulfate ions (SO_4^{2-}) from the sulfuric acid combine with the bivalent positive lead at both electrodes so that lead sulfate (PbSO_4) is formed as a discharge product at each electrode (Fig. 15).

The reaction equation of the discharging procedure is:



Both electrodes have now returned to their initial state: the chemical energy stored in the cell has been transformed back into electrical energy by the discharge process.

An overview of the battery discharging procedure is shown in Table 2.



2 Overview of the discharge processes

	Positive electrode	Electrolyte	Negative electrode
Lead cell, charged	Active material: lead dioxide (PbO_2 , brown)	High-density sulfuric acid (H_2SO_4)	Active material: lead (Pb , metallic gray)
Current drain Electron stream from negative electrode (via consumers) to positive electrode	Electron absorption reduces the tetravalent lead dioxide (Pb^{4+}O_2) to bivalent positive lead ions (Pb^{2+}), which combine with the residual sulfuric acid (sulfate ions SO_4^{2-}) to form the light-colored lead sulfate (PbSO_4).	The oxygen (O_2) of the lead dioxide (PbO_2) of the positive electrode forms water with the liberated positively charged hydrogen ions (H^+ , H_3O^+) of the sulfuric acid; it becomes diluted.	Electron emission oxidizes the neutral metallic lead (Pb) to bivalent positive lead ions (Pb^{2+}), that combine with the residual sulfuric acid (sulfate ions SO_4^{2-}) to form light-colored lead sulfate (PbSO_4).
Lead cell discharged	Lead sulfate (PbSO_4) from the $\text{Pb}^{2+} + \text{SO}_4^{2-}$ ions	Low-density sulfuric acid	Lead sulfate (PbSO_4) from the $\text{Pb}^{2+} + \text{SO}_4^{2-}$ ions

Table 2

Battery designs

Conventional and low-maintenance batteries

Conventional batteries are hardly ever installed in new vehicles any more, since they are not maintenance-free and their fluid level has to be checked at regular intervals. Since the gassing voltage (voltage at which the gassing process occurs) is lower than it is in any of the other popular battery versions, the gassing process starts sooner, and more charge gas escapes (see section entitled “Electrochemical processes in the lead cell”). For this reason, water has to be added through the sealing plug at regular intervals.

Low-maintenance batteries are slightly better. Because of the low water consumption ($< 4 \text{ g/A} \cdot \text{h}$), the maintenance intervals are longer, and they then depend on the type of vehicle use and the battery installation location.

Characteristics

Conventional batteries

Conventional batteries have ridges on the inner floor surface of the case, on which the plate feet rest. The spaces between the element rests, and below the bottom edges of the plates, form sediment chambers which serve to accumulate the solid material which, in the course of time, is sloughed off from the plates and falls to the bottom of the case. The sediment, which contains lead and is electrically conductive, accumulates in the sediment chambers without contacting the plates. Short-circuits are thus avoided. The design with ridges and sediment chamber is needed in conventional batteries, since there are plate separators between the plates that do not surround them at the bottom.

Low-maintenance batteries

However, low-maintenance batteries use envelope-type separators. Because of their envelope shape they surround the positive

or negative plate completely. Material particles that become detached drop to the bottom inside the envelope and can therefore not cause short circuiting. No ridges are therefore needed on the base of the battery - the bottom of the case is smooth. This means that the available plate surface can be increased (more current drain possible) and the plates rest on their entire underside (greater stability).

Plate grid alloy material

Lead-antimony alloy (PbSb)

In order to improve the casting capability of the thin lead grids (extremely important in high-performance starter batteries), accelerate the hardening process and give the lead plates the stability that is required when driving (*Deep-cycle resistance*), the lead of the grid consists of a lead-antimony alloy (PbSb). In other words, the antimony acts as a hardener (this is where the term “hard lead” comes from). However, during the battery’s service life the antimony is increasingly separated out due to positive-grid corrosion. It migrates to the negative plate, passing through the electrolyte and the separators on the way, and “poisons” it by forming local voltaic couples. These local voltaic couples increase the self-discharge of the negative plate and reduce the gassing voltage. Both cause increased water consumption when overcharging occurs, which encourages the release of antimony. This self-energizing mechanism leads to a continuous drop in power throughout the battery’s service life. The reduced charging current leads to charging problems, particularly in winter. The battery does not achieve the required state of charge, and the electrolyte has to be checked at frequent intervals.

The self-discharging of the negative plates that is caused by the antimony content of 4 to 5 % in the grid lead is therefore one of the main causes of failure in conventional starter batteries. Depending upon driving conditions, water consumption ($> 4 \text{ g/A} \cdot \text{h}$) resulting from increased

gassing in older batteries used to mean the battery had to be checked every four to six weeks. The proportion of antimony in the positive plate is lower in low-maintenance batteries (< 3.5 %). This slows down the increase in the self-discharging of the battery during the course of its service life.

Since batteries with lead-antimony alloy are extremely deep-cycle resistant, they are mainly used in commercial vehicles and taxis. Batteries for motorcycles are also based on antimony technology, since frequent use during good weather and long park-up times during winter require extremely deep-cycle resistant batteries.

However, the advantages of deep-cycle resistance are frequently counteracted by grid corrosion during service life. This is caused by the fact that grids with lead-antimony are subject to more corrosion than the lead-calcium alloy or the even better lead-calcium-silver alloy that are described below.

However, since maintenance-free batteries and not extremely deep-cycle resistant batteries are required for the majority of passenger cars, antimony batteries are seldom used in new vehicles.

Maintenance-free battery (as per EN)

Plate grid alloy material

The negative grid of maintenance-free batteries (hybrid batteries) consists of a lead-calcium alloy (PbCa) - with added silver in some versions - and the positive grid is made from an antimony alloy (PbSb).

Calcium takes over from antimony as the hardener for the negative plates. Calcium is electrochemically inactive under the potential conditions that exist in lead storage batteries. This means that negative plate poisoning is avoided and self-discharging is prevented. However, what is more important is the high gassing voltage which is stable over the usage period, and the associated water consumption, which is lower compared to lead-antimony alloy.

Another advantage of a hybrid battery is ease of manufacture. The negative grids with the calcium alloy can usually be manufactured using a simple drawing process, and the positive grids, which are subjected to greater mechanical stresses through corrosion, are produced in antimony alloy using a more complicated casting process. However, due to the antimony content, hybrid batteries can seldom meet the high demand for low water consumption in passenger cars (< 1 g/A · h). Only completely maintenance-free batteries, in which both positive and negative grids are made of a lead-calcium alloy, are capable of meeting these demands.

Under the high cyclical loading conditions encountered in taxis in densely populated areas, in town buses, and in delivery vehicles, maintenance-free starter batteries have come to the forefront due to the advantages inherent in their separator concept (envelope-type separators) and the reliable protection against failure that they provide.

Characteristics

The water loss of a maintenance-free battery is lower over its entire service life than a conventional battery (< 4 g/A · h, usually < 2 g/A · h). The liquid level only needs to be checked during the normal service intervals in the workshop. Porous polyethylene foil which is resistant to the effects of oxidation and acid is used as the separator material. This is formed into envelopes which surround (and separate) the positive and negative plates in the plate block.

Other features are:

- ▶ Labyrinth one-piece cover with central gas vent. This minimizes water consumption caused by evaporation and prevents electrolyte from escaping if the battery is briefly tipped over.
- ▶ Frits provide protection from backfiring when sparks are present: this means

that they prevent escaping charge gas from being ignited by an external influence (naked flame) or igniting and flashing back into the battery.

- ▶ The terminal posts are protected from inadvertent short circuiting by caps.
- ▶ The cover plate over the plug trough covers the sealing plugs and prevents the accumulation of dirt and moisture.
- ▶ The bottom of the case is smooth on the inside because envelope-type separators are used. The plates reach down to the floor of the case (= increased plate surface) with which they are in contact along their complete length (increased stability).
- ▶ Microporous envelope-type separators prevent the loss of active material, as well as the formation of short circuits at the bottom and at the side edge of the plates. The mean pore diameter of the envelope-type separators is smaller than that of conventional plate separators by a factor of 10, thereby effectively preventing short-circuits as well as providing reduced electrical resistance.

Maintenance-free Bosch starter batteries comply with the minimum performance values stipulated by DIN, and also fulfill the following requirements:

- ▶ No impairment of performance data and charging characteristics as a result of water consumption.
- ▶ The performance data and charging characteristics remain consistent during the entire service life of the battery.
- ▶ After exhaustive discharge and a subsequent park-up period under vehicle electrical system conditions, the battery can be charged again.
- ▶ In case of seasonal operation without intermediate charging (but with disconnected ground cable), no reduction in usable life is to be expected compared to all-year-round operation.
- ▶ The battery must be able to be stored for long periods after filling.

Completely maintenance-free battery

Completely maintenance-free batteries have a longer service life for extremely long distance driving and are more resistant to continuous overcharging. This is achieved by means of a further development of the plate alloy.

Plate grids

Lead-calcium-silver alloy (PbCaAg)

The higher performance of modern car engines combined with more compact and streamlined body shapes has resulted in an increase in the average engine-compartment temperature. This change in conditions also has an effect on the starter battery, as a result of which the most recent development in battery design has seen the use of an improved lead alloy for the grids of the positive plates. In addition to having a lower calcium content and a higher tin content, they also contain a proportion of silver (Ag). This lead-calcium-silver alloy (PbCaAg) has a refined grid structure and has proven itself to be extremely durable, even under the influence of high temperatures that accelerate corrosive decay. This is true when destructive overcharging occurs with high electrolyte density and also during (equally undesirable) periods of disuse with low electrolyte density.

The alloy material for the negative plates is a lead-calcium alloy. These batteries are also antimony-free.

Characteristics

The optimized geometry of the grid structure with optimized electric conductivity allows better use to be made of the active material. The connection (center tab) for the cell connector ensures that the grid plates inside the battery casing are uniformly fixed. This method provides the possibility of creating plates that are about 30 % thinner (but more stable) than maintenance-free batteries and, as a result, of increasing their number. That makes it

possible to increase cold-starting power without sacrificing quality.

Versions with the “robust design” have shorter and thicker positive plates with a stable frame, meaning that a high volume of electrolyte is needed above the plates. Consequently, they are always covered by electrolyte and thus protected against corrosion. These properties generally make this type of battery tougher in practical applications.

The completely maintenance-free battery does not require electrolyte level monitoring and does not usually provide a facility for doing this. It is completely sealed apart from two ventilation openings. As long as the vehicle electrical system is operating normally (i.e. constant voltage limited to a maximum level), water decomposition is reduced to such an extent ($< 1 \text{ g/A} \cdot \text{h}$) that the electrolyte reserves above the plates will last for the entire service life of the battery. A completely maintenance-free battery also has the advantage of extremely low self-discharging. This makes it possible to store the fully charged battery for several months after delivery.

Because of the low self-discharging, all completely maintenance-free batteries are filled with electrolyte in the manufacturing plant. This avoids dangerous spillages in workshops or on dealer premises when mixing and adding electrolyte.

If a completely maintenance-free battery is recharged outside the vehicle electrical system, the charge voltage must not exceed 2.3 to 2.4 volts per cell, since overcharging with constant current or using chargers with a W characteristic (constant resistance) leads to water decomposition (gassing).

The completely maintenance-free batteries that are currently on the market have a safety labyrinth cover with side ventilation openings that prevent electrolyte from escaping if the battery is tilted at angles of up to 70° and, with the frits,

also provide protection against backfiring. Sealing plugs are no longer required.

The Bosch battery with silver alloy in the positive plates has a 20 % longer service life than conventional batteries. There are more plates per cell because the plates are thinner. The resulting greater surface area increases the starting power by 30 % in comparison with conventional batteries. This type of battery also has a “power control system” which is located in the battery cover and indicates the state of charge by displaying different colors.

- ▶ Green: state of charge is OK
- ▶ Dark gray: battery needs to be recharged. After recharging, the indicator changes back to “green”
- ▶ White: battery defective, must be replaced

For applications in commercial vehicles there are Bosch batteries with silver alloy which offer the advantages of the completely maintenance-free, passenger car starter battery (“Bosch TECMAXX”). The complete freedom from maintenance that has cost benefits that should not be underestimated in the commercial vehicle area is combined with new battery cover technology: a new type of labyrinth cover prevents the electrolyte from escaping. The use of central degassing instead of degassing via the plugs means that a frit can be installed, which prevents backfiring into the inside of the battery from external flames or sparks.

Completely maintenance-free batteries are used in almost all new vehicles nowadays. They make maintenance intervals longer and have proven themselves millions of times over.

AGM technology

AGM batteries (Absorbent Glass Mat, i.e. batteries with electrolyte bound in fiberglass mats) have proven themselves in cases where greater demands are made of the vehicle battery. These batteries differ from batteries with free electrolyte by the fact that the electrolyte is bound in a fiberglass mat that is located between the positive and negative plates instead of separators.

The battery is separated from the environment by valves (airtight). By means of internal circulation inside the battery, the oxygen that occurs at the positive electrode due to gassing is used up again, the amount of hydrogen that is produced is suppressed and the amount of water loss is therefore minimized. This circulation is made possible because of the formation of small gas channels between the positive and negative plates, through which the oxygen is transported. The valves only open in the event of considerable overpressure. The sealed AGM battery therefore has extremely low water loss and is therefore completely maintenance-free.

This technology also has other benefits. The mat is flexible, meaning that the plate set can be installed under pressure. Pressing the mat onto the plates reduces the shedding and separation effect of the active material considerably. This produces a load throughput that is usually up to three times greater than that of comparable starter batteries. This battery type also has the advantage that even if the battery case is destroyed as a result of an accident, for example, electrolyte does not usually escape, because it is bound in the fiberglass mat. No electrolyte escapes from the battery, even if it is turned 180° for long periods. High cold start currents are achieved because of the high porosity of the fiberglass mat.

Another advantage of the AGM battery is that electrolyte stratification is prevented. When a battery with free electrolyte is be-

ing cyclically charged and discharged, an electrolyte density gradient gradually builds up from top to bottom. This is because when the battery is charged, electrolyte with a higher concentration occurs at the plates, drops to the bottom because of its greater specific density and collects there, whereas the electrolyte with lower concentration remains in the upper part of the battery cell. Among other things, this electrolyte stratification reduces both the battery capacity and the service life. The electrolyte stratification effect occurs to varying extents in all batteries with free electrolyte. In AGM batteries, however, electrolyte stratification is prevented by absorbing the electrolyte in fiberglass mats.

When an installation location is being chosen for an AGM battery it must be ensured that high temperatures are avoided, since the thermal capacity is lower than it is for a battery with free electrolyte.

Gel-electrolyte batteries

Another type of sealed and therefore completely maintenance-free battery uses a multi-component gel in which the electrolyte is incorporated, instead of fiberglass mats. Here too, the internal gas circuit prevents gassing and therefore water consumption. This means that this battery is also completely maintenance-free.

The battery cell sealing plugs incorporate a safety valve that releases pressure in the event of continuous overcharging. These gel-electrolyte batteries only have a self-discharge rate of 2 % per month at 20 °C.

The short, thick plates and the gel electrolyte also guarantee high deep-cycle resistance. It is also completely sealed and completely spill-proof. This means that the electrolyte cannot escape, even if the battery is rotated 180°.

Completely maintenance-free batteries for motorcycles

Leak-proof batteries with mat technology are becoming increasingly popular. These batteries do not use gel technology. The battery is filled at the required time from an electrolyte bottle which is included in the sale. After initial filling, the electrolyte is bound in mat-like separators. The gas circulation inside the battery prevents gassing. If the battery is overcharged with high charge voltages over along period, gassing may occur in spite of the mat technology. In this case the gas escapes via a safety valve. After initial filling, the battery is sealed so that electrolyte cannot escape, even if the battery is tilted or rotated 180° for short periods.

Batteries for special applications

It is impossible to completely cover the wide range of different operating conditions which can be encountered in the field with one single standard battery. Such a battery would be far too large for normal operation and far too expensive.

Batteries with additional starting power are required for extremely low temperatures in cold countries. The starting temperatures are often below -20 °C. These batteries feature an increased number of thinner plates and separators.

The situation in tropical climates is just the opposite, since there is a risk of electrolyte "thickening" because of increased water consumption (electrolysis and evaporation). In temperate and cold zones, there is no need to change the electrolyte's density from its fully-charged freezing limit of -68 °C. In tropical regions, however, the electrolyte density must be reduced.

In the industrial and commercial sectors (bus, taxi, ambulance, delivery van, etc.), the fact that the vehicle is repeatedly driven only short distances means that the current taken from the battery is correspondingly higher. In addition, there are further cyclical loadings during high power demands when the vehicle is sta-

tionary. These are caused by the air-conditioning system, the lighting system, fans, electrohydraulically operated platform lifts, cooling units, auxiliary heaters etc.

As well as cyclical loading, batteries in all-terrain vehicles, commercial vehicles, construction machinery, towing vehicles and agricultural/forestry vehicles must be able to withstand severe vibration and impacts on slopes, construction sites or off-road conditions.

Deep-cycle resistant battery

Due to their design characteristics, starter batteries are poorly suited for applications in which frequent, extreme discharges occur (so-called cyclic loading), since high rates of wear occur at the positive plates due to the separation and "shedding" (sedimentation) of the active material. In the commercial area (e.g. in commercial vehicles), the battery is subjected to extremely high loading because of the short distances traveled between engine starts, and the associated high power consumption. Because of continuous power consumption, the battery can become extremely discharged and may not be adequately recharged by the alternator. Additional loading can also be caused by high power consumption whilst the vehicle is stationary by the fan, the air conditioning system, the auxiliary heater, the lighting system, the car radio, radio equipment, etc. In these cases, it is often advisable to use an AGM battery. If an AGM battery cannot be used, a deep-cycle resistant starter battery with free electrolyte is an alternative. It can be "deep"-discharged more often than a normal battery without impairing its service life.

In the deep-cycle resistant starter battery, separators with additional fiberglass mats are used to give the positive material the extra support needed to prevent premature sludge formation. The service life of such a battery, measured in charge/discharge cycles, is approximately double that of a standard battery.

Vibration-proof battery

In the vibration-proof battery, an anchor of cast resin and/or plastic prevents the plate blocks from moving inside the battery case. According to specified standards, this type of battery must withstand 20 hours of sinusoidal vibration at a frequency of 22 Hz and a maximum acceleration of 6 g. Requirements are therefore about 10 times greater than for standard batteries. Vibration-proof batteries are installed primarily on commercial vehicles, construction machinery and tractors used on building sites, on ski slopes, and for off-road applications in agriculture, and forestry.

Heavy-duty battery

Commercial vehicles have a considerable number of additional electrical consumers such as a lift platform, radio equipment, a television or a coffee machine. The dry-charged (i.e. the battery is ready for use after filling with electrolyte) heavy duty battery (HD battery) is maintenance-free in accordance with EN and includes a combination of measures for making the battery deep-cycle resistant and vibration-proof. It ensures a reliable power supply even under conditions of high continuous loading caused by a large number of electrical consumers. This type of battery is installed in commercial vehicles which are subject to high levels of vibration and cyclical loading.

The HD-Extra battery features even more far-ranging characteristics. These are needed in order to withstand unusually high levels of loading:

- ▶ Extreme cold-start reliability (up to 20 % higher starting reserve)
- ▶ Extremely long service life
- ▶ Extremely vibration-proof (100 % more than EN)
- ▶ Extremely deep-cycle resistant (four times greater than standard batteries)

Battery for extended current output

This battery shares the basic design of the deep-cycle resistant version. However, it has thicker plates, though there are less of them. No low-temperature test current is specified for these batteries, since they are not suitable for starting. Their starting power is considerably lower (by approximately 35 to 40 %) than starter batteries of the same size.

These batteries are used in cases of heavy cyclic loading, sometimes even for traction purposes (drive battery). One example of this type of use is in forklift trucks, which do not need any starting power but have to be recharged at frequent intervals. These batteries are also used to provide the drive energy for small drive units (e.g. in street sweepers and wheelchairs), and the energy for signaling systems, construction-site lighting, boats, ancillaries and for leisure time and hobby applications.

The proportion of antimony makes Bosch drive and lighting batteries with liquid electrolyte particularly deep-cycle resistant. The negative effects of the antimony are limited to a tolerable level.

Battery characteristics

The European standard EN 50 342 and national standards define the characteristics and test procedures for starter batteries. Although these tests are suitable for evaluating and monitoring the quality of new starter batteries, they do not claim to fully represent the wide range of load demands which can be encountered in the field.

One of the characteristics of a chemical battery is that the amount of current that can be drained (capacity) depends on the magnitude of the discharge current I_E . This means that the higher the current that is drained, the smaller the available capacity at the defined cutoff voltage. In order to be able make a comparison between starter batteries at all, their capacity is related to the strength of discharge current which the battery can deliver during a discharge time of 20 hours at a defined cutoff voltage (10.5 V) (nominal capacity K_{20}).

Cell voltage

The cell voltage U_z is the difference between the potentials which are generated between the positive and negative plates in the electrolyte. These potentials depend on the plate materials and on the electrolyte and its concentration. The cell voltage is not a non-variable figure but depends on

the state of charge (electrolyte density) and the electrolyte temperature.

Nominal voltage

The nominal voltage U_N of a cell for lead storage batteries has been defined by standards (DIN 40 729) to a value of 2 V. The nominal voltage of the complete battery results from multiplying the individual cell nominal voltages by the number of cells connected in series. According to the EN 50 342 the nominal voltage for starter batteries is 12 V.

Open-circuit and steady-state voltage

The open-circuit voltage is the voltage across the unloaded battery. After completion of the charging and discharging processes, due to polarization and diffusion, it changes to become a final value which is referred to as the steady-state voltage U_0 (Fig. 16). The steady-state voltage is the actual voltage obtained by multiplying the number of series-connected cells by the cell voltage U_{Z0} . For six cells, the following applies:

$$U_0 = U_{Z01} + U_{Z02} + \dots + U_{Z06} \approx 6 \cdot U_{Z0}$$

Similar to the cell voltage, the steady-state voltage is also dependent upon the state of charge and the electrolyte temperature. No conclusions can be drawn about the state of charge from a voltage that has been measured immediately after charging or dis-

Fig. 16 Steady-state voltage of a battery

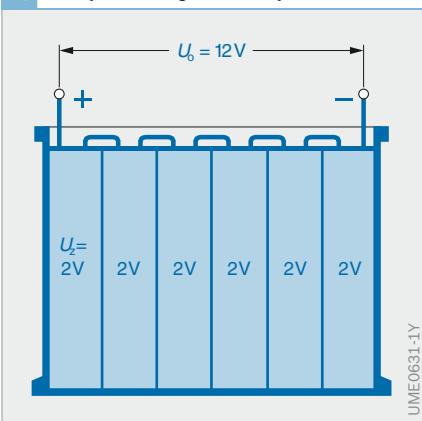
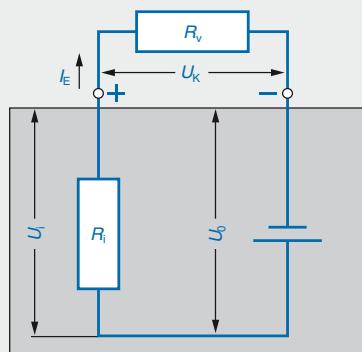


Fig. 16
 U_z Cell voltage
 U_0 Steady-state voltage

- I_E Discharge current
- R_i Internal resistance
- R_v Consumer resistance
- U_0 Steady-state voltage
- U_k Terminal voltage
- U_i Voltage drop across internal resistance

Fig. 17 Battery voltages



UME0633-2Y

charging procedures. It may take several days for a steady-state voltage to arise which can then be used to evaluate the state of charge. The electrolyte density is a more suitable indicator of the battery charge level.

Internal resistance R_i

The internal resistance R_i of a cell is made up of a number of individual resistances. Essentially these are the contact resistance R_{i1} between the electrodes and the electrolyte (polarization resistance). Then there is the resistance R_{i2} of the electrodes (plates with separator) to the electrode flow and the resistance R_{i3} of the electrolyte to the ion current. And when a number of cells are connected in series, the resistances of the individual cell connectors R_{i4} must be added.

This results in $R_i = R_{i1} + R_{i2} + R_{i3} + R_{i4}$.

As the number of plates increases (greater surface area), the internal resistance of the cell decreases. In other words, the greater the capacity of a cell, the smaller the internal resistance (with the same plate thickness). On the other hand, the more the battery discharges, and at low temperatures (sulfuric acid becomes more viscous), the internal resistance increases again R_i .

The internal resistance of a 12 volt starter battery comprises the series connection of the internal resistances of the individual cells, together with the resistances presented by the internal connecting elements (cell connectors and plate connectors). The resistance for a fully charged 50-Ah battery at 20 °C is 5 to 10 mΩ; with a charge status of 50 % and at -25 °C, it increases to about 25 mΩ. It is a characteristic variable for the starting response. The internal resistance of the battery in combination with other sources of resistance in the starter-motor circuit determines the speed at which the engine is turned over when being started.

Terminal voltage U_K

The terminal voltage U_K is the voltage measured between the two terminal posts of a

battery. It is dependent on the open-circuit voltage and the voltage drop U_i across the battery's internal resistance R_i (Fig. 17):

$$U_K = U_0 - U_i \text{ with } U_i = I_E \cdot R_i$$

If a discharge current I_E is taken from a battery by a consumer with a load resistance of R_L , a lower terminal voltage is measured than in the no-load state. This is due to the battery's internal resistance. If a current I_E flows through the cells, a voltage drop U_i takes place across R_i which increases together with the current. Since internal resistance is a function of temperature and the state of charge (among other things), this means that a loaded battery's terminal voltage drops at low temperatures and when it is inadequately charged.

Information about the state of charge and degree of wear of a battery that is under load can be derived by measuring its terminal voltage.

Gassing voltage

The gassing voltage according to DIN 40729 is the charge voltage at which a battery starts to gas significantly. This will result in water loss from the battery and the risk of gas detonation. According to DIN VDE 0510, a gassing voltage of 2.40 to 2.45 volts per cell applies as a rough limit depending on the particular battery design. For 12 volt batteries, this voltage limit is therefore between 14.4 and 14.7 volts, depending on the electrolyte temperature. In order to reduce water loss when driving and also ensure that rapid recharging is possible, temperature-dependent controller characteristic curves must be used. For completely maintenance-free batteries these stipulate a maximum value of 16 V at temperatures that are significantly lower than 0 °C and about 13.5 V at temperatures that are significantly above 30 °C. This takes into consideration the fact that the charging capability of the lead storage battery is inhibited at low temperatures, which means that the charge voltage has to be increased in this case. The charge voltage must be set to a lower level at high tempera-

tures in order to restrict water loss and battery corrosion. The latter has a greater effect at high temperatures and high voltages.

A charge voltage of 14.1 V (2.35 V/cell) for a maximum charging time of 48 hours is specified for maintenance-free, sealed gel batteries.

Capacity

Available capacity K

The capacity K is the quantity of electricity which the battery can deliver under specified conditions. It is the product of current intensity and time (ampere hours Ah). The quantity of active material that is used and the quantity of electrolyte essentially determines the capacity of the battery. For high powers (such as are needed when starting an IC engine), the active material must have large internal and external surfaces (large number of large-area plates). The large internal surface area is produced during electrochemical pretreatment known as "forming". However, the battery's capacity is not a fixed parameter, but depends, among other things, on the following influencing variables (Figs. 18 and 19):

- ▶ Discharge current strength
- ▶ Density and temperature of the electrolyte
- ▶ Discharging process as a function of time (capacity is higher when a pause is made during discharge, than when discharge is continuous)
- ▶ Battery age (due to the loss of active material from the plates, capacity decreases as the battery approaches the end of its service life) and
- ▶ Degree of battery electrolyte stratification

The discharge current strength is extremely important. The greater the discharge current strength, the smaller the available capacity. In the example shown in Fig. 19, a discharge current of 2.2 amps means that the available 44 Ah capacity can be used for up to 20 hrs. With a mean starter current of 150 amps at a temperature of 20 °C and a

discharge time of approximately 8 minutes, the available capacity drops to approximately 20 Ah. The reason for this is that with a small discharge current, the electrochemical processes slowly penetrate deep into the pores of the plates and the electrolyte on the outside of the plates can also be used (approx. 50 %), whereby the conversion mainly takes place on the surface of the plates using the electrolyte that is in the pores when discharging takes place at a higher current.

Effect of temperature on capacity

The capacity and the discharge voltage of a battery increase together with the temperature because of, among other things, the lower viscosity of the electrolyte and the lower internal resistance that this causes. However, as the temperature drops, they also reduce because the chemical processes are then less efficient.

Therefore, when selecting a starter battery, the capacity should not be too low. Otherwise, with a wrongly dimensioned battery, there is the danger that at very low temperatures, an internal-combustion engine will not be cranked fast enough and long enough to start it. This fact is demonstrated in Figure 20:

Curve 1a shows the starter speeds as a function of the temperature with the battery

18 Available capacity as a function of temperature and discharge current

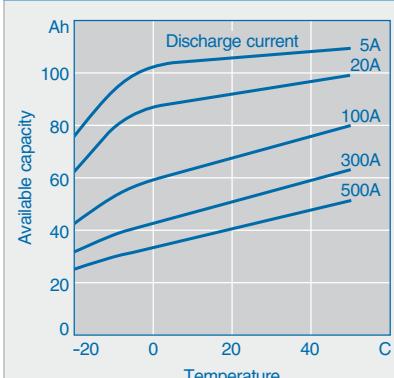


Fig. 18

Battery: 12 V, 100 A·h
(in relation to discharge time of 20 hrs and 100 % state of charge)

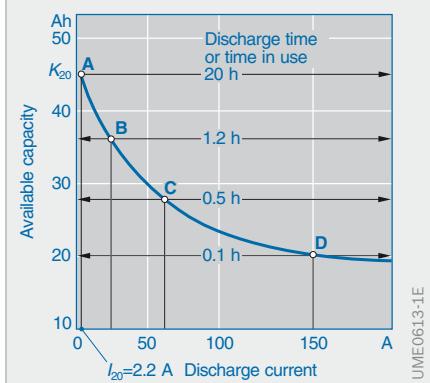
20 % discharged (curve 1b is for a heavily discharged battery). Curve 2 shows the minimum starting speed required by an internal-combustion engine. This engine speed is relatively high at low temperatures because of the high frictional resistance in the vehicle engine and the transmission (due to greater lube oil viscosity, for example).

The intersection S_1 of curves 1a and 2 shows the cold start limit (limit temperature) of the 20 % discharged battery. This means that starting is no longer possible at lower temperatures or with a lower battery charge because the output that can be produced by the battery and the starter is less than the starting power that is required by the internal-combustion engine. The more the battery is discharged, the more the cold-start limit (intersection S_2) is shifted to higher temperatures.

Nominal capacity K_{20}

The nominal capacity K_{20} is the battery's rated electrical charge in ampere hours (Ah). According to EN 50 342, this is the electrical charge that can be taken from the battery within 20 h at a fixed discharge current of I_{20} until the specified cutoff voltage of 10.5 V at $(25 \pm 2)^\circ\text{C}$ is reached. The discharge current I_{20} is the current that is allocated to the nominal capacity, and must be delivered by the battery

19 Available battery capacity as a function of discharge current strength (battery: 12 V 44 Ah)



during the total discharge period:

$$I_{20} = K_{20}/20 \text{ h}$$

The nominal capacity is a measure for the energy which can be stored by the battery in the as-new condition. It depends on the quantity of active material and upon the electrolyte's specific gravity. For instance, a new 44 Ah battery can be discharged at a current of 2.2 A for at least 20 hours ($44 \text{ Ah}/20 \text{ h} = 2.2 \text{ A}$) until the cut-off voltage of 10.5 V is reached. The nominal capacity must be taken into consideration in the design of the continuous consumers in the vehicle electrical system (Fig. 19).

20 Influence of temperature on starter speed and the engine's minimum start speed

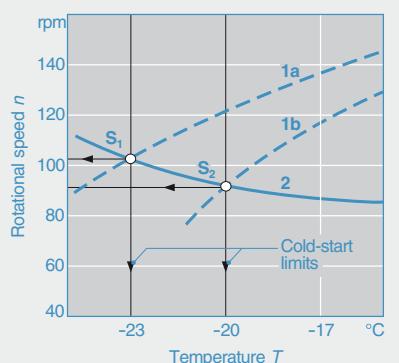


Fig. 19

- A 20-hour discharging
- B Ignition and lighting
- C In addition, fan, windshield and rear-window heating, fog lamps, wipers and radio
- D Mean starter current

21 Discharge of a 12 V battery with a low-temperature test current I_{CC} at -18°C and at 27°C

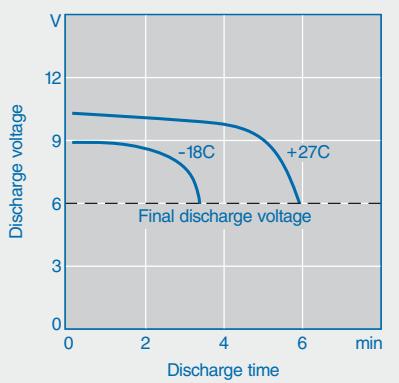


Fig. 20

- Example:
- 1a Starter speed
Battery discharged by 20 %
 - 1b Starter speed
Battery heavily discharged
 - 2 Minimum engine start speed
 - S_1, S_2 Cold-start limit

Low-temperature test current I_{CC}

The low-temperature test current I_{CC} (previously I_{KP}) indicates the battery's current output capability at low temperatures. According to EN 50 342, the battery terminal voltage at I_{CC} and -18 °C and 10 seconds after discharging has started, must be at least 7.5 V (1.25 V per cell). Other details relating to discharge period are specified in the standard. The determining factors of the short-term behavior indicated by I_{CC} are the number of plates, the plate surface area, the plate spacing and the separator material.

A variable that indicates the starting response is the internal resistance R_i . For -18 °C and a full 12-V battery the following applies: $R_i = 4,000/I_{CC}$, whereby I_{CC} in amps must be used. The internal resistance R_i is measured in the units mΩ.

The internal resistance of the battery in combination with other sources of resistance in the starter-motor circuit determines the speed at which the engine is turned over when being started. However, the low-temperature test current is defined according to a variety of different testing conditions in different countries so that it is not always possible to make a direct comparison of the cited figures.

For an automotive battery which must provide the electrical energy for the starter, startability at low temperatures is usually more important than the capacity. Since it refers to low-temperature current delivery, the low-temperature test current is therefore a measure for startability. It is extremely dependent on the total surface area (number of plates and plate area) of the active material. The bigger the contact surface there is between the lead material and the electrolyte, the greater the amount of current that can be drained in short bursts. Plate spacing and separator materials are important influencing variable which affect the rapidity of the chemical processes in the electrolyte and which also determine the low-temperature test current.

Type designations

The design specifications and designations of the various types of starter battery are defined by standards in order to make the products of different manufacturers interchangeable (compatibility). Bosch batteries are generally labeled with the following information:

- ▶ Characteristics according to EN standard
- ▶ European type number (ETN) and general safety instructions for handling batteries
- ▶ Bosch part number (TTNR number)
- ▶ Customer search number KSN (specially for Bosch Silver)

Characteristic quantities

The battery characteristics defined by European Standard EN 50 342 characterize the specifications/properties of a starter battery. The most important starter battery rating parameters are

- ▶ Nominal voltage (e.g. 12 Volts),
- ▶ Nominal capacity (e.g. 44 Ah) and
- ▶ Low-temperature test current (e.g. 360 amps).

In the USA, a code such as an SAE (Society of Automotive Engineers) number, and in Japan, a JIS (Japanese Industrial Standard) number is used.

European type number (ETN number)

The European type number (ETN) replaced the DIN number in Germany in 1998. It indicates the voltage, capacity and low-temperature test current of the battery concerned.

Example: 5 44 059 036

Code number for battery type:

The first digit of the ETN is a code number for the battery voltage (in the above example, "5" for 12 V). The various code numbers used indicate the following:

- | | |
|---------|--------------------------|
| 1 to 4: | 6-Volt batteries |
| 5 to 7: | 12-Volt batteries |
| 8: | Specialized batteries |
| 9: | Small traction batteries |

Code number for battery capacity:

The second and third digits of the ETN are code numbers for the battery capacity (20 hours) in Ah (44 for 44 Ah in the above example). For capacities over 100 Ah, the first digit is increased by 1 for each 100 Ah (in the range 5 to 7).

Sequence number

The fourth, fifth and sixth digits of the ETN form a serial number (in the above example 059). Various items of information about the battery can be read off on the basis of this number using a list (e.g. vibration resistance).

Code number for low-temperature test current:

The seventh, eighth and ninth digits of the ETN form a code number for the low-temperature test current to EN. The figure represents one tenth of the current (in the above example, 036 for 360 A).

Bosch part number (TTNR number)

The alphanumeric Bosch part number (TTNR number) consists of a Bosch combination number for the battery type and a Bosch-coded ETN.

Example: 0 093 S 544 1N

Code number for battery type:

The second and third digits of the TTNR number indicate whether the battery contains antimony or not. 09 corresponds to antimony-free batteries and 18 corresponds to batteries with antimony.

Code number for voltage and capacity:

Digit 6 of the TTNR specifies the battery voltage, and digits 7 and 8 specify the capacity. Digits 1 to 3 have the same coding as the ETN.

Customer search number (KSN number)

For Bosch Silver batteries, there is a customer search number (KSN) which makes it easier for customers to identify the correct battery for a particular vehicle model.

Practical and laboratory battery testing

Various laboratory endurance tests are described in EN 50 342. The standard also describes charge acceptance testing, water consumption testing and vibration resistance testing. Car manufacturers also frequently require other testing to be performed. The charging and discharging cycles of the battery are simulated at extreme temperatures, followed by an engine start. One example is the J240 test, which is oriented to the American SAE standard. It tests the service life of a battery at high temperatures (75 °C).

Useful life

Various claimed battery characteristics can be tested in the laboratory. However, the interaction of the battery characteristics for achieving optimum usage in the vehicle can only be tested in the field. Batteries are therefore tested in driving mode. One example is the taxi test in Las Vegas. Special demands are made of the battery in Las Vegas. High temperatures occur because of the climate, and the battery is also subjected to high cyclic loading due to the nature of taxi driving. The usage time of lead-calcium batteries is 1.4 times longer than conventional batteries, and 3 times longer for lead-calcium-silver alloy.

Self-discharge

The self-discharge of the positive and negative plates is inherent in the principle of the lead storage battery. The effects of temperature and other factors lead to the battery becoming electrically "empty" after a given period of time. Considering conventional starter batteries, antimony poisoning leads to an increase of the self-discharge reaction at the negative plates, and the rate of discharge climbs considerably along with increasing battery age. In the field, this means that after a six-month storage period at room temperature, new, conventional-type starter batteries only

have a state of charge of approximately 65 %. This corresponds to an electrolyte density of 1.20 kg/l. Under certain circumstances, old batteries drop to this value within a few weeks. The state of charge of maintenance-free starter batteries after six months is 90 %. The corresponding electrolyte density is 1.26 kg/l. It takes 18 months for the state of charge to reach 65 % (electrolyte density $\rho = 1.20 \text{ kg/l}$) (Fig. 22).

Because the lead-calcium grids have a purer alloy system, completely maintenance-free batteries discharge at a considerably slower rate. The low self-discharge rate of the positive plate and negative plate remains constant during the entire usage period. The self-discharge phenomenon is of particular significance for vehicles used only during specific seasons (agriculture, forestry, and construction sites). It is also important, however, for second cars and mobile homes which are not driven in winter, or only rarely. It also applies to new vehicles which, since there is no stop in production, are manufactured but have to be parked for long periods due to seasonal fluctuations in sales trends or the transportation time to the salesroom (export vehicles).

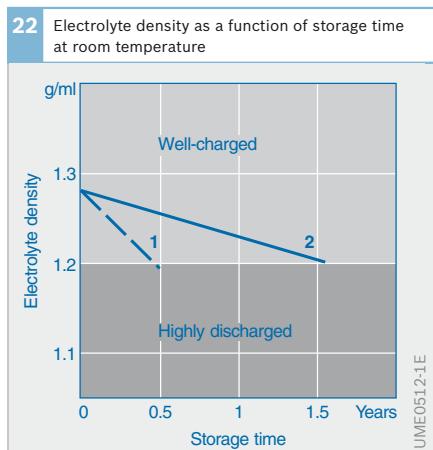
Starting power

Completely maintenance-free starter batteries with lead-calcium-silver technology have 30 % more starting power than conventional batteries. Basically speaking, this is due to the envelope-type separators with their low specific resistance, and to the increase in plate surface due to the omission of the sediment chamber.

Furthermore, thanks to the lead-calcium alloy used for the plate grids, the maintenance-free battery's starting power remains practically unchanged for years. It does not drop to below the setpoint value for new batteries until towards the end of its useful life. Whereas the completely maintenance-free starter battery is still above the setpoint value after 75 % of its useful life, the conventional battery drops far below this setpoint value considerably sooner (at approx. 40 % of useful life). In practice, it has already lost about a third of its original starting power after only 75 % of its useful life (Fig. 23).

Current draw

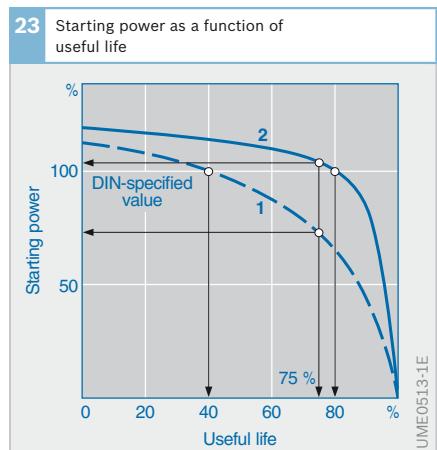
Low-antimony and antimony-free batteries behave almost identically during the EN 50 342 current draw test. Completely maintenance-free starter batteries have major advantages when charging with regulators that take the battery temperature

**Fig. 22**

- 1 Conventional starter battery (PbSb)
- 2 Maintenance-free starter battery (PbCa)

Fig. 23

- 1 Conventional starter battery (PbSb)
- 2 Maintenance-free starter battery (PbCa)



into consideration and can exceed the voltage of 14.5 V (higher gassing voltage, less water loss). In practice, the differences between the most commonly used regulator characteristic curves (see section “Battery/gassing voltage characteristics”) are small, apart from the better current draw of maintenance-free batteries with lead-calcium alloy with a state of charge of less than 50 %. This is due to the phenomenon applying to every battery, namely the fact that the lower the temperature, the higher the charging voltage required for a given state of charge. Since its gassing curve is higher during charging, the completely maintenance-free battery stores the increased charging current without losses due to gassing. In other words it achieves a higher state of charge and therefore provides more starting power.

Resistance to overcharge

Overcharge is a decisive factor with regard to battery life. It applies for instance to vehicles which have very high annual mileages, as well as to courier and agricultural vehicles, construction-site machinery, and long-distance haulage trucks. In such cases, the battery is fully charged, the engine turns at high speed, and the alternator only supplies a few consumers. Under these circumstances, the charging current leads to overcharging, corrosion, and loosening of the active material. In laboratory tests, when these conditions are simulated at an electrolyte temperature of 40 °C and a charge voltage of 14 or 16 V, the maintenance-free starter battery has a considerably longer service life than a battery containing antimony.

Resistance to exhaustive discharge

To check the battery's resistance to exhaustive discharge, lamps are used to discharge it completely, after which it is left 4 weeks in the short-circuit condition. The battery must then be recharged under vehicle electrical system conditions. It must still be operational and it must only display certain specified reductions in perfor-

mance. For example, a loss of battery capacity that is smaller than a specified limit.

Water consumption

When they are new, starter batteries with and without antimony show a far lower water consumption than the maximum 6 g/Ah as stipulated by DIN. As a rule, the lead-calcium battery has a long-term figure of less than 1 g/Ah.

The fact that the gassing voltage remains at its high initial level throughout the battery's complete useful life is responsible for the water-consumption figures for completely maintenance-free starter batteries being so favorable. High gassing voltage leads to minimum water decomposition.

This confines the electrolyte checks to

- ▶ Every 15 months or 25,000 km for all low-maintenance batteries, and
- ▶ Every 25 months or 40,000 km for completely maintenance-free batteries (as per EN)

Summary

Completely maintenance-free batteries have the following benefits.

- ▶ The charge voltage only exceeds the gassing voltage at high temperatures. This means that gassing rarely occurs, meaning that topping up with distilled water is unnecessary throughout the battery's useful life.
- ▶ Maintenance errors such as forgetting to top up with distilled water or filling with contaminated water cannot occur.
- ▶ The risk of skin contact with sulfuric acid and the damage that this causes is avoided.
- ▶ Longer service life and durability.
- ▶ Consistently high starting power.
- ▶ Better short-distance driving performance.
- ▶ Better performance capability in all temperature ranges.
- ▶ Cheaper maintenance and care.
- ▶ Possibility of locating the battery in areas that are difficult to access in the vehicle.

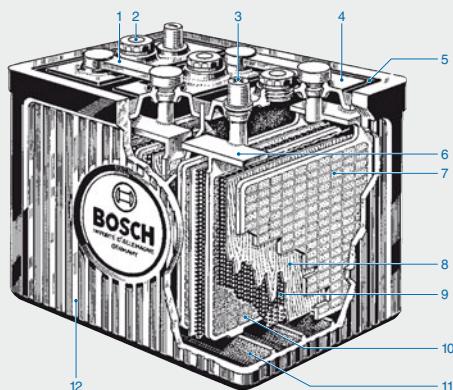
► History of the battery

When considering the history of the battery, a number of renowned scientists and inventors deserve particular credit. Above all, such personalities as Luigi Galvani (1789), Alessandro Graf Volta (around 1800), Johan Ritter (around 1800), Gaston Planté (1859) and Camille Faure were decisive in driving on the development of the battery.

At the end of the 19th century grid plates were already being manufactured. Their basic principles are still found in today's lead batteries. In other words, basically speaking the lead battery has hardly changed up to the present day: It still contains cells, grid plates, and sulfuric acid. If we examine the battery more closely, we find that not only has the energy density increased immensely, but also that the previously used materials (in some cases, separators and the housing were made of wood) have been superseded to a great extent by plastic, and complete freedom from maintenance is today standard practice for starter batteries. In exceptional cases, these can achieve a service life comparable to that of the vehicle itself.

▼ Starter battery: 1951 version

- 1 Joining bar
- 2 Sealing plug
- 3 Terminal post
- 4 Cell cover
- 5 Sealing compound
- 6 Plate strap
- 7 Negative plate
- 8 Wooden separator
- 9 Hard-rubber separator
- 10 Positive plate
- 11 Cell connector
- 12 Battery case



The history of batteries in figures

- 1905: The first batteries were fitted in a motor vehicle (at first only for lighting purposes)
- 1914: The first starter battery was fitted in a motor vehicle
- 1922: The first Bosch motorcycle batteries were installed, and 4 years later the first Bosch battery charger was introduced
- From 1927 onwards, Bosch also developed automotive starter batteries, and mass production started nine years later

After the Second World War, the development of Bosch automotive batteries was marked by the

- Introduction of plastics in battery manufacture (e.g. polystyrene in 1955; polypropylene in 1971)
- The improvement of individual battery components (e.g. the folded-rib separator in 1956; the one-piece battery cover for 6-V batteries in 1964 and for 12-V batteries in 1966; direct cell connectors in 1971; expanded-metal technology for the negative grid in 1985) and
- The production of special-type batteries (e.g. "deep-cycle resistant" in 1969; "low-maintenance" in 1979; "vibration-proof" in 1980; "maintenance-free" in 1982, and "completely maintenance-free" in 1988)

Battery maintenance

Maintenance and care of batteries

Batteries are usually filled in the as-delivered condition, i.e. initial filling with electrolyte is no longer required. Filling may only be needed for certain battery types such as motor cycle batteries and commercial vehicle batteries. In these cases, the operating instructions must be followed.

In many cases, maintenance is not longer needed, and may even no longer be possible, e.g. for maintenance-free batteries such as the Bosch Silver. These batteries lose so little water that topping up with distilled water is not required.

Batteries that do not require maintenance can be detected by the fact that the battery plugs are no longer accessible.

Electrolyte density and state of charge

The electrolyte density is the major indicator for the battery's state of charge. Table 1 shows a number of density figures and the associated freezing thresholds at different states of charge.

Electrolyte density and operating temperature

High temperatures have the effect of accelerating chemical processes in the battery. However, the increase in temperature not only increases the capacity and starting system power, it also has a greater negative effect on the plates (active material falls away and grid corrosion occurs). The self-discharge rate is also accelerated.

Electrolyte density and freezing point

The more the battery is discharged, the more the electrolyte is diluted. This leads to the freezing point rising to more unfavorable temperatures. The electrolyte in a fully charged battery with a specific density of 1.28 kg/l has a freezing point of -60 to -68 °C. However, a discharged battery with a specific density of 1.04 kg/l has a freezing point of -3 to -11 °C and can freeze when outside temperatures are low (see Table 1).

A battery with frozen electrolyte can only deliver a very low current and cannot be used for starting. A polypropylene battery case remains stable even with frozen electrolyte. Since the fluid does not fully crystallize out, it is unlikely that the housing will fracture. A frozen battery is not to be charged because the frozen electrolyte then starts to swell. The battery must first thaw out before it can be recharged.

Measuring the electrolyte density

The electrolyte density is seldom measured in workshops because maintenance-free batteries are used in the majority of vehicles. A battery syringe is occasionally used to check the electrolyte density of conventional batteries. This is used to draw electrolyte up into the glass tube. The hydrometer (a float with a scale) measures the density of the electrolyte and the measurement can be read off the scale.

Modern acid refractometers use the refraction of the electrolyte to determine the density. Just a small drop of electrolyte is required to perform this measurement. This device can also be used to check the frost protection of the coolant and the washing system water.

On conventional batteries, the electrolyte level should be regularly checked, and if necessary the battery topped-up with distilled water to the "Max" mark. Before the cold season starts, the state of charge should be checked by measuring the electrolyte density. If this is below 1.20 kg/l, the battery should be recharged.

Table 1

- ¹⁾ At 20 °C:
the electrolyte density drops by approximately 0.01kg/l for every 14K that the temperature rises and vice-versa when the temperature drops
- ²⁾ Lower value: high electrolyte utilization; higher value: low electrolyte utilization

1 Electrolyte values of the diluted sulfuric acid		
State of charge	Electrolyte density in kg/l ¹⁾	Freezing threshold in °C
Charged	1.28	-68
Semi-charged	1.16/1.20 ²⁾	-17 to -27
Discharged	1.04/1.12 ²⁾	-3 to -11

The electrolyte density of maintenance-free batteries cannot be measured, nor does the battery need to be topped up with distilled water. The Bosch Silver battery provides an indication of the electrolyte density and therefore the charge via the "Power Control System" (Fig. 1). If the indicator in the transparent window is green, the battery is adequately charged. If the indicator is dark, the electrolyte density and therefore the state of charge is too low and the battery needs recharging. If the window is light, the electrolyte has dropped below the minimum level and the battery must be replaced.

Battery storage

The following storage times are specified for new batteries on the aftermarket:

- Unfilled: unlimited
- Filled, conventional: 3 (max. 6) months
- Filled, completely maintenance-free:
18 months

If the battery is to be stored for longer periods, it must be regularly recharged following the normal charging procedure. Batteries must be stored in a cool and dry state and in a good state of charge. The older a battery gets, the less time it can be stored. As far as possible, the battery is to be trickle-charged with a very low current. If the battery is to remain in the vehicle

during the storage period, its ground cable is to be disconnected.

Battery charging

If it is impossible for the alternator to charge the battery adequately, a battery charger must be used. This applies when the battery has been out of use for a long time, or directly before it is removed from the vehicle and put into storage.

Charging methods

Normal charging

Generally speaking, normal charging takes place using a battery charge current of I_{10} , which corresponds to 10 % of the battery's nominal capacity:

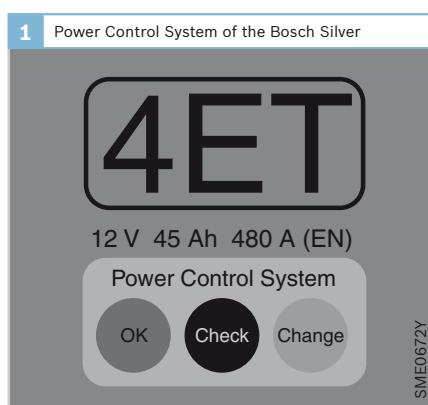
$$I_{10} = 0.1 \cdot K_{20} \cdot A/Ah.$$

Depending upon the process used, the charging time can be up to 14 hours.

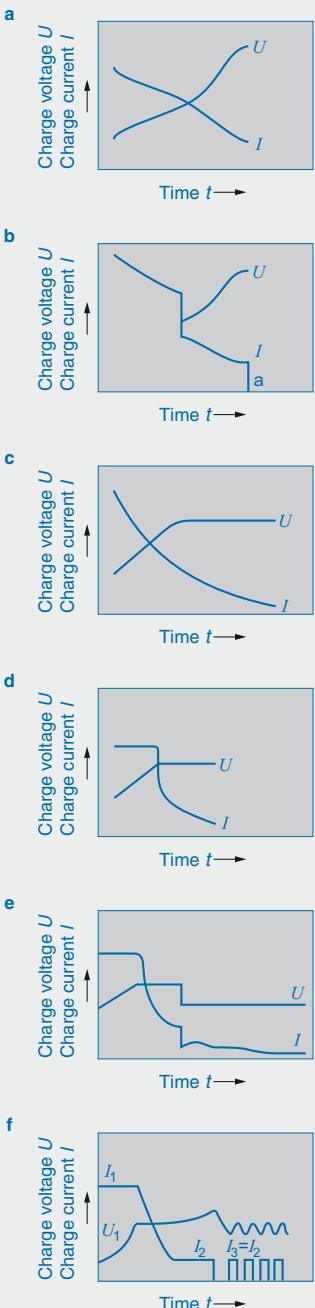
Boost charging

Boost charging can bring a discharged intact battery back up to about 80 % of its nominal capacity so that it can handle loadings which are typical in automotive applications. A high charging current can be used without problems below the gassing voltage - e.g. in the range of the numerical value for the nominal capacity (relative battery charge current $I_1 = K_{20} \cdot A/Ah$). Once the gassing voltage is reached, the boost charge must either be switched off or changed to normal charging.

The gassing voltage depends on the design of the battery, its age and the electrolyte temperature. If the charge voltage exceeds this value during charging, the battery starts to gas. This leads to water loss in the battery and the production of oxyhydrogen gas. Controlled chargers therefore restrict the charge to a typical 14.4 V (2.4 V/cell) with a cold battery and 13.8 V (2.3 V/cell) with a warm battery.



2 Charging characteristics



Trickle charging

To compensate for the self-discharge losses in stored batteries (for instance when caravan or mobile-home batteries are stored during the winter), the battery is left connected to a battery charger for an extended period of time, whereby charging current is limited to 1 mA/Ah.

Floating-mode operation

With floating-mode operation, the battery charger and the consumers are permanently connected to the battery. This means that energy is drawn from the battery by consumers during charging. The electronics of the charger prevent the battery from becoming overcharged.

Charging characteristics

There are various methods for charging the battery, each of which is characterized by its own charging characteristics (DIN 41 722):

- W Constant resistor (battery charge current drops when the charge voltage increases)
- U Constant charge voltage
- I Constant charging current
- a Automatic shut-off
- e Automatic restart
- o Automatic switch-over to other characteristic curve

The different characteristic curves can also be combined. For instance:

- W Like W characteristic curve but charge voltage remains constant above a given value (e.g. just below the gassing voltage)
- IU Constant charging current up to a value from which voltage remains constant and the charging current falls
- WoW Switch from one W characteristic curve to another

Fig. 2

- a W charging characteristic (normal charge)
- b WoWa charging characteristic
- c WU charging characteristic
- d IU charging characteristic (boost charge)
- e IUoU charging characteristic
- f $I_1U_1I_2a_1a_3I_3\dots$ charging characteristic

With the W charging characteristic (Fig. 2a) the charging current is determined by the charge circuit resistance and the driving voltage difference according to Ohm's law ($I = \Delta U/R$). Since the charge voltage slowly increases during charging, the driving voltage difference becomes smaller as does the battery charge current.

The W charging characteristic is easiest to achieve, i.e. it leads to cheaper chargers. However, the uncontrolled end of charging and the long charging time taken to fully charge the battery are a disadvantage. The battery charge current drops long before the gassing voltage has been reached.

Both disadvantages are avoided by using the IU charging characteristic (Fig. 2d). A constant high battery charge current I is maintained until the final charging voltage U is reached. This method permits a high level of charge in a short time whereby an overcharge is avoided.

With the IUoU charging characteristic (Fig. 2e) the system U (2.3 to 2.4 V per cell) is permanently switched to a lower voltage (2.23 V per cell) when the final charge voltage is reached (trickle charge).

Battery overcharging is also prevented on equipment with a limited charge voltage (WU characteristic curve, Fig. 2c) or which automatically switches to weaker W charging characteristics when a limit stress is reached (Fig. 2b) or stops charging altogether (Wa characteristic curve).

The $I_1U_1I_2aI_3aI_3\dots$ charging characteristic (Fig. 2f) starts like IU charging. As soon as the battery charge current drops below a certain limit in the U-phase, the system switches to recharging with I_2 . This is for a limited time and with a limited voltage. Batteries with defined electrolyte (mat or gel technology) are fully charged, and normal starter batteries with free electrolyte ("wet" batteries) undergo a defined gassing phase with electrolyte mixing. The final trickle charge ($I_3aI_3\dots$) charges at approx. 1A/100 Ah until an upper limit stress is reached, and then shuts off. As soon as the

battery voltage has reached a lower limit stress by means of self-discharging, the recharging current I_3 starts again.

Safety requirements

In order to avoid the risk of accidents the charger must have reliable potential separation between the 230V mains and the accessible charging terminals. Additional reverse-polarity protection prevents battery short-circuiting and the destruction of the battery charger if the battery terminals are wrongly connected.

Battery testers

The condition of installed starter batteries can also be checked with battery testers. They mainly measure and evaluate the high-current capability of the battery. The Bosch BAT121 battery tester also makes it possible to print out the result of the test on the built-in thermal printer.

The battery tester is connected to the battery via cables. The low-temperature test current of the battery is selected on the tester, then the test can be started. The digital display shows the following information when the test is complete:

- ▶ The available starting power as a percentage of the input value
- ▶ The battery voltage
- ▶ The evaluation "good" or "replace"
- ▶ A charge recommendation if a low state of charge has been detected

Battery chargers

Highly sensitive electronic components such as airbags, car telephones, car radios and electronic control units must be protected from voltage peaks when the battery is being charged. Previously, the battery had to be disconnected from the vehicle electrical system. However, when modern electronic chargers are being used, the battery can be charged with the consumers connected (floating-mode operation). Since charging is faster and there is no danger to the vehicle's electronic consumers, this is of course a considerable

help in the vehicle repair shop because:

- ▶ Expensive battery removal or disconnection is not required
- ▶ Stored data from the car radio, electronic control units, the telephone, the on-board computer etc. is retained
- ▶ Electrical consumers (airbags, control units etc.) are protected
- ▶ No damage caused by erroneous operation

Electronic charger

Bosch electronic chargers supply an output voltage that is free of damaging voltage peaks. Batteries can therefore be charged while they are installed in the vehicle. The equipment is protected from overcharging, overcurrents and polarity reversal.

The BML2415 electronic charger charges the battery with a WU charging characteristic. The battery charge current is infinitely adjustable. The charger is suitable for trickle charging and floating-mode operation. Discharged batteries are charged at a lower current to begin with, then the current is increased.

The BAT415 electronic charger is suitable for charging conventional batteries and batteries with fixed electrolyte (gel battery or mat battery). This is made possible by the microprocessor-controlled $I_1U_1I_2al_3al_3\dots$ charging characteristic. The currents I_1 , I_2 and I_3 are optimally adapted to the battery by entering the battery capacity in Ah.

Rapid-start charger

The BSL2470 rapid-start charger BSL2470 has enough power reserves to rapidly charge 12 V and 24 V batteries and provide mains-powered start assistance. Electrical components in the vehicle electrical system are protected from damage during charging and starting. The battery charge current is infinitely adjustable, even deeply discharged batteries can be charged.

The BSL2470 rapid-start charger charges using the WU charging characteristic, and

the SL24100E uses the WoWa charging characteristic.

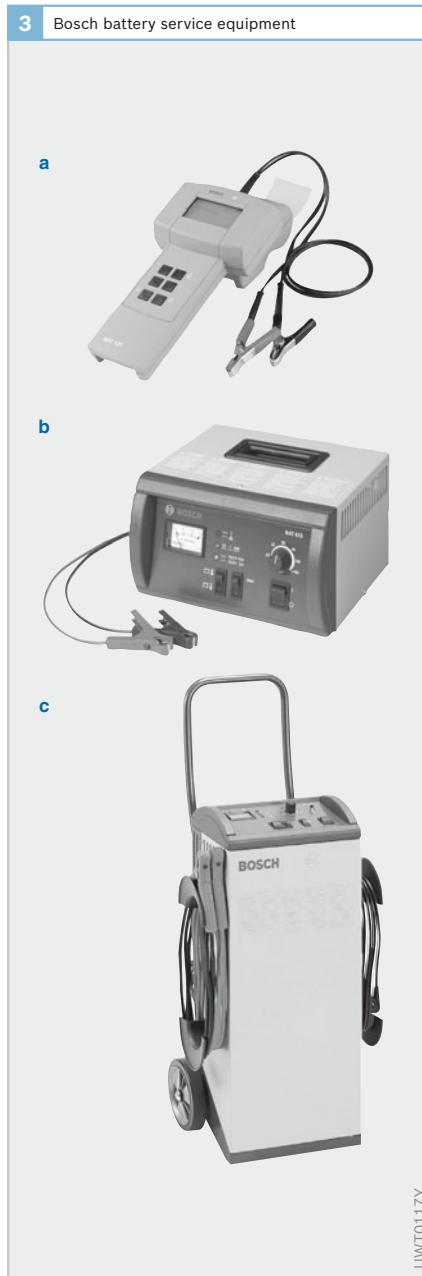


Fig. 3

- a BAT121 battery tester
- b BAT415 electronic charger
- c BSL2470 rapid-start charger

Important! Start assistance is only permitted on vehicles for which the manufacturer has not imposed limitations or bans on start assistance in the operating instructions.

Start assistance with jumper cables

The battery from another vehicle can be used to provide start assistance. This procedure should only be used if the batteries are installed in both vehicles and the manufacturer's instructions are followed.

For efficient starting assistance, only standardized jumper cables (DIN 72 553) with a conductor cross section of at least 16 mm² must be used for petrol engines and 25 mm² for diesel engines. Both of the batteries (or battery chargers) which are to be connected must have the same nominal voltage. The following operations are required:

- ▶ Determine the cause of battery discharge. If the fault is in the electrical system, do not use start assistance since this can damage the battery charger, or the battery and alternator of the vehicle providing the assistance.
- ▶ Connect the positive terminal of the discharged battery to the positive terminal of the external power source.
- ▶ Connect the negative terminal of the external power source to a bare metallic surface away from the battery (such as the engine ground strap) on the vehicle receiving assistance.
- ▶ Check that the battery jumper cables are firmly attached (good electrical contact).
- ▶ Start the vehicle with the sound battery. After a brief pause also start the disabled vehicle.
- ▶ Once start assistance is complete, disconnect the cable clamps in the reverse order.

Malfunctions

Faults in the battery

Malfunctions caused by internal battery faults (such as short-circuits due to separator wear or loss of active material, open

circuit in cell or plate connectors) are irreparable and the battery must be replaced. A cell short-circuit is recognizable from the battery voltage, which is too low by approx. 2 V. When an open circuit occurs in the cell connectors, low currents can often be taken from the battery and it can also be charged, but even when fully charged, the voltage collapses when an attempt is made to start the engine.

Faults in the vehicle electrical system

If no battery defect can be found but the battery remains continuously overcharged (high water consumption, battery voltage permanently above 14.5 V with engine running) or is extremely discharged (no starting power, low electrolyte density in all cells, battery voltage less than approx. 12.3 V, battery voltage continuously less than approx. 13.9 V with engine running), there is a fault in the vehicle electrical system. Reasons for faults may be defects or malfunctions in the following components:

- ▶ Alternator (permanent exhaustive discharge, starting no longer possible)
- ▶ V-belt (missing alternator drive)
- ▶ Regulator (fluctuation in lamp brightness when engine is accelerated, loss of battery water)
- ▶ Shutoff relay (consumers remain on when the engine has been switched off)
- ▶ Accessories (e.g. radio, clock, alarm system) consuming excessive no-load current

In many cases it may be worth checking the alternator voltage if a battery frequently shows a low state of charge.

Sulfation

If a battery is left discharged for a long period, the fine crystalline lead sulfate that occurs when the battery is discharged may turn into large crystals under unfavorable circumstances. This is difficult or even impossible to revert. Such a battery is then described as "sulfated".

Sulfation is a result of careless maintenance. It causes an increase in the battery's internal resistance which impedes the chemical processes and makes charging more difficult.

A sulfated battery gets very hot when charged with a "W" charging characteristic. The charge voltage rises sharply as soon as charging starts. If sulfation is only slight, the lead sulfate is broken down slowly and the charge voltage drops steadily. As soon as the lead sulfate has been broken down completely (regenerated), the voltage climbs again the same as it does when charging a serviceable, non-sulfated, battery (Fig. 4).

Troubleshooting

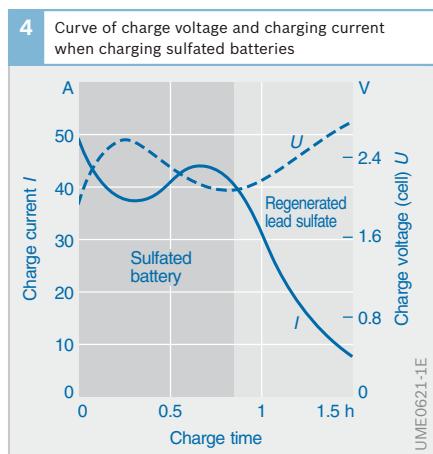
If the starter battery fails to start the engine, this can be due to insufficient charge or to a battery defect. If a battery's state of charge is less than 50 %, startability at very low temperatures (-20 °C) can no longer be taken for granted. The battery's exact state of charge can be determined by measuring the electrolyte density and also by precise measurement of the steady-state voltage. This makes it extremely easy to differentiate between a poorly charged battery and a defective but properly charged battery. A state of charge of 50 %

is approximately equivalent to a steady-state voltage of 12.3 V. Excessive water consumption, and/or charging immediately before the test will falsify the results in the direction of a more favorable verdict. The charge on the surface of the plates that is falsifying the result can be removed by brief discharging to the extent of 5 % of the nominal capacity.

Safety instructions

Short-circuits (e.g. with the tool) can create sparks and cause burning. For this reason the earthing strap must be disconnected before starting work on the electrical system or in the vicinity of the battery (after all consumers have been switched off). Particular care must be taken when connecting/disconnecting battery-charger cables or jumper cables in order to avoid a short-circuit. The following basic rules concerning safety must be followed when working with batteries:

- ▶ When handling sulfuric acid or when topping-up batteries which are not maintenance-free with distilled water, always wear protective goggles and rubber gloves
- ▶ Do not fill up with electrolyte to above the MAX mark
- ▶ Do not tilt batteries to extreme angles from the vertical for long periods
- ▶ Due to the risk of electrolytic-gas detonation, smoking and naked flames are not allowed and sparks must be avoided when batteries are being charged (connect and disconnect cables in the prescribed order with the battery charger switched off)
- ▶ Battery-charging rooms must be well ventilated



Alternators

Motor vehicles have an alternator to generate electrical energy for supplying power to electrical consumers such as the starter, ignition and fuel-injection system and electronic control units, etc. The alternator also charges the battery if it generates more current than the consumers need. **Alternator output power, battery capacity and the power requirement of the electrical consumers must be matched to each other to ensure that sufficient current is supplied to the vehicle electrical system in all operating conditions and that the battery is always adequately charged.**

Electrical power generation in the vehicle

Three-phase alternators

The availability of reasonably priced power diodes as from around 1963, paved the way for Bosch to start with the series production of three-phase alternators. Thanks to its greater capacity factor (ratio of energy generated to mass), higher efficiency and wider engine-speed range in comparison to the direct-current generator, the three-phase synchronous generator is able to deliver power while the internal-combustion engine is idling and meet the increasing power requirements in the motor vehicle. At engine idle speed, a three-phase alternator already delivers at least a third of its rated output. As the battery and electrical consumers have to be supplied with direct current, the alternating voltage generated by the three-phase alternator is rectified.

Since the alternator speed can be matched to that of the engine by means of a conversion ratio, this means that the battery remains at a high charge level even in unfavorable conditions (high proportion of idle phases, low outside temperatures).

To be able to supply the battery and consumers with a constant voltage, the alternator is equipped with a voltage regulator.

Three-phase alternators are designed for charge voltages of 14 V (for passenger cars) and 28 V (for commercial vehicles with a 24 V vehicle electrical system).

Influencing variables

Engine speed

An alternator's efficiency (energy generated per kg mass) increases with rotational speed. This factor dictates as high a conversion ratio as possible between engine crankshaft and alternator. However, it must be ensured that the maximum permissible speed of the alternator is not exceeded at maximum engine speed. For passenger cars, typical values for the turns ratio are between 1:2.2 and 1:3, and for commercial vehicles up to 1:5.

Temperature

The losses suffered in the conversion of mechanical energy to electrical energy cause the components of the alternator to heat up. A fresh-air supply or liquid cooling are suitable measures for reducing component temperatures.

Vibrations and oscillations

Depending on installation conditions and the engine's vibration patterns, the alternator could be subjected to vibration accelerations of between 500 and 800 m/s². This places heavy loads on the mountings and the components of the alternator. Critical natural frequencies inherent in the alternator design must be avoided.

Other impacts

The alternator is also subjected to such detrimental influences as spray water, dirt, oil, fuel mist, and possibly road salt. The alternator must be suitably protected from corrosion to prevent leakage paths from forming between conducting parts.

Requirements

Essential requirements for the automotive alternator include:

- ▶ Supplying all connected consumers with DC voltage
- ▶ Providing power reserves for rapidly charging the battery and keeping it charged, even when continuous consumers are switched on
- ▶ Maintenance of a constant alternator voltage throughout the entire ranges of engine speed and load conditions
- ▶ Robust design capable of withstanding external stresses, such as vibration, high ambient temperatures, temperature cycles, dirt, moisture
- ▶ Long service life comparable with that of the vehicle engine
- ▶ Low weight
- ▶ Installation-friendly dimensions
- ▶ Low running noise
- ▶ High efficiency

Operating principle of the alternator

Electromagnetic induction

When an electric conductor moves relative to the lines of force of a magnetic field, a voltage is induced in the conductor.

According to Faraday's law, the induced voltage U_{ind} is greater the faster the speed v of vertical motion in relation to the lines of force and the greater the magnetic flux B through the conductor cross section:

$$U_{\text{ind}} \sim |\vec{v} \times \vec{B}|$$

To generate an alternating current, a conductor loop is turned between the north and south pole of a permanent magnet. With a uniform turning of the conductor loop, the curve of the induced voltage is sinusoidal. The induced voltage can be tapped at the ends of the conductor loop using sliprings and carbon brushes.

Alternating current (AC) flows as soon as the circuit is closed (Fig. 1).

1 Induced single-phase alternating voltage

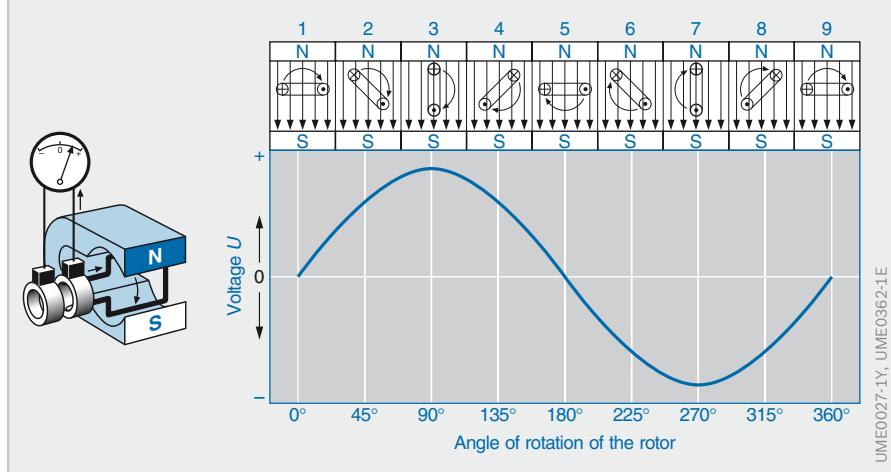


Fig. 1
Voltage curve generated during one full revolution of a winding rotating in a magnetic field. The position of the rotor on the left corresponds to position 3.
UME0027-1Y, UME0362-1E

Excitation field

The magnetic field for inducing the voltage (excitation field) can be generated by permanent magnets. Due to their simplicity, these have the advantage of requiring only a minimum of technical outlay. They are used for small alternators (e.g. bicycle dynamos).

Substantially higher, and at the same time controllable, currents can be generated if the excitation field is established by electromagnets. The electromagnet essentially comprises a coil (excitation winding) through which an excitation current flows; the number of windings, together with the strength of the excitation current, determines the magnetic field strength. The magnetic field can be amplified by means of a magnetizable iron core in the electromagnet.

By changing the excitation current, it is possible to regulate the magnetic field and thus, too, the value of the induced voltage.

If an external source of energy (e.g. a battery) supplies the excitation current, this is known as external excitation. Self-excitation is where the excitation current is tapped from the generated alternator current.

Three-phase alternators

The fundamental components of an alternator are the stationary stator (Fig. 2, Item 4) and the rotating rotor (5) inside it. Discussions up to this point have centered on an alternator with static excitation field and rotating armature winding in which the load current is generated. For automotive three-phase alternators though, the winding system is in the stationary stator

2 Compact alternator: Section drawing (example)

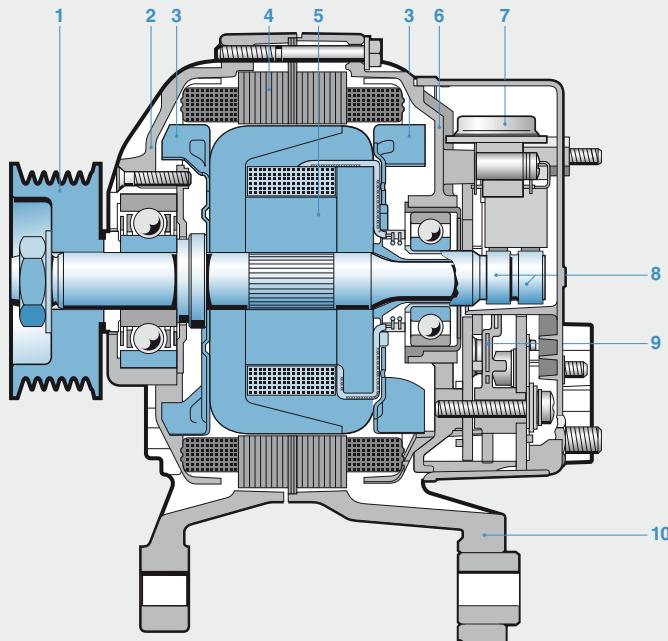


Fig. 2

- 1 Pulley
- 2 Drive end shield
- 3 Inboard ventilator
- 4 Stator
- 5 Rotor
- 6 Slipring end shield
- 7 Electronic voltage regulator with brush holder
- 8 Sliprings
- 9 Rectifier
- 10 Swivel arm

(hence the term, stator winding). The poles of the magnet together with the excitation winding are now situated on the rotating part, the rotor. The rotor's magnetic field builds up as soon as excitation current flows through the excitation winding.

With the three-phase alternator, the armature contains three identical windings which are offset from each other by 120° . The start points of the three windings are usually designated u, v, w, and the end points x, y, z. Due to the 120° spatial offset between the windings, the sinusoidal alternating voltages generated in them are also out-of-phase by 120° (time offset). The resulting alternating current is known as a three-phase current.

For unconnected windings, six electrical connections would be required for the stator winding (Fig. 3a). The number of current lines is reduced to three by linking the three windings in a star connection (Fig. 3b) or in a delta connection (Fig. 3c).

In the case of the star connection, the ends of the 3-phase windings are joined to form a star point. Without a neutral conductor, the sum of the 3 currents at any instant time is always 0.

Rectification of the alternating voltage

The alternating voltage generated by the alternator has to be rectified because the supply to the battery and the electronics in the vehicle electrical system requires direct current.

Half-wave rectification

A rectifier diode only allows electric current to pass in one direction and blocks it in the other. The negative half waves of the induced voltage are thus suppressed and only positive half waves are allowed to pass. The result is a gapped DC voltage (Fig. 4a)

Full-wave rectification

Full-wave rectification (Fig. 4b) is used to exploit both the positive and negative half waves for current generation. This is also known as two-wave rectification. Due to the diodes with opposite bias at each of the two stator connections (3), the alternating current has access to a conductive diode during the positive half wave and the negative half wave. The bridge circuit with four diodes transforms the alternating voltage into a pulse-controlled DC voltage (without gaps).

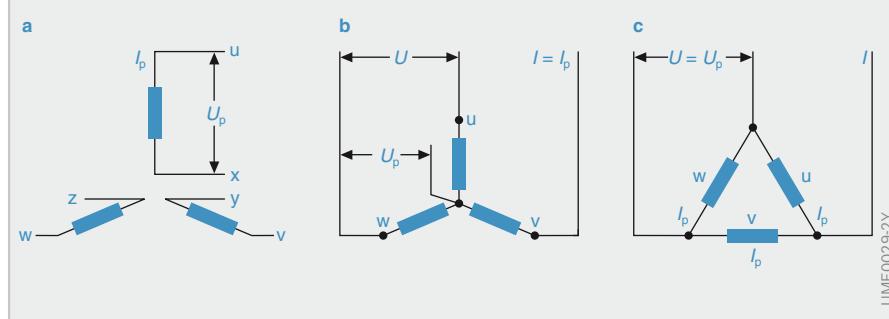
Fig. 3

a Windings not connected

b Star connection; alternator voltage U and phase voltage U_p (partial voltage) differ by the factor $\sqrt{3} = 1.73$. The alternator current I is equal to the phase current I_p . $U = \sqrt{3} \cdot U_p$; $I = I_p$

c Delta connection; alternator voltage U is equal to the phase voltage U_p . The alternator current I and the phase current I_p differ by the factor $\sqrt{3} = 1.73$. $U = U_p$; $I = \sqrt{3} \cdot I_p$

3 Connection of the three windings



Rectification of a three-phase alternating voltage

The alternating voltages generated in the three windings of the three-phase alternator (Fig. 5) are rectified by six diodes in an 3-phase bridge circuit. Two power diodes are connected to each phase winding, one diode to the positive side (positive diode to terminal B+) and one to the negative side (negative diode to terminal B-).

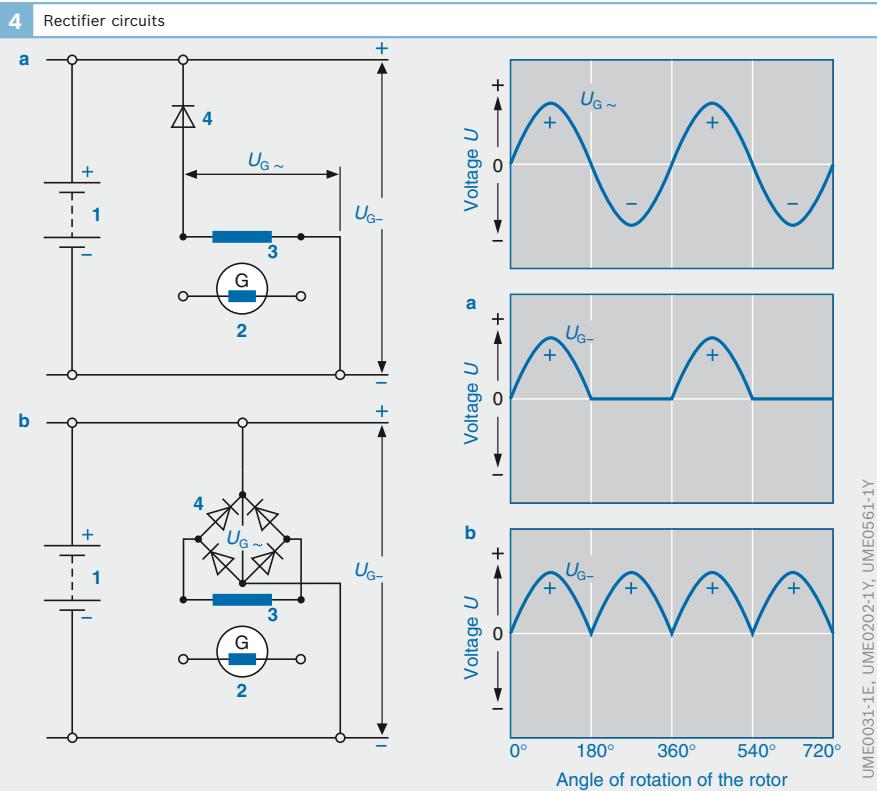
The positive half-waves pass through the positive-side diodes, and the negative half-waves through the negative-side diodes (Fig. 5b).

With full-wave rectification of the three phases using the B6 bridge circuit, the positive and negative half-wave envelopes are added to form a rectified alternator voltage with a slight ripple (Fig. 5).

This means that the direct current (DC) which is taken from the alternator at terminals B+ and B- to supply the vehicle electrical system is not smooth but has a slight ripple. This ripple is further smoothed by the battery, which is connected in parallel to the alternator, and by any capacitors that may be present in the vehicle electrical system.

Reverse-current block

The rectifier diodes in the alternator not only rectify the alternator and excitation voltage, but also prevent the battery discharging through the 3-phase winding in the stator.



With the engine stopped, or with it turning too slowly for self-excitation to take place (e.g. during cranking), without the diodes, battery current would flow through the stator winding. With respect to the battery voltage, the diodes are polarized in the reverse direction so that it is impossible for battery-discharge current to flow. Current flow can only travel from the alternator to the battery.

Rectifier diodes

The positive and negative diodes are completely identical in their operation. The knurled metal casing of the positive diodes acts as a cathode, the casing of the negative diode as an anode. The positive diodes are pressed into the positive heat sink that is conductively connected to the

B+ battery terminal. As an anode, the metal casing of the negative diode pressed into the negative heat sink is conductively connected to ground (B-). The diode wire terminations are connected to the ends of the stator winding. Most power diodes in use today are Zener diodes (Z-diodes). These also serve to limit the voltage peaks which occur in the alternator due to extreme load changes (load-dump protection).

Generally, three-phase alternators are not provided with reverse-polarity protection. If battery polarity is reversed (e.g. when starting with the assistance of an external battery), this will destroy the alternator diodes as well as endanger the semiconductor components in other equipment.

5 3-phase bridge circuit

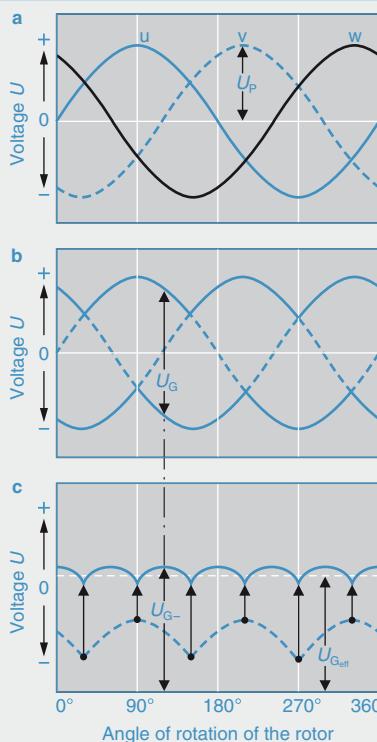
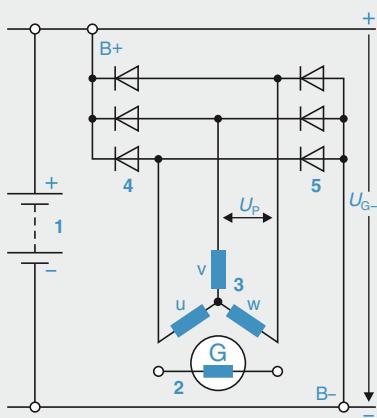


Fig. 5
 a Three-phase alternating voltage
 b Alternator voltage, formed by the envelope curves of the positive and negative half-waves
 c Rectified alternator voltage
 U_P Phase voltage
 U_G Voltage at the rectifier (negative not grounded)
 U_{G-} Alternator DC voltage (negative grounded)
 U_{Geff} Effective value of the DC voltage

- 1 Battery
- 2 Excitation winding
- 3 Stator winding
- 4 Positive diodes
- 5 Negative diodes

The three-phase alternator's circuits

Standard-version three-phase alternators have the following three circuits:

- ▶ Preexcitation circuit (external excitation using battery current)
- ▶ Excitation circuit (self-excitation)
- ▶ Alternator circuit or main circuit

Preexcitation circuit for alternators with excitation diodes

Before self-excitation of the alternator can be initiated, a voltage must be induced in the stator that can drive the excitation current. While the rotor does have some residual magnetism when de-energized, this is not sufficient for operation by self-excitation. The alternator must therefore be excited independently at the start of operation. This is achieved by the preexcitation circuit, which is supplied by the battery.

When the ignition or driving switch (Fig. 6, Item 4) is operated, the battery current I_B first of all flows through the alternator indicator lamp (3), through the excitation winding in the rotor (1d), and through the voltage regulator (2) to ground. In the

rotor, this battery current serves to pre-excite the alternator.

Self-excitation is initiated when the alternator voltage is greater than the voltage drop at both diodes present in the excitation circuit (one a conductive excitation diode, the other a power negative diode: $2 \times 0.7 \text{ V} = 1.4 \text{ V}$). The rotation speed at which self-excitation is initiated is the self-excitation speed. It is above the 0-ampere speed. The 0-ampere speed is the speed at which the induced voltage of the self-excited alternator has exactly the same value as the vehicle system voltage, so the alternator current is zero. Since the charge-indicator lamp increases the resistance in the preexcitation circuit in relation to that of the excitation circuit (with self-excitation), a greater excitation current flows in self-excitation mode that induces a greater voltage (at the same speed) in the alternator stator. The output of the charge-indicator lamp therefore influences the self-excitation speed. Standard outputs of alternator indicator lamps are 2 W for 12 V systems and 3 W for 24 V systems.

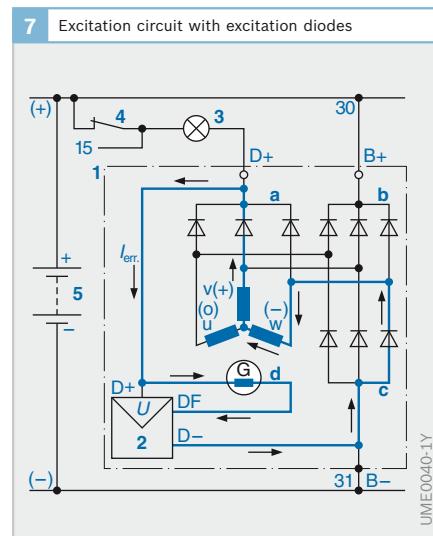
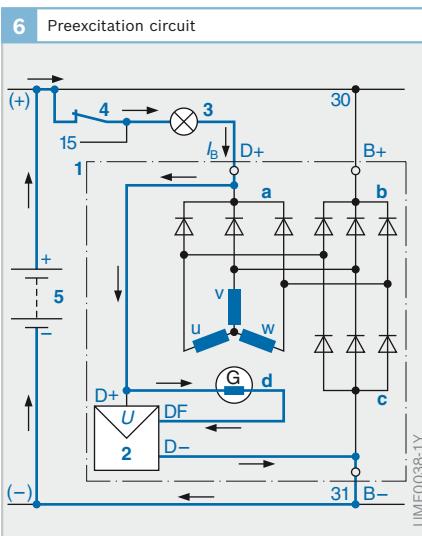


Fig. 6 and Fig. 7

- 1 Alternator
- 1a Excitation diodes
- 1b Positive-plate diodes
- 1c Negative-plate diodes
- 1d Excitation winding
- 2 Regulator
- 3 Alternator indicator lamp
- 4 Ignition switch
- 5 Battery

Preexcitation with alternators with multifunction controller (alternators with excitation diodes)

When the multifunction controller (see “Voltage regulation” section) receives the “ignition ON” signal through the L connection, it switches on the preexcitation current at the fixed on/off ratio (approx. 20 %, “controlled preexcitation”). When the rotor starts to turn, the regulator registers a voltage at the phase connection V, whose frequency it uses to calculate the alternator speed. A cutting-in speed is set in the regulator, and as soon as this is reached, the regulator switches fully through the output stage (on/off ratio 100 %) so that the alternator starts to deliver current to the vehicle electrical system.

Excitation circuit

During alternator operation, it is the task of the excitation current I_{err} to generate a magnetic field in the excitation winding of the rotor so that the required alternator voltage can be induced in the stator windings. Since three-phase alternators are “self-excited”, the excitation current must be tapped off from the stator winding.

Alternators with multifunction controllers (without exciter diodes, see “Voltage regulation” section) draw their excitation current directly from terminal B+ (Fig. 8). The excitation current flows through the positive power diodes, multifunctional regulator, carbon brushes, sliprings, and excitation winding to ground (B-).

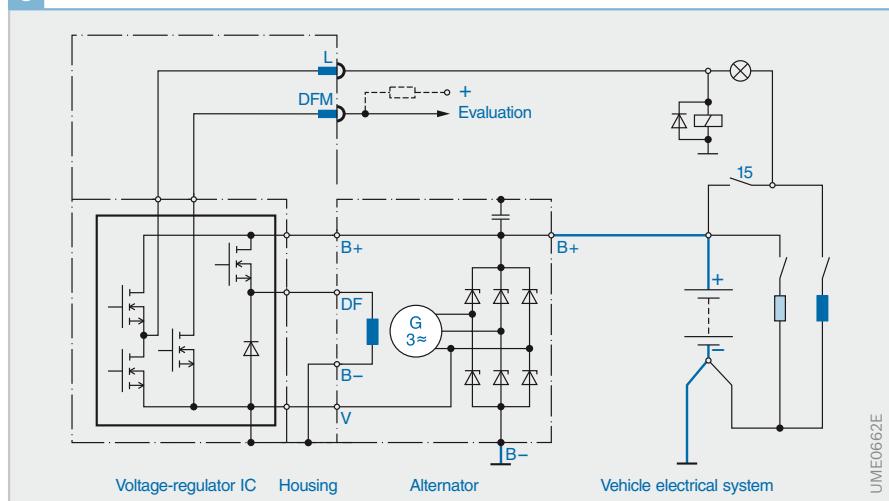
Older versions of the alternator still have three excitation diodes at terminal D+, which, with the three negative diodes at terminal B-, form a B6 bridge circuit for rectifying the excitation current (Fig. 7). The excitation current flows through the excitation diodes, carbon brushes, sliprings, and excitation winding to terminal DF of the regulator, and from terminal D- of the regulator to ground (B-).

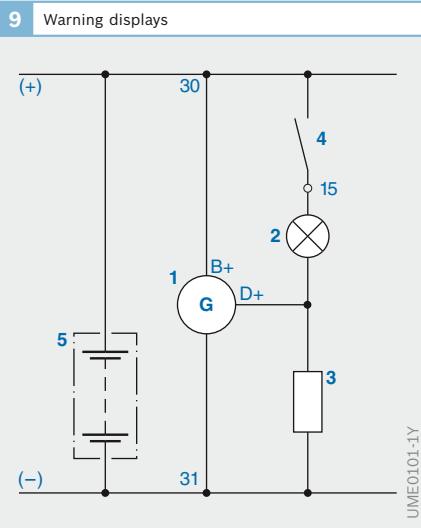
In both cases, the excitation current flows from B- back to the stator winding through the negative power diodes.

Warning displays

If the alternator indicator lamp fails to go out even at high speeds, there is a fault in the alternator itself, at the regulator, in the wiring, or at the V-belt. By installing an

8 Excitation circuit without excitation diodes for alternator with multifunction controller



**Fig. 9**

- 1 Alternator
- 2 Alternator indicator lamp
- 3 Resistor
- 4 Ignition switch
- 5 Battery

additional resistor (only in the case of alternators with excitation diodes; Fig. 9), it is possible to cause the alternator indicator lamp also to light up in the event of an interruption in the excitation circuit.

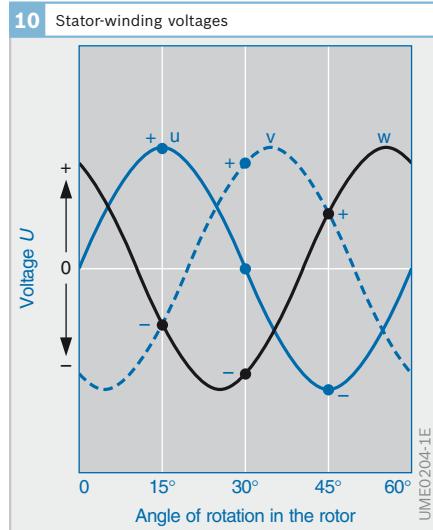
Alternator circuit

The alternating voltage induced in the three phases of the three-phase alternator

must be rectified by the power diodes in the bridge circuit before it is passed on to the battery and to the consumers.

The alternator current I_G , flows from the three windings and through the respective power diodes to the battery and to the consumers in the vehicle electrical system. In other words, the alternator current is divided into battery-charging current and consumer current.

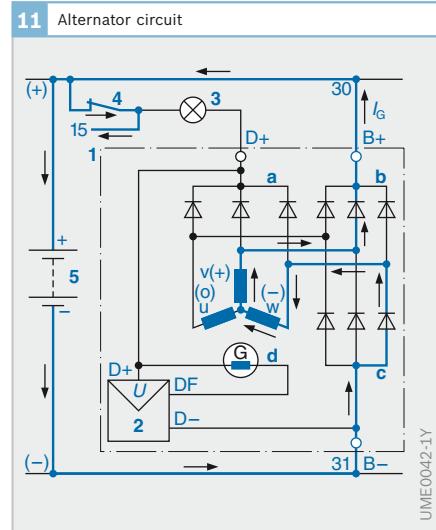
In Figure 10, the curves of the stator-winding voltages are shown as a function of the angle of rotation of the rotor (in this example, a rotor with six pole pairs): with an angle of rotation of 30° , the voltage in relation to the star point at the end of winding v is positive, for winding w it is negative, and for winding u it is zero. The resulting current curve is shown in Figure 11: the current flows from the end of winding v and through the positive diodes to alternator terminal B+ from where it flows through the battery, or the consumers, to ground (alternator terminal B-) and via the negative diodes to winding end w. At a 45° angle of rotation, current from the v and w winding ends takes the same path to winding end u. In this case, there is voltage present across all of the phases.

**Fig. 10**

Voltage curves as a function of the angle of rotation of a rotor with 6 pole pairs.

Fig. 11

- 1 Alternator
- 1a Excitation diodes
- 1b Positive-plate diodes
- 1c Negative-plate diodes
- 1d Excitation winding
- 2 Regulator
- 3 Alternator indicator lamp
- 4 Ignition switch
- 5 Battery



Voltage regulation

Purpose of voltage regulation

Presuming constant excitation current, the alternator voltage is dependent upon the alternator's speed and loading. The regulator has the job of maintaining the alternator voltage, and therefore also the vehicle system voltage, at a constant level across the engine's complete speed range, independent of electrical load. To this end, the voltage regulator controls the level of excitation current and thus the size of the magnetic field in the rotor as a function of the voltage generated in the alternator. In this way, the regulator keeps the vehicle system voltage constant, protects voltage regulation and prevents the battery from being overcharged or discharged during vehicle operation. The voltage-regulation tolerance zone for vehicle electrical systems with 12 V battery voltage is around 14 V, and for systems with 24 V battery voltage around 28 V. Whenever the voltage generated by the alternator is below the regulator response voltage, the voltage regulator remains switched off - the regulator output stage is on (on/off ratio 100 %).

The voltage-regulator characteristics (alternator voltage as a function of temperature) are matched to the battery's chemical characteristics. At low temperatures, the alternator voltage is somewhat higher to improve battery charging in winter. The input voltages of the electronic equipment are borne in mind here. At higher temperatures, on the other hand, alternator voltage is reduced slightly in order to prevent battery overcharge and gassing in summer. Figure 12 shows an example of a characteristic. The voltage level is 14.5 V with a slope of -10 mV/K (temperature compensation).

Principle of voltage regulation

The regulator switches off the output stage if the voltage exceeds the upper setpoint value (Fig. 13). As a consequence of the inductance of the field coil, the excitation current at first flows on through the free-wheeling diode; excitation is weakened and, as a result, the alternator voltage falls. If the alternator voltage falls below the setpoint value in this scenario, the regulator switches the excitation current back on. Excitation increases as does the alternator voltage. When the voltage exceeds the specified upper value again, the control cycle is repeated. Since these control cycles all take place within a matter of milliseconds, the alternator mean voltage is regulated effectively in line with the defined characteristic.

The relationship between switch-on and switch-off times is decisive for the magnitude of the resulting mean excitation current. At low rotational speeds, the "On" time is relatively long and the "Off" time is short. The excitation current is interrupted only briefly and its mean value is high. On the other hand, at high rotational speeds the "On" time is short and the "Off" time long. Only a low excitation current flows.

When the regulator switches to the "Off" status, upon interruption of the excitation current a voltage peak is induced in the

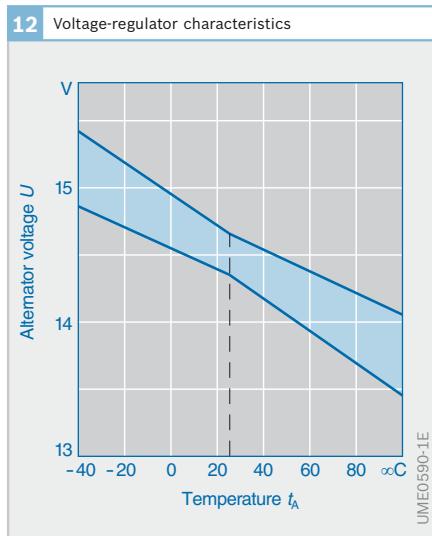


Fig. 12
Permissible tolerance band for the alternator voltage (14 V) as a function of the alternator's intake-air temperature.

excitation winding due to self-induction. To prevent the occurrence of these voltage peaks, a free-wheeling diode is connected in the regulator parallel to the alternator's excitation winding. As soon as the excitation winding is interrupted, the free-wheeling diode "takes over" the excitation current and permits it to decay.

The three-phase alternators of today are equipped with electronic voltage regulators as standard. By contrast, the electromagnetic regulator is in practice only ever used as a spare part.

Electromagnetic voltage regulators

The contact regulators used for three-phase alternator applications are of the single-element type (Figs. 14 and 15). That is, regulators with a voltage-regulator element comprising an electromagnet (3), a regulating armature, and a regulating contact (4). Alternator current flows through the electromagnet and the electromagnet controls the regulating armature, which in turn closes or opens the excitation circuit by means of a regulating contact.

Voltage is regulated in the following way with the single-element, single-contact regulator (Fig. 14): the regulating armature is acted on by the magnetic force of the electromagnet dependent on the alternator current on the one hand, and by the spring force of a suspension and adjustment spring on the other. As soon as the alternator voltage exceeds the setpoint value, the electromagnet pulls in the armature and opens the contact (position *b*). This switches a resistor into the excitation circuit which reduces the excitation current and with it the alternator voltage. If the alternator voltage falls below the set-point value, the magnetic force is reduced again. The spring force prevails and closes the contact (position *a*). This opening and closing cycle is repeated continually.

The single-element double-contact regulator (Fig. 15) operates with a second pair of contacts which permit 3 switching positions. The regulating resistor is short-circuited in position *a* and a high excitation current flows. In position *b*, the regulating resistor and the excitation winding are connected in series and the excitation

13 Regulation of excitation current I_{err}

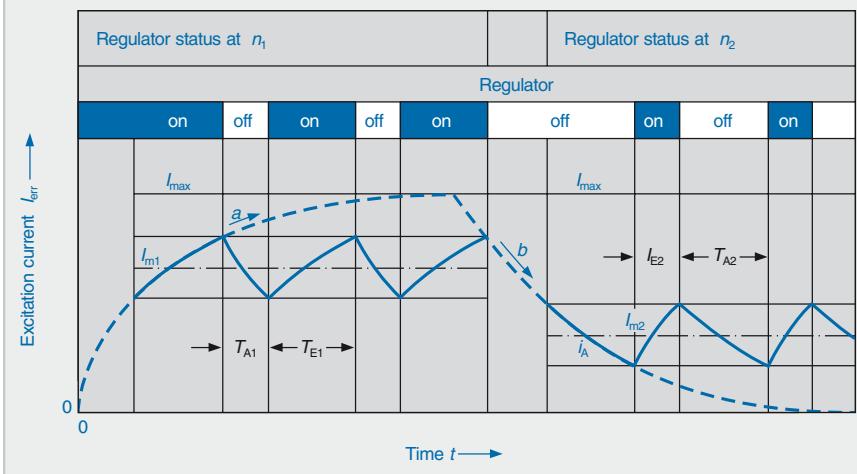


Fig. 13

The relationship between on-time T_E and off-time T_A is decisive for the magnitude of the mean excitation current I_m . The excitation current rises along curve *a*, and decays along curve *b*.

current is reduced. In position c, the excitation winding is short-circuited and the excitation current drops to zero (the time constant is independent of the excitation winding's inductance and reactance).

Electronic voltage regulators

Characteristics

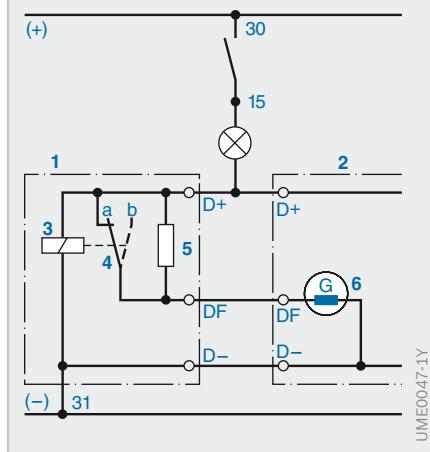
Electronic voltage regulators are used solely with three-phase alternators. Thanks to its compact dimensions, its low weight, and the fact that it is insensitive to vibration and shock, this regulator can be integrated directly in the alternator.

Whereas the first transistor regulators were built from discrete components, modern-day versions all use hybrid and monolithic technology.

The transistor regulator's essential advantages are:

- ▶ Shorter switching times which permit closer control tolerances
- ▶ Maintenance-free (no wear)
- ▶ High switching currents
- ▶ Spark-free switching, which causes no radio interference
- ▶ Insensitive to shock, vibration, and climatic effects
- ▶ Temperature compensation

14 Circuit diagram of a single-element, single-contact voltage regulator



- ▶ Compact construction allows direct mounting on the alternator, irrespective of alternator size

Method of operation

The operating concept is the same for all electronic-regulator types. An explanation is given here with the EE transistor regulator as an example.

The actual value of the alternator voltage between terminals D+ and D- is registered by a voltage divider (R1, R2, and R3). A Zener diode is switched in parallel with R3 and functions as the alternator's set-point generator. A partial voltage proportional to the alternator voltage is permanently applied to this diode.

The regulator remains in the "On" state as long as the actual alternator voltage is below the setpoint value (Fig. 16a). The Z-diode's breakdown voltage has not yet been reached at this point, i.e. no current flows to the base of transistor T1 through the branch with the Z-diode. T1 is in the blocking state. With T1 blocked, a current flows from the exciter diodes via terminal D+ and resistor R6 to the base of transistor T2 and switches T2 on. Terminal DF is now connected to the base of T3 by the switched transistor T2.

15 Circuit diagram of a single-element, double-contact voltage regulator

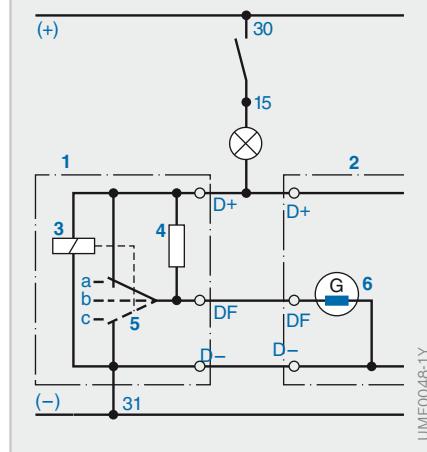


Fig. 14

- 1 Regulator
- 2 Alternator
- 3 Electromagnet
- 4 Regulating contact
- 5 Regulating resistor
- 6 Excitation winding (G)

Fig. 15

- 1 Regulator
- 2 Alternator
- 3 Electromagnet
- 4 Regulating resistor
- 5 Regulating contact
- 6 Excitation winding (G)

This means that T3 always conducts when T2 is conductive. T2 and T3 are connected as a Darlington circuit and form the regulator's power stage. The excitation current I_{err} flows through T3 and the excitation winding and increases during the "On" period, causing a rise in the alternator voltage U_G . At the same time, the voltage at the setpoint generator also rises. The regulator assumes the "Off" state as soon as the actual alternator voltage exceeds the setpoint value (Fig. 16b).

The Z diode becomes conductive when the breakdown voltage is reached. A current flows from D+ through the resistors R1, R2 and into the current path containing the Z-diode to the base of transistor T1. Thus T1 also becomes conductive. As a result, the voltage at the base of T2 is practically 0 in relation to the emitter, and transistors T2 and T3 (power stage) block. The excitation circuit is open-circuited, the excitation decays, and the alternator voltage falls as a result. As soon as the alternator voltage drops under the setpoint value,

16 Circuit diagram of a type EE transistor voltage regulator

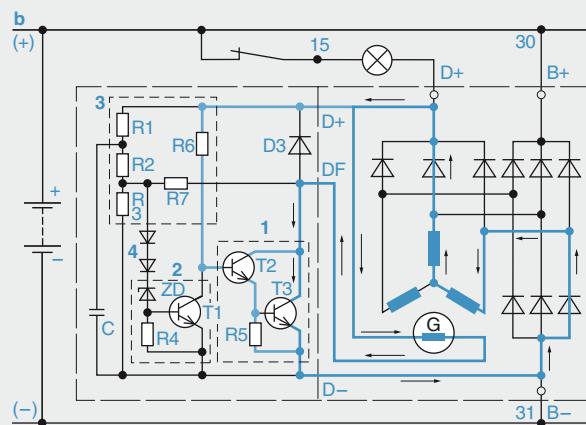
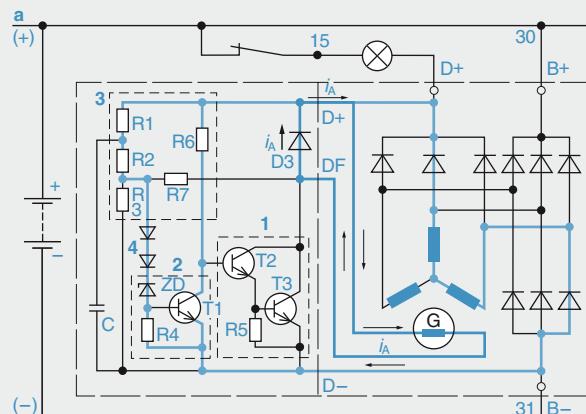


Fig. 16

- a Excitation current switched off by T3
- b Excitation current switched on by T3
- 1 Power stage
- 2 Control stage
- 3 Voltage divider
- 4 Temperature compensation diodes
- C Voltage-smoothing capacitor
- D3 Free-wheeling diode

and the Z diode switches to the blocked state, the power stage switches the excitation current on.

When the excitation current is interrupted, a voltage peak would be induced due to the excitation winding's self-induction which could destroy transistors T2 and T3. To prevent this, the free-wheeling diode D3 is connected in parallel to the alternator's excitation winding. Upon the excitation winding being interrupted, the free-wheeling diode "takes over" the excitation current and prevents the occurrence of voltage peaks.

The control cycle in which the current is switched on and off by connecting the excitation winding alternately to the alternator voltage or short-circuiting it with the free-wheeling diode is repeated periodically. The on/off ratio depends on the alternator speed and the applied load.

The ripple on the alternator DC voltage is smoothed by capacitor C. Resistor R7 ensures the rapid, precise switch-over of transistors T2 and T3, as well as reducing the switching losses.

Hybrid regulators

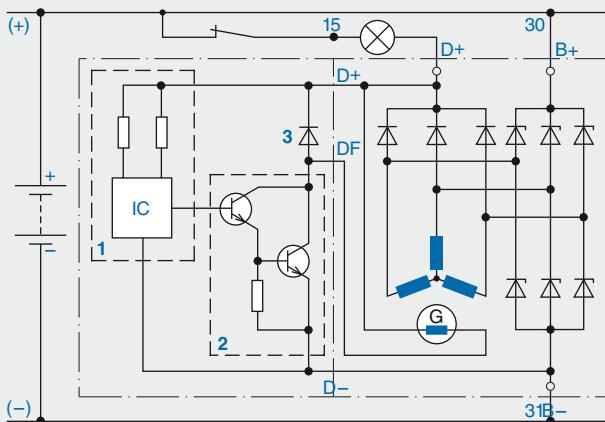
The transistor regulator using hybrid technology comprises a hermetically encapsulated case, in which are enclosed a ceramic substrate, protective thick-film resistors, and a bonded integrated circuit (IC) incorporating all the open- and closed-loop control functions.

The power components of the output stage (Darlington transistors and the free-wheeling diode) are soldered directly onto the metal socket in order to ensure good heat dissipation. The electrical connections are via glass-insulated metal pins.

The regulator is mounted on a special brush holder and directly fastened to the alternator without wiring. Due to the Darlington circuit in the power stage (two transistors), there is a voltage drop of about 1.5 V in the current-flow direction.

Figure 17 shows the circuit diagram for an three-phase alternator with a hybrid voltage regulator. The hybrid regulator's advantages can be summed up as follows: compact construction, low weight, few components, few connections and high reliability even in the extreme operating

17 Circuit diagram of an alternator equipped with type EL hybrid electronic voltage regulator



UME0056-1Y

Fig. 17

- 1 Control stage using thick-film techniques, with resistors and IC
- 2 Power stage (Darlington stage)
- 3 Free-wheeling diode

conditions encountered in automotive applications. Normally, hybrid regulators using conventional diodes are used with compact-diode-assembly alternators (see "Alternator versions" section).

Monolithic regulators

The monolithic regulator was developed from the hybrid regulator. The functions of the hybrid regulator's IC (Integrated Circuit), power stage, and free-wheeling diode have been incorporated on a single chip. The compact construction with fewer components and connections enabled reliability to be even further improved. Since the output stage is in the form of a simple power stage, the voltage drop in the current-flow direction is only 0.5 V.

Monolithic regulators in combination with rectifiers (with Zener diodes) are used in compact alternators.

Multifunction controller

Monolithic regulators can have a multitude of additional functions. These multifunction controllers are fitted in all new compact alternators today. They draw excitation current directly from B+, eliminating the need for excitation diodes.

Additional functions of the multifunction controller in comparison to conventional regulators are:

- ▶ Controlled preexcitation
- ▶ "Alternator turning" detection through monitoring of the phase connection
- ▶ Emergency regulation in the event of open circuit in the line at terminal L
- ▶ Over-temperature protection by means of a controlled reduction of excitation current if the limit temperature is exceeded
- ▶ Load-response functions: slow, continued regulation of the excitation current after switching to the load of the vehicle electrical system so that the engine has enough time when idling to balance out the increasing drive torque of the alternator.

The power loss associated with the alternator indicator lamp in the instrument cluster is often excessive. It can be reduced by using an LED display instead, for example. Multifunctional voltage regulators permit the triggering of lamp bulbs as well as of LED display elements in the instrument cluster.

Ovvoltage protection

The starter battery's low internal resistance generally suppresses all the voltage peaks occurring in the vehicle electrical system.

Nevertheless, it is often advisable to install overvoltage protection as a precautionary measure against faults in the vehicle electrical system on vehicles for the transport of hazardous materials.

Reasons for overvoltage

Overvoltage may occur in the vehicle electrical system as the result of:

- ▶ Regulator failure
- ▶ Failed or disconnected battery
- ▶ Switching off of consumers with a predominantly inductive load
- ▶ Loose connections

The overvoltages are voltage peaks that last a matter of milliseconds. The maximum voltage peaks are in the region of 350 V. Overvoltages are also generated when the line between battery and alternator is open-circuited with the engine running (this happens when an outside battery is used as a starting aid), or when high-power consumers are switched off. For this reason, under normal driving conditions, the alternator is not to be run without the battery connected.

Under certain circumstances though, short-term or emergency operation without battery is permissible. This applies to the following situations:

- ▶ Driving of new vehicles from the final assembly line to the parking lot
- ▶ Loading onto train or ship (the battery is installed shortly before the vehicle is delivered to the customer)
- ▶ Service work, etc.

With towing vehicles and agricultural tractors it is also not always possible to avoid operation without the battery connected.

Types of protection

Overvoltage protection can be achieved by various measures.

Protection by Zener diodes (Z diodes)

Z diodes can be used in place of the rectifier power diodes. They limit high-energy voltage peaks.

Furthermore, Z diodes function as a central overvoltage protection for the remaining voltage-sensitive consumers in the vehicle electrical system. The limiting voltage of a rectifier equipped with Z diodes is 25 to 30 V for an alternator voltage of 14 V, and 50 to 55 V for an alternator voltage of 28 V. Compact three-phase alternators are always equipped with Z diodes.

Surge-proof alternators and regulators

The semiconductor components in surge-proof alternators have a higher electric-strength rating. For an alternator voltage of 14-V, the electric strength of the semiconductors is at least 200 V, and for a 28-V alternator voltage it is 350 V.

In addition, a capacitor is fitted between the alternator's B+ terminal and ground which serves for intensified interference suppression.

The surge-proof characteristics of such alternators and regulators only protect these units, they provide no protection for other electrical equipment in the vehicle.

Overvoltage-protection devices

(only for 28 V alternators)

These are semiconductor devices which are connected to the alternator terminals D+ and D- (ground). In the event of voltage peaks, the alternator is short-circuited through its excitation winding. Primarily, overvoltage-protection devices protect the alternator and the regulator, and to a lesser degree the voltage-sensitive components in the vehicle electrical system.

18 Circuit diagram of a (non-automatic) overvoltage-protection device for a 24-V vehicle electrical system

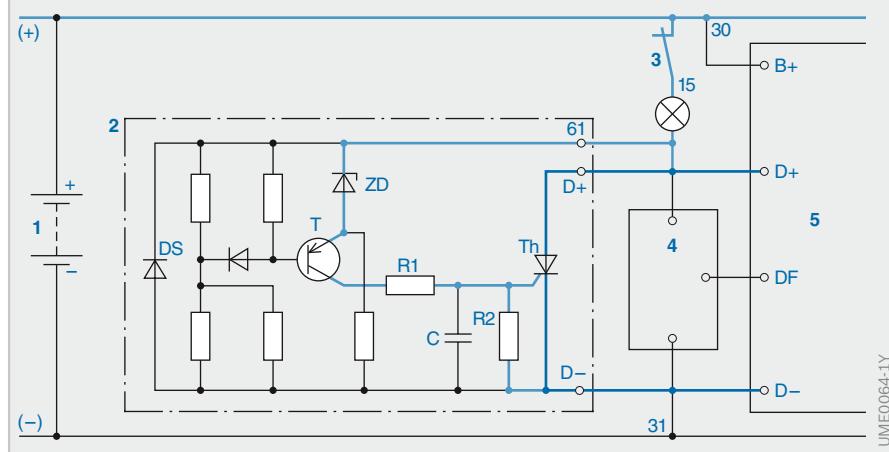


Fig. 18

- | | |
|---|--------------------------------------|
| 1 | Battery |
| 2 | Overvoltage-
protection
device |
| 3 | Driving switch |
| 4 | Regulator |
| 5 | Alternator |

Non-automatic overvoltage-protection device

This type of overvoltage-protection device is connected directly to the terminals D+ and D- on T1 alternators, e.g. in buses and heavy trucks (Fig. 18). The unit responds to voltage peaks and overvoltages at these terminals which exceed its response threshold of 31 V. At this point, thyristor Th becomes conductive. The Z diode ZD acts as a setpoint generator. The necessary response delay is regulated by resistors R1 and R2 along with capacitor C. The unit requires only milliseconds to short circuit the regulator and alternator via terminals D+ and D-. The thyristor assumes responsibility for the short-circuit current.

Meanwhile, current from the battery triggers the alternator indicator lamp to alert the driver. The thyristor remains active, reverting to its off-state only when the short-circuit current has stopped flowing, i.e. after the ignition has been switched off, or the engine and alternator come to rest.

The unit will not provide overvoltage protection if the wires at terminals D+ and D- are reversed.

The alternator indicator lamp does not indicate this fault. To ensure a reliable display, a safety diode DS is connected between terminals D+ and D-. If the connections are reversed, this diode is poled to allow current flow, and the alternator indicator lamp remains on continuously.

Overvoltage-protection device with automatic starting mechanism

This type of protection device is designed for use with T1 alternators (Fig. 19). The unit incorporates two inputs, D+ and B+, which react to different voltage levels and with varying response times. Input D+ provides rapid overvoltage protection, as on the device described above.

The input B+, responds only to defects at the voltage regulator, while the alternator voltage continues to climb until it reaches the unit's response voltage of approx. 31 V. The alternator then remains shorted until the engine is switched off. Input B+ is thus designed to prevent consequential damage.

Fig. 19 Simplified circuit diagram of an automatic overvoltage-protection device for a T1 alternator

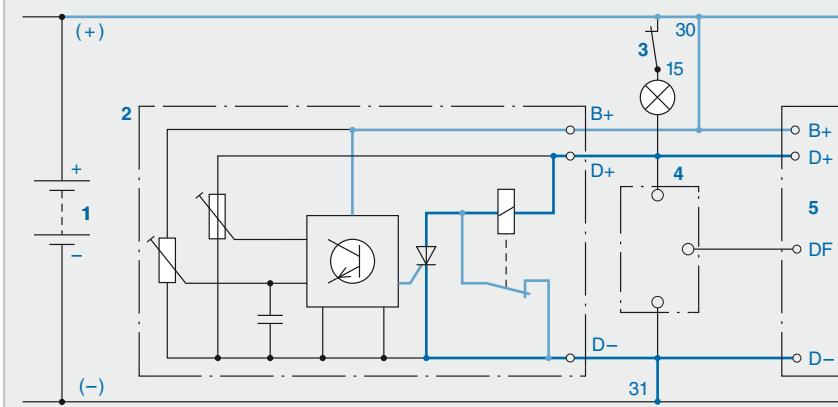


Fig. 19

- 1 Battery
- 2 Consequential-damage protection device
- 3 Driving switch
- 4 Regulator
- 5 Alternator

Voltage peaks which can be generated by the alternator itself when loads are switched off ("load dump") cannot damage other consumers in the system because the alternator is immediately short-circuited. In addition, the vehicle electrical system is protected against consequential damage if the regulator is faulty.

Consequential-damage protection device
This protection device is designed for use with the Double-T1 alternator with two stators and two excitation systems (Fig. 20). While the overvoltage-protection devices short-circuit the alternator, the consequential-damage protection unit functions as a kind of backup regulator, even with the battery out of circuit during emergency operation. Provided that the alternator's speed and the load factor allow it, the protection unit maintains a mean alternator voltage of approx. 24 V, thereby enabling emergency operation.

The consequential-damage protection device responds to battery operation and a faulty, short-circuited regulator by interrupting the alternator's excitation circuit for approx. 2 seconds after the alternator output passes the response threshold of

30 V. The unit's relay contact then assumes a backup voltage-control function by operating as a contact regulator. When the system is operated with the battery out of circuit (emergency operation without battery), the unit reacts to voltage peaks of 60 V or more which last for more than 1 ms.

The alternator indicator lamp flashes to indicate that the system is working in emergency operation mode. The system does not charge the battery, as the mean voltages in this mode are very low. The maximum operating time of the substitute regulating function is approx. 10 hours. The consequential-damage protection device will then have to be replaced.

Characteristic curves

Alternator performance

Due to the constant transmission ratio between alternator and engine, the alternator must be able to operate across a wide engine-speed range. The characteristic performance of the alternator is shown by its characteristics.

The alternator full-load curve relates to a constant alternator voltage and a defined ambient temperature.

20 Circuit diagram of a consequential-damage protection device for a Double-T1 alternator

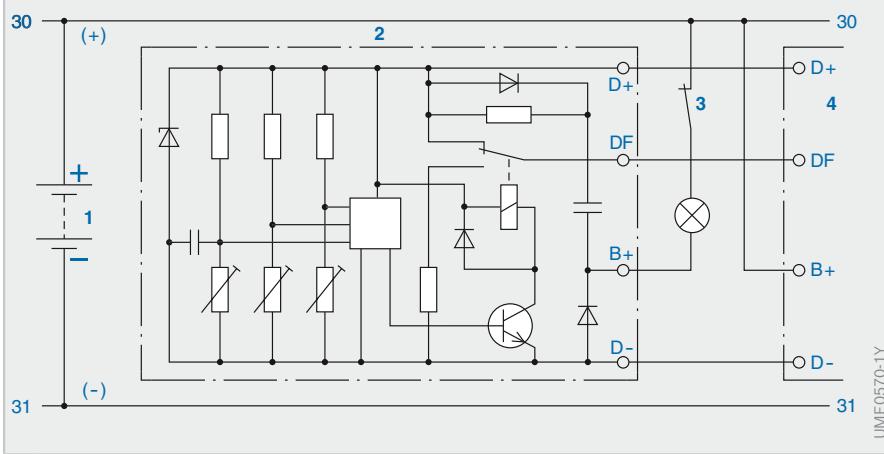


Fig. 20

- 1 Battery
- 2 Consequential-damage protection device
- 3 Driving switch
- 4 Alternator with voltage regulator

Current characteristic

Defined speed points of the alternator are used to describe the current characteristic (Fig. 21).

0-Ampere speed (n_0)

The 0-ampere speed is the speed at which the alternator reaches its nominal voltage without delivering power (approx. 1,000 rpm). The alternator can only deliver power at higher speeds.

Speed at engine idle (n_L), current at engine idle (I_L)

The alternator speed at engine idle is given as a range in the diagram because its value depends on the turns ratio between the alternator and the engine. In the case of

compact-diode-assembly alternators, it is defined as $n_L = 1,500$ rpm; for compact alternators, $n_L = 1,800$ rpm due to the usually higher turns ratio.

The current generated by the alternator at engine idle speed (idle current) must at least be sufficient to supply the permanently active consumers.

Rated speed (n_N), rated current (I_N)

The rated current is the full-load current of the alternator that it delivers at rated speed $n_N = 6,000$ rpm, at alternator voltage $U_G = 13.5$ V and at ambient temperature $T_U = 23$ °C.

Maximum speed (n_{\max}), maximum current (I_{\max})

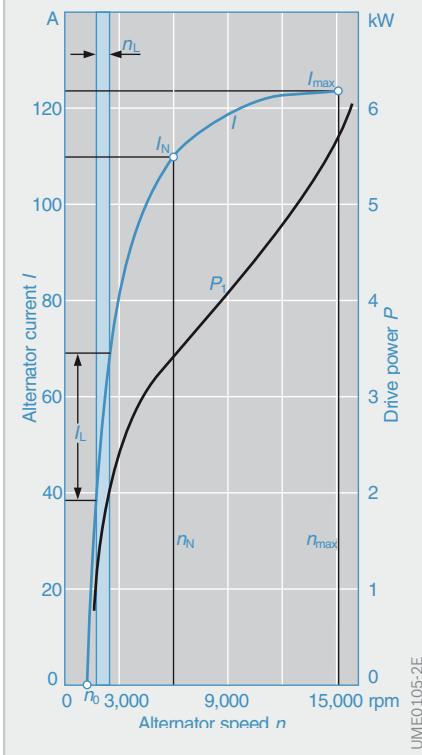
The maximum current I_{\max} is the highest current that the alternator is able to generate. At high speeds, the effect of the reverse magnetization field generated by the load current prevents the characteristic from climbing any higher. The speed of the alternator is first and foremost limited by its type of construction. The maximum speed of compact alternators is in the region of 18,000 to 22,000 rpm, 15,000 to 18,000 rpm for compact-diode-assembly alternators and between 8,000 and 15,000 rpm for commercial-vehicle alternators.

Characteristic of drive power

The characteristic of power input is used as a basis for designing the drive belt. The characteristic shows the maximum power transferred from the engine to the alternator across the engine's speed range. The alternator's drive power and power output can also be used to calculate its efficiency.

The characteristic of drive power is flat in the middle engine-speed range and increases sharply at higher speeds.

21 Typical characteristics for an three-phase alternator



Power losses

Efficiency

The maximum efficiency of a modern, air-cooled alternator is around 70 % at full load. In automotive applications, the alternator mainly operates in the part-load range, in which an efficiency of up to 75 % is achieved.

Presuming the same loading, using a larger, heavier alternator means that operation can take place across a more favorable portion of the efficiency program map (Fig. 22). On the other hand, the increased efficiency of the larger alternator can be offset by an accompanying increase in fuel consumption.

Alternator circuits

Parallel-connected power diodes

With large alternators with rated currents of over 180 A, the six power diodes of the three-phase-current bridge circuit could be damaged by overheating. For this reason, large alternators have two or more power diodes per phase connected in parallel (Fig. 24, Item 2, two diodes in this example). The alternator current is divided between the diodes connected in parallel

so that a commensurately weaker current flows through the individual diodes.

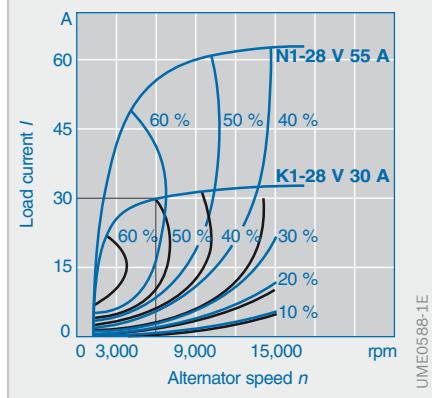
Auxiliary diodes at the star (neutral) point

On three-phase alternators with star-connected stator windings, the ends of the windings are joined at a single point, the star or neutral point. Thanks to two additional (auxiliary) diodes connected as power diodes between the neutral point and positive or negative terminal (Fig. 24, Item 3), it is possible to make use of the third-harmonic voltage induced in-phase in all three paths. The induced voltage of the third harmonic reaches vehicle system voltage above an alternator speed of approx. 3,000 rpm. Beyond this speed, the auxiliary diodes contribute to the increase in power and efficiency. However, this additional proportion also increases the ripple of the alternator voltage.

Operation of alternators in parallel

Alternators of the same voltage can be connected in parallel to meet demands for high power. The voltage regulators concerned must have the same regulating characteristics, and their characteristics must be identical.

22 Efficiency program map for K1 and N1 alternators



23 Loss distribution in an three-phase alternator

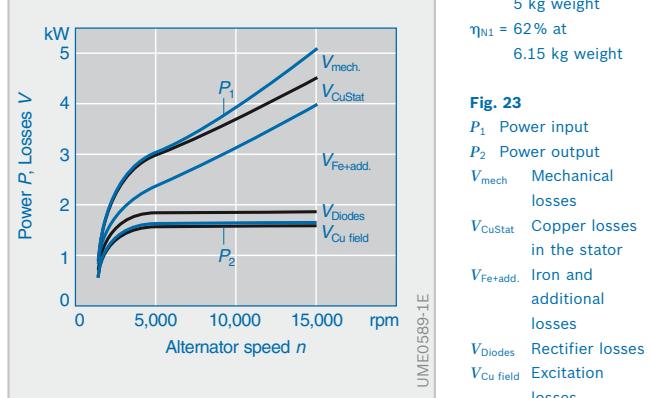


Fig. 22

I	= 30 A
U	= 28 V
n	= 6,000 rpm
η_{K1}	= 50 % at 5 kg weight
η_{N1}	= 62 % at 6.15 kg weight

Fig. 23

P_1	Power input
P_2	Power output
V_{mech}	Mechanical losses
V_{CuStat}	Copper losses in the stator
$V_{\text{Fe+add.}}$	Iron and additional losses
V_{Diodes}	Rectifier losses
$V_{\text{Cu field}}$	Excitation losses

Terminal W

Terminal W (Fig. 24, Item 5) provides a pulse-controlled DC voltage (half-wave-rectified alternating voltage) which can be used for measuring engine speed. The frequency depends on the number of pairs of poles and the speed of the alternator.

$$f = p \cdot n / 60$$

f Frequency (pulses per second)

p Number of pole pairs (6 on Size G,
K and N; 8 on Size T)

n Alternator speed (rpm)

Interference-suppression measures

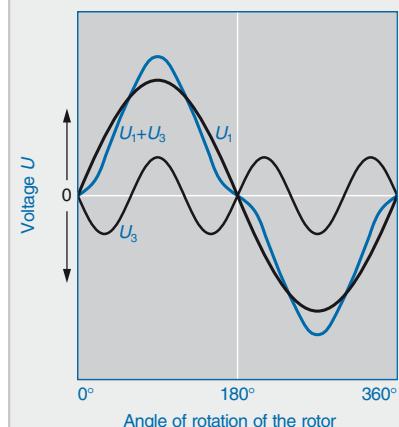
The electromagnetic fields of the alternator and other electrical consumers of a motor vehicle may interfere with other electrical and electronic equipment.

If a 2-way radio, car sound system, or car telephone etc. is operated in the vehicle itself or in the vicinity, it is necessary to install intensified interference suppression for alternator and regulator. Alternators are generally therefore fitted with a suppression capacitor. In the case of older compact-diode-assembly alternators, the suppression capacitor can be retrofitted on the outside of the slinging end shield.

On compact alternators, it is already integrated in the rectifier.

Older versions of the contact regulator are combined with an interference-suppression filter or are replaced by an interference-suppressed version. Transistor regulators do not require additional suppression measures.

25 Voltage with third harmonic



UME0221-1E

24 Circuit diagram of a T1 alternator with special circuit variants

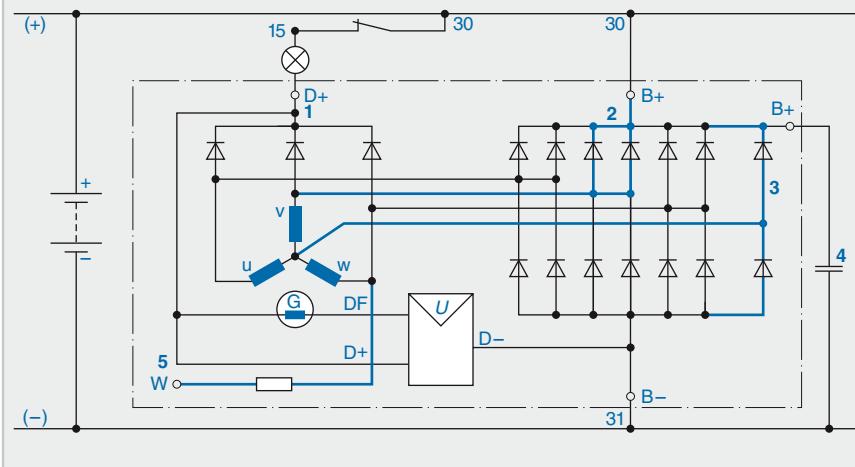


Fig. 24

- 1 Excitation diodes
- 2 Parallel-connected power diodes
- 3 Auxiliary diodes at the star (neutral) point
- 4 Suppression capacitor
- 5 Terminal W (interference-suppressed)

Fig. 25

- U_1 Phase voltage (fundamental wave)
 U_3 Third-harmonic voltage

UME0090-1V

Alternator designs

Design criteria

Various different operating conditions, and the power ranges of the vehicle types and engines concerned, led to the development of a number of basic models which will be described in the following.

The following data are decisive for alternator design:

- ▶ Vehicle type and the associated operating conditions
- ▶ Speed range of the engine with which the alternator is to be used
- ▶ Battery voltage of the vehicle electrical system
- ▶ Current draw of consumers
- ▶ Environmental loading imposed on the alternator (heat, dirt, dampness, etc.)
- ▶ Available installation space, dimensions

These criteria are decisive in the electrical dimensioning required for the alternator, i.e.

- ▶ The alternator voltage (14 V/28 V)
- ▶ The maximum power output (voltage × current intensity)
- ▶ The maximum current

Design of the compact alternator

The housing of the alternator comprises the drive and slings end shield (Fig. 26), between which the stator is clamped. The rotor shaft runs in bearings in each end shield.

The stator (3) of the alternator consists of mutually insulated, grooved laminations which are pressed together to form a solid laminated core. The turns of the three-phase stator winding are embedded in the grooves.

26 Design of the compact alternator

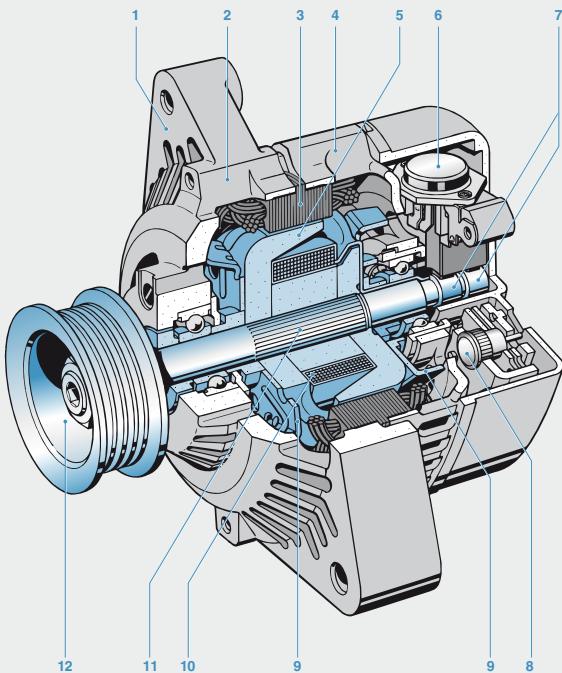


Fig. 26

- | | |
|----|--|
| 1 | Housing |
| 2 | Drive end shield |
| 3 | Stator |
| 4 | Slings end shield |
| 5 | Magnet pole |
| 6 | Electronic voltage regulator with brush holder |
| 7 | Slipsprings |
| 8 | Rectifier |
| 9 | Ventilators |
| 10 | Excitation winding |
| 11 | Rotor shaft |
| 12 | Pulley |

The rotor gives the claw-pole alternator its name: it comprises two oppositely-poled pole-wheel halves whose two claw-shaped pole fingers mesh with each other in the form of alternating north and south poles (Fig. 27).

The number of poles which can be realized in practice is limited. Whereas on the one hand, a low number of poles leads to a low machine efficiency, on the other hand, the more poles there are, the higher is the magnetic leakage. For this reason, such alternators are designed as 12-pole or 16-pole machines depending upon the power range.

The pole-wheel halves overlap the ring-coil-shaped excitation winding located on the pole body (Fig. 28). The excitation current reaches the excitation winding through carbon brushes which are mounted in the slippings end shield and pressed against the slippings by springs.

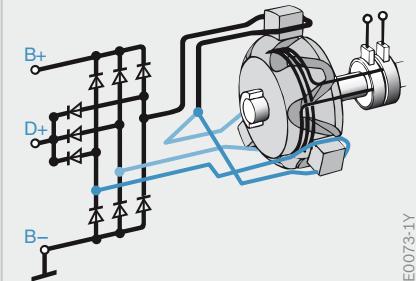
The magnetic flux flows through the pole body and the left-hand pole half and its pole fingers, across the air gap to the stationary stator lamination with stator winding, from where it flows back to the pole body through the right-hand pole half and completes the magnetic circuit.

The compact alternator is cooled by two ventilators (Fig. 26, Item 9) fitted on the drive and slippings sides of the rotor. Each ventilator draws in cooling air from the front end of the alternator and forces out the (warm) air radially (double-flow ventilation). These two small ventilators generate considerably less noise than the single large ventilator in a compact-diode-assembly alternator. In addition, they are designed for higher speeds (maximum speed: 18,000 to 22,000 rpm). These two special features permit a higher conversion ratio between crankshaft and alternator which means that for the same speed and size compact alternators can generate up to 25 % more power.

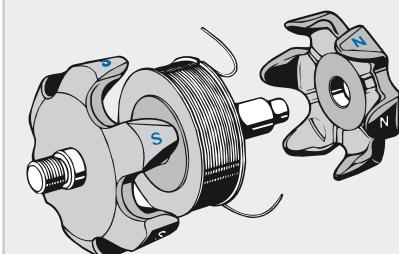
The pulley for the belt drive is also mounted on the rotor shaft. Alternator rotors can be rotated in either direction. The ventilator must be designed for clockwise or counterclockwise rotation.

The electronic voltage regulator forms a single unit with the brush holder (6), provided it is intended to be directly mounted on the alternator. In some cases, the electronic voltage regulator on commercial vehicles is secured to a protected part of the body independently of the alternator and connected to the brush holder by electrical plug-in connections.

27 Basic structure of a claw-pole alternator with slippings



28 Active components of a 12-pole claw-pole rotor



UME0073-1Y

Design of the compact-diode-assembly alternator

The claw-pole alternator of compact-diode-assembly design (Fig. 29) differs from the compact alternator first and foremost in the large ventilator (2), which is located between belt pulley (1) and the bowl-shaped alternator body. The external ventilator with a maximum speed of 12,000 to 18,000 rpm draws cooling air axially through the housing (single-flow ventilation).

Other differences with the compact-diode-assembly alternator are the larger sliprings (7) and the position of the rectifier inside slipring end shield (6).

Double-T1 alternator

The Double-T1 alternator consists of two electrically and mechanically coupled T1-size alternators in a common housing (Fig. 31) The electronic voltage regulator

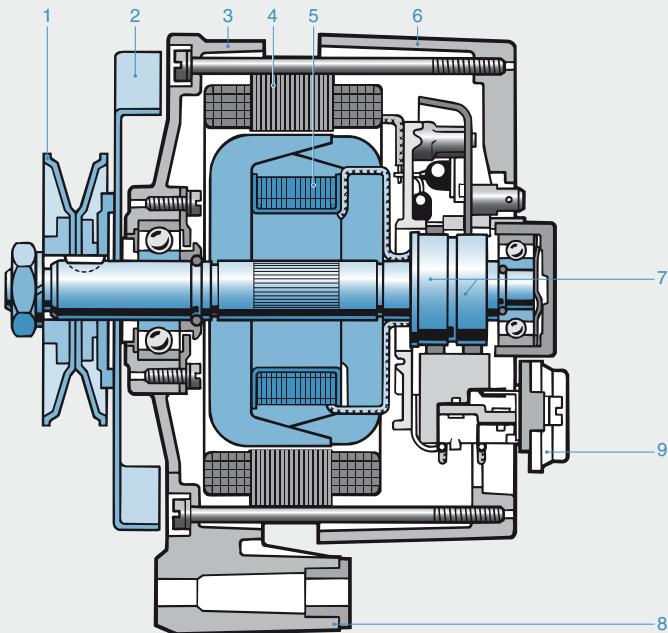
is installed inside the alternator. The electrical circuit of the two stators and the two excitation systems is illustrated in Figure 30. Between D+ and D- there is a 100 watt resistor, which causes the alternator indicator lamp to light up in case of an open-circuit field.

The Double-T1 alternator (DT1) is used in buses that place high power demands made on the alternator due to rising demands for more comfort.

Windingless-rotor alternators without sliprings

A typical feature of the windingless rotor (Fig. 32) is that the housing with the stator core (4), the heat sinks with power diodes, the attached transistorized voltage regulator (8), and the internal pole with excitation winding (5) all belong to the stationary part of the alternator, so sliding contacts are not required.

29 Claw-pole alternator with compact diode assembly: section drawing



UME0074-1Y

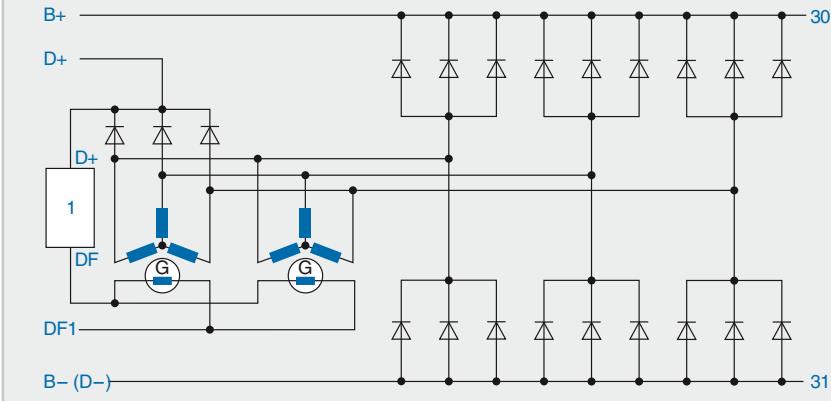
Fig. 29

- 1 Belt pulley
- 2 Ventilator
- 3 Drive end shield
- 4 Stator core
- 5 Excitation winding
- 6 Slipring end shield
- 7 Sliprings
- 8 Swivel arm
- 9 Regulator

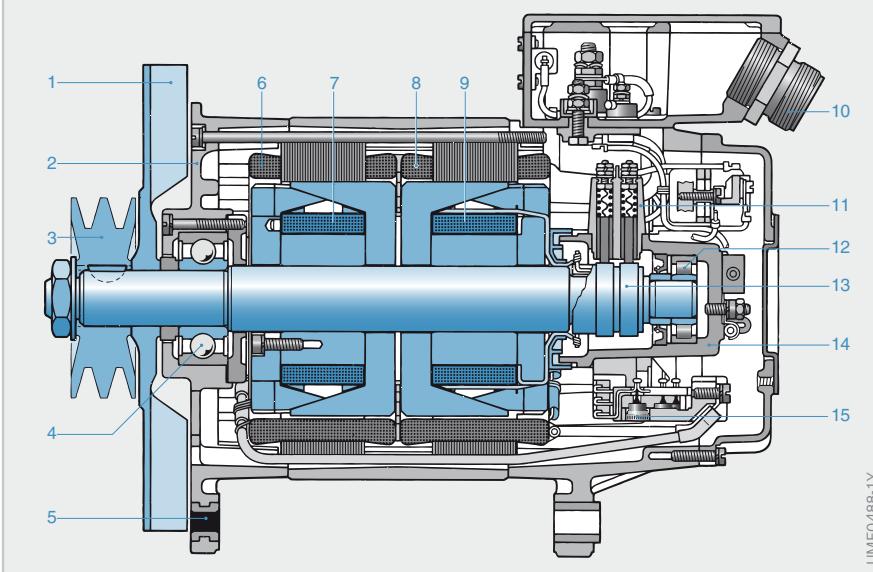
The rotating part merely comprises the rotor (6) with pole wheel and conductive element. Six pole fingers of the same polarity form a single north or south pole-finger crown. Only one of the two pole-finger

crowns is directly connected to the rotor shaft. The two crowns form claw-pole half sections, and are retained by a non-magnetic ring positioned below the mutually-engaged pole fingers.

30 Schematic circuit diagram of a Double-T1 alternator with two stators and two excitation systems



31 Cutaway view of a DT1 Double-T1 alternator with two stators and two excitation systems



In the process, the magnetic flux travels from the rotating rotor's pole core, through the stationary internal pole to the conductive element, from whose pole finger it then flows to the stationary stator core. The magnetic circuit closes in the rotor's pole core via the oppositely poled claw half. In comparison to a slipring rotor, the magnetic flux must cross two additional air gaps between the rotating pole wheel and the stationary internal pole.

With the omission of the slipring/carbon-brush system wear component, these alternators are particularly suitable for applications requiring a long service life in which the alternator is subjected to heavy loads, e.g. for construction machines or railway alternators.

The alternator with windingless rotors is also available as a liquid-cooled version. A considerable reduction of noise is achieved here in comparison to air-cooled alternators because the alternator is completely encapsulated and there is no ventilator to produce aerodynamic noise. Engine coolant flows around the complete jacket and rear of the alternator casing. The electronic components are mounted on the drive end shield.

32 Sectional drawing of type N3 windingless-rotor alternator

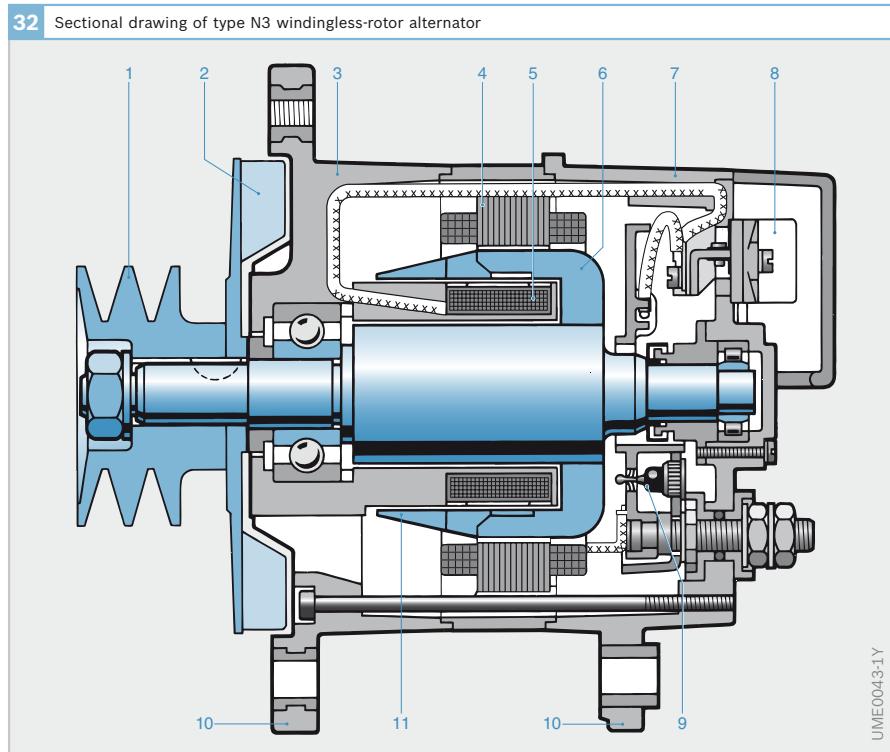


Fig. 32

- 1 Double-groove pulley
- 2 Ventilator
- 3 Drive end shield with stationary internal pole
- 4 Stator core
- 5 Stationary excitation winding
- 6 Windingless rotor
- 7 Rear end shield
- 8 Attachment transistor regulator
- 9 Power diode
- 10 Swivel arm
- 11 Conductive element

Salient-pole alternators

Salient-pole alternators are required in special applications which are characterized by extreme power demands (as in tour buses) (Fig. 33). The rotor is equipped with individual magnetic poles, each of which is provided with its own field winding. This layout allows the stator to be substantially longer (relative to its diameter) than would be possible with a claw-pole alternator. This means that higher outputs can be achieved without

increasing diameter. The achievable maximum speeds, however, are lower in comparison to claw-pole versions because the rotor windings are not as well protected against centrifugal forces as they are with claw-pole alternators.

The losses in the regulator are also greater (in comparison to the claw-pole alternator) due to the substantially greater excitation current. The regulator is therefore fitted separately in a special housing where it can be cooled more effectively.

33 Cutaway view of a type U2 salient-pole alternator

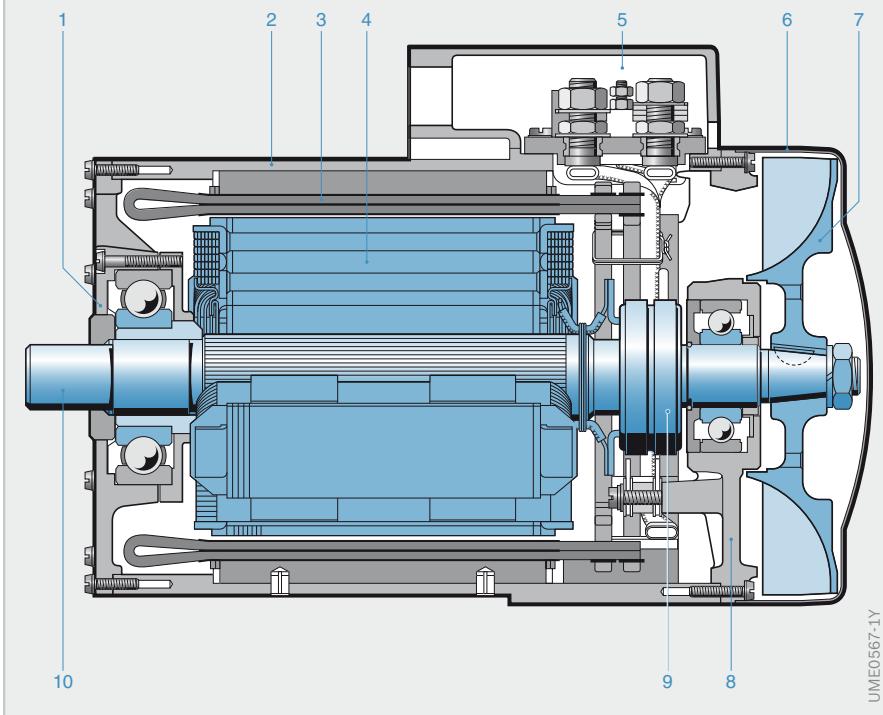
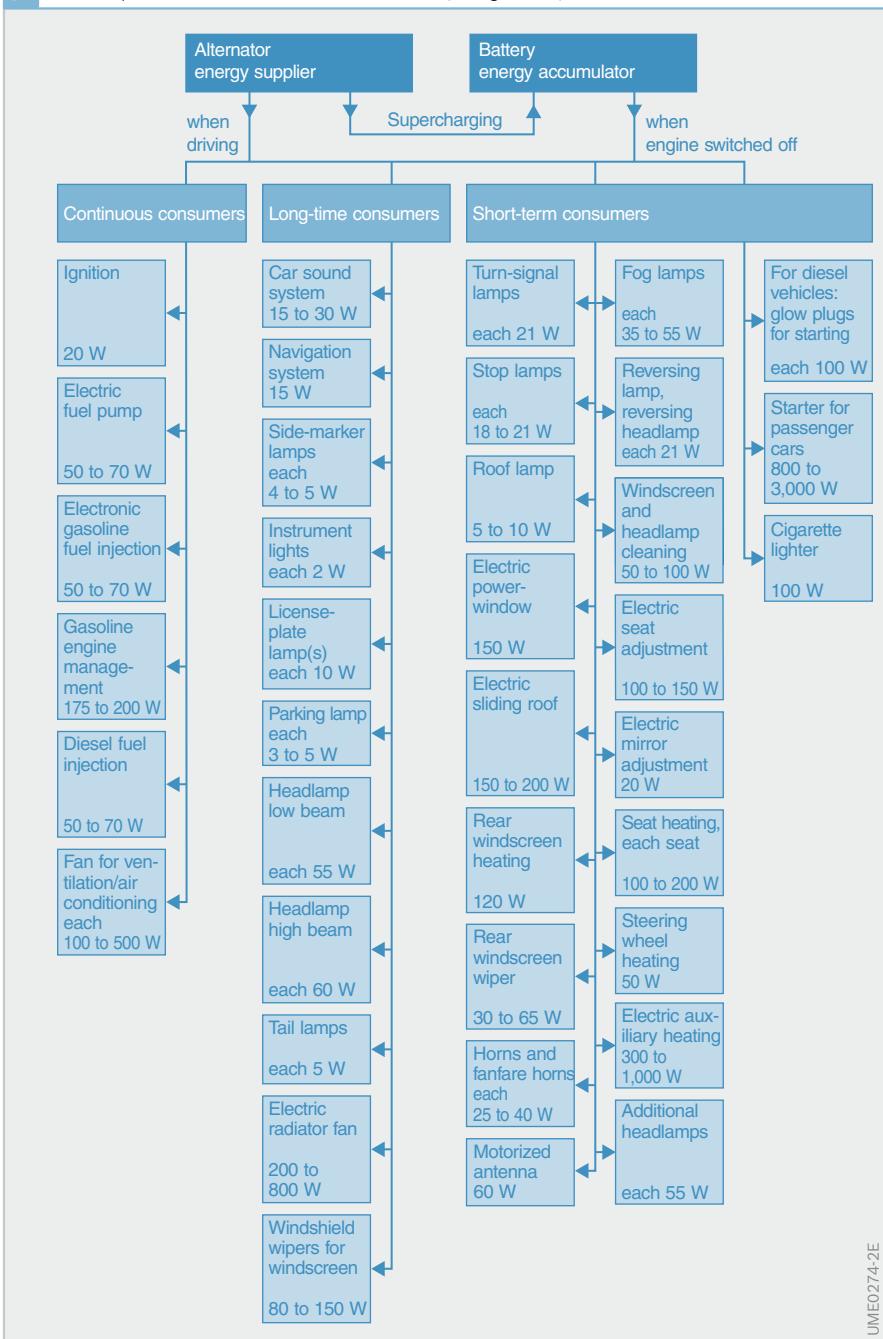


Fig. 33

- 1 Drive end shield
- 2 Housing
- 3 Stator winding
- 4 Rotor
- 5 3-phase connection (remote rectifier and voltage regulator)
- 6 End cap
- 7 Ventilator
- 8 Slipping end shield
- 9 Slipping
- 10 Drive shaft

34 Power requirements of the consumers in the vehicle (average values)



Starting systems

Internal-combustion engines must be cranked by a starter at a minimum speed before they can supply sufficient energy in sustained operation from the combustion cycles to cover the momentary requirements for the compression and gas exchange cycles. When an engine is first started, the bearing surfaces are not adequately lubricated so that high levels of friction have to be overcome when cranking the engine.

Overview

Electric motors (direct current, alternating current and three-phase AC motors), and also hydraulic and pneumatic motors, are used to start internal-combustion engines. The electrical direct current series motor is particularly suitable as a starter motor, as it develops the high initial torque required to overcome the engine's initial resistance to rotation and to accelerate the masses in the engine.

The energy required for starting is generally drawn from the battery, which also supplies other electrical components in the vehicle electrical system.

The torque from the starter is more often than not transferred by way of a pinion and a ring gear on the engine, but in some cases a V-belt, toothed belt or chain is used, or it is applied directly to the crankshaft.

In order to start the engine, the starter-motor pinion (Fig. 1, Item 7) engages in the ring gear on the engine (12). The ring gear has typically approx. 130 teeth (103 to 144 on passenger cars, 110 to 160 on commercial vehicles) and is fitted on the engine flywheel on vehicles with manual transmission, while on vehicles with automatic transmission, it is fitted on the torque-converter housing. When in its initial position, the starter pinion, which typically has 10 teeth (8 to 10 in passenger cars, 9 to 13 in commercial vehicles) is disengaged from the ring gear, but remains only a few millimeters away from it. When the driver turns

the ignition key to the start position, a mechanical connection is first established between the starter motor and the ring gear (meshing), so as to then be able to transfer the torque of the starter motor, which then starts running, to the engine.

Because of the high conversion ratio between the starter pinion and the ring gear, the starter may be designed for high speeds and low torque. This allows the dimensions and the weight of the starter to be kept small.

Starter

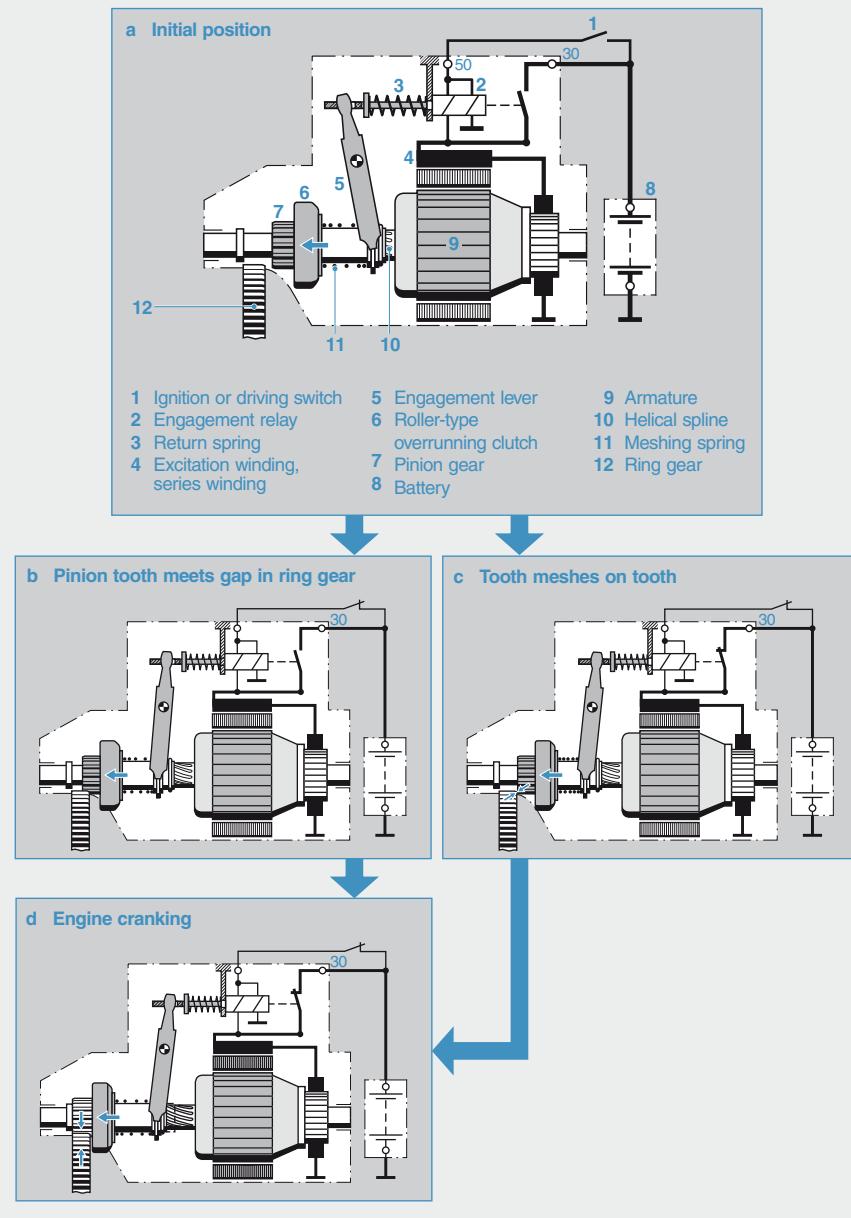
Working phases of the pre-engaged-drive starter

The pre-engaged-drive starter has established itself the worldwide standard for vehicles. The meshing travel consists of thrust and helical travel.

Meshing

The ignition lock in the start position closes the ignition/starting switch (Fig. 1a, Item 1), which triggers the engagement relay (2). The magnetic field created by the relay coil energizes the relay armature, thus operating the engagement lever (5) so that the pinion gear (7) is moved forwards and comes into contact with the ring gear (12) (pinion travel). A helical spline (10) on the shaft turns the pinion slightly against the starter's driving direction in the pre-meshing process, thus making the meshing process easier.

Under ideal circumstances, one tooth of the pinion will encounter a space between the teeth in the ring gear, thereby establishing the coupling between the starter and the internal-combustion engine (Fig. 1b). During the movement of the pinion into the ring gear, the engagement relay armature closes the starter motor's main circuit at the contact bridge. The starter motor thus starts to rotate so that its drive shaft and pinion gear drive the ring gear, with the result that the engine also starts to turn (Fig. 1d).

1 Working phases of the pre-engaged-drive starter

This ideal meshing of the pinion with the ring gear is not, however, the normal case, as the 0.4 mm backlash between the teeth leaves very little leeway for smooth meshing.

In many cases (approx. 50 to 80 % of starts, depending on the geometrical conditions), there is a clash in which the teeth of the pinion and those of the ring gear block each other (Fig. 1c). The pinion can then not move any further outwards and mesh with the ring gear unless it is rotated. However, in this case too, the engagement relay armature continues to be drawn further into the coil so that the engagement lever compresses the meshing spring (Fig. 1a, Item 11). Consequently, the pinion is pressed with increasing force against the ring gear.

As it reaches the end of its travel, the engagement relay closes the switch in the starter-motor main circuit. As a result, the pinion starts to turn and at some point reaches a position where it will mesh with the ring gear. The pre-stressed meshing spring now forces the pinion forwards via

the helical spline (helical travel) until it meshes fully in the ring gear. The pinion is pushed into the ring gear by the screw effect of the helical spline in conjunction with the rotary motion of the starter motor. The helical spline prevents transmission of torque to the engine until the pinion is fully meshed in the ring gear.

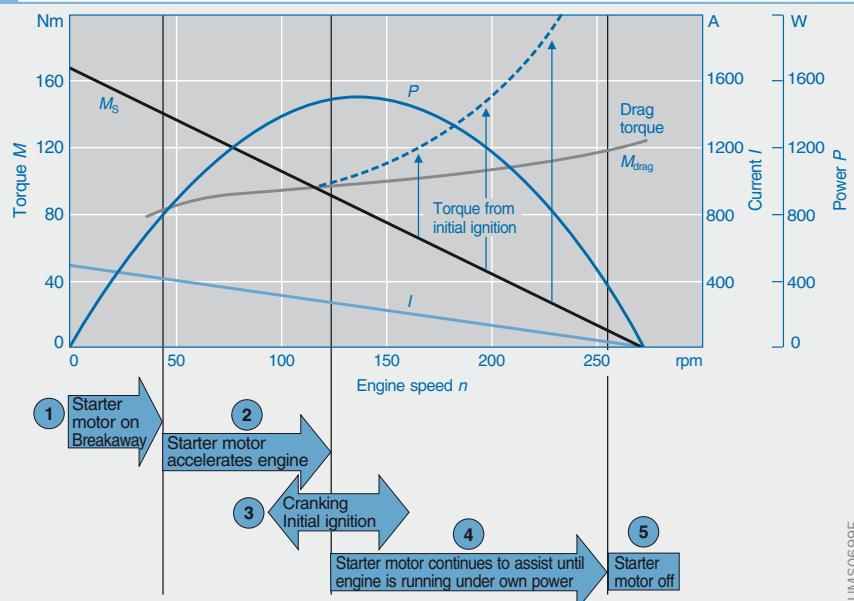
Overloading the gear teeth must be avoided in commercial vehicles because of the high torques which are transferred. Appropriate meshing principles ensure that there is sufficient overlap of pinion and ring gear teeth before the starter motor delivers its full power.

Breakaway and cranking

Once the pinion is meshed, the starter transfers the torque to the engine crank-shaft.

The torque delivered by the starter is at its highest when first switched from standstill and decreases continuously as speed increases. The surplus torque provided

2 Phases of the starting procedure



by the starter motor at the beginning of the starting sequence initially overcomes the static friction in the engine bearing surfaces (breakaway) and then accelerates the moving masses in the engine and ancillaries to the turnover speed. This is characterized by the intersection of the engine and starter torque curves (Fig. 2), i.e. at higher speeds, the starter-motor torque is no longer sufficient to continue driving the engine. The turnover speed must therefore be higher than the minimum cranking speed for a successful start.

Starting and overrunning

A warm engine will normally take only two revolutions of the crankshaft to start. A cold engine may have to be cranked for a little longer (up to a minute is not uncommon). If, for whatever reason, the fuel system needs to be purged of air when the engine is started, it may have to be turned over for as long as 20 to 30 seconds before it starts.

The engine-management system uses sensors to detect the position of the crankshaft within the first two revolutions (initiation of the fuel-injection system). Their signals are used to selectively deliver fuel to the cylinder that will be the next to perform its compression stroke. This ensures a rapid starting sequence with low levels of exhaust emissions.

As soon as fuel is injected/ignited, the internal-combustion engine starts to generate its own torque and, therefore, to increase its speed of rotation, although this is not sufficient for starting in all cases. To attain the air-fuel mixture required by the gasoline engine or, as the case may be, the spontaneous-ignition temperature required by the diesel engine for sustained operation, the starter may have to continue to assist the internal-combustion engine to reach the minimum sustained operation speed. After only a few ignition cycles, the engine accelerates so sharply that the starter can no longer match its

speed and is outrun. At this point, it is essential that the roller-type overrunning clutch disengages the pinion from the starter-motor drive shaft. By doing so, it protects the starter motor against overspeed and prevents excessive wear.

As soon as the driver releases the ignition key, the solenoid-switch circuit is broken. The return spring moves the relay armature back towards its initial position as a result of which the primary-circuit switch opens. The demeshing spring completes the retraction of the pinion, disengaging it from the ring gear. Demeshing is here supported by the rotary motion of the pinion as it returns over the helical spline. The starter motor freewheels to a standstill and the entire mechanism returns to its initial position.

Switch-off function

The relay armature is connected to the engagement lever in such a way that there is a certain amount of free travel or "play". If the engine fails to achieve sustained operation (e.g. due to lack of fuel), the starting sequence must be aborted. At that point, the pinion is fully engaged, and pinion and ring gear are under maximum load.

When the power supply to the relay is switched off, there must be sufficient relay armature free travel to allow it to open the primary-circuit switch. If this were not the case, the engagement lever would hold the relay armature in place. The primary-circuit switch would remain closed and it would not be possible to stop the starter motor.

On commercial-vehicle starters, the geometrical constraints demand that a cutout spring (Fig. 3) is used instead of the free-travel arrangement explained above. The force of this spring when the pinion is in its initial position overcomes the force of the relay armature return spring and pushes the armature back against the stop on the lever. When the relay armature is

pulled in, on the other hand, the force of the armature return spring must be sufficient to compress the cutout spring enough for the contact bridge of the switch contacts to be lifted away.

Engagement relay

Relays are used to switch a high current by means of a relatively low control current. The starter current can be as high as 1,500 A on vehicles and as much as 2,500 A on commercial vehicles. As contacts for currents of this level are heavily loaded, the use of a power relay is an absolute requirement. The relay is actuated by a relatively low control current (relay current) of approx. 30 A for vehicles and up to approx. 80 A in commercial vehicles. A simple mechanical switch (ignition/start-ing switch, starter button) or a mini-relay operated by the engine control unit can then be used to switch the engine on.

The engagement relay (Fig. 4) built into the starter is the combination of an engagement solenoid and a relay. It fulfills two functions:

- ▶ It advances the drive pinion so that it meshes with the engine's ring gear
- ▶ It closes the contact bridge which completes the starter's primary electric circuit

The magnetic core (4) is firmly fixed to the housing and protrudes into the solenoid winding (2, 3) from one side, while the movable relay armature (1) protrudes from the other side. The distance between the magnetic core and the armature when at rest represents the total travel of the relay armature. The solenoid housing, magnetic core and relay armature together form the magnetic circuit.

The solenoid force created when the current is switched on draws the armature into the winding. That armature movement firstly causes the engagement lever to displace the pinion along its longitudinal axis, and secondly forces the contact bridge (8) against the main current contacts (6).

The relay winding in most designs consists of a pull-in winding and a hold-in winding (Fig. 4, Items 2 and 3, and Fig. 5, Items 4a and 4b). This division of tasks is particularly effective with regard to thermal stress capacity and the magnetic forces which can be achieved.

In the initial position (no current applied to the relay), the air gap between armature and magnetic core is relatively large. Only a high magnetomotive force can overcome meshing resistance.

The air gap closes up as the armature retracts and the solenoid force increases

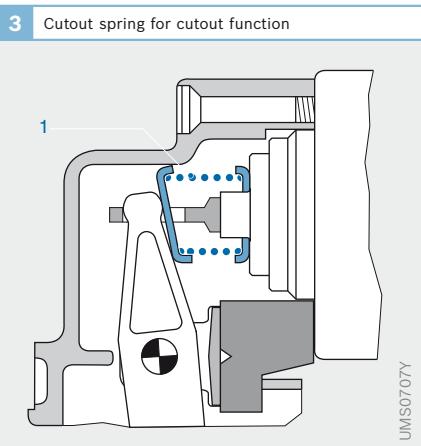
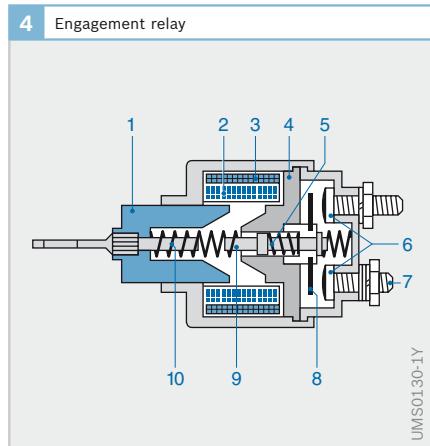


Fig. 3

- 1 Relay with cutout spring

Fig. 4

- 1 Relay armature
- 2 Pull-in winding
- 3 Hold-in winding
- 4 Magnetic core
- 5 Contact spring
- 6 Contacts
- 7 Electrical connection
- 8 Contact bridge
- 9 Switch shaft (split)
- 10 Return spring



significantly. When the relay armature is fully retracted, i.e. when there is practically no residual air gap, the solenoid force of the hold-in winding on its own is sufficient to hold the armature in position until the starting sequence is completed. The pull-in winding is short-circuited across the main-circuit switch and the ignition switch when the armature is pulled in.

It is important that the pull-in winding and the hold-in winding have the same number of turns. Otherwise, the solenoid switch might adopt a self-locking characteristic during switching off due to the supply of the windings, which would then be connected in series, in reverse through terminal 45. Having the same number of turns in each winding ensures that the magnetic fields of the two coils through which the current now flows in opposite directions cancel each other out and the relay switches off reliably.

Since the starter causes a substantial voltage drop in the vehicle electrical system when its primary current is switched on, the hold-in winding must be dimensioned so as to be capable of reliably holding the armature in place even at supply voltages significantly below half the nominal battery voltage. Otherwise, the relay switch might under certain circumstances repeat-

edly switch on and off in rapid succession, resulting in damage to the switch contacts.

Return springs between the individual components ensure that the switch opens again when the solenoid is switched off and the relay armature returns to its initial position.

All starter terminals (terminals 50, 30 and 45) are usefully incorporated within an assembly on the cap of the engagement relay. The engagement relay cap is usually made of duromer material which provides the required mechanical loadability even at high levels of thermal stress (up to 180 °C for short periods) and thus ensures that the engagement relay and all its components are held together reliably. In order to ensure reliable retraction and switching of the relay even at high temperatures, there are special winding designs that can also withstand temperatures of up to 180 °C.

The primary-circuit switch contacts on car starters generally consist of steel studs with riveted copper contacts that provide reliable switching characteristics and minimal switch impedance. The relay switches for commercial-vehicle starter motors are manufactured as one-piece components from a high-strength copper alloy in order to provide high current capacity. A conical form of contact is increasingly being used to reduce bounce, and thus the wear, on the contacts in commercial vehicle starters.

Depending on the type of wiring harness used, the control-circuit terminal (terminal 50) may take the form of a screw terminal or some type of plug-in contacts which may or may not be sealed.

Engagement relays for vehicles normally have a round plug connector to connect the cap to the relay winding. This allows a closed-surface cap design without riveted or soldered connections or the holes they would require. In addition, the standard arrangement provides for a seal between the cap and the relay housing.

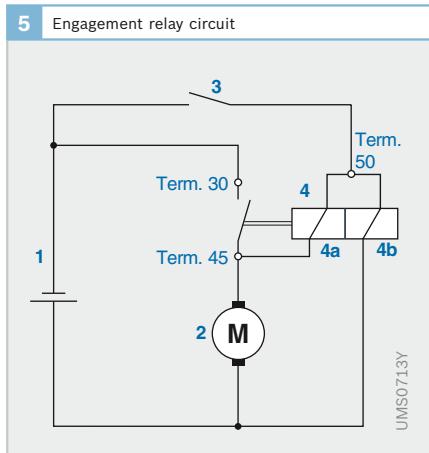


Fig. 5

- 1 Battery
- 2 Starter
- 3 Ignition lock
- 4 Engagement relay
- 4a Pull-in winding
- 4b Hold-in winding

For situations requiring special sealing arrangements, the relay armature can be protected by a flexible rubber boot (Fig. 6, Item 2) in order to prevent the ingress of humidity from the starter's pinion-engaging drive. This measure is primarily adopted for applications in which it is not possible to entirely prevent the entry of water into the drive end shield. Another area where this armature boot may be used are those installation positions where the engagement relay is pointing downwards, with the result that moisture can collect in the drive end shield cover.

Some large commercial-vehicle starters do not have an integral engagement relay; instead, the engagement solenoid for pinion advance and the control relay for electrical switching stages are separate from one another.

One-way clutch

In all starter designs, the torque is transmitted from the starter motor to the pinion through a one-way clutch (overrunning clutch). The task of the one-way clutch is to transmit the drive to the pinion when the drive shaft is driven and to disengage the pinion from the drive shaft once the pinion is rotating more quickly. The one-way clutch therefore prevents the armature of the starter motor from being

accelerated to excessive speeds once the engine has started.

Roller-type overrunning clutches for starters can be either frictional (roller and multi-plate overrunning clutch) or positively interlocking (spur-gear overrunning clutch) in design.

Roller-type overrunning clutch

Pre-engaged-drive starters normally have roller-type overrunning clutches (Fig. 7). The driver and clutch shell (3) is connected to the drive shaft by means of a helical spline. The frictional connection between the central cylindrical shaft of the pinion and the clutch shell of the driver surrounding it is formed by cylindrical rollers (5) that are able to move within the roller race (4).

When the mechanism is at rest, the springs (7) force the rollers into the constricted space between the roller race in the clutch shell and the pinion shaft (6). Because of the constriction, the rollers jam between the surfaces of the roller race and the pinion shaft, so that when the starter motor starts running, the pinion is turned at the same time.

Once the overrunning process commences, the friction between the pinion shaft and the rollers pushes the rollers against the spring force into the wider part of the roller race; the force of the springs

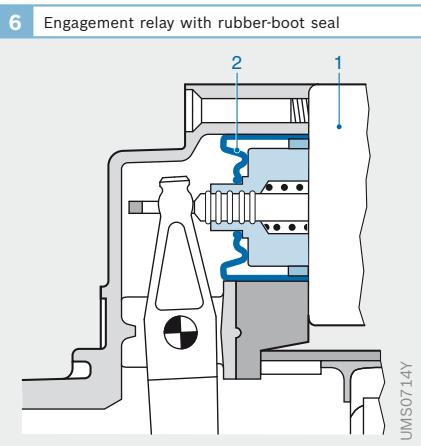
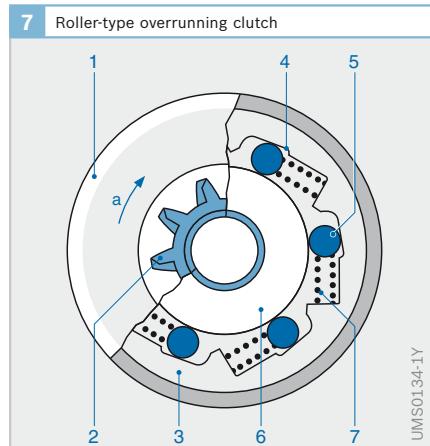


Fig. 6

- 1 Engagement relay
- 2 Rubber boot

Fig. 7

- 1 Plug cap
- 2 Pinion
- 3 Driver and clutch shell
- 4 Roller race
- 5 Cylindrical roller
- 6 Pinion shaft
- 7 Spring
- a Direction of rotation



ensures that the rollers are held in contact with the pinion shaft and roller race at all times. The resulting overrunning torque is relatively small and has very little effect as far as the idle speed of the starter-motor armature is concerned.

Spur-gear overrunning clutch

A spur-gear overrunning clutch (Fig. 8) is fitted on Type JE and KE sliding-gear starters in combination with a mechanical two-stage pinion-engaging drive. The roller-type overrunning clutch transfers positive torque by means of a spur gear.

During the overrunning process, the ring gear on the engine's flywheel is driving the pinion (1) which is connected to the clutch race (4) by a spur gear. Due to the sawtooth form of the spur gear, the clutch race is forced inwards along the helical gear towards the starter motor when the pinion is overrunning. Separation of the pinion and clutch race is assisted by three flyweights (2) which apply an axial force on the clutch race via a tapered thrust ring (3).

Multi-plate overrunning clutch

The multi-plate overrunning clutch is used on larger sliding-gear starters (KB, QB, QF, TB, TF). The frictional connection between armature and starter pinion is achieved by means of a circular laminated core (Fig. 9,

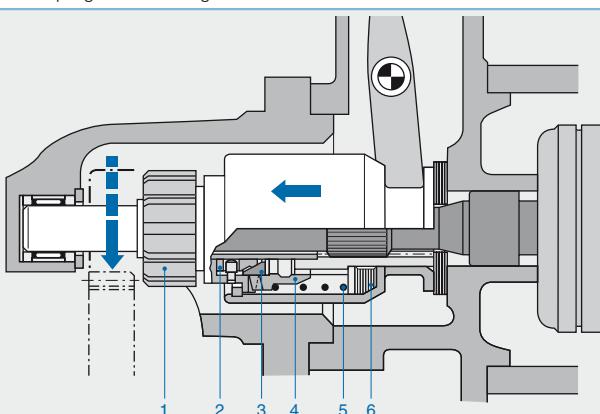
Item 3). The plates are alternately engaged by driver cams with the drive flange (1) and with the clutch race (4). The individual plates are mounted so that they can move in the axial direction, but cannot be twisted radially against the drive flange or clutch race. The external drive flange is firmly connected to the armature shaft. The clutch race, on the other hand, can be rotated helically on the drive spindle's helical spline (7).

Frictional connection

In order that the multi-plate overrunning clutch can use friction to transmit the drive from the motor to the pinion, there has to be a certain amount of pressure holding the plates against one another (Fig. 9a). When the clutch is in its initial position, the laminated core (3) is compressed by a light, pre-tensioned spring force on a shaft spring (on K starters; 8) or by studs and springs (on Q and T starters) to such an extent that the friction created is able to drive the clutch race (4).

Once the pinion has meshed and reached the limit of its outward travel, full frictional connection comes into operation for starting. Because the pinion is held stationary while the armature shaft is rotating, the clutch race slides outwards on the helical

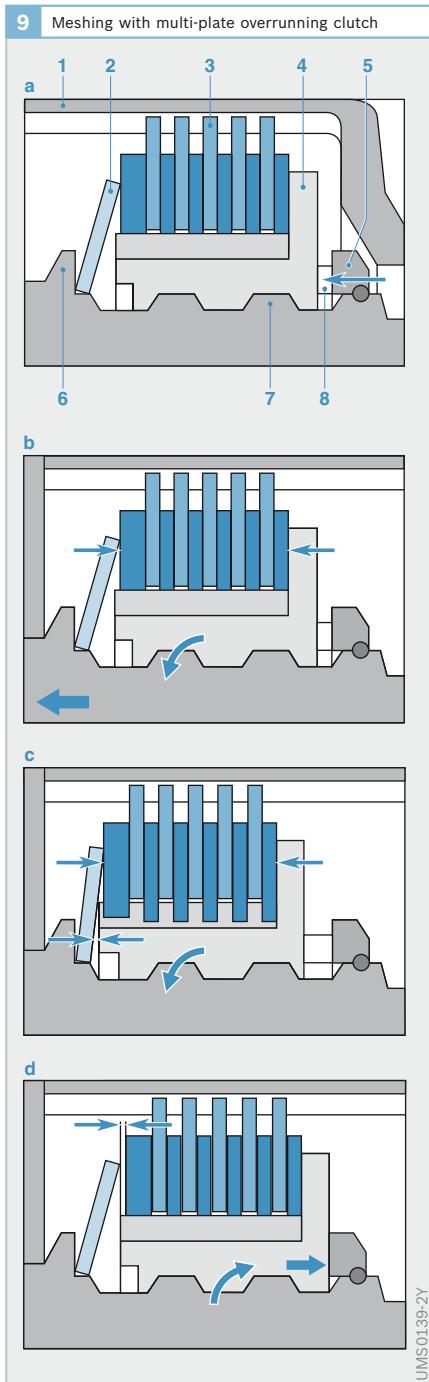
8 Meshing drive with spur-gear overrunning clutch



UMS0140-1Y

Fig. 8

- 1 Pinion with spur gear teeth
- 2 Flyweights
- 3 Internally tapered thrust ring
- 4 Clutch race with spur gear
- 5 Spring
- 6 Rubber bumper



spline (7) of the output shaft and presses against the disk spring. This further compresses the laminated core and increases the pressure between the individual plates. Compression continues to increase until the friction between the plates is sufficient to transmit the torque required to start the engine. Frictional connection is transmitted as follows: armature - drive flange - outer plates - inner plates - clutch race - drive spindle - pinion (Fig. 9b).

Torque limitation

The increasing plate pressure due to the helix effect of the clutch race, and the resulting torque transmitted, are limited by the fact that, on reaching the maximum load, the end face of the clutch race comes into contact with the disk spring. In doing so, it uses its end face to press the disk spring against the stop collar (6) of the output shaft (Fig. 9c). This prevents the laminated core being compressed any further. The multi-plate overrunning clutch then functions as an overload clutch since the plates will slip if the maximum power transmission rating, and therefore the maximum tightening torque, is exceeded.

Removal of frictional connection (overrunning)

When the engine flywheel accelerates due to ignition pulses or the engine starting, the pinion starts to turn faster than the starter motor. As a result, the direction of force transmission is reversed and the clutch race is twisted along the helical spline towards the armature until it meets the stop ring (Fig. 9d, Item 5). The disk spring is released and no longer presses against the laminated core. The plates consequently separate from one another and the frictional connection between them is removed.

Direct-drive and reduction-gear starters

In a conventional direct-drive starter motor, the starter pinion rotates at the same speed as the armature. The one-way clutch and the pinion are mounted directly on the

starter motor's armature shaft. In order to be able to deliver the high torque required for cold starting, the motor has to be relatively large - and therefore heavy - with this type of design. For automotive applications, therefore, this design is only usual for power levels < 1 kW, while in commercial vehicle applications, all more recent starter types in the RE/HE(F) series are reduction-gear starters.

Thanks to the use of a planetary gear, the same torque is achieved on the reduction-gear starter with a smaller and faster rotating electric motor. Depending on the design, this can allow a weight saving of 30 to 40 %.

Planetary gearing systems generally offer the advantage of combining compact overall dimensions with high gearing ratios. If the engagement geometry is favorable, this allows the achievement of high torque output without generating high levels of noise. Furthermore, they do not generate any external transverse forces so that the armature-shaft and pinion-shaft bearings are not subjected to high stresses even at high power outputs.

The planetary gearing system (Fig. 10) used in starters has a fixed internal gear (3). The drive is supplied via the

sun gear (2) which is attached to the armature shaft of the electric motor. The planet gears (1) are hence in engagement with both the sun gear and the internal gear. As the planet gears orbit around the sun gear, their bearing-shaft journals drive the drive shaft which carries the helical spline and one-way clutch.

The transmission ratio between the armature and the pinion is variable across a broad range from approx. 3 : 1 to 6 : 1 depending on the design of the planetary gear. This allows optimum matching of the starter to the characteristics of the engine and the vehicle's electrical network.

Lower transmission ratios allow higher engine speeds for warm starting, while a higher conversion ratio enables the starter to cope with extreme cold-starting conditions and result in lower power consumption by the starter when turning the engine over.

The reduction-gear starter offers further benefits. For instance, the high armature speed obtained as a result of the high conversion ratio creates a flywheel effect which helps to turn the internal-combustion engine through top dead center cylinder positions and smooth out engine-speed fluctuations. As a result, the engine has a higher momentary speed precisely at the point when fuel is injected or at ignition, and this has a positive effect on the fuel-injection pattern (and therefore the ease with which the engine starts) and exhaust-gas values.

On internal-combustion engines with smaller numbers of cylinders, this flywheel effect also has the advantage that the high peak-torque levels of the individual cylinders can be reliably overcome with a relatively low amount of starting power.

End shield type

There are two possible arrangements for the pinion end bearing (Fig. 11).

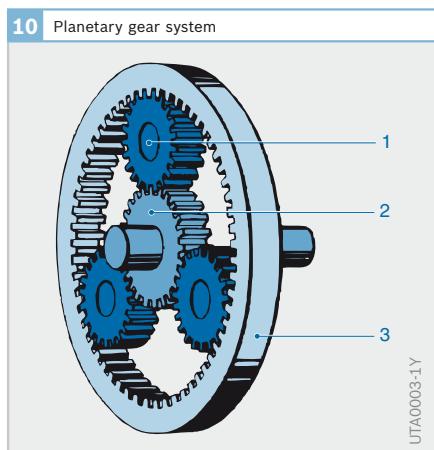


Fig. 10

- 1 Planet gear
- 2 Sun gear
- 3 Internal gear

- ▶ Most starters have a drive end shield which has a “mouth” close to the engine flywheel. This mouth carries the bearing hub and has an opening near the flywheel. The bearings used can be plain or roller bearings (in some cases sealed). The meshing drive (pinion and roller-type overrunning clutch) is open towards the ring gear.
- ▶ For applications where the pinion-engaging drive would be subjected to high levels of dirt, moisture or clutch-plate dust, the closed-shield design is preferable. The pinion bearing surfaces are completely enclosed within the drive end shield. Shaft seals provide such effective protection that these starters are suitable for even the most unfavorable conditions. In order to be able to cope with the higher bearing loads and resist wear and tear throughout the full service life of the motor, the pinion bearing surfaces are suitably reinforced.

Closed-shield starters also have advantages in terms of fitting because they do not require a mounting flange that is shaped to match the “mouth” opening. Consequently, they allow more scope for choice of fitted position and the space requirements in the bell housing are smaller.

Nevertheless, the following features prevent the general use of this design:

- ▶ The bearing forces in the drive end shield and the effective radius of friction are about twice as great as for an open-shield starter, with the result that friction-related power losses are greater (rolling bearings can help to alleviate this problem and also improve wear and noise characteristics).
- ▶ Bearings, seals and elongated starter pinion design result in higher costs.
- ▶ The overall length of the starter increased by this design.
- ▶ The gearing ratios, and therefore the noise characteristics, are less favorable.

Other types of starter motor

Starters for commercial vehicles with electrical two-stage pinion-engaging system

Various starters in the higher performance range for commercial vehicles (HEF95-L, HEF95-M) are equipped with an electrically operated, two-stage pinion-engaging system (Fig. 12) which permits secure meshing while minimizing ring gear wear. The first stage of the sequence, with the starter motor turning gently, only involves starter pinion meshing; at this stage the starter is not cranking the internal-com-

11 Pinion bearing on open-shield starter and on closed-shield starters

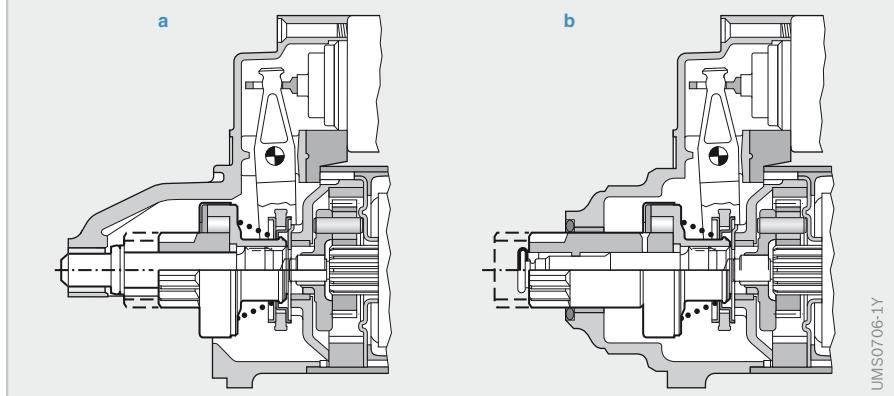
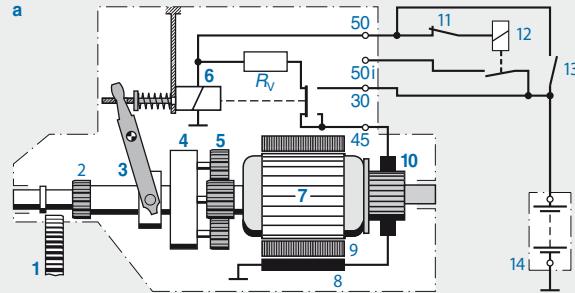


Fig. 11

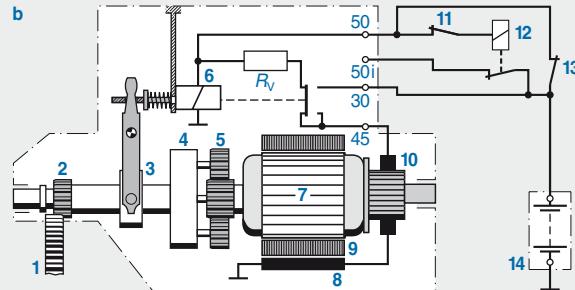
- 1 Open-shield starter
- 2 Closed-shield starter

12 Electrical two-stage engaging system
Initial position

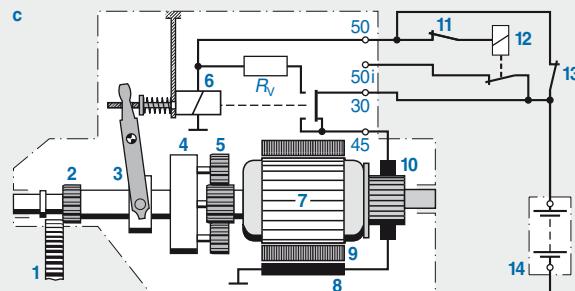
Starter disconnected from power supply

**Pinion tooth meets gap in ring gear**

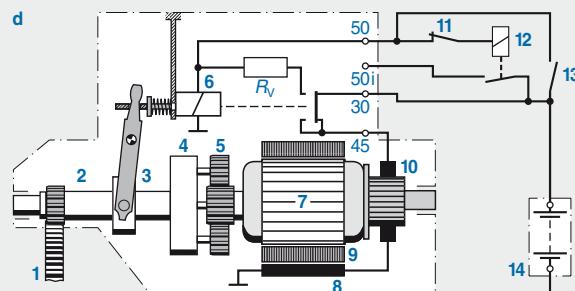
Starter switched on, favorable meshing position (1st switching stage)

**Tooth meshes on tooth**

Starter switched on, meshing is difficult (1st switching stage). Meshing spring ensures that meshing occurs. Primary current switched on.

**Engine cranking**

End position (2nd switching stage). Starter at full torque.

**Fig. 12**

- 1 Ring gear
- 2 Pinion
- 3 Engagement lever
- 4 Roller-type overrunning clutch
- 5 Planetary gear
- 6 Engagement relay
- 7 Armature
- 8 Excitation winding
- 9 Pole shoe
- 10 Commutator
- 11 Thermostatic switch
- 12 Type IMR pilot solenoid
- 13 Ignition/starting switch
- 14 Battery

bustion engine. The main circuit is only closed in the second stage. Two-stage meshing is implemented through a pilot solenoid and an engagement relay. The meshing spring is built into the engagement relay.

1st switching stage (preliminary stage)

The voltage signal from the ignition lock or an electronic control unit initially switches the pilot solenoid which is capable of switching the high current of approx. 180 to 200 A (in 24 V starters).

The current then flows through a series resistor or a pull-in and hold-in winding. First this causes the engagement lever to move the one-way clutch with the pinion shaft in the axial direction, and second, the electric motor is started softly. As a result, the pinion can normally be fully meshed in the ring gear before the maximum current strength is applied to the motor by the closure of the primary-circuit switch.

2nd switching stage (main stage)

When the pinion has successfully meshed with the ring gear, the relay armature disconnects the circuit through the series resistor or the pull-in winding by means of an NC connector shortly before it reaches its fully retracted position. A few milliseconds later the primary current is completed and the starter begins to generate its full torque output.

In some cases, the pinion may not engage with the ring gear if the two gears are not aligned favorably. In this case, the meshing spring on the relay armature makes sure that the primary current is completed before the pinion is fully engaged with the ring gear. Single-stage meshing is the consequence.

Sliding-gear starters for commercial vehicles with electrical pinion rotation

Sliding-gear starters incorporating motor-assisted pinion rotation (Types KB, QB, QF, TB, TF, Fig. 13) are used to start

large-scale internal-combustion engines. They have a two-stage electrical pinion-engagement system in order to reduce the loads on the pinion and the ring gear.

1st switching stage (preliminary stage)

In the first switching stage, therefore, the starter pinion is moved outwards along the motor axis and simultaneously slowly rotated in order to allow smooth meshing with the ring gear.

When the starting switch is operated, the control relay and the hold-in winding of the engagement solenoid are energized. The control relay then immediately completes the pull-in winding circuit. The engaging armature then pushes the pinion against the engine ring gear by means of the pinion-engaging axle and the drive spindle.

At the same time, the shunt winding which initially is connected in series with the starter armature winding is energized. It acts in conjunction with the engagement solenoid pull-in winding as a series resistor for the starter armature winding. This circuit limits the armature current so that the starter armature can only produce a small amount of torque and, therefore, only turns very slowly.

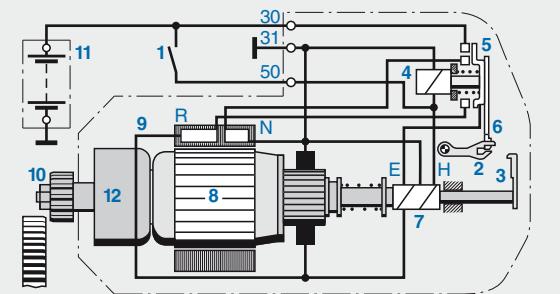
If the pinion teeth are not in a position where they can mesh with the ring gear when the two gears meet, the pinion continues to rotate until it is able to mesh with the ring gear. The pinion and ring-gear teeth can sometimes meet in such a way that the pinion can neither rotate nor slide outwards into engagement with the ring gear (known as tooth-on-tooth); in such cases, the starting sequence must be aborted and repeated.

2nd switching stage (main stage)

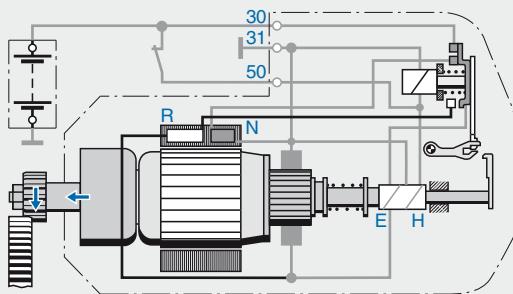
Immediately before the pinion reaches the end of its outward travel, a release lever lifts up a tripping lever and releases the control-relay contact bridge. A pre-loaded spring can then abruptly close the switch

from power supply

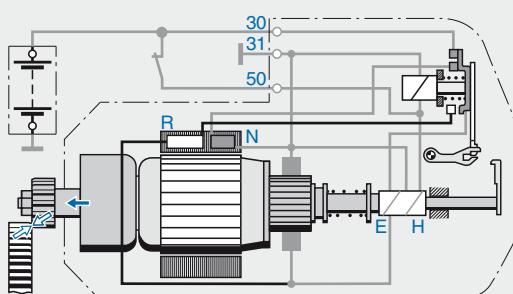
Initial position
Starter disconnected



Pinion tooth meets gap in ring gear
Starter switched on, favorable meshing position (1st switching stage)



Tooth meshes on tooth
Starter switched on, meshing not possible (1st switching stage). Start attempt must be repeated.



Engine cranking
Final position (2nd switching stage). Starter at full torque.

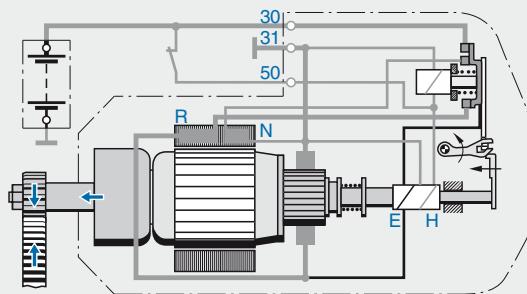


Fig. 13

1 Ignition or driving switch
2 Inhibitor pawl
3 Release lever
4 Control relay
5 Contact bridge
6 Stop
7 Engagement solenoid
8 Armature
9 Excitation winding
10 Pinion
11 Battery
12 Multi-pile overrunning clutch
E Pull-in winding
H Hold-in winding
N Shunt winding
R Series winding

UMS0161-1Y

contacts. The full current is then applied to the starter motor so that it can generate its maximum torque output in order to turn over the internal-combustion engine by means of the multi-plate overrunning clutch and the pinion.

Sliding-gear starters for commercial vehicles with mechanical pinion rotation

Sliding-gear starters with mechanical pinion rotation (Types JE and KE) are used to start large-scale internal-combustion engines. They have a two-stage mechanical pinion-engagement sequence in order to reduce the loads on the pinion and the ring gear. The most important functional component of these starters is the spur-gear overrunning clutch.

1st pinion-engagement stage

When the ignition/starting switch is operated, the engagement relay initially moves the pinion-engagement lever against the action of the return spring. The engagement lever slides the spur-gear overrunning clutch assembly in a straight line along the straight-tooth gearing until it meets the ring gear. If the pinion happens to be in a position where it can mesh with the ring gear, it will continue to move outwards to the full extent of its linear travel.

2nd pinion-engagement stage

If the pinion is not in a position where it can mesh with the ring gear, the remaining components of the spur-gear overrunning clutch continue to move outwards on a linear path towards the ring gear. The action of the clutch-race helix forces the pinion to rotate in its working direction while simultaneously extending the meshing spring of the roller-type overrunning clutch. The pinion teeth slip past the teeth of the ring gear to the next gap in which the pinion can mesh under the force of the stretched meshing spring.

The pinion and ring-gear teeth can sometimes meet in such a way that the pinion can neither rotate nor slide outwards into engagement with the ring gear (known as tooth-on-tooth); in such cases, the starting sequence must be aborted and repeated. When this happens, the helical spline in the spur-gear overrunning clutch forces the starter armature to rotate in the opposite direction to its normal direction of rotation (armature reversal) when the overrunning clutch is being moved by the engagement lever. The next time the starting sequence is performed, the pinion is then in a better position relative to the ring gear and is able to mesh.

Spur-gear overrunning-clutch and relay are precisely matched to each another so that the relay contact (primary current contact) do not close until meshing is completed. Consequently, the starter cannot develop its full torque until that point.

Starting systems

Starting systems for vehicles

Vehicle starting systems are generally equipped with pre-engaged-drive starters up to a rated output of approx. 2.5 kW. The nominal voltage everywhere is 12 V. These can start gasoline engines up to approx. 7 liters engine displacement and diesel engines up to approx. 3 liters. The starter performance range depends on the combustion system: a diesel engine needs a higher performance starter than a gasoline engine of the same engine capacity.

The circuit of vehicle starting systems is generally a relatively simple design. The internal-combustion engine is near the driver, who consequently can generally also monitor the starting process by ear. Once the engine is started, it can be heard to run so that it is not likely that the starter will be unintentionally switched on again, meshing the starter pinion in an engine-flywheel ring gear that is already rotating. This means that no equipment

to protect and monitor the starting process is generally required in a vehicle. In many vehicle models, an ignition/starting switch with additional start-repeating block is fitted, to prevent inadvertent operation of the starter.

In newer luxury class vehicles, automatic control (switching on/off) with engine speed detection has recently become standard.

Starting systems for gasoline-engine vehicles

The starting system is one of the systems controlled via a generally multistage ignition switch. The ignition system is switched on before the "Start" switch position is reached, as without its intervention it is not possible to start the gasoline engine and to achieve sustained operation. The ignition process continues after the starter is switched off and allows sustained operation of the gasoline engine.

Starting systems for diesel-engine vehicles

The pre-glow system must be switched on before the starting process can commence. The pre-glow system has a combined driv-

ing-glow start switch, which can be connected directly to the starter once the pre-heating time is complete. Once the starting process is completed, the diesel engine's pre-glow system is switched off along with the starter.

Starting systems for commercial vehicles

Medium-sized commercial vehicles with diesel engines up to approx. 12 liters engine displacement have starting system with 12 V or 24 V nominal voltages.

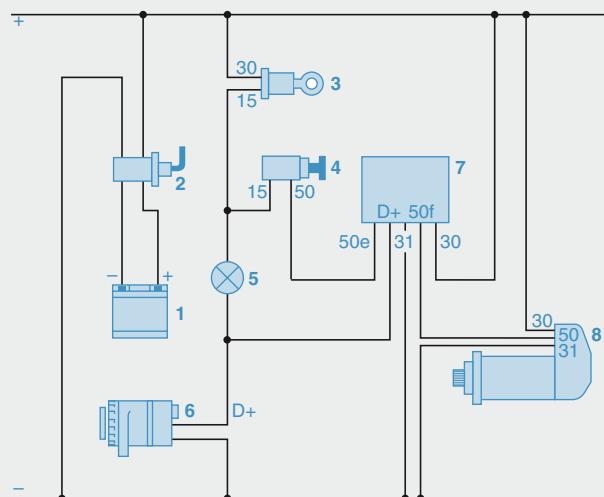
Heavy commercial vehicles with diesel engines up to approx. 24 liters engine displacement use only 24 V starting systems, supplied by two 12 V batteries connected in series.

Occasionally, mixed 12/24 V systems with a 12 V vehicle electrical system and a 24 V starter voltage are used.

Starting systems with start-locking unit

Starting system on vehicles in which the starting process cannot be clearly monitored by ear (e.g. on busses with a rear

14 Circuit for a starting system with an electronic start-inhibit relay



UMS0187-1Y

Fig. 14

- 1 Battery
- 2 Battery master switch
- 3 Driving switch
- 4 Ignition/starting switch
- 5 Alternator indicator lamp
- 6 Alternator
- 7 Electronic start-inhibit relay
- 8 Starter

engine), require more complex circuitry, as they require effective protection for the starter and engine ring gear.

A starting system with an electronic start-inhibit relay (Fig. 14) protects the starting system in a number of ways:

- ▶ Switching-off once start is successful
- ▶ Inhibiting start if the engine is already running
- ▶ Inhibiting start if the engine is still running down
- ▶ Inhibiting start after failed start, if the engine was unable to sustain operation

In the last two cases, it is only possible to attempt starting again after an inhibition time specified in the relay. This function is nowadays more and more frequently integrated into the engine control unit.

Starting system with 12/24 V battery changeover

Various commercial vehicles have mixed 12/24 V installations. In such cases, all electrical components, with the exception of the starter, including the alternator are designed for nominal 12 V operation. By contrast, the starter is run at a nominal

voltage of 24 V. This permits the power output required for starting large engines.

12/24 V systems are equipped with a battery changeover relay for this purpose. The two 12 V batteries on the vehicle electrical system are connected in parallel in normal vehicle operation or when the engine is stopped to supply the consumers, providing a voltage of 12 V.

When the ignition/starting switch is operated, the battery changeover relay automatically switches the two batteries to series connection temporarily for the starting process, so that a voltage of 24 V is applied across the starter terminals.

All other electrical components continue to be supplied with 12 V.

After the ignition/starting switch is released, the starter is switched off and the batteries are connected in parallel again. While the internal-combustion engine is running, the 12 V alternator charges the batteries (terminal B+).

Special starting systems for commercial vehicles

Special starting systems are used in large commercial vehicles (large, rear-engined tour busses, special vehicles with under-

Fig. 15 Circuit for a 12 or 24 V starting system for individual drive of a sliding-gear starter KB or QB model with start-inhibit and start-repeating equipment

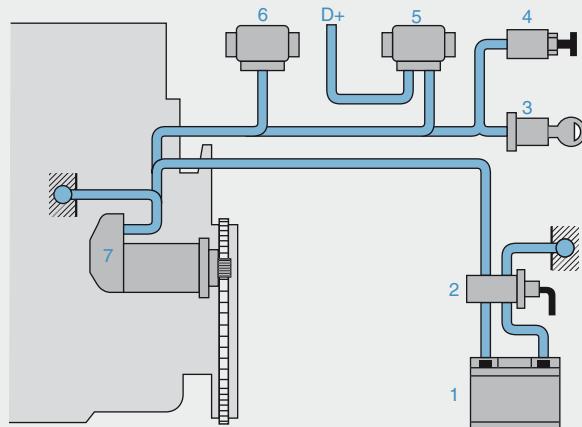


Fig. 15

- | | |
|----|-------------------------|
| 1 | Battery |
| 2 | Battery master switch |
| 3 | Driving switch |
| 4 | Ignition/starter switch |
| 5 | Start-inhibit relay |
| 6 | Start-repeating relay |
| 7 | Starter |
| D+ | Terminal on alternator |

floor mounted engine, etc.), diesel tractor units, ships and stationary engines. The different operating conditions often demand complex starting systems with specially coordinated protection and monitoring relays combined in various different ways. These relays control the starting process and also permit simultaneous start-up of two starters being used in parallel operation. The law requires a battery master switch on many larger electrical systems in commercial vehicles; this switch can be used to isolate the vehicle electrical system from the battery when the engine is not running.

Of the great variety of special starting systems for various applications, two particularly significant systems are described below.

Starting system with start-repeating equipment

In starting systems with remote control or indirect starter actuation (e.g. in stationary installations, diesel traction units and also in isolated cases in rear-engined commercial vehicles) a start-repeating relay is used, especially if it is not possible to

determine directly whether a starting attempt has been successful.

The start-repeating relay does not respond if the starter pinion has meshed successfully. In the event of a malfunction (tooth-to-tooth meshing), however, it interrupts the unsuccessful starting attempt in order to prevent thermal overloading of the starter. It repeats the starting procedure automatically until the starter pinion meshes in the ring gear and the contact for the starter current is switched on.

The start-inhibit relay also included in the circuit protects the starter against erroneous starting if the engine is already running or is still running. This circuit is used only for sliding-gear starters with electrical two-stage switch-on.

Starting system (12 V or 24 V) with starter double relay for parallel operation

Very powerful starters would be required if they were to run alone to start very large internal-combustion engines. For space reasons, it is better to use two smaller starters instead of one large one. The two starters must drive the engine's ring gear simultaneously in parallel operation so that the engine reaches the necessary

16 HEP109 series parallel starting system circuit

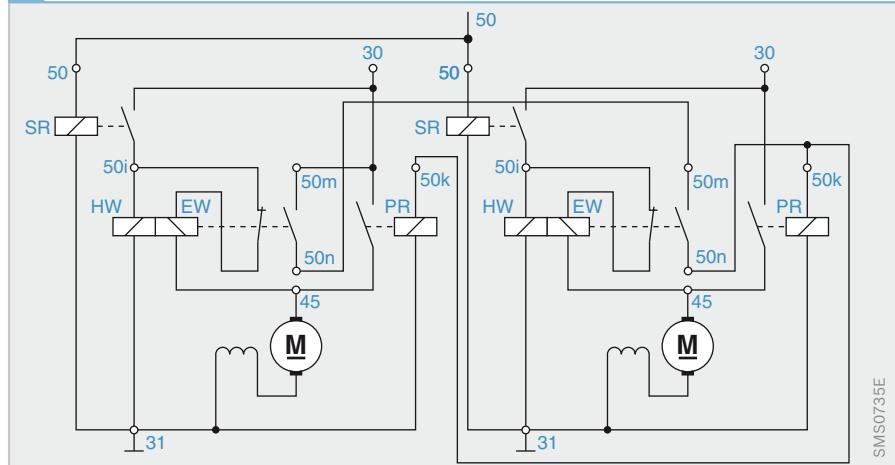


Fig. 16
SR Starting relay
HW Hold-in winding
EW Pull-in winding
PR Power relay
M DC motor

cranking speed. If the power supply is adequate, almost twice the starter power of a single unit may be achieved if the two starters are connected in parallel.

In addition to the start-inhibit relay and the start-repeating relay, a starter double relay is connected into the starting system in low voltage (12 V or 24 V) parallel starting systems. The starter double relay ensures that the full starter current is only switched on after both starters are fully meshed. This means that the two starters develop their full torque simultaneously and are uniformly loaded.

Starter motors for parallel operation have an additional terminal connector for this purpose.

HEP109 series parallel starting system
A parallel starting system for 24 volt vehicle electrical systems has been developed for internal-combustion engines requiring a starting power of approx. 10 to 25 kW (Fig. 16). These high powers are necessary in construction machinery, large tractors and in stationary engines, for instance. Depending on the ambient temperature during a cold start and the amount of additional hydraulic loads when the engine is started, the system can be necessary from engines with as little as 9 l engine displacement. In the standard application at a cold start temperature of -20 °C, it is designed for diesel engines up to 72 l engine displacement and gasoline engines up to 144 l engine displacement.

The system is based on two or three coupled HEP109 series reduction-gear starters, the closed end shield design of which increases the resistance to environmental influences.

Tooth-to-tooth engagement is excluded by the design of the HEP109 series starters. This means that the extra circuitry for the start-repeating relay can be omitted. The functionality of the starter double relay is integrated into the design of the

individual parallel starters. Furthermore, in a double or triple set-up, only identical HEP109 series starters are used, as the corresponding coding can be applied as an option via a five-pin plug.

The service life is almost doubled, by comparison with a single direct starter, thanks to the ideal division of the mechanical and electrical loads.

Automatic starting systems

The high demands on the latest generation of vehicles with regard to convenience, safety, quality, and low noise levels have resulted in an increase in the use of automatic starting systems. An automatic starting system differs from a conventional one by virtue of two additional components:

- A ballast relay (Fig. 17, Item 2)
- A control unit in the vehicle (e.g. an engine control unit, 3) which controls the starting sequence.

The driver then no longer directly controls the starter solenoid-switch circuit; instead the ignition key is used to send a request signal to the control unit, which then performs a series of checks before initiating the starting sequence. The following checks, for example, are possible:

17 Automatic starting system

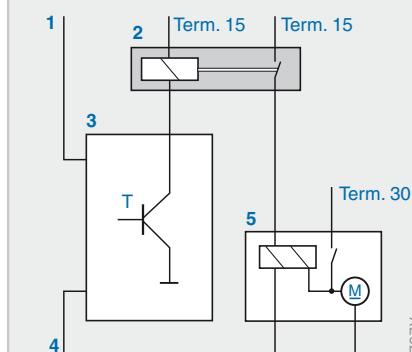


Fig. 17

- 1 Start signal from driver
- 2 Ballast relay
- 3 Control unit
- 4 Park/neutral position/clutch signal
- 5 Starter

- ▶ Is the driver authorized to start the vehicle? (theft-deterrence feature)
- ▶ Is the engine at a standstill? (in quiet vehicles or even in buses or coaches, the driver can hardly hear the engine inside the vehicle)
- ▶ Is the battery sufficiently well charged for starting, taking account of the engine temperature?
- ▶ For automatic transmissions, is the selector lever in neutral or, for on manually shifted transmissions, is the clutch disengaged, or no gear selected?

When the check has been completed successfully, the control initiates the start.

A current is supplied to the winding of the ballast relay to allow this. The ballast relay closes the circuit between the battery (terminal 30) and the starter relay.

The starting sequence then proceeds as previously described.

During the starting sequence, the engine speed (which is detected by the control unit) is compared with a reference figure for the speed at which the engine is running under its own power (the reference speed may also be variable according to engine temperature). Once the engine reaches sustained operation speed, the control unit switches the starter off. As a result, a short starting time and a reduced noise output is achieved and starter wear is minimized.

This process can also be used as the basis to implement "start-stop" operation. This involves automatically switching the internal-combustion engine off when the vehicle is stationary and restarting it when required. The results are considerable fuel savings - particularly in city traffic. At present, reduction-gear starters used in this type of system in conjunction with a high-quality ring gear can perform over 200,000 starts.

Design

Marginal conditions

The most important marginal conditions that must be taken into consideration in the design of the starter are:

- ▶ The starting temperature, i.e. the lowest engine and battery temperature at which it must still be possible to start the engine
- ▶ The cranking resistance of the engine, i.e. the torque required at the crank-shaft, including that required to accommodate auxiliary loads
- ▶ The required minimum engine speed at minimum starting temperature
- ▶ The conversion ratio between starter and crankshaft
- ▶ The nominal voltage of starting system
- ▶ The characteristics of the starter battery
- ▶ The resistance in the supply leads between battery and starter and the contact resistances of terminals and switching elements (circuit breakers, etc.)
- ▶ The speed/torque characteristics of the starter
- ▶ The maximum permissible voltage drop in the vehicle electrical system (the continued functioning of the engine electronics must be ensured)

18 Minimum starting temperature

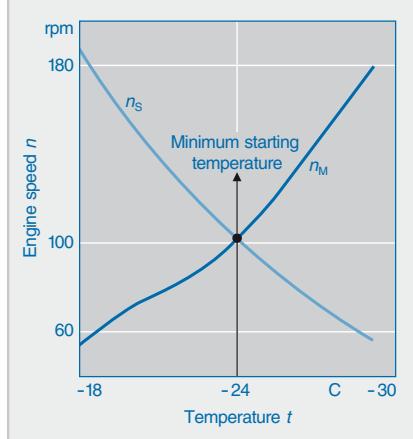


Fig. 18
 n_s Engine speed achieved by starting system
 n_M Minimum speed required for starting

The starter cannot be viewed in isolation because of these marginal conditions. As a component of the overall system consisting of the engine and its ancillary systems, vehicle electrical system including battery and wiring, and the starter itself, the latter must be matched to the other components.

Minimum starting temperature

The degree of frictional resistance that has to be overcome is heavily dependent on the viscosity of the engine lubricant and thus also on engine temperature. At low temperatures the frictional resistance it can be as much as two or three times greater than at normal engine operating temperature. As the temperature falls, the minimum cranking speed required for obtaining satisfactory mixture formation in a gasoline engine or compression ignition a diesel engine furthermore increases. Consequently, the starter must produce significantly greater power output for cold starting than when the engine is warm.

On the other the vehicle can only make a reduced voltage or a low current strength available, as the internal resistance of the lead-acid batteries normally used increases significantly as the temperature falls. As a result, the power of the starting system is lower at lower temperatures.

The increasing frictional resistances in the engine on the one hand, and the declining power of the starting system on the other, mean that the achievable speed decreases and under certain circumstances may not be sufficient to start the engine at low temperatures. For this reason, the engine can only be successfully started above a specific temperature, the minimum starting temperature (Fig. 18).

In Europe, starting systems are generally designed for the minimum starting temperatures listed in Table 1, with the values taking into consideration the varying climatic conditions of the country in which the engine is used.

Cranking resistance

The cranking resistance, i.e. the torque required to turn the engine over, depends primarily on the engine capacity and the viscosity of the engine oil (and hence on the engine temperature). The design of the engine, the number of cylinders, the ratio of stroke to bore, the compression ratio, the mass of the moving engine components and the nature of their bearings, as well as the additional drag from clutch, transmission and auxiliary drive systems also have an effect. In general, the mean cranking resistance increases with engine speed in the case of gasoline engines. With diesel engines, on the other hand, the resistance may decrease after reaching a peak at about 80 to 100 rpm because of the energy recuperated from the relatively large amount of compression work.

Minimum cranking speed

The minimum cranking speed varies according to engine type, start-assist measures also have an effect in diesel engines. Table 2 shows some empirical values.

Criteria

The combination of required torque at minimum starting temperature and minimum cranking speed provides the

Table 1 Minimum starting temperatures

Engine type for	Minimum starting temperature
Cars	-18°C to -28°C
Trucks, busses	-15°C to -32°C
Tractor vehicles	-12°C to -15°C
Drive and peripherals engine in ships	-5°C
Diesel locomotives	+5°C

Table 2 Typical minimum cranking speeds at -20 °C

Engine type	Engine speed
Gasoline engine	60 rpm to 100 rpm
Direct-injection diesel engine without start assistance	80 rpm to 200 rpm
Direct-injection diesel engine with start assistance	60 rpm to 120 rpm

Table 1

Table 2

required starter power output, which is the initial criterion for selection of starter-motor type.

Rated output

The rated output of the starter is a characteristic determined on the test bench, which depends on the marginal conditions for the starting procedure. The figures quoted by Bosch relate to cold starting with the most powerful battery permitted with a charge level of "20 % discharged" at a temperature of -20 °C. The impedance of the starter cable is 1 mΩ. The power is measured at the engine-flywheel ring gear.

Rated-output figures are commonly quoted on the basis of quite different marginal conditions. Some other testing requirements and standards are based on a temperature of + 20 °C and centric torque-loading of the pinion so that the losses on the engine-flywheel ring gear are ignored. Finally, the assumed battery characteristics are also different. Consequently, the representation of power specifications for starters common today is quite unstandardized.

Actual output

The actual output output of a starting system essentially depends on the internal resistance of the battery and the power supply cable. The lower the internal resistance of the battery, the greater its power and hence also the power output of the starter. The battery must be rated such that it is capable of supplying the necessary current at the minimum starting temperature and maintaining it for a sufficiently long period under unfavorable conditions.

The battery and starter must be matched in terms of rating. Starter motors are designed for a maximum and a minimum battery size. If powered by a smaller battery, the actual output of the starting system will be less than the rated output. As long

as the cold-start requirements and the operational reliability of the assemblies relevant for starting (electronic control units, relays) are met, this is technically permissible.

If the starter motor is powered by a larger battery, the power will exceed the rated value. This can lead to overloading of the mechanical components, increased wear and thermal overload. In the case of permanent-magnet starters, it can cause partial demagnetization of the magnets and irreversible loss of torque output. Consequently, the specified battery size should not be exceeded.

Transmission ratio

Selection of the correct starter power output (as a composite value of torque and speed of rotation) is not sufficient for dimensioning the starter. The starter speed and torque characteristics must, rather, match the requirements of the engine on which it is used. The transmission ratio between the starter pinion and the engine crankshaft is therefore used as an additional matching quantity. It can be varied within limits by changing the number of teeth on the starter pinion. However, much greater scope is offered by the use of an internal reduction gear in the starter which allows for broader variation of starter-motor characteristics.

Nominal voltage

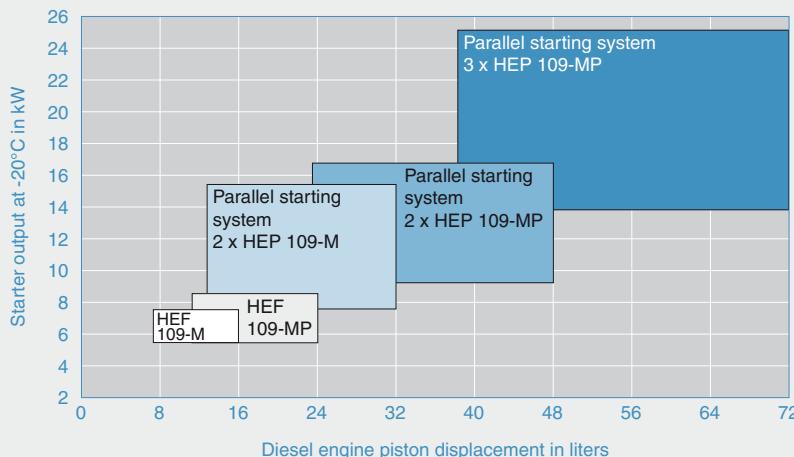
The nominal voltage of starting systems is normally fixed. For cars, it is now universally 12 V, while for commercial vehicles in Europe it is 24 V. In the USA, commercial vehicles also mostly have 12-volt electrical systems. On vans, construction and earthmoving vehicles, and small generator and marine engines, both 12-volt and 24-volt systems are common. Stationary engines, railway locomotives and specialized vehicles can have special systems rated 36 to 110 V.

Overview of the types of starters

Types of starter

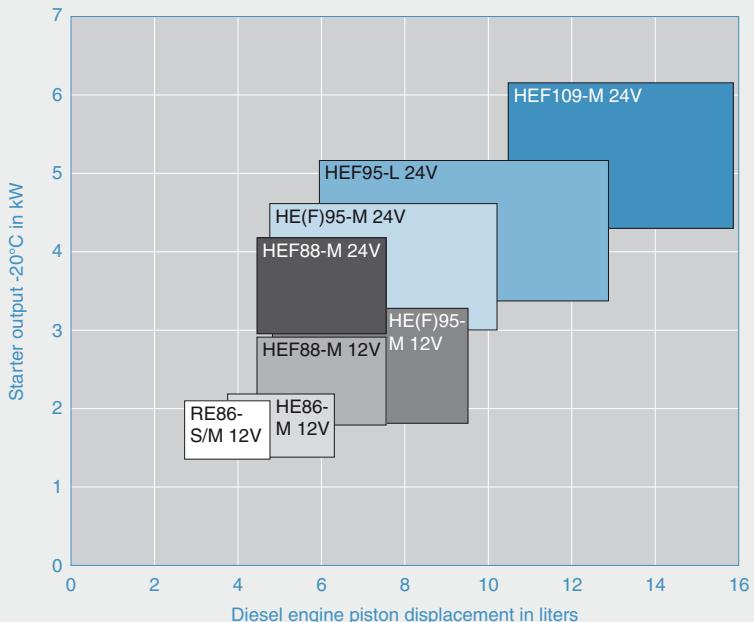
Figures 19, 20 and 21 provide an overview of the various types of starter.

19 Starter types for commercial vehicles (large engines)



SMS0734E

20 Starter types for commercial vehicles

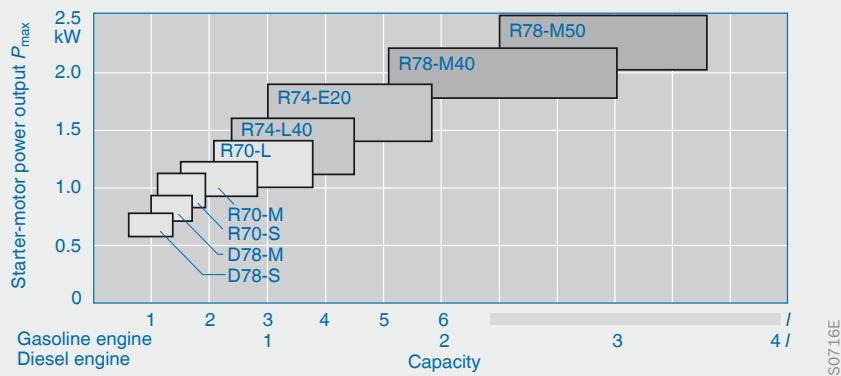


SMS0733E

Type designation key

The designation of a starter provides information regarding the series, the excitation system, the bearing type, the pole housing diameter and the armature core length. Bosch starter motors are classified according to the type designation key shown in Figure 22.

21 Starter types for passenger cars



NMS0716E

22 Starter designations (Bosch)

Series	Excitation system	Bearing type	Pole housing diameter	Armature core length	Option	Optional: voltage, direction of rotation
R: reduction gear up to Ø86	-: permanent	-: mouth	86 / 88	S: shortest	P: boosted type	12 V (R)
H: reduction gear from Ø86	E: electrical 6 pole	F: FAS	95	M: medium		24V (L)
D: direct	X: electrical 4 pole	P: FAS parallel starter	109	L: long		36 V
C: compact				E: longest		

SMS0715-1E

Electromagnetic compatibility (EMC) and interference suppression

Electromagnetic compatibility (EMC) consists of two elements. One is understood as the ability of a device to continue providing reliable service when exposed to electromagnetism from external sources. The second aspect focuses on electromagnetic fields generated by the same device; these should remain minimal in order to avoid creating interference that would impinge upon the quality of radio reception, etc. in the vicinity.

Today's vehicle contains a multitude of systems whose functions are implemented by electrical or electronic components. These same functions, if present at all, used to be carried out fully or predominantly mechanically. With the exception of two-way transceivers used in special-purpose applications, a car's radio was once the only device in which signal reception was of any importance.

In contrast, today's vehicles feature a host of devices that rely on the reception of electromagnetic radio waves. Mobile phones, navigation systems, theft-deterrent systems with radio remote control, and integrated fax and PC units are now being installed and used in vehicles.

This trend has led to a commensurate increase in the importance of suppressing interference and ensuring reception in the vehicle of the operationally vital radio waves.

EMC ranges

Transmitter and receiver

All vehicle systems must remain impervious to electromagnetic radiation emitted from such external sources as extremely powerful radio transmitters. In other words, there must be no threat to the motor vehicle's operational integrity, and malfunctions representing a potential source of driver irritation are unacceptable. Another consideration is that stationary receivers should remain unaffected by passing traffic. Both considerations are governed by national and international codes (EU Directives, German StVZO).

Electrical and electronic components

Motor vehicles contain an extensive array of electrical and electronic components including servo and fan motors, solenoid valves, electronic sensors and control units with microprocessors. These devices must rely on a single vehicle electrical system for their power supply. It is thus vital to avoid mutual interference and feedback phenomena generated by one or several systems so that these do not cause malfunction.

On-board electronic systems

Mobile communications equipment – such as the radio – also exists within an interlinked environment including all of the vehicle's electronic systems. Every device is powered via the same vehicle electrical system, and its reception antennas are located in the immediate vicinity of potential radio-interference sources. These considerations make it imperative that strict limits are imposed on the levels of interference emitted by on-board electronic systems. Compliance with official regulations is essential, and it is important to maintain interference-free reception in the vehicle even when conditions are well below optimal.

EMC between different systems in the vehicle

Shared vehicle electrical system

All of the motor vehicle's electrical systems rely on a single shared vehicle electrical system. Because the electric lines leading to the individual systems are frequently combined within a single wiring harness, feedback pulses can easily travel from one system to the I/O ports of its neighbor (Fig. 1).

This transfer of interference can be in the form of signal pulses (abrupt, steep jumps in current and voltage) generated during the switching on and off of various electrical components such as electric motors, solenoid valves and actuators. Yet another source is the ignition system's high-tension circuit. Similar to other interference signals (for instance the ripple on the power supply), these signal pulses can propagate through the wiring harness. These interference pulses then proceed

to the I/O ports of adjacent systems either directly, through shared conductors such as the power supply (galvanic coupling), or indirectly, through capacitive and inductive coupling stemming from electromagnetic emissions.

Galvanic coupling

Currents for two different electric circuits (such as a solenoid valve's trigger loop and the electric circuit for assessing sensor data) flowing through a single conductive path (common ground through the vehicle chassis, etc.) will both generate a voltage owing to the consistent resistance in the shared conductor (Fig. 2a, next page). Continuing with this sample scenario, the voltage produced by radio-interference source u_1 has the effect of a supplementary signal voltage in signal circuit 2, and could lead to erroneous interpretation of the sensor signal. One remedy is to use separate return lines for each electric circuit (Fig. 2b, next page).

1 Mutual interference between two systems as transmitted through the shared vehicle electrical system (A) and wiring harnesses (B and C)

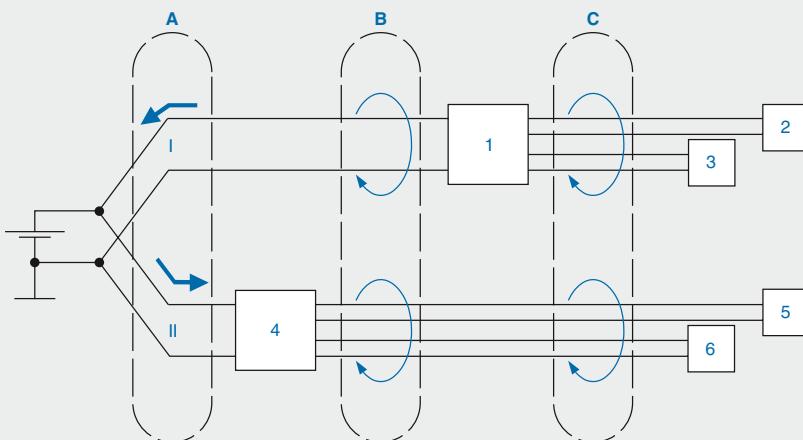


Fig. 1
System I:
1 Control unit
2 Actuator
3 Sensor

System II:
4 Control unit
5 Actuator
6 Sensor

Capacitive coupling

Capacitive coupling allows variable-periodicity signals such as pulse voltages and sinusoidal alternating voltages to produce interference and crosstalk in adjacent electric circuits, even without the existence of a direct physical link (Fig. 3). The potential level of capacitive (interference) voltage is proportional to such factors as the closeness of the neighboring conductor paths and the rise rate of the pulse-shaped voltage shifts (or the frequency of the alternating voltage). The first step is to separate the conductive paths while at the same time extending the ramp periods during which signals rise and fall (or to limit the frequency of the alternating voltages to the absolute minimum required for the function).

Fig. 2

- a Electric circuit with shared return conductor path
- b Electric circuit with separate return lines

u_1, u_2 Voltage source

Z_i Internal resistor

Z_a Terminal resistor

Fig. 3

- 1 Electric circuit 1
- 2 Electric circuit 2

u_1 Voltage source

Z_i Internal resistor

R_E Input resistance

C_E Input capacitance

$C_{1,2}$ Capacitance

between two

conductor paths

u_s Radio interference

voltage

Galvanic coupling of interference signals

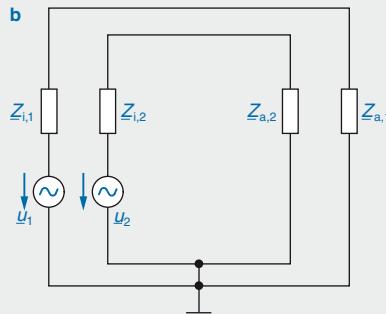
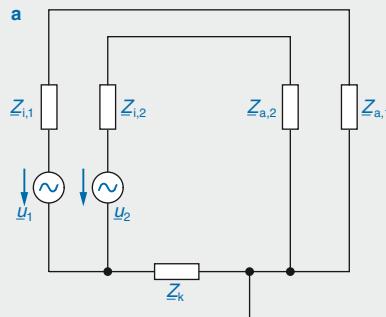


Fig. 4

- 1 Electric circuit 1
- 2 Electric circuit 2

u_1 Voltage source

u_2 Voltage source

Z_i Internal resistor

Z_a Terminal resistor

L_1, L_2 Inductance of

conductors

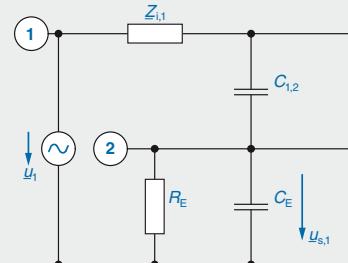
$M_{1,2}$ Inductive coupling

voltage

Inductive coupling

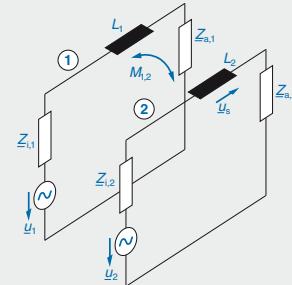
Currents recurring with variable periodicity in one conductor can induce voltage pulses in adjoining circuits. These voltage pulses then generate current in the secondary circuit (Fig. 4). This is the inductive principle exploited in transformer design. One prime factor defining susceptibility to overcoupling is the signal's rise and fall time (or alternating voltage frequency), reflecting the situation encountered with capacitive coupling. Also significant is the effective mutual inductance, as determined by such factors as the size of the wires and their relative routing. Strategies for avoiding inductive coupling include minimizing the dimensions of circuit wires, keeping critical circuits as far apart as possible from each other, and the avoidance of parallel conductive paths. The tendency toward inductive interference is especially pronounced in circuits carrying low-frequency signals (e.g. coupling into loudspeaker wiring).

Capacitive coupling of interference signals



UAE0683-1Y

Inductive coupling of interference signals



UAE0684-1Y

Pulsation in the vehicle electrical system

One strategy for dealing with interference pulses in the vehicle electrical system entails limiting the amplitude of the interference emanating from the radio-interference sources. On the other hand, the susceptible devices (affected electronic components) are designed for insensitivity to pulses of specific shapes and amplitudes. The initial step was to list and classify the pulses encountered within vehicular electrical systems (Table 1). Special-purpose pulse generators can be used to generate the test patterns defined in Table 1 as a basis for assessing resistance to these interference waves in susceptible devices. Both the test pulses and the test technology are codified in standards (DIN 40 839, Section 1; ISO 7637, Section 1) including definitions of the measurement technology for evaluating emissions of pulse-pattern interference. Classifications based on pulse amplitude levels facilitate effective definition of radio-interference sources and the susceptible devices (interference receptors) within each vehicle.

It would thus be possible to specify Class II for all of the radio-interference sources

within a vehicle while designing all the susceptible devices (e.g. control units) to comply with and exceed - by a certain safety margin - Class III. If suppressing interference at the source proves to be cheaper or to involve less technical complexity than reducing sensitivity at the receptors, the logical response might be to shift the definitions to Classes I/II. If the scenario is inverted, with the shielding of potential susceptible devices as the cheaper and simpler solution, then a move to Classes III/IV is warranted.

Because numerous wires are combined within a single wiring harness, each individual conductor is potentially susceptible to inductive and capacitive interference. Although reduced in intensity, the resulting voltage pulses in adjacent supply lines can then appear as spurious signals at the input ports and control outputs in neighboring systems. The test technology for simulating crosstalk interference within wiring harnesses (as defined in DIN 40 839, Section 3 and ISO 7637, Section 3) uses a standardized substitute wiring layout (capacitive clip) with a defined wiring capacitance. Test pulses are fed into this

1 Mutual interference within power supply							
Test pulses as defined in DIN 40839, Section 1				Max. pulse amplitude classes			
Pulse pattern	This is caused by	Internal resistance	Pulse duration	I	II	III	IV
1	Shutdown of inductive consumers	10 Ω	2 ms	-25 V	-50 V	-75 V	-100 V
2	Shutdown of motorized consumers	10 Ω	50 μs	+25 V	+50 V	+75 V	+100 V
3a	Steep overvoltages	50 Ω	0.1 μs	-40 V	-75 V	-110 V	-150 V
3b				+25 V	+50 V	+75 V	+100 V
4	Voltage curve during starting operation	10 mΩ	to 20 s	12 V -3 V	12 V -5 V	12 V -6 V	12 V -7 V
5	Alternator load dump ¹⁾	1 Ω	to 400 ms	+35 V	+50 V	+80 V	+120 V

Table 1

¹⁾ "Load dump", i.e. when the alternator is feeding high current to the battery and the connection to the battery is suddenly interrupted.

layout and through the specimen's wiring harness to produce overcoupling in the signal and control lines. The effects of low-frequency oscillations within vehicular electrical systems can be simulated by producing the desired signals with a signal generator and projecting these into the wiring harness through a current clamp. This process reflects the procedure described above by serving as the basis for the correct balance between the amplitude of radiated interference pulses and the resistance to interference of the potentially susceptible devices (receptors).

Effects on mobile radio reception of high-frequency signal feedback in the vehicle electrical system

Along with the pulses and other interference signals described above, high-frequency signals constitute yet another source of undesired interference within the vehicle electrical system. These signals can stem from periodic switching operations, as found in high-intensity ignition systems, at the commutators in DC motors and from the CPU clock signals that are generated in microprocessor-equipped control units. The interference these sig-

Fig. 5

- a As a function of time
- b As a function of frequency

T Period duration

 T_r Rise time T_i Pulse duration $f_0 = T^{-1}$ Fundamental wave f_g Fundamental frequencies f_{\min} Periodic minima

H Envelope curve

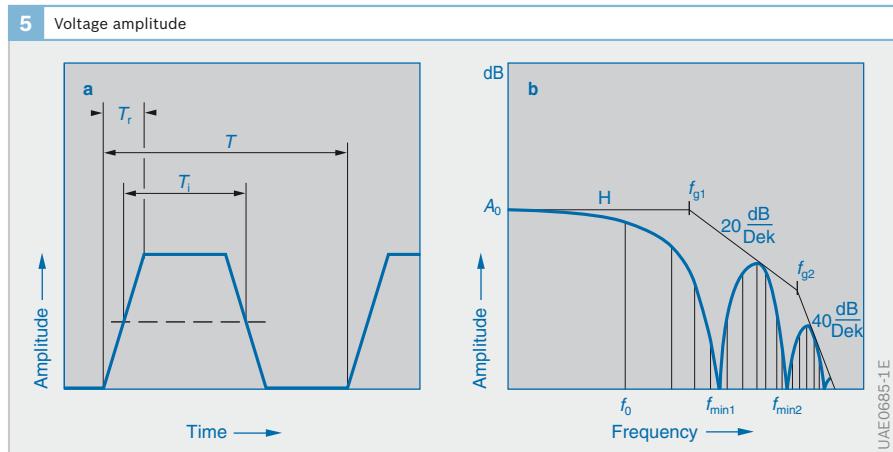
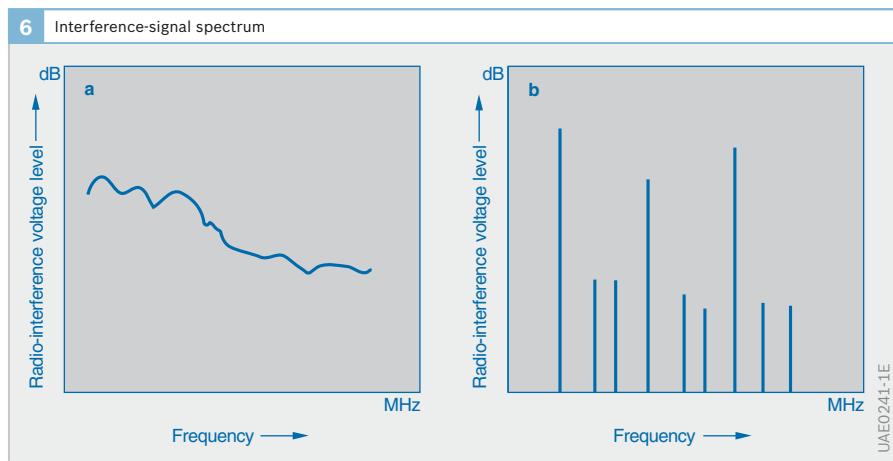


Fig. 6

- a Wide-band interference
- b Narrow-band interference



nals can induce in mobile communications receiver equipment may impair reception or even render it impossible.

Spectrum

Pulsation within on-board electrical systems is usually based on the analysis of current and voltage progressions as function of time (Fig. 5a). Amplitudes at specific frequencies are generally viewed as the prime criterion for evaluating interference signals affecting radio reception (Fig. 5b). Under standard conditions, the interference signals encountered in the automotive environment are rarely isolated sinusoidal waves with single amplitudes recurring at consistent frequencies. Much more common are superimpositions consisting of numerous oscillation components with a variety of amplitudes and frequencies. The "spectrum" for an interference signal is a portrayal of amplitude versus frequency designed to facilitate evaluation of interference potential in individual wavebands (Figs. 6a and 6b). Table 4 (refer to section "Interference-suppression categories") lists the most significant wavebands for automotive applications.

Interference test technology divides interference signals into wide-band and narrow-band interference: a spectrum with a continuous, progressive curve (Fig. 6a) is indicative of wide-band interference, and identifies the corresponding radio-interference source as a wide-band interferer. A contrasting pattern composed of isolated spikes forming a so-called line spectrum indicates narrow-band interference stemming from narrow-band emitters (Fig. 6b).

Initially, the classifications in either category are arbitrary: among the factors that determine whether interference is emanating from a "wide-band" or "narrow-band" radio-interference source are the reception properties of the potentially susceptible devices as reflected in the characteristics of the test equipment employed to monitor the emissions.

In the case of radio interference this equipment is in the form of a selective laboratory receiver or spectral analysis unit. Similar to a radio tuner therefore, the tester thus measures the signal amplitude only within a specific narrow frequency band (receiver bandwidth). The test progresses through the entire relevant frequency range using a procedure analogous to the station search function in standard radios: while maintaining the initial bandwidth, the laboratory receiver proceeds through a range of frequencies, either at graduated intervals or in a continuous progression.

Interference signals recurring at frequencies lower than the test bandwidth produce the continuous signal pattern indicative of wide-band interference. If the frequency is higher than the test bandwidth, the test monitor will also pick up gaps in the spectrum, and the result will be the line spectrum that indicates narrow-band interference.

Electric motors are a typical source of wide-band interference. Commutation frequencies in electric motors are located around just a few 100 Hz, with the exact figure depending on the motor's number of poles and its speed. At a test bandwidth of 120 kHz (bandwidth corresponding to an FM radio tuner receiver bandwidth) this produces a continuous spectrum. At the same test bandwidth, a 2 MHz cycle frequency (of the kind that might be encountered in a microprocessor-equipped control unit) will generate a completely different spectrum, this time with the spiked line spectrum typical of narrow-band interference (interference signal recurs at frequency greater than test bandwidth).

While all electric motors - fans, windshield wipers, servo units and fuel-supply pump, etc. - join the alternator as typical emitters of wide-band interference, yet another potential emitter is the high-tension ignition circuit. In addition, low-frequency clock signals stemming from devices such as switching elements can produce wide-

band interference. The list of narrow-band interference emitters includes the micro-processors in control units as well as all other devices using high-frequency control signals.

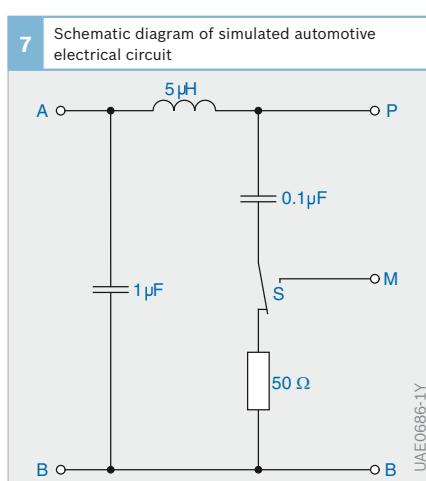
Interference signals picked up by radio receivers may also be of the so-called conducted interference type, that is they may propagate through wiring and conductors (such as the radio's power supply) or they can enter the receiver's I/O ports via inductive and capacitive radiation in the wiring harness. However, interference usually enters through the antenna, either through coupling in the antenna cable, or because the antenna picks up the electromagnetic fields generated by radio-interference sources. One particularly effective transmission antenna is the wiring harness. Other factors influencing the amplitude of the interference signals picked-up by the receiver are the vehicle-body structure and the type and location of the antenna.

Fig. 7

Terminals:
P-B Test
A-B Power supply
M-B Radio-interference monitor
S Switch
B Reference ground (metal sheet, shielding for simulated vehicle electrical system)

Measuring incident interference

Test techniques (DIN 57 879/VDE 0879, Sections 2 and 3; CISPR 25) are stipulated for monitoring interference transmitted through wiring and antennas. Research on individual components is carried out using



2 Interference-suppression levels: permissible radio-interference voltage limits for individual frequency ranges in dB μ V for wide-band (B) and narrow-band (S) interference as defined in CISPR 25 (DIN/VDE 0879-2)

Interference-suppression levels	Interference-suppression levels									
	0.15 to 0.3 MHz (LW)		0.53 to 2.0 MHz (MW)		5.9 to 6.2 MHz (SW)		30 to 54 MHz		70 to 108 MHz (VHF)	
B	S	B	S	B	S	B	S	B	S	
1	100	90	82	66	64	57	64	52	48	42
2	90	80	74	58	58	51	58	46	42	36
3	80	70	66	50	52	45	52	40	36	30
4	70	60	58	42	46	39	46	34	30	24
5	60	50	50	34	40	33	40	28	24	18

3 Permissible radio interference-field strength in dB μ V/m for interference-suppression levels in individual frequency ranges according to DIN/VDE 0879, Section 2, or CISPR 25 for wide-band interferers, measured with quasi-peak detector (B), and narrow-band interference, measured with peak detector (S).

Interference-suppression level	Interference-field strength level																	
	0.15 to 0.3 MHz (LW)		0.53 to 2.0 MHz (MW)		5.9 to 6.2 MHz (SW)		30 to 54 MHz		68 to 87 MHz		76 to 108 MHz		142 to 175 MHz		380 to 512 MHz		820 to 960 MHz	
B	S	B	S	B	S	B	S	B	S	B	S	B	S	B	S	B	S	
1	83	61	70	50	47	46	47	46	36	36	36	42	36	36	43	43	49	49
2	73	51	62	42	41	40	41	40	30	30	30	36	30	30	37	37	43	43
3	63	41	54	34	35	34	35	34	24	24	24	30	24	24	31	31	37	37
4	53	31	46	26	29	28	29	28	18	18	18	24	18	18	25	25	31	31
5	43	21	38	18	23	22	23	22	12	12	12	18	12	12	19	19	25	25

Table 3

laboratory setups and standardized test setups in shielded chambers. Interference is monitored with a laboratory receiver. Line lengths and other geometric dimensions are specified. The power supply must be provided by a precisely defined vehicle electrical system. Voltage sources are also precisely defined, with laboratory specimens being powered from simulated vehicle electrical systems (Fig. 7).

Interference-suppression categories

Similar to pulsation in the vehicle electrical system, narrow- and wide-band interference factors are classified by interference-suppression category in order to facilitate selection and design for specific applications. Within this classification system, demands on sporadically active radio-interference sources are less stringent than the requirements placed on components in continuous operation, such as the alternator. The permissible CISPR 25 (DIN/VDE 0879-2) radio interference levels are listed in Table 2. Table 3 indicates the permissible interference-field strengths for radiation measurements with antennas.

Narrow-band interference of the kind generated by CPU clock signals in control units poses an especially acute problem for radio reception. These interference signals are always present (the control

units are usually switched on along with the ignition). Also of significance is the fact that the radio receiver cannot distinguish these spurious signals from useful signals, making reception of weak transmissions impossible. This situation is reflected in the definition of interference-suppression categories. In a given interference-suppression category, narrow-band interferences are assigned lower permissible interference levels than wide-band interferences.

Because individual vehicle configuration also has a substantial effect on the quality of broadcast reception, the interference-suppression data derived from laboratory testing must be confirmed in a practical automotive environment. The test involves measuring antenna voltage at the end of the antenna cable to which the radio receiver will subsequently be connected.

CISPR 25 also prescribes limits for radio interference voltage as determined using this procedure (Table 4). It contains voltage levels defined to reflect the unfavorable reception conditions as encountered in motor vehicles, where the strengths of useful signals not only fail to exceed just a few mV, but also suffer from substantial fluctuation owing to vehicle motion and the multiple reception paths generated by signal reflections.

4 Limits defined for permissible radio interference voltage at vehicle antenna in dB μ V					
Frequency range	Frequency	Continuous wide-band interference	Sporadic wide-band interference	Narrow-band interference	
	MHz	QP-B	b	QP-B	b
LW	0.14 to 0.30	9	22	15	28
MW	0.53 to 2.0	6	19	15	28
SW	5.9 to 6.2	6	19	6	19
2-way transceivers	30 to 54	6 (15*)	28	15	28
2-way transceivers	70 to 87	6 (15*)	28	15	28
VHF	87 to 108	6 (15*)	28	15	28
2-way transceivers	144 to 172	6 (15*)	28	15	28
C-network car phone	420 to 512	6 (15*)	28	15	28
D-network car phone	800 to 1000	6 (15*)	28	15	28

Table 4
 QP-B Quasi-peak detector relays aural impression produced by interference
 B Wide-band interference emitter with peak detector indicates maximum level
 S Narrow-band interference emitter with peak detector indicates maximum level
 * Limit values for high-voltage ignition

Interference from electrostatic charges

ESD, or electrostatic discharge, can damage or even destroy electronic components. The voltages produced in the event of this discharging can be anywhere up to several volts, which in turn also generate very high, pulsed currents. For this reason, suitable measures are required to prevent the destructive effect or, better still, the discharges themselves. Electronic components that come into contact with vehicle passengers are particularly endangered.

ISO TR 10605 defines test procedures for examining the effects of electrostatic discharge and the immunity to interference that components exhibit both in the laboratory and when installed in the vehicle. The test is conducted with a suitable ESD test-pulse generator, usually in the shape of a pistol. The generator produces the high-voltage pulses that are then coupled into the test specimen.

EMC between the vehicle and its surroundings

Since early 1996, an official regulation has placed mandatory limits on radiation by the vehicle of interference affecting stationary radio reception, while at the same time defining suppression requirements during exposure to external electromagnetic fields. This ordinance (European Directive 95/54/EC) superseded an earlier directive of more limited scope, which only limited interference emissions while defining the homologation certification procedure governing EMC in motor vehicles.

Emitted interference

Electromagnetic emissions radiated by motor vehicles must remain within the limits for narrow and wide-band signals (Fig. 1) to ensure that they do not interfere with radio and television broadcasts or private transmissions. These limits are specified in the Directive 95/54/EC cited above as well as in VDE 0879, Section 1 and the CISPR 12 standard. These regulations contain detailed descriptions of the test procedure, which employs antennas set up at defined distances from the vehicle (10 meters, 3 meters).

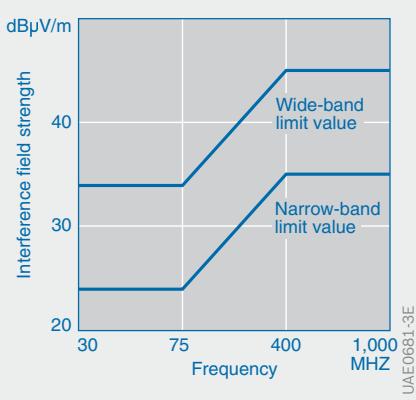
The specifics of the test procedure are described in the regulations quoted.

In practice, it is the ignition system that is usually responsible for the greatest amount of interference radiation. However, since a comprehensive range of measures have been taken to ensure good radio reception in the vehicle, the emitted interference is limited to such an extent that compliance with the legal limit values is usually comfortably achieved.

Incident radiation

As a vehicle travels through the high-intensity electromagnetic field (near field) immediately surrounding a high-intensity emissions source, radiation penetrates through gaps and apertures in the vehicle

1 Wide- and narrow-band limit values for vehicles as monitored at a distance of 10 m



body and interferes with on-board electrical systems. The intensity of this effect (incident radiation) varies according to component locations, body configuration and the wiring harness.

Vehicle measurements

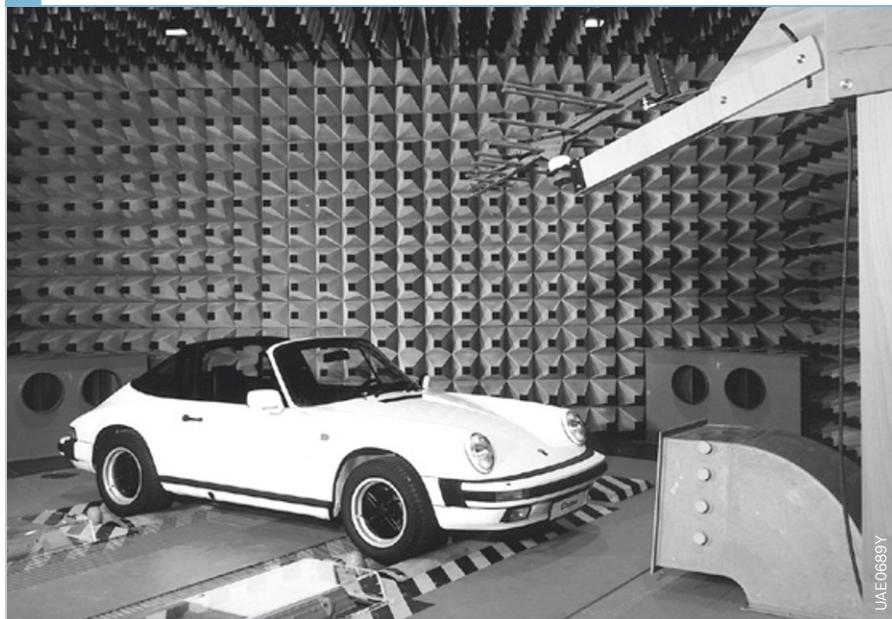
At one time the procedure for verifying that various electronic systems would continue to provide satisfactory performance under exposure to high-intensity fields involved driving the motor vehicle to a number of different broadcast transmitter locations. Now, special test chambers are available for this purpose.

These chambers must be surrounded (shielded) with a metallic casing to prevent the electromagnetic field generated during testing from radiating to the outside. In addition, the chambers must be equipped with absorption materials to inhibit formation of standing waves (nodes and antinodes), which would lead to major variations in field strength between the various measuring points.

The anechoic chamber is used for investigating the overall performance of vehicle electrical systems under real-world operating conditions. The Bosch anechoic chamber (Fig. 2) is equipped with devices capable of generating high-frequency fields within a frequency range extending from 10 kHz to 18 GHz, with a maximum field strength of $E_{\max} = 200 \text{ V/m}$. Since these field strengths represent a health hazard, the test vehicle is operated via remote control from within a shielded control room, while testing is monitored by video cameras. The chamber is screened with metallic sheeting, and non-conductive materials (wood and plastic) line the interior, as metallic substances could foster spurious readings. To prevent reflections and inhibit the formation of standing waves, walls and ceiling are covered by pyramid-shaped absorption elements made of graphite-filled polyurethane foam.

Vehicle testing proceeds on a chassis dynamometer (rolling road) capable of accommodating simulated speeds of up to

2 Measuring resistance to incident radiation of vehicle electrical systems in the EMC anechoic chamber



200 km/h. A fan is also present to direct up to 40,000 m³/h (corresponding to a wind speed of 80 km/h) of air over the vehicle.

Among the advantages that distinguish indoor testing from the old open-air tests near transmitters is that the former method allows latitude for considerable variations in both frequency and field strength. This facilitates evaluation of vehicular resistance to incident radiation under a wide range of conditions, and not at just a few available points. The ability to gradually increase field strengths right up to each electronic device's operational limits furnishes information on safety margins.

Specifics defining the test procedures used in determining radiation resistance in the overall vehicle are described in DIN 40 839, Section 4. ISO 11452, Sections 1 – 4 contain similar descriptions as elements within a more extensive compilation of special-purpose test procedures.

Laboratory measurements

Although the data garnered from assessments of incident radiation on the vehicle as a whole are invaluable, the disadvantage of this type of testing is that it can only be performed at an advanced stage in the vehicle and electronic system design process. Only very limited latitude is available for responding to indications of inadequate resistance to incident radiation uncovered at this stage.

This explains the necessity to determine at an early stage how an electronic system will operate during subsequent use in its automotive environment: prompt information allows effective recourse to remedial action when needed. Various test procedures have emerged for obtaining the required information.

In the first of the three tests described below, interference waves propagating through a conductor are coupled into the wiring harness of the system under investigation. As the suitability of these test arrangements for evaluating the frequency range beyond > 400 MHz is restricted,

a different procedure is used for this range. The alternate procedure relies on antennas to project electromagnetic fields into standardized bench setups.

The specifics of the methods outlined below are defined in DIN 40 839, Section 4, and ISO 11452, Sections 1 – 7 (which also contains additional process descriptions detailing less widespread methods).

All these methods furnish a precise picture of a system's resistance to incident radiation. This can then be used to implement improvements while the development phase is still in progress. Thanks to this inestimable advantage it is now impossible to imagine development projects without these laboratory measurements as flanking measures.

It must be taken into consideration that the design of the electrical system in itself is not the only important factor. Because installation conditions in the vehicle and the routing of the wiring harness can also exercise a decisive influence on ultimate levels of interference resistance, all earlier results still need to be confirmed on a production vehicle in the anechoic chamber.

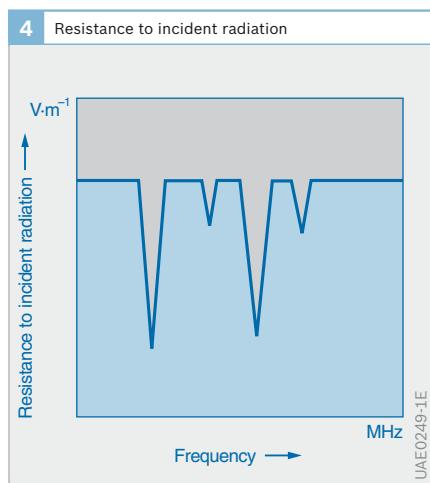
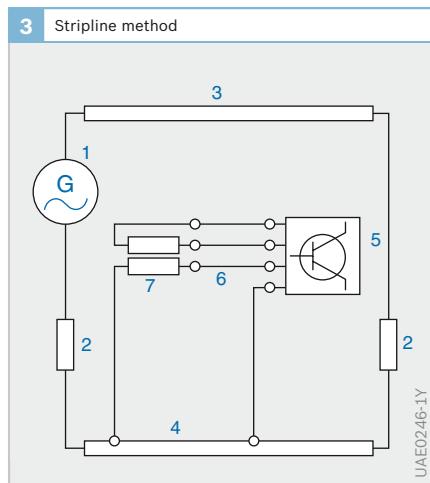
Stripline method (Figs. 3 and 4)

The designation, "stripline", refers to a conductor in the form of a strip. This conductor has a length of 4.1 m and a width of 0.74 m. It is arranged with a clearance of 0.15 m above the conductive sheet (counter-electrode). A high-frequency alternator serves as the source of a transverse magnetic wave generated between the stripline and the counter-electrode, continuing until it reaches the terminal resistor. The stripline dimensions have been selected to minimize the likelihood of reflections occurring during wave propagation, thus ensuring constant field-strength amplitudes relative to frequency.

A typical system setup might include a control unit, wiring harness and peripheral devices (sensors and actuators). These are set up at a height half way be-

tween the base plate and the stripline, with the wiring harness in alignment with the waves' direction of propagation.

The field strength between the planes at a given frequency is then increased until the system malfunctions or until a maximum specified level is reached. If the increments are small enough, repeating this process at progressively altered frequencies will produce a detailed diagram portraying resistance to incident radiation as a function of frequency (Fig. 4).



Bulk current injection method (Fig. 5)

Bulk current injection (BCI) is an immunity test commonly used in the automotive and aircraft industry. As with the stripline method, the system being tested is arranged above a conductive sheet (counter-electrode). A current clamp attached to the wiring harness injects transformer current into its individual wires. The vectorial sum of these currents corresponds to the total current radiated by the clamp. While testing using the stripline method proceeds through a range of increasing field strengths, BCI relies on progressively higher current flows as it continues until the system malfunctions or until current reaches the specified terminal value.

TEM cell (Fig. 6)

As with the stripline method, a transverse electromagnetic field (TEM) is generated here between a strip conductor and a counter-electrode. The TEM counter-electrode, however, is an enclosed housing instead of a metal plate. This leads to another distinguishing feature: TEM test benches do not need to be set up within specially shielded chambers.

Fig. 3

- 1 High-frequency alternator
- 2 Resistor
- 3 Stripline
- 4 Counter-electrode (conductive sheet or cell)
- 5 System under test
- 6 Wiring harness
- 7 Peripherals (sensors, actuators)

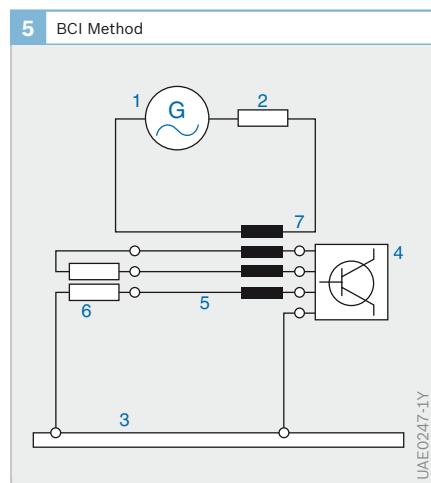


Fig. 4

Determined using stripline, BCI or TEM cell method

Fig. 5

- 1 High-frequency alternator
- 2 Resistor
- 3 Counter-electrode (conductive sheet or cell)
- 4 System under test
- 5 Wiring harness
- 6 Peripherals (sensors, actuators)
- 7 Current clamp

A further difference lies in the fact that only the test specimen per se (for example: a control unit) is subject to electromagnetic exposure. The peripheral equipment is located outside the TEM cell and is connected to the test specimen via a short wiring harness arranged at right angles to the direction of propagation of the electromagnetic waves.

The measurement procedure is consistent with that of the stripline method. Here, too, the field strength is increased until the system malfunctions or until a maximum specified level is reached.

Incident radiation via antenna

Once again similarities with the stripline method are encountered, as the test specimen, including control unit, wiring harness and peripherals, is set up on a base plate. The wiring harness is laid out at a defined distance from this base plate. An antenna located at a specified distance from this assembly generates an electromagnetic field for projection into the entire test setup. Again, field strength is raised until the system malfunctions or until field strength reaches the maximum value specified for testing.

Guarantee of immunity and interference suppression

As early as the planning and conception phases, EMC requirements for immunity to and suppression of radio interference have to be considered as part of every development project focusing on any electronic system or component. The finished products must incorporate EMC measures as an integral design element.

EMC in the electronic control unit (ECU)

The first EMC requirement reflected in any ECU is to select microprocessor cycle frequencies that are as low as possible while simultaneously making every effort to keep the signal transit curves as flat as possible. EMC characteristics join standard operational aspects as an essential criterion governing the selection of components (integrated circuits). Whereas the components' resistance to interference must be maximal, its interference emissions should be minimal. On printed-circuit boards designed for optimal EMC performance, all components with pronounced sensitivity to interference, and all of those with the potential to cause it,

6 TEM-cell test setup

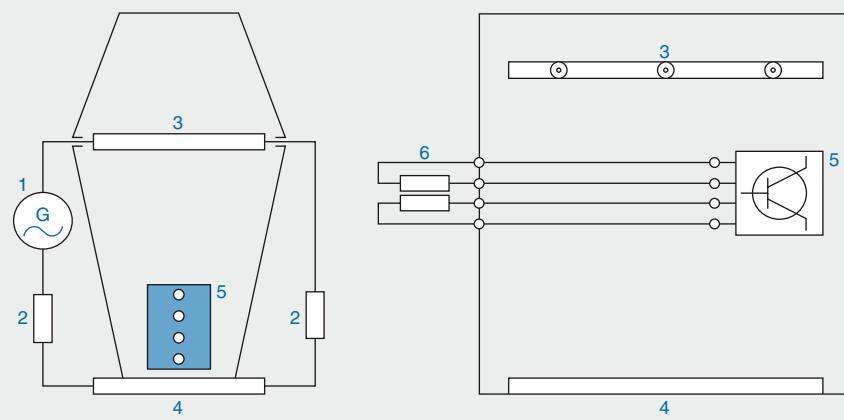


Fig. 6

- 1 High-frequency alternator
- 2 Resistor
- 3 Stripline
- 4 Counter-electrode (conductive sheet or cell)
- 5 System under test
- 6 Peripherals (sensors, actuators)

are decoupled from the connected wiring harness. This is done by locating the components as far as possible from the plug connections.

Suppression components, usually capacitors designed for high-frequency applications, place the required restrictions on potentially disruptive interference. These interference-suppression capacitors are located either directly within the integrated circuits or in the plug-connector area. The suppression elements in the plug-connector area combine with housings featuring optimal conductive properties (shielded casing) to prevent high-frequency radiation in the immediate vicinity from reaching the inside of the ECU. This ensures that signals produced outside the unit do not cause interference inside the unit. On the other side of the coin, high-frequency signals produced inside the unit do not cause undesirable interference in the surrounding environment.

Electric motors and other electromechanical components

Standard practice in the design of electric motors reflects the procedures used for ECUs and sensors by incorporating interference suppression in the design process right from the initial development stages. A sample problem is the interference produced by the brush arcing during current reversal in motors with commutator control. This can severely disrupt radio reception. Suitable suppression elements (capacitors and throttles) limit the interference. Motors are designed with optimal operation of these interference suppressors in mind.

The voltage pulses produced during switching operations in electromagnetic actuators are kept to an acceptable level by suitable elements, such as quench resistors, in the circuit.

High-voltage ignition

The ignition system's high-voltage circuit represents a potential source of substantial interference to radio reception.

One standard solution is to use spark plugs and cable sockets with noise-suppression resistors. While the resistor element is usually located at the end of the ignition cables, in modern (multiple-coil) systems it can be integrated within the dedicated ignition coils plugged directly onto the spark plugs for the individual cylinders. In using resistors, the object is to find a balance between the demand for ignition energy and the required level of interference suppression.

Retrofit interference suppression

As made clear in the above sections, EMC design considerations must be carefully aligned and balanced against operational requirements. Subsequent remedial measures superimposed on the original design at a later date usually entail substantial effort, and should be avoided in production vehicles.

Should the interference-suppression measures incorporated in electrical components prove inadequate for dealing with the conditions encountered in specific applications (in special-purpose emergency vehicles, etc.), supplementary action can be implemented to deal with the problem. Options include installation of filters as well as supplementary shielding for components and wiring.

These options should be exercised with great care, as field modifications to electronic components can lead to operational malfunctions.

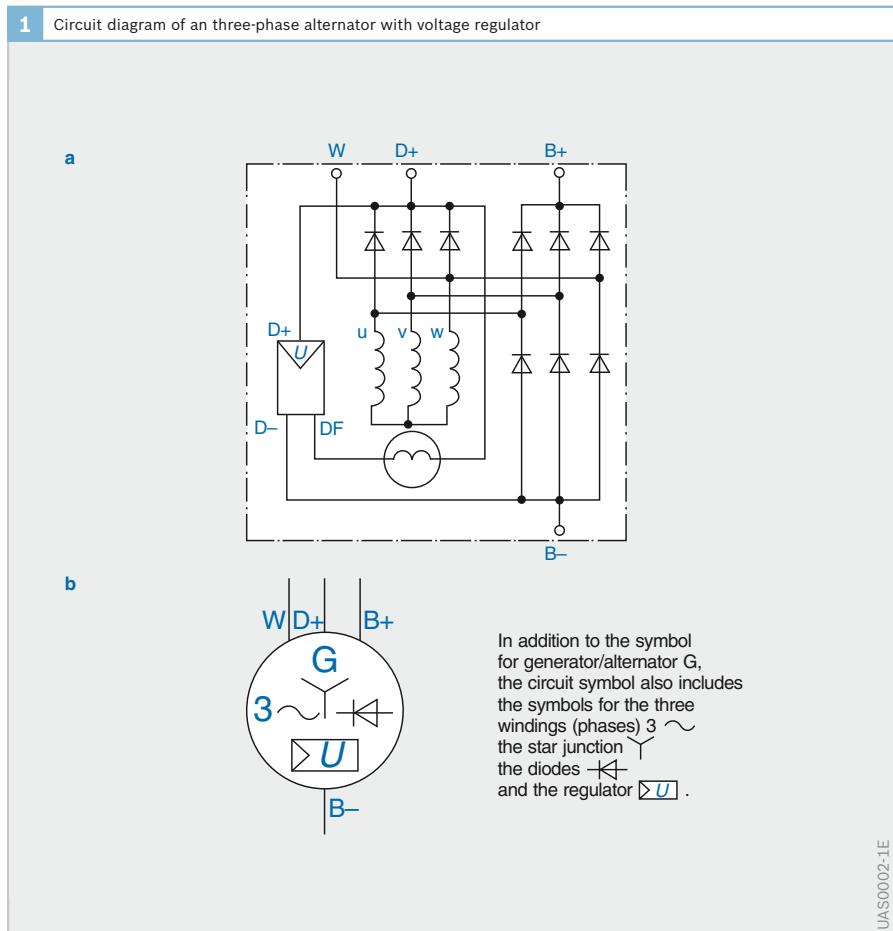
Symbols and circuit diagrams

The electrical systems in vehicles contain a wide array of electric and electronic devices for open and closed-loop engine-management systems as well as numerous comfort and convenience systems. Only by using expressive symbols and circuit diagrams is it possible to provide an overview of the complex circuits in the vehicle's electrical system. Circuit, schematic and terminal diagrams are a help during troubleshooting. They also facilitate field installation of accessories, and furnish support for trouble-free installations and modifications on the vehicle's electrical equipment.

Circuit symbols

The circuit symbols shown in Table 1 are a selection of the standardized circuit symbols relevant to automotive electrical systems. With a few exceptions, they conform to the standards of the International Electrotechnical Commission (IEC).

The European Standard EN 60 617 (Graphical Symbols for Electrical Circuit Diagrams) corresponds to the international standard IEC 617. It exists in three official versions (German, English and French). The standard contains symbol

**Fig. 1**

a With internal circuitry
b Circuit symbols

elements, signs and, in particular, circuit symbols for the following areas:

General applications	Part 2
Conductors and connectors	Part 3
Passive components	Part 4
Semiconductors and electron tubes	Part 5
Production and conversion of electrical energy	Part 6
Switchgear, controlgear and protective devices	Part 7
Measuring instruments, lamps and signaling devices	Part 8
Telecommunications, switching and peripheral equipment	Part 9
Telecommunications, transmission equipment	Part 10
Architectural and topographical installation plans and circuit diagrams	Part 11
Binary components	Part 12
Analog elements	Part 13

Requirements

Symbols are the smallest components of a circuit diagram, and are the simplest way to represent electrical devices and their component parts. They illustrate how a device operates, and are used together with circuit diagrams to illustrate how technical sequences proceed. Symbols do not indi-

cate the shapes and dimensions of the devices they represent, nor do they show the locations of their terminal connections. This abstract representation format represents the only practicable option for illustrating how the devices are connected using a circuit diagram.

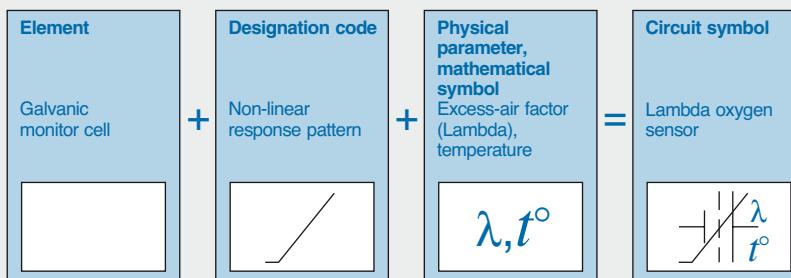
Every symbol should satisfy the following criteria: it must be easy to remember and identify, easy to understand, easy to draw and should clearly indicate the type of device which it represents.

Symbols comprise symbol elements and designations (Fig. 2). Designation codes can take the form of letters, numbers, mathematical and special-purpose symbols, abbreviations of units, characteristic curves, etc.

If a circuit diagram showing the internal circuitry of a device becomes too complex, or if the function of the device can be illustrated without showing all of the details, the circuit diagram for this specific device can be replaced by a single symbol (without internal circuitry, refer to Figs. 1b and 2).

Simplified representations are usually used for integrated circuits, with their typically high levels of compactness (synonymous with high levels of functional integration within an individual component).

2 Sample composition of a circuit symbol: the Lambda oxygen sensor



Representation

Symbols show circuits in their passive (base) state, unaffected by physical parameters such as the application of current, voltage or mechanical force. Other operating states, i.e. any condition that varies from the base state defined above, is indicated by a juxtaposed double arrow (Fig. 3).

Symbols and connecting lines (electrical and mechanical linkage elements) feature a single standard line width.

In order to keep the connecting lines as straight as possible and to avoid crossed lines, symbols can be rotated by 90° increments or shown as mirror images as long as their significance remains unaltered. The direction of continuation conductors may be freely selected. Tangential connections may be shown as emerging from any convenient point on the symbols; the sole exceptions are resistors (connections only at the narrow ends) and the terminals for electromechanical drives (connections only at the long sides, Fig. 4).

Branches may be shown either with or without a node point. No electrical connection is made at junctions without a node point (Fig. 5). There is no mandatory format for illustrating terminals on electrical devices. Terminals, plugs, sockets and threaded connections are identified by symbols only at those points relevant for installation and removal. All other junctions are represented by dots.

In the case of assembled representation, actuators with a common drive are illustrated as responding to this motive force by all moving in the direction indicated by the dashed line (---) which represents the mechanical linkage (Fig. 6).

3 Operating state of the symbol deviating from the base position

	Base position	Deviation
NO contact		
Electro-mechanical drive		

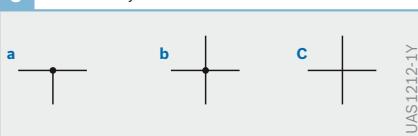
UAS1236-1Y

4 Terminals

At resistor	
Electro-mechanical drive	

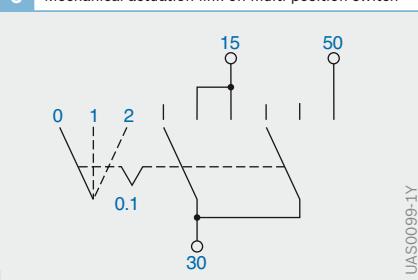
UAS1236-1Y

5 Branch and junctions



UAS1212-1Y

6 Mechanical actuation link on multi-position switch



UAS0099-1Y

Fig. 5

- a Branch with electrical connection
- b Junction with electrical connection
- c Junction without electrical connection

1 Selected circuit symbols		
Connections	Mechanical functions	
Conductor; conductor junctions (with/without connection)	Switching positions (base position: solid line)	Variability, not intrinsic (external), general
Shielded wire		Variability, intrinsic, in response to a physical quantity, linear/nonlinear
Mechanical actuation link; electrical conductor (retrofitted)		Adjustability, general
Junctions (with/without connection)	Manual activation, with sensor (trip), thermal (bimetal strip)	Switches
Connection, general; separable connection (if portrayal is required)	Detent; non-automatic/automatic return in the direction of the arrow (push-button)	Push-button switch, NO/NC contact
Plug-in connection; socket; plug; 3-pin plug connection	Activation, general (mech., pneum., hydraul.); piston drive	Detent switch, NO/NC contact
Ground (housing or vehicle ground)	Actuation at engine speed n , pressure p , quantity Q , time t , temp. t°	Changeover contact, make before break or break before make

1 Selected circuit symbols (continued)

Switches	Various components	
Three-position switch with three contact modes (e.g. turn-signal indicator switch)	Drives with one winding	Resistor
NO/NC contact	Drive with two windings acting in same direction	Potentiometer (with three terminals)
Double-make contact	Drive with two opposed windings	Heating resistor, glow plug, flame plug, screen defroster
Multiple-position switch	Electrothermal drive, thermal relay	Antenna
Cam-lobe switch (e.g. ignition contact breakers)	Electrothermal drive, tractive solenoid	Fuse
Thermal switch	Solenoid valve (closed)	Permanent magnets
Initiator trigger	Relay (drive and switch), example: NC contact without delay and delayed NO contact	Winding, inductive

Table 1
(continued)

1 Selected circuit symbols (continued)

Various components	Devices for automotive applications	
Positive temperature coefficient (PTC) resistor	Dotted/dashed line to delineate or frame logically associated circuit components	Battery
Negative temperature coefficient (NTC) resistor	Shielded device, frame connected to ground	Plug connection
Diode, general, current flows toward the apex of triangle		Lamp, headlamp
PNP transistor NPN transistor E = Emitter (arrow indicates flow direction) C = Collector B = Base	Control units	Horn, fanfare horn
Light emitting diode (LED)	Display element in general; voltmeter; clock	Heated rear windshield (heating resistor in general)
Hall alternator	Tachometer; temperature gauge; speedometer	Switch, general, without function lamp

UAS1232-1E

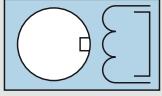
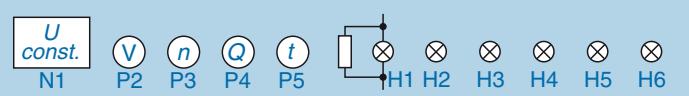
Table 1
(continued)

1 Selected circuit symbols (continued)**Devices for automotive applications**

Pressure switch	Spark plug	Engine with blower, fan
Relay, general	Ignition coil	Starter motor with engagement relay (with/without internal circuitry)
Solenoid valve, injection valve (injector), cold-start valve	Ignition distributor, general	
Thermo-time switch	Voltage regulators	Wiper motor (one/two wiper speeds)
Throttle-valve switch	Three-phase alternator with voltage regulator (with/without internal circuitry)	Intermittent-wiper relay
Rotary actuator	Electric fuel-supply pump, hydraulic pump motor	Car radio
Auxiliary-air valve with electrothermal drive		

Table 1
(continued)

1 Selected circuit symbols (continued)**Devices for automotive applications**

Loudspeaker	Piezoelectric sensor	Rpm sensor
		
Voltage stabilizer	Resistive position indicator	ABS wheel-speed sensor
		
Inductive-type sensor, reference-mark controlled	Air-flow sensor	Hall sensor
		
Flasher, pulse generator, interval relay	Air-mass meters	Converter, transformer (quantity, voltage)
		
Lambda oxygen sensor (not heated/heated)	Flow sensor, fuel-gauge sensor	Inductive-type sensor
		
	Thermostatic switch, temperature sensor	
		
Instrument cluster (dashboard)		
	 N1 P2 P3 P4 P5 H1 H2 H3 H4 H5 H6	

UAS1234-1E

Table 1
(continued)

Circuit diagrams

Circuit diagrams are idealized representations of electrical devices, rendered in the form of symbols. Such diagrams also include illustrations and simplified design drawings as needed (Fig. 7). A circuit diagram illustrates the relationship between the various devices and shows how they are connected to each other. A circuit diagram may be supplemented by tables, graphs, or descriptions. Circuit diagrams vary according to the intended application (e.g. showing circuit operation) and the selected representation mode. A "legible" circuit diagram will meet the following requirements:

- ▶ The representations must reflect the applicable standards; explanations should be provided for any exceptions
- ▶ Current paths should be specially arranged so that current flow or mechanical action takes place from left to right and/or from top to bottom

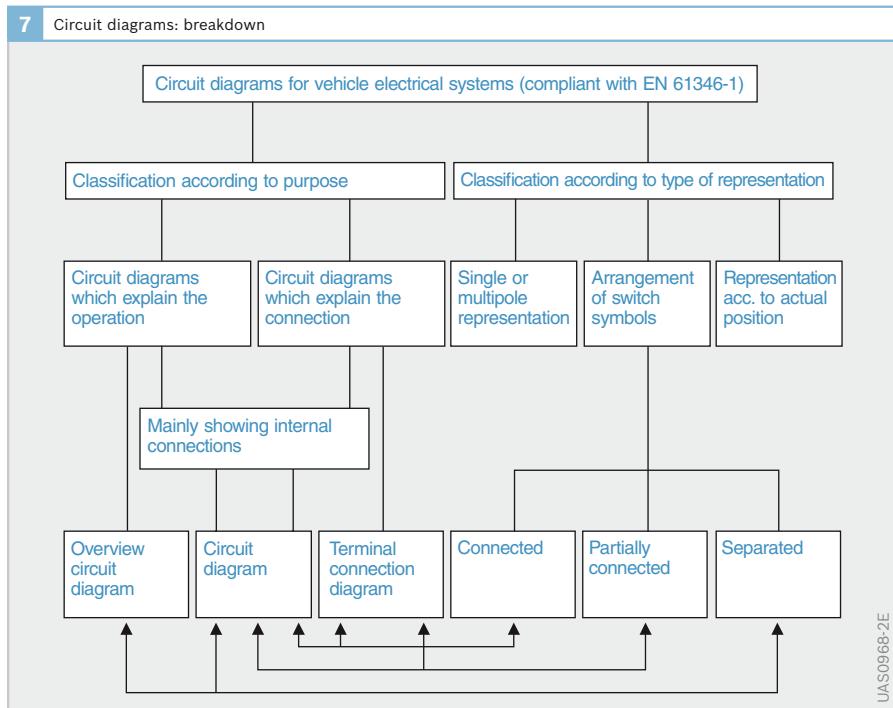
In automotive electrical systems, block diagrams are used to provide a quick overview of circuit and device functions. They are usually unipolar and also dispense with representations of internal circuitry components. The circuit diagrams in their various permutations (as defined by differences in symbol arrangements) provide a detailed diagram of the circuit. As they illustrate how the circuit operates, they are suitable for use as a reference for repair operations. The terminal diagram (with equipment terminal locations) is used by service facilities in replacing defective electrical equipment and when installing supplementary equipment.

Depending upon the type of representation, we distinguish between:

- ▶ Unipolar and multipolar representation and (according to symbol arrangement)
- ▶ Assembled representation, semi-assembled representation, detached representation, and topographical (positionally

7

Circuit diagrams: breakdown



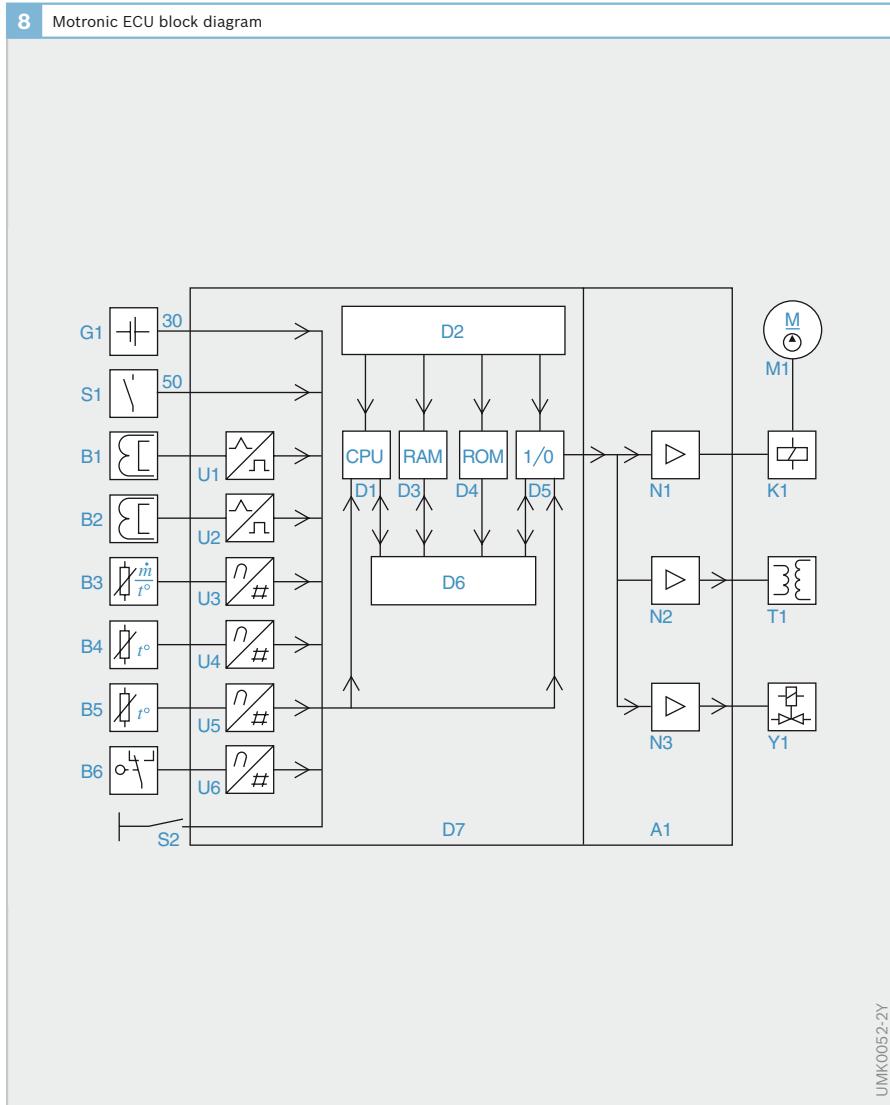
correct) representation. One circuit diagram may employ all of the above forms of representation.

Block diagram

The block diagram is a simplified representation of a circuit showing only the most significant elements (Fig. 8). It is designed to furnish a rapid overview of function,

structure, layout and operation of an electrical system, or part of it. This format also serves as the initial reference for understanding more detailed circuit diagrams.

Squares, rectangles and circles together with attendant symbols based on EN 60 617, Section 2 are employed to illustrate the devices. Wiring is usually shown in single-pole form.



- Fig. 8**
- A1 ECU
 - B1 Engine-speed sensor
 - B2 Reference-mark sensor
 - B3 Air-mass meter
 - B4 Intake-air temperature sensor
 - B5 Engine-temperature sensor
 - B6 Throttle-valve switch
 - D1 Microprocessor (CPU)
 - D2 Address bus
 - D3 User memory (RAM)
 - D4 Program data memory (ROM)
 - D5 I/O
 - D6 Data bus
 - D7 Microcomputer
 - G1 Battery
 - K1 Pump relay
 - M1 Electric fuel-supply pump
 - N1 to N3 Power-output stages
 - S1 Ignition switch
 - S2 Program map selector
 - T1 Ignition coil
 - U1 and U2 Pulse shapers
 - U3 to U6 Analog/digital converters
 - Y1 Fuel injector
- UMK0052-2Y

Circuit diagram

A circuit diagram is the detailed representation of a circuit. By clearly depicting individual current paths it also indicates how an electrical circuit operates. In the circuit diagram, presentation of individual circuit components and their spatial relationship to each other must not interfere with the clear, logical and legible presentation of circuit operation. Figure 9 shows the circuit diagram for a starter, in the form of assembled (composite) and detached (exploded) representations.

The circuit diagram must contain the following:

- Wiring
- Device designation (EN 61 346, Section 2)
- Connection designation or terminal designation (DIN 72 552)

The circuit diagram may also include:

- Comprehensive illustrations including internal circuitry, to facilitate testing, trouble-shooting, maintenance and replacement (retrofitting)
- Reference codes to assist in finding symbols and installation locations, especially in detached representation diagrams

9 Two different ways of showing the circuit diagram for a Type KB starter motor for parallel operation

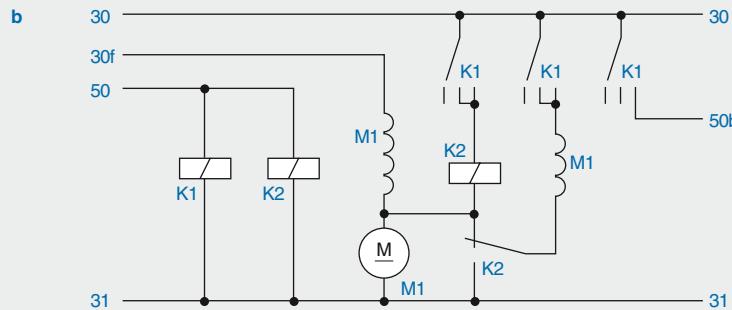
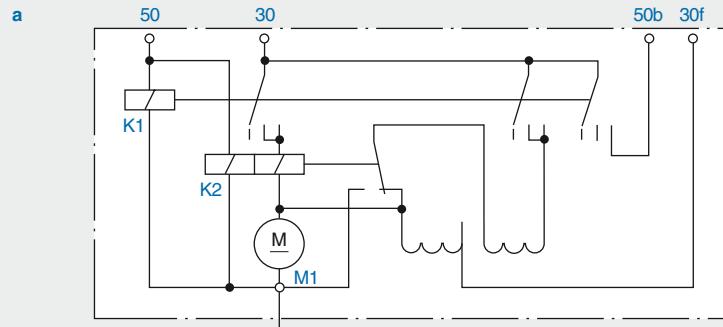


Fig. 9

- a Assembled representation
- b Detached representation
- K1 Control relay
- K2 Engagement relay, hold-in winding, pull-in winding
- M1 Starter motor with series and shunt winding

Circuit representation

Circuit diagrams usually use multipolar component connections. In accordance with EN 61 346, Section 1, symbols can be represented in the following ways, all of which may be combined within the same circuit diagram.

Assembled (composite) representation

All parts of a device are shown directly next to one another, and mechanical linkage of one part to another is indicated by a double line or broken connecting lines (dashes). This format may be employed to depict simple circuits of relatively limited complexity without unduly hindering clarity (Fig. 9a).

Detached (exploded) representation

Symbols for the elements within electrical devices are shown separately in displays designed to show with maximum clarity the various routes taken by the current. No attempt is made to orient the symbols for individual devices and component parts in their actual spatial relationship to each other. Priority is assigned to arranging the symbols so that the individual current paths are as clear and free of crossovers as possible.

Primary purpose: to indicate function and operation of a circuit.

A system of symbols defined in EN 61 346, Section 2 can be used to indicate the relationships between the individual components. Each separately illustrated device symbol includes the code for the device. If it is necessary in the interests of clarity and comprehension, a section of the circuit diagram should be set aside for the complete and assembled representation of devices which have otherwise been shown in the detached form (Fig. 9b).

Topographical representation

This type of representation places the symbol in a position that completely or partly corresponds to its location within the device or component.

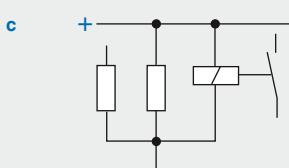
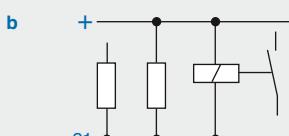
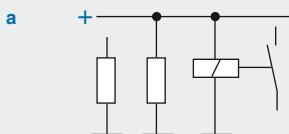
Ground (earth) symbols

For the sake of conceptual simplicity, most vehicles employ a single (hot) conductor layout, relying on the metallic body to conduct the return current. Designers resort to insulated return wiring either when restraints prevent using the body for satisfactory ground connections or when voltages in excess of 42 volts are being handled.

All terminals represented by the ground symbol are mutually connected electrically through the component (housing) or vehicle ground.

All components with a ground symbol must be mounted on the vehicle ground to

10 Ground representation



UAS1222-1Y

Fig. 10

- a Individual ground symbols
- b Common ground connection
- c With common ground point

which they must have a direct electrical connection.

Figure 10 depicts several options for showing connections to ground.

Current paths and conductors (wiring)

Circuits should be arranged to be clear and easy to follow. When possible, individual current paths should indicate a force-transfer direction from left to right and/or from top to bottom, as well as being straight and free of crossovers and changes in direction. They should also be parallel to the border of the circuit diagram.

When a number of conductors run parallel to each other, they are grouped in sets of three with spaces between the groups.

Lines of demarcation, borders

Dot/dash demarcation or border lines are used to separate individual components within a circuit as an indication of functional and/or structural relationships.

In illustrations of automotive electrical systems, these alternating dots and dashes represent a non-conductive border around a device or circuit component. The line will not always correspond with the component housing and does not indicate

ground. In high-tension circuits this outer line is frequently combined with the protective conductor (PE), also represented as a broken dotted line.

Interruptions, codes, destination reference

For clarity, connecting lines (conductors and lines denoting mechanical linkage) can be interrupted if they would otherwise extend too far within the circuit diagram. Only beginning and end of the connecting line are shown. The association of these interrupted points must be clearly recognizable. The code and/or destination reference may be used for this.

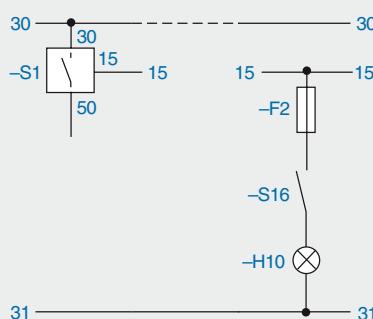
The codes on associated ends of the open circuit match. The following may be used as codes:

- ▶ Terminal designations (DIN 72552),
Figure 11a
- ▶ Indication of the function
- ▶ Identification using alphanumeric symbols

The destination reference is shown in parentheses to avoid confusion with the code; it consists of the section number of the target (Fig. 11b).

11 Broken-line designation using terminal destinations

a



b

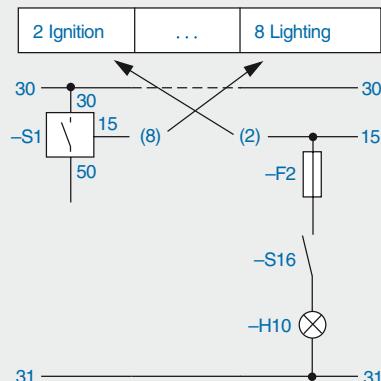


Fig. 11

- a By terminal designation, e.g. terminal 15
- b With destination reference, e.g. in section 8 and 2

Section identification

The section-identification code along the upper border of the diagram is used for locating circuit sections. This designation can be in one of three forms:

- ▶ Consecutive numbers at equal intervals from left to right (Fig. 12a)
- ▶ Indication of the content of the circuit sections (Fig. 12b)
- ▶ Or a combination of the two (Fig. 12c).

Designation codes

Devices, parts or symbols are labeled in circuit diagrams with a letter and a serial number as defined in EN 61 346, Section 2. This code is located to the left of or underneath the symbol.

The prefix to designate the type of device (as specified in the standard) can be omitted as long as this does not lead to ambiguity.

In nested devices, one device is a component part of another, e.g. starter M1 with built-in engagement relay K6. The device identifier is then: - M1 - K6.

12 Different methods of section identification

a

1	2	3	4	5	6	7	8	...
---	---	---	---	---	---	---	---	-----

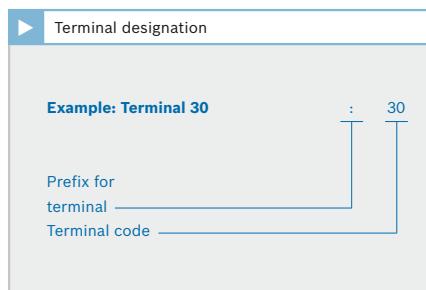
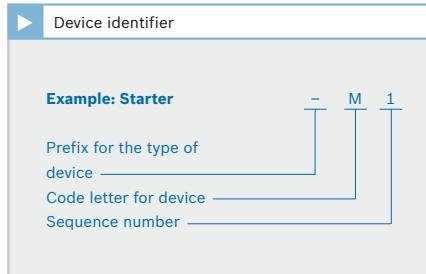
b

1 Power supply	2 Starting system	3 Ignition
----------------	-------------------	------------

c

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

1 Power supply	2 Starting system	3 Ignition
----------------	-------------------	------------



Identification of related symbols in detached views: each individual symbol is shown separately and all symbols for a particular device are assigned a code which is the same as that used for the device itself.

Connection designations (such as those defined in DIN 72 552) must be placed outside the symbol and, if boundary lines are present, outside these lines if possible.

For horizontal current paths: the data associated with the individual symbol is provided beneath the circuit symbol concerned. The terminal code is above the connecting line, just outside the symbol proper.

For vertical current paths: the data applying to the individual symbol are provided to its left. The terminal code is just outside the symbol. If the type is horizontal, the code is provided next to the connecting line on the symbol's right; if the type is vertical, it is on the left.

Fig. 12

- a With consecutive numbers
- b With section indication
- c With a combination of a and b

Terminal diagram

The terminal diagram shows the terminal locations of electrical devices. It also illustrates the external (and internal as required) connections (lines) at these points.

Representation

Individual electrical devices are illustrated using squares, rectangles, circles, symbols or illustrations, and their locations may correspond to their installed positions. The connections are represented by circles, dots, plug connectors, or simply by the connecting line. The following conventions govern the methods of representation used in automotive electrical systems:

- ▶ Assembled, circuit symbols complying with EN 60 617 (Fig. 13a)
- ▶ Assembled, pictorial representation of the device (Fig. 13b)
- ▶ Detached, representation of the device with symbol, terminals with destination references; wiring color codes optional (Fig. 14a and Table 2)
- ▶ Detached, pictorial representation, including terminals with destination references; wiring color codes optional (Fig. 14b)

2 Color coding for electrical conductors
(in accordance with DIN 47002)

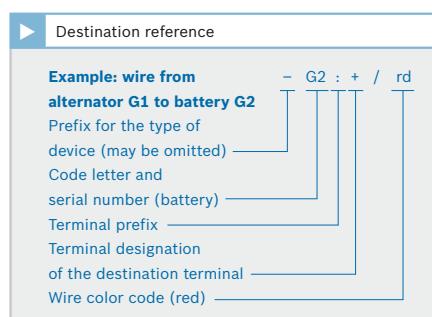
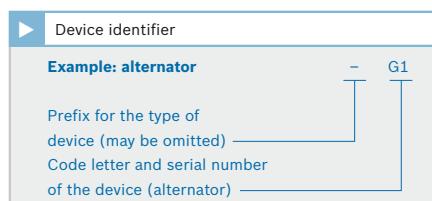
bl	blue	gn	green	bk	black
br	brown	or	orange	tq	turquoise
ye	yellow	pi	pink	vi	violet
gr	gray	rd	red	wh	white

Table 2

Labeling

Devices identified in accordance with EN 61 346, Section 2. Terminal connectors and plugged connections are identified with the terminal designations present on the device (Fig. 13).

Detached views dispense with continuous connecting lines between devices. All conductors leaving a device have a destination reference (EN 61 346, Section 2), consisting of the code for the target device and the terminal designation with - if necessary – statement of the wiring color code as specified in DIN 47 002 (Fig. 15 and Table 1).



13 Terminal diagram, assembled representation

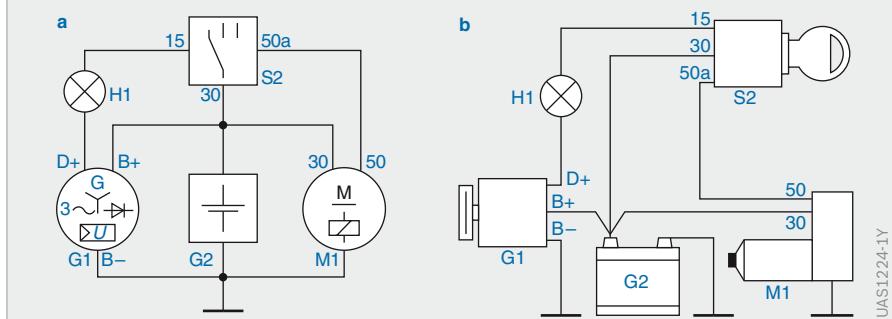
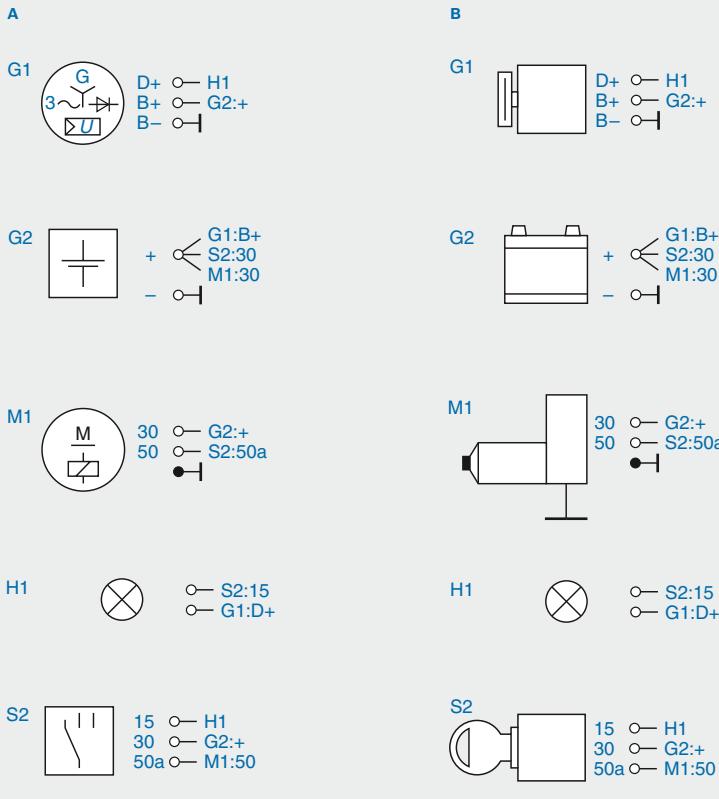


Fig. 13

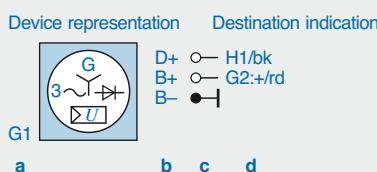
- a With symbols
- b With devices

14 Terminal diagram, detached view



- Fig. 14**
- a With circuit symbols and destination references
 - b With devices and destination references
 - G1 Three-phase alternator with voltage regulator
 - G2 Battery
 - H1 Charge indicator lamp
 - M1 Starter motor
 - S2 Ignition switch
 - XX Device ground on vehicle chassis
 - YY Terminal for ground connection
 - :15 Conductor potential, e.g. terminal 15
- UAS1225-1Y

15 Device identifier (example: alternator)



UAS1055-1E

- Fig. 15**
- a Device identifier (code letter and sequence number)
 - b Terminal designation on device
 - c Device to ground
 - d Destination reference (code letter plus serial number/terminal designation/wire color code)

Block diagram

Bosch has responded to the requirements associated with trouble-shooting on complex, networked systems with a self-diagnosis function by developing system-specific circuit diagrams. Bosch makes block diagrams for further systems in a great number of motor vehicles available in ESI[tronic] (Electronic Service Information). This provides automotive repair shops with a useful tool for locating faults or wiring retrofit equipment. Figure 17 shows the block diagram for a door-locking system.

The representations in the block diagrams diverge from those in the standard circuit diagrams by relying on US symbols with supplementary legends (Fig. 16). These legends include component codes (for instance: "A28" for theft-deterrence system), as shown in Table 3, as well as the color codes for wiring (Table 4). Both tables can be accessed in ESI[tronic].

Table 3

3 Explanation of component codes	
Position	Description
A1865	Electric power-seat system
A28	Theft-deterrence system
A750	Fuse/relay box
F53	Fuse C
F70	Fuse A
M334	Fuel-supply pump
S1178	Warning-buzzer switch
Y157	Vacuum actuator
Y360	Actuator, door, right front
Y361	Actuator, door, left front
Y364	Actuator, door, right rear
Y365	Actuator, door, left rear
Y366	Fuel filler-flap actuator
Y367	Actuator, lock, luggage compartment, trunk lid

Table 4

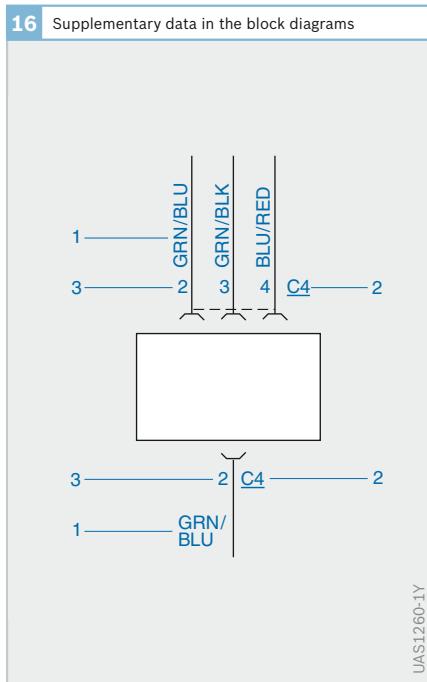
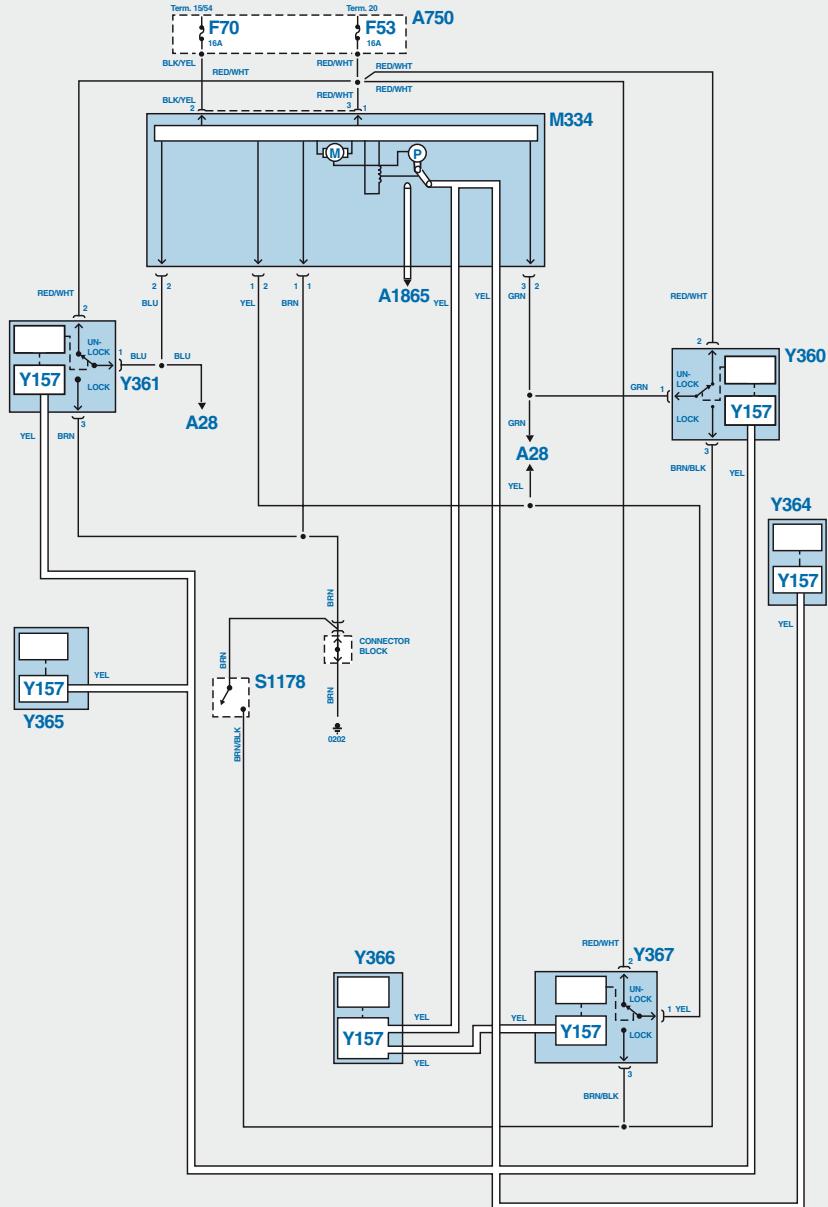


Fig. 16

- 1 Wire color
- 2 Connector number
- 3 PIN number
(dashes between PINs indicate that all PINs are part of the same plug)

4 Wire color code chart	
Position	Description
BLK	Black
BLU	Blue
BRN	Brown
CLR	Transparent
DK BLU	Dark blue
DK GRN	Dark green
GRN	Green
GRY	Gray
LT BLU	Light blue
LT GRN	Light green
NCA	No Color Assigned
ORG	Orange
PNK	Pink
PPL	Purple
RED	Red
TAN	Tan
VIO	Violet
WHT	White
YEL	Yellow

17 Block diagram for a door-locking system (example)



Block diagrams are classified according to system circuits, with further divisions by subsystem as indicated (Table 5). Classification of system circuits reflects the standard ESI[tronic] practice as used for other systems, according to which they are assigned to one of four assembly groups:

- Engine
- Body

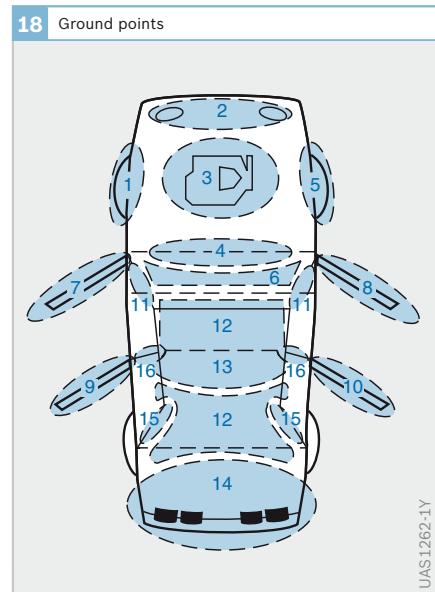
5	System sub-circuits
1	Engine management
2	Starter/charging circuit
3	Heating and air conditioning
4	Radiator fan
5	ABS
6	Vehicle-speed controller/tempomat
7	Power-window unit
8	Central locking system
9	Instrument panel
10	Wiper/washing system
11	Headlamps
12	Exterior lighting
13	Power supply
14	Ground allocation
15	Data line
16	Shift lock
17	Theft-deterrence system
18	Passive safety and security systems
19	Electric antenna
20	Alarm system
21	Heated screen/mirror
22	Supplementary safety and security systems
23	Interior lighting
24	Power-assisted steering
25	Mirror adjuster
26	Soft-top controls
27	Horn
28	Trunk, trunk lid
29	Seat adjustment
30	Electronic damping
31	Cigarette lighter, socket
32	Navigation
33	Transmission
34	Active bodywork components
35	Vibration damping
36	Mobile phone
37	Radio/sound system
38	Immobilizer (drive-away protection)

Fig. 18

- 1 Left front fender
- 2 Forward section
- 3 Engine
- 4 Bulkhead
- 5 Right front fender
- 6 Footwell panel or instrument panel
- 7 Left front door
- 8 Right front door
- 9 Left rear door
- 10 Right rear door
- 11 A-pillars
- 12 Passenger compartment
- 13 Roof
- 14 Rear section
- 15 C-pillars
- 16 B-pillars

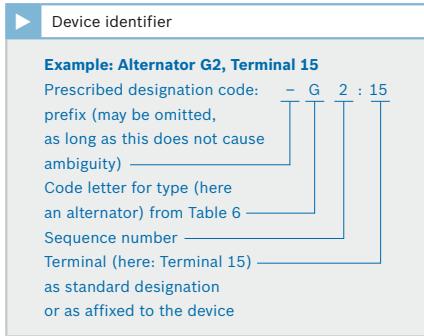
- Suspension and
- Drivetrain

It is important to be aware of the grounding points, particularly when fitting additional accessories. For this reason, ESI[tronic] includes the vehicle-specific location diagram for the grounding points (Fig. 18) in addition to the block diagrams.

**Table 5**

Designations for electrical devices

The designation defined in EN 61 346, Section 2 (Table 6) serves as the basis for unambiguous and internationally-recognized labeling conventions for assemblies, components, etc., as represented in circuit diagrams by circuit symbols. It appears next to the circuit symbol and consists of a sequence of defined prefixes, letters and numerals.



6 Identification codes for electrical devices

Identification code	Type	Examples
A	System, assembly, component group	ABS control unit, radio, two-way radios, mobile phone, alarm system, equipment assembly, ignition trigger box, electronic control unit, cruise control
B	Converter of non-electrical to electrical quantities and vice-versa	Reference-mark sensor, pressure switch, fanfare horn, horn, Lambda oxygen sensor, loudspeaker, air-flow sensor, microphone, oil pressure switch, sensors of all types, ignition triggering device
C	Capacitor	All types of capacitors
D	Binary component, memory	On-board computer, digital equipment, integrated circuit, pulse counter, magnetic tape recorder
E	Various devices and equipment	Heating system, air-conditioning system, lamp, headlamp, spark plug, ignition distributor
F	Protective equipment	Initiator trigger (bimetal), reverse-polarity guard, fuse, current protection circuit
G	Power supply, alternator	Battery, alternator, battery charger
H	Monitor, indicator, signaling device	Audible alarm, function lamp, flasher indicators, turn-signal lamp, brake-pad wear indicators, stop lamp, high-beam indicator, alternator-charge indicator, indicator lamp, indicator, oil pressure monitor, optical indicator, signal lamp, buzzer
K	Relay, contactor	Battery relay, turn-signal flasher, flasher relay, engagement relay, starting relay, hazard warning light flasher
L	Inductance	Choke coil, coil, winding
M	Motor	Blower motor, fan motor, pump motor for ABS/TCS/ESP hydraulic modulators, windshield washer/wiper motor, starter motor, servomotor
N	Regulator, amplifier	Regulator (electronic or electromechanical), voltage stabilizer

Table 6

6 Identification codes for electrical devices (continued)

Identification code	Type	Examples
P	Measuring device	Ammeter, diagnosis socket, rev counter, pressure display, tachograph, measuring point, test point, speedometer
R	Resistor	Sheathed-element glow plug, flame plug, heating resistor, NTC resistor, PTC resistor, potentiometer, regulating resistor, series resistor
S	Switches	Switches and push-buttons of all kinds, ignition contact breaker
T	Transformer	Ignition coil, ignition transformer
U	Modulator, converter	DC converter
V	Semiconductor, electron tube	Darlington transistor, diode, electron tube, rectifier, semiconductors of all kinds, varicap, transistor, thyristor, Zener diode
W	Transmission path, conductor, antenna	Vehicle antenna, shielding, shielded cable, all types of cable, wiring harness, (common) ground conductor
X	Terminal, plug, plug-in connection	Terminal stud, electrical connections of all types, spark-plug connector, terminal, terminal strip, electric wiring coupling, cable tie, plug, socket, multipoint connector, (multiple) plug connection, distributor connector
Y	Electrically operated mechanical equipment	Permanent magnet, (solenoid) injection valve, solenoid clutch, electromagnetic brake, electric air valves, electric fuel-supply pump, solenoid, electric start valve, transmission control, tractive solenoid, kickdown solenoid valve, headlamp range control, ride-height control valve, switching valve, start valve, door lock, central locking system, auxiliary-air device
Z	Electric filter	Interference suppressor, suppression filter, filter network, clock

Table 6 (continued)

Terminal designations

The system of standard (DIN 72 552) terminal designations prescribed for use in automotive applications has been designed to facilitate correct connection of devices and their wiring, with emphasis on repairs and replacement.

The terminal designations (Table 7) are not wire designations, as devices with differing terminal codes can be connected to the opposite ends of a single wire.

For this reason, the terminal designations need not be written on the wires.

In addition to the terminal designations listed, designations according to DIN VDE standards may also be used on electrical machines. Multi-pin plug connections large enough to exhaust the range provided by DIN 72 552 are allocated consecutive numbers or letters, avoiding any characters to which the standard has already assigned a specific function.

7 DIN 72552 terminal designations	
Terminal	Definition
1	Ignition coil, ignition distributor Low voltage
1a	Ignition distributor with two isolated electric circuits To ignition contact breaker I
1b	To ignition contact breaker II
2	Short-circuit terminal (magneto ignition)
4	Ignition coil, ignition distributor High voltage
4a	Ignition distributor with two isolated electric circuits From ignition coil I, terminal 4
4b	From ignition coil II, terminal 4
15	Switched plus downstream from battery (output from ignition [start] switch)
15a	Output on series resistor to ignition coil and to the starter
17	Glow plug switch Start
19	Preglow
30	Line from battery positive terminal (direct)
30a	Battery changeover relay 12/24 V Input from battery II positive
31	Return line from battery negative or ground (direct)
31 b	Return line to battery negative or ground via switch or relay (switched negative)
31a	Battery changeover relay 12/24 V Return line to battery II negative
31c	Return line to battery I negative
32	Electric motors Return line ¹⁾
33	Main terminal ¹⁾
33a	Self-parking
33b	Shunt field
33f	For second reduced-rpm operation
33g	For third reduced-rpm operation
33h	For fourth reduced-rpm operation
33L	Rotation to left (counterclockwise)
33R	Rotation to right (clockwise)
45	Starter motor Separate starting relay, output Starter: input (primary current)
45 a	Dual starters, parallel activation Starting relay for pinion-engagement current Starter I output
45 b	Starter I and II input Starter II output
48	Terminal on starter and start-repeating relay for monitoring the starting process
49	Turn-signal flasher (pulse generator) Input
49 a	Output
49 b	Output to second flasher circuit
49 c	Output to third flasher circuit

Table 7

¹⁾ Polarity reversal terminal 32/33 possible

7 DIN 72552 terminal designations (continued)

Terminal	Definition	Terminal	Definition
Starter motors 50	Starter control (direct)	58	Side-marker lamp, tail lamp, license plate lamp, and instrument lamps
Battery changeover relay 50a	Output for starter control	58b	Tail lamp mode selection on single-axle tractors
Starter control 50 b	Parallel operation of two starter motors with sequence control	58c	Trailer gladhand assembly for single-strand tail lamp with fuse in the trailer
Starting relay for sequence control of engagement current in parallel operation of two starter motors 50 c 50 d	Starter I input at starting relay Starter II input at starting relay	58d	Adjustable instrument illumination, tail and side-marker lamp
Start-inhibit relay 50 e 50 f	Input Output	58 L	Left
Start repeating relay 50 g 50 h	Input Output	58 R	Right, license-plate lamp
Alternator 51 51 e	DC voltage at rectifier DC voltage at rectifier with choke coil for daylight operation	Alternator (magneto alternator) 59	AC voltage output rectifier input
Trailer signals 52	Supplementary signal transmission from the trailer to the towing vehicle	59 a	Charging armature output
53 53 a 53 b 53 c 53 e 53 i	Wiper motor, input (+) Wiper (+), self-parking Wiper (shunt winding) Electrical windshield washer pump Wiper (brake winding) Wiper motor with permanent magnet and third brush (for higher speed)	59 b	Tail lamp armature output
55	Fog lamps	59 c	Stop lamp armature output
56 56 a 56 b 56 d	Headlamps High beam with indicator lamp Low beam (dipped beam) Headlamp-flasher contact	61	Alternator charge indicator
57 57 a 57 L 57 R	Motorcycle parking lamp (also for passenger car, truck etc. in some export markets) Parking lamp Parking lamp, left Parking lamp, right	Tone sequence controller 71 71 a 71 b	Input Output to horns 1 and 2 (bass) Output to horns 1 and 2 (treble)
58	Alarm switch (rotating beacon)	72	Radio, cigarette lighter
59	Loudspeaker	75	Door valve control
60		77	
61	Trailer signals	54	Trailer gladhand assembly and light combination, stop lamp
62		54 g	Pneumatic valve for retarder on trailer, electromagnetically actuated
63		Switch, NC contact and changeover contact 81 81 a 81 b	Input First output (NC contact side) Second output (NC contact side) NO contact
64		82	Input
65		82 a	First output
66		82 b	Second output
67		82 z	First input
68		82 y	Second input
69			Multiple-position switch

Table 7
(continued)

7 DIN 72552 terminal designations (continued)

Terminal	Definition	Terminal	Definition
	Switch, NC contact and changeover contact (continued)		Three-phase alternators
83	Input	U, V, W	Three-phase terminals
83a	Output (position 1)		
83b	Output (position 2)	C	Turn-signal lamp (turn-signal flasher)
83L	Output (left position)	C0	Indicator lamp 1
83R	Output (right position)	C2	Main terminal for turn-signal flasher not connected to turn-signal lamp
	Current relay	C3	Indicator lamp 2
84	Input: drive and relay contact	L	Indicator lamp 3 (e.g. for dual trailer operation)
84a	Output: drive	R	Left-side turn-signal lamp
84b	Output: relay contact		Right-side turn-signal lamp
	Switching relay		
85	Output: drive (negative winding end or ground)		
	Input: drive		
86	Start of winding		
86a	Start of winding or first winding		
86b	Winding tap or second winding		
	Relay contact for NC contact and changeover contact		
87	Input		
87a	First output (NC contact side)		
87b	Second output		
87c	Third output		
87z	First input		
87y	Second input		
87x	Third input		
	Relay contact for NO contact		
88	Input		
	Relay contact for NO contact and changeover contact (NO side)		
88a	First output		
88b	Second output		
88c	Third output		
	Relay contact for NO contact		
88z	First input		
88y	Second input		
88x	Third input		
	Alternator and voltage regulator		
B+	Battery positive terminal		
B-	Battery negative terminal		
D+	Generator positive terminal		
D-	Generator negative terminal		
DF	Generator field winding		
DF1	Generator field winding 1		
DF2	Generator field winding 2		

Table 7
(continued)

Index of technical terms

Technical terms

0-ampere speed, alternator, 452
3-phase bridge circuit, 439

A

ABS control loop, 48
Absolute humidity, 295
Absolute rotating-speed measurement, 269
Acceleration sensors, 271
Accelerator-pedal sensors, 328
AC firing, airbag, 68
Active material, 403
Active speed sensors, 310
Active wheel-speed sensor, 317
Actuators, 354
Adaptive cruise control (ACC), 54
Adaptive frontlighting system (AFS), 42
Adaptive rearlighting system (ARS), 42
Addressing, MOST, 116
Addressing, network, 74
Administrative functions, MOST, 118
AGM technology, 415
Air system, Motronic, 15
Alternator, 434
Alternator circuit, 442, 453
Alternator designs, 455
AMR sensors, 311
Anemometer, 291
Anisotropic magnetoresistive sensors, 248
Antenna system, ACC, 59
Antilock brake system (ABS), 47
Application layer, 78
Application protocols, 146
Arbitration, 97, 149
Architecture, 155
Architecture of electronic systems, 152
ASIC modules, 182
ASSP modules, 182
Asynchronous machines, 362
Automatic starting systems, 480
Automotive networking, 82
Autosar Initiative, 160
Axe sensors, 335

B

Back-bias magnet, 314
Barber pole sensor, 249
Basic CAN, 102
Bathtub pattern, 221
Battery-cable terminals, 404
Battery case, 402

Battery changeover, 391
Battery chargers, 430
Battery charging relay, commercial-vehicle electrical system, 393
Battery-cutoff relay, commercial-vehicle electrical system, 393
Battery designs, 411
Battery maintenance, 427
Battery management, 388
Battery master switch, commercial-vehicle electrical system, 392
Battery relay, commercial-vehicle electrical system, 393
Battery status recognition, 389
Battery testers, 430
Belt-force limiters, 63
Bernoulli's equation, 290
Bi-Litronic projection, 40
Bi-Litronic reflection, 39
Bimorph, 274
Bipolar transistors, 178
Bit stuffing, 101
Block diagram, 509, 516
Bluetooth, 122
Bluetooth architecture, 128
Bluetooth versions, 123
Bolometer, 306
Boost charging, 428
Booster, 369
Braking-force distribution, recuperative braking, 378
Breakdown voltage, 172
Bulk current injection method, 497
Bus access method, 74
Bus driver, 134
Bus guardian, 135
Bus systems, 92
Bus topology, 71

C

Calibration process, sensors, 210
CAN bus, 92
CAN controller, 102
CAN protocol, 96
Capacitance, 173
Capacitive coupling, 488
Capacitor motor, 364
Capacitors, 173
Capacity, battery, 420
CARTRONIC® concept, 158
CCD sensors, 309
Cell connectors, 404
Cell voltage, 407, 418
Characteristic curve type, sensors, 216
Characteristics, alternator, 451
Characteristic curves, battery, 418

Charge-balance calculation, vehicle electrical system, 386
Charging characteristics, 429
Charging methods, starter battery, 428
Charging procedure, starter battery, 408
Circuit diagrams, 508, 510
Circuit symbols, 500
Classification, hybrid drives, 368
Classification, starter battery, 402
Classification of bus systems, 85
Claw-pole alternator, 456
CMOS technology, 309
Coils, 174
Communication controller, 134
Communication layer, 78
Compact-diode-assembly alternator, 457
Compact alternator, 455
Compact ignition coil, 22
Completely maintenance-free battery, 413
Concentration sensors, 294
Consequential-damage protection device, alternator, 451
Contact systems, plug-in connections, 398
Content-based addressing, 96
Continuity equation, 290
Control algorithms, ACC, 56
Controlled system, ABS, 49
Controlled variables, ABS, 50
Control mechanisms, 78
Control units, 196
Control unit software, 204
Conventional batteries, 411
Cooperative recuperative brake system, 377
Coriolis acceleration, 270
Cornering lights (Europe), 43
Course setting, ACC, 54
Cranking resistance, 482
Current characteristic, three-phase alternator, 452
Current generation in the vehicle, 434
Cyclic redundancy check, 100

D

Damping, 272
Data processing, control unit, 196
Data processing, EDC, 26
Data transfer, MOST, 115
Data transfer rate, 83
Deep-cycle resistant battery, 416
Development methods, mechatronics, 164
Diagnosis interfaces, 144
Diagnostic protocol, 145

- Diaphragm-type sensors, 276
 Differential Hall sensor, 242, 311
 Diodes, 176
 Direct-contact sensors, temperature sensors, 299
 Direct-current machines, 361
 Direct-drive starters, 470
 Discharge procedure, starter battery, 409
 Discrete semiconductor components, 176
 Doping, 169
 Doping processes, 186
 Doppler effect, 58
 Drag torque increase, 376
 Drag torque simulation, 376
 Drive concepts, hybrid drives, 366
 Dual-spark ignition coils, 21
 Dual-winding rotary actuator, 357
 Dynamic RAM, 203
 Dynamic segment, 139
 Dynamic system characteristic curve, vehicle electrical system, 386
- E**
 Echo-sounding process, 256
 EC motors, 363
 EDC, overview of functions, 28
 EDC systems, torque-controlled, 30
 Eddy-current sensors, 235, 287
 EE system architectures, 157
 Electrical energy accumulators for hybrid vehicles, 380
 Electrical energy management, 388
 Electrical energy supply, 384
 Electrical machines, 360
 Electrodynamic actuators, 357
 Electrodynamic principle, 355
 Electrolyte density, 427
 Electromagnetic actuators, 356
 Electromagnetic compatibility, 486
 Electromagnetic induction, alternator, 435
 Electromagnetic principle, 355
 Electromagnetic voltage regulators, 444
 Electromechanical actuators, 354
 Electronic charger, 431
 Electronic components, 168
 Electronic diesel control, 24
 Electronic stability program (ESP), 46
 Electronic voltage regulators, 445
 Emitted interference, 494
 Energy accumulator for hybrid systems, 381
 Energy supply in the vehicle, 434
 Engagement relay, 466
- Engine-speed sensors, 310
 Engine management system, EDC, 31
 EOL programming (End of Line), 198
 Epitaxy, 187
 Error types, sensors, 218
 Event control, 78
 Excitation circuit, alternator, 441
 Excitation current, alternator, 436
 Excitation diodes, alternator, 440
 Excitation field, alternator, 436
 Excitation winding, alternator, 436
 Exhaust valve, 48
 External excitation, alternator, 436
- F**
 Failure quota, 219
 Failure rate, 219
 Faraday's law, 435
 Fault-tolerant average algorithm, 141
 Fiber-optic sensors, 224
 Field-effect transistors, 180
 Film circuits, 194
 Finger-type sensor, 349
 Finite state machine, hybrid drives, 374
 FlexRay, 132
 FlexRay controller, 134
 FlexRay protocol, 136
 Flowmeters, 288
 Flow profile, 289
 Fluid-mechanical actuators, 359
 Force sensors, 279
 Frequency-hopping method, 123
 Frequency modulation, 57
 Frictional connection, 469
 Front airbag, 63
 Fuel-injection calculation, 28
 Fuel-injection calculation, EDC, 29
 Fuel-injection control, EDC, 28
 Fuel injector, 16
 Fuel-level sensor, 234
 Fuel system, Motronic, 15
 Full CAN, 103
 Full-wave rectification, 437
 Functional structure, 155
- G**
 Gage factor, 282
 Galvanic coupling, 487
 Galvanomagnetic effect, 240
 Gas-discharge lamps, 33
 Gas inflators, airbag, 68
 Gas sensors, 294
 Gassing, 409
 Gassing voltage, 419
 Gateway, 87
 Gel-electrolyte batteries, 415
 Giant magnetoresistive sensors, 253
- Global positioning system, 257
 GMR multilayers, 254
 GMR sensors, 311
 GMR spin valves, 254
 Gradient correction, 141
 Gradient sensors, 266
- H**
 Half-differential short-circuiting ring sensor, 237
 Half-wave rectification, 437
 Hall angular-position sensors, 243, 329
 Hall phase sensors, 312
 Hall rod sensors, 312
 Hall sensors, 266
 Hall switch, 240
 Hall vane switches, 265
 Halogen lamp, 32
 Headlamp leveling control, 40
 Heavy-duty battery, 417
 Hella sensor, 238
 High-pin-count plug-in connections, 397
 High-pressure fuel injector, 18
 High-pressure sensors, 326
 High beam, 36
 Hot-film air-mass meter, 292, 336
 Hot-wire air-mass meter, 292
 Hybrid circuits, 194
 Hybrid controller, 370
 Hybrid drives, 366
 Hybrid topologies, 73
 Hydraulic modulator, 47
- I**
 iBolt™ force sensor, 344
 Ignition, 20
 Ignition coils, 22
 Illuminance, 35
 Imaging sensors, 307, 308
 Impact-pressure gage, 290
 Incident interference, 492
 Incident radiation, 494
 Incident radiation via antenna, 498
 Individual control, ESP, 46
 Inductance, 174
 Inductive coupling, 488
 Inductive ignition system, 20
 Inductive speed sensors, 310
 Inductive-type sensors, 261
 Intake valve, 48
 Integration level, 181
 Interference immunity, 83
 Interference suppression, 486
 Interference-suppression categories, 493
 Intrinsic conductivity, 170

- K**
- K line, 146
 - KWP 2000, 145
 - KWP 71, 145
- L**
- Laser diode, 185
 - Lead-antimony alloy, 411
 - Lead-calcium-silver alloy, 413
 - Lead-calcium alloy, 412
 - Lead cell, 407
 - Light-emitting diode (LED), 35, 184
 - Lighting systems, adaptive, 42
 - Lighting technology, 32
 - Light sensor, 347
 - LIN bus, 106
 - LIN protocol, 108
 - Lithium ion system, 382
 - Lithium-polymer battery, 382
 - Litronic, 38
 - Longitudinal effect, 273
 - Low beam, 36
 - Low-maintenance batteries, 411
 - Low-pin-count plug-in connections, 398
 - Low-temperature test current, 422
 - Luminous flux, 35
 - Luminous intensity, 35
- M**
- Macrotick, 140
 - Magnetically inductive-type sensors, 235
 - Magnetoelastic principle, 280
 - Magnetoresistive sensors, 247
 - Magnetostatic sensors, 240, 265
 - Magnetostriction, 280
 - Main headlamps (Europe), 36
 - Main headlamps (North America), 37
 - Maintenance-free battery, 412
 - Market volumes of electrics/electronics, 11
 - Master-slave, 75
 - Mat technology, battery, 416
 - Maximum current, alternator, 452
 - Maximum speed, alternator, 452
 - McMess, 146
 - Mechatronics, 162
 - Meshing, starter, 462
 - Mesh topology, 73
 - Microbending effect, 224
 - Microcontroller, 200
 - Micro Hybrid, 369
 - Micromechanical bulk silicon acceleration sensors, 342
 - Micromechanical pressure sensors, 324
- N**
- Micromechanical yaw-rate sensors, 320
 - Microsystem technology, 225
 - Microtick, 140
 - Mild Hybrid, 369
 - Minimum starting temperature, 482
 - Minislots, 139
 - Modulation method, Bluetooth, 124
 - Moisture measurement, 294
 - Moment of ignition, 21
 - Monolithic integrated circuits, 181
 - MOST application layer, 119
 - MOST bus, 112
 - Motor-vehicle bulbs, specifications, 44
 - Motronic engine-management system, 13
 - Movable magnet, 243
 - Multi-plate overrunning clutch, 469
 - Multifunction controller, 448
 - Multimaster, 75
 - Multiplex applications, 86
- O**
- n-type doping, 169
 - Nernst principle, 348
 - Networking, 70
 - Network management, LIN, 110
 - Network nodes, 93
 - Network organization, 74
 - Network topology, 70
 - Nickel metal hydride systems, 381
 - Nominal capacity, battery, 421
 - Nominal voltage, battery, 418
 - NTC resistors (thermistors), 178
- P**
- Overtoltage protection, alternator, 448
 - Overtoltage-protection devices, 449
 - Oxygen-concentration cell, 348
 - p-n junction, 171
 - p-type doping, 170
 - Parallel hybrid drive, 367
 - Passive components, 172
 - Passive (inductive) wheel-speed sensor, 316
 - Pencil coil, 23
 - Phase sensor, 312
 - Photodiode, 184
 - Photolithography, 188
 - Photoresistor, 183
 - Phototransistor, 185
 - Photovoltaic cell, 184
 - Physical layer, 76
 - Piconet, 124
 - Piezoelectric “tuning-fork” yaw-rate sensor, 323
 - Piezoelectric acceleration sensors, 343
 - Piezoelectric effect, 297
 - Piezoelectric knock sensors, 339
 - Planar Lambda oxygen sensor, 351
 - Planar wide-band Lambda oxygen sensor, 352
 - Plate blocks, 403
 - Plug-in connections, 396
 - Position control, 272
 - Position sensors, 232
 - Position sensors for transmission control, 332
 - Potentiometer sensors, 233
 - Power classes, Bluetooth, 124
 - Power generation in the vehicle, 434
 - Powerline communication, 161
 - Preeexcitation, alternator, 441
 - Preeexcitation circuit, alternator, 440
 - Pressure sensors, 276
 - Printed-circuit boards (pcb), 189
 - Propagation time measurement, 57
 - Protocol layers, 96
 - Pseudo Hall sensors, 249
 - PTC resistors (thermistors), 178
 - Pulsation, 288
 - Pulse wheels, 318
 - PWM signals, 198
- R**
- Radar signal processing, 59
 - Rain sensor, 347
 - Random-access memory, 203
 - Range, lighting technology, 35

- Ranging radar, 57
 Rapid-start charger, 431
 Rated current, alternator, 452
 Rated speed, alternator, 452
 Read-only memory, 203
 Real-time applications, 85
 Real-time capability, 84
 Rectifier diodes, 177, 439
 Recuperative brake system, 376
 Reduction-gear starters, 470
 Redundancy, 221
 Regenerative braking, 376
 Relative humidity, 295
 Reliability, 219, 222
 Resistance to exhaustive discharge, starter battery, 425
 Resistance to overcharge, starter battery, 425
 Resistive sensors, 299
 Reverse-current block, 438
 Ring topology, 72
 Roller-type overrunning clutch, 468
 Rollover protection systems, 65
 rpm sensors, 259
- S**
 Salient-pole alternators, 460
 Scatternet, 125
 Seat belt, 62
 Seat-belt pretensioners, 62
 Section identification, 513
 Self-discharge, starter battery, 423
 Self-excitation, alternator, 436, 440
 Self-excitation speed, alternator, 440
 Semiconductor components, 176
 Semiconductor memories, 202
 Semiconductor resistors, 178
 Sensor, 208
 Sensor classification, 216
 Sensor measuring principles, 232
 Sensor model, 209
 Sensor types, 310
 Separators, 404
 Serial hybrid drive, 366
 Series characteristics, 361
 Shear effect, 274
 Short-circuiting ring sensors, 236
 Short-stroke linear motor, 357
 Shoulder-belt tightener, 62
 Shunt characteristic, 361
 Side airbag, 64
 Signal conditioning, control unit, 197
 Signal processing, control unit, 197
 Silicon semiconductor resistors, 302
 Single-phase alternating-current machines, 364
 Single-phase asynchronous motors, 364
 Single-spark ignition coils, 21
 Single-winding rotary actuator, 357
 Single-wire line, 94
 Sintered-ceramic NTC resistors, 300
 Sleep mode, 136
 Sliding-gear starters for commercial vehicles, 474
 Smart sensor, 209
 Software architecture, 158
 Speed sensors, 259
 Spinning current, 241
 Split-power hybrid drive, 368
 Spur gear overrunning clutch, 469
 Starter batteries, 400
 Starter motors, 462
 Starting power, starter batteries, 424
 Starting systems, 476
 Star topology, 71
 Startup, 136
 Static RAM, 203
 Static segment, 139
 Steady-state voltage, 418
 Steering-angle sensors, 330
 Strain gage pick-off, 277
 Strain gage principle, 282
 Strain-gage resistor, 277
 Strain-gage strips, 277, 282
 stripline method, 496
 Strong Hybrid, 369
 Sulfation, 432
 Surface micromechanical acceleration sensors, 340
 Surface-mounting method, 191
 Switch-off function, starter, 465
 Symbol window, 140
 Synchronous machines, 363
 System blocks (EDC), 24, 25
- T**
 Tangential sensors, 267
 TCS interventions, 51
 TEM cell, 497
 Temperature sensors, 298, 327
 Terminal designations, 521
 Terminal diagram, 514
 Terminal posts, 404
 Terminal voltage, 419
 Thermal acceleration sensors, 275
 Thermal radiators, 32
 Thermocouples, 303
 Thermopile sensor, 306
 Thick-film resistors, 302
 Thin-film metallic resistors, 301
 Three-phase alternating voltage, 438
 Three-phase alternator, 434, 436
 Three-phase machines, 362
 Throttle device, 15
 Throttle-valve position sensor, 234
 Through-fitting method, 191
 Timer control, 79
 Time synchronization, 141
 Tolerance requirements, sensors, 218
 Torque control, 30
 Torque motor, 357
 Torque sensor, 346
 Torque sensors, 279, 284
 Traction-control system (TCS), 50
 Transistors, 178
 Transmission agent, MOST, 114
 Transmission speed sensors, 313
 Transverse effect, 274
 Trickle charging, 429
 Two-battery vehicle electrical system, 390
 Two-step Lambda oxygen sensors, 348
 Two-wave rectification, 437
 Two-wire line, 94
 Type designations, battery, 422
 Types of protection, alternator, 449
- U**
 Universal motors, 364
- V**
 Vacuum bulb, 32
 Value creation structure, 11
 Variable-capacitance diode, 177
 Vehicle dynamics controller, 46
 Vehicle electrical systems, commercial vehicles, 391
 Vehicle electrical systems, hybrid drives, 380
 Vehicle system architecture, 155
 Vibration-proof battery, 417
 V model, 166
 Voltage distribution, ignition, 21
 Voltage regulation, alternator, 443
 Voltage-regulator characteristics, 443
- W**
 Wakeup, 136
 Water consumption, starter battery, 425
 Wave-propagation sensors, 256
 Wheel-speed sensors, 316
 Windingless rotor, alternator, 457
 Wiring harnesses, 394
- Z**
 Zener diode, 177

Abbreviations**A**

A/C: Air Conditioner
 A/D: Analog/Digital converter
 AAS: Activation Arc Suppression
 ABC: Active Body Control
 ABS: Antilock Brake System
 AC: Alternating Current
 ACC: Adaptive Cruise Control
 ACL: Asynchronous Connection Less
 ACS: Automatic Child Seat Detection
 ADC: Analog/Digital Converter
 AFH: Adaptive Frequency Hopping
 AFM: AntiFerroMagnet
 AFS: Adaptive Frontlighting System
 AGM: Absorbent Glass Mat
 ALWR: Automatic headlamp range control (German: Automatische Leuchtwellenregulierung)
 AMA: Active Member Address
 AMR: Anisotrop Magneto Resistive
 AMS: Application Message Service
 APB: Automated Parking Brake
 ARS: Angle of Rotation Sensor
 ASC: Active Suspension Control
 ASG: Automatic transmission (German: Automatisches Schaltgetriebe)
 ASIC: Application Specific Integrated Circuit
 ASIS: Active Shift Strategy
 ASSP: Application Specific Standard Product
 AT: Attention Sequence
 AT: Automatic Transmission
 ATF: Automatic Transmission Fluid
 Autosar: AUTomotive Open Systems ARchitecture

B

BCI: Bulk Current Injection
 BD: Bus Driver
 BDC: Bottom Dead Center
 BG: Bus Guardian
 BGE: Bus Guardian Enable
 Bit: Binary digit
 BLDC: BrushLess DC Motor
 BM: Bus Minus
 BNEP: Bluetooth Network Encapsulation Protocol
 BP: Bus Plus
 BSR: Battery Status Recognition
 BSS: Byte Start Sequence

C

CAC: Channel Access Code
 CAD: Computer Aided Design
 CAL: Computer Aided Lighting
 CAN: Controller Area Network
 CAS: Collision Avoidance Symbol
 CBG: Central Bus Guardian
 CC: Communication Controller
 CCD: Charge Coupled Device
 CD: Compact Disc
 CDMA: Code Division Multiple Access
 CHI: Controller Host Interface
 CISPR: Special Committee for Electromagnetic Interference (French: Comité International Spécial des Perturbations)
 CMOS: Complementary Metal Oxide Semiconductor
 CMS: Control Message Service
 COP: Coil on Plug
 CPU: Central Processing Unit
 CR: Common Rail
 CRC: Cyclic Redundancy Checksum
 CVG: Coriolis Vibrating Gyros
 CVSD: Continuous Variable Slope Delta Modulation
 CVT: Continuously Variable Transmission

D

DAC: Device Access Code
 DAC: Digital Analog Converter
 DC: Direct Current
 DF: Speed sensor (German: Drehzahlfühler)
 DGPS: Differential GPS
 DI: Direct Injection
 DIN: German Institute for Standardization (German: Deutsches Institut für Normung)
 DKG: Double-clutch transmission (German: Doppelkupplungsgetriebe)
 DRAM: Dynamic RAM
 DRO: Dielectric Resonance Oscillator
 DRS-MM: Yaw-rate sensor, micromechanical (German: Drehratesensor, mikromechanisch)
 DSP: Digital Signal Processor
 DTS: Dynamic Trailing Sequence
 DVD: Digital Versatile Disc

E

E-ECU: Engine Electronic Control Unit
 EAB: Solenoid-operated shutoff valve (German: Elektrisches Abstellventil)
 EAS: Electronic Active Steering
 EBS: Extended Byte Sequence
 EBS: Electronic Battery Sensor
 ECU: Electronic Control Unit
 EDC: Electronic Diesel Control
 EDR: Enhanced Data Rate
 EE: Electrics/Electronics
 EEM: Electrical Energy Management
 EEPROM (E²PROM): Electrically Erasable Programmable Read Only Memory
 EGAS: Electronic throttle control (German: Elektronisches Gaspedal)
 EGR: Exhaust-Gas Recirculation
 EL: ElectroLuminescence foil
 EMC: ElectroMagnetic Compatibility
 EMM: Electrical Energy Management
 EN ECU: Vehicle Electrical System Electronic Control Unit
 EOL: End of Line
 EPROM: Erasable Programmable Read Only Memory
 ESD: ElectroStatic Discharge
 ESI: Electronic Service Information
 ESP: Electronic Stability Program
 ETN: European Type Number
 EU: European Union

F

FEC: Forward Error Correction
 FES: Frame End Sequence
 FET: Field-Effect Transistor
 FFT: Fast Fourier Transformation
 FH/TDD: Frequency Hop/ Time Division Duplex
 FHSS: Frequency Hopping Spread Spectrum
 Flash-EPROM: Flash-Erasable Programmable Read Only Memory
 FMCW: Frequency Modulated Continuous Wave
 FMVSS: Federal Motor Vehicle Safety Standard
 FOT: Fiber Optic Transceiver
 FSK: Frequency Shift Keying
 FSS: Frame Start Sequence
 FTDMA: Flexible Time Division Multiple Access
 FTM: Fault-tolerant average (German: Fehlertoleranter Mittelwert)

G

GAP: Generic Access Profile
 GFSK: Gaussian Frequency Shift Keying
 GMR: Giant Magneto Resistive
 GOEP: Generic Object Exchange Profile
 GPS: Global Positioning System

H

HCI: Host Controller Interface
 HD: Heavy Duty
 HDEV: High-pressure fuel injector (German: Hochdruck-Einspritzventil)
 HDL: Hardware Description Language
 HEV: Hybrid Electrical Vehicle
 HF: High-Frequency
 HFM: Hot-Film Air-Mass Meter
 HHC: Hill Hold Control

I

I/O: In-/Output
 IAC: Inquiry Access Code
 IC: Integrated Circuit
 ICT: In-Circuit Test (printed-circuit-board manufacture, production process)
 IDE: Identifier Extension Bit
 IDI: InDirect Injection
 IEEE: Institute of Electrical and Electronics Engineers
 IMC: Integrated Magnetic Concentrator
 INIC: Intelligent Network Interface Controller
 ISM: Industrial Scientific Medicine
 ISO: International Organization for Standardization

J

JIS: Japanese Industrial Standard

K

KSN: Customer search number (German: Kundensuchnummer)

L

L2CAP: Logical Link Control and Adaptation Protocol
 LAN: Local Area Network
 LAP: Lower Address Part
 LCD: Liquid Crystal Display
 ldf: LIN Description File
 LED: Light Emitting Diode

LIN: Local Interconnect Network

Litronic: Light-Electronics
 LMP: Link Manager Protocol
 LSB: Least Significant Bit
 LSI: Large Scale Integration
 LSU: Planar wide-band Lambda oxygen sensor (German: Lambda-Sonde Universal)
 LTCC: Low Temperature Cofired Ceramic

M

MAC: Media Access Control
 MAMAC: MOST Asynchronous Medium Access Control
 MC, μC: MicroController
 MEMS: MicroElectro-Mechanical System
 MHP: MOST High Protocol
 MM: MicroMechanics
 MOS: Metal Oxide Semiconductor
 MOST: Media Oriented Systems Transport
 MSB: Most Significant Bit
 MSC: Message Sequence Chart
 MSI: Medium Scale Integration
 MST: MicroSystem Technology
 MTS: Media Access Test Symbol

N

NAP: Non-significant Address Part
 NIC: Network Interface Controller
 NIT: Network Idle Time
 NRZ: Non-Return to Zero
 NTC: Negative Temperature Coefficient

O

OBEX: OBject EXchange
 OC: Occupant Classification
 ODB: Offset Deformable Barrier crash
 OSI: Open System Interconnection

P

PAN: Personal Area Network
 PAS: Peripheral Acceleration Sensor
 PC: Personal Computer
 PCB: Printed-Circuit Board
 PCI: Protocol Control Information
 PCM: Pulse Code Modulation
 PDA: Personal Digital Assistant
 PES: Poly-Ellipsoid System
 PLC: PowerLine Communication
 PLL: Phase Locked Loop
 PMA: Parked Member Address

POC: Protocol Operation Control

POF: Plastic Optical Fiber
 ppm: parts per million
 PPP: Point-to-Point Protocol
 PPS: Peripheral Pressure Sensor
 PROM: Programmable Read Only Memory
 PSK: Phase Shift Keying
 PTC: Positive Temperature Coefficient
 PTFE: Polytetrafluoroethylene
 PVDF: Polyvinylidenefluoride
 PWM: Pulse-Width Modulated
 PZT: Lead Zirconate Titanate

R

RADAR: RAdiation Detecting And Ranging
 RAM: Random Access Memory
 RBH: Replaceable Bulb Headlamp
 RFCOMM: Radio Frequency COMMunication
 RHD: Right Hand Drive
 ROM: Read Only Memory
 ROM: Roll Over Mitigation
 RTR: Remote Transmission Request
 RWG: Control-rack travel sensor (German: Regelweggeber)

S

SAE: Society of Automotive Engineers
 SAF: Synthetic AntiFerromagnet
 SAW: Surface Acoustic Wave
 SCO: Synchronous Connection Oriented
 SDAP: Service Discovery Application Protocol
 SDP: Service Discovery Protocol
 SEFI: SEquential Fuel Injection
 SEI: Software Engineering Institute
 SIG: Special Interest Group (Bluetooth)
 SiO₂: Silicon Dioxide
 SLIO: (Serial Linked Input/Output)

SMD: Surface Mounted Device
 SMM: Surface MicroMechanics
 SMT: Surface Mount Technology
 SOC: State Of Charge (battery)
 SoC: System on a Chip
 SOF: State Of Function (battery)
 SOH: State Of Health (battery)
 SPP: Serial Port Profile
 SRR: Substitute Remote Request
 SSI: Small Scale Integration
 STP: Shielded Twisted Pair
 StVZO: German version of FMVSS/CUR

T

TAS: Torque Angle Sensor
 TCS: Traction-Control System
 TCS BIN: Telephony Control Protocol Specification – BI^Nary
 TDC: Top Dead Center
 TDD: Time Division Duplex
 TDMA: Time Division Multiple Access
 TEM: Transverse ElectroMagnetic Field
 TIM: Twist Intensive Mounting
 TPO: True Power On
 TSS: Transmission Start Sequence
 TTNR: Bosch part number
 (German: Typenteilenummer)

U

UAP: Upper Address Part
 UART: Universal Asynchronous Receiver/Transmitter
 UIS: Unit Injector System
 UDS: Unified Diagnostic Services
 UPS: Unit Pump System
 UTP: Unshielded Twisted Pair
 UV: UltraViolet

V

VCSEL: Vertical Cavity Surface Emitting Laser
 VCO: Voltage-Controlled Oscillator
 VDE: Association of German Electrical Engineers (German: Verband der Elektrotechnik Elektronik Informationstechnik)
 VHD: Vertical Hall Devices
 VHDL: Visual Hardware Description Language
 VLSI: Very Large Scale Integration

W

WAN: Wide Area Network
 WD: WatchDog
 WLAN: Wireless Local Area Network
 WUP: WakeUp Pattern
 WUS: WakeUp Symbol

Z

ZD: Zener Diode