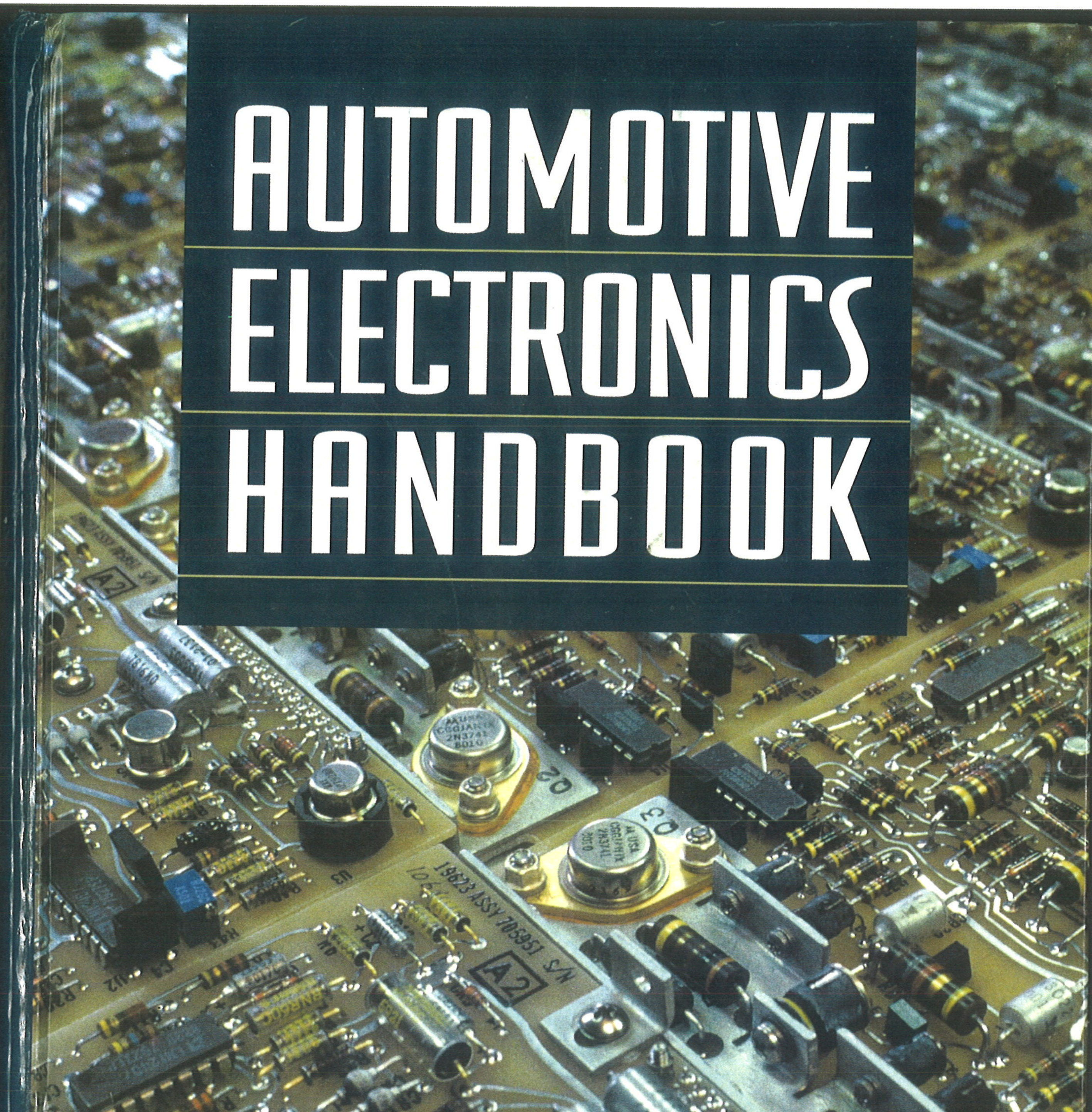


AUTOMOTIVE ELECTRONICS HANDBOOK

RONALD JURGEN

AUTOMOTIVE ELECTRONICS HANDBOOK

Ronald K. Jurgen Editor in Chief

McGraw-Hill, Inc.
New York San Francisco Washington, D.C. Auckland Bogotá
Caracas Lisbon London Madrid Mexico City Milan
Montreal New Delhi San Juan Singapore
Sydney Tokyo Toronto

Related McGraw-Hill Books of Interest

Handbooks

Avallone and Baumeister • MARK'S STANDARD HANDBOOK FOR MECHANICAL ENGINEERS
Benson • AUDIO ENGINEERING HANDBOOK
Brady • MATERIALS HANDBOOK
Chen • COMPUTER ENGINEERING HANDBOOK
Considine • PROCESS/INDUSTRIAL INSTRUMENTS AND CONTROL HANDBOOK
Coombs • PRINTED CIRCUITS HANDBOOK
Coombs • ELECTRONIC INSTRUMENT HANDBOOK
Di Giacomo • DIGITAL BUS HANDBOOK
Fink and Beaty • STANDARD HANDBOOK FOR ELECTRICAL ENGINEERS
Fink and Christiansen • ELECTRONICS ENGINEERS' HANDBOOK
Ganic • McGRAW-HILL HANDBOOK OF ESSENTIAL ENGINEERING INFORMATION
Harper • ELECTRONIC PACKAGING AND INTERCONNECTION HANDBOOK
Harper and Sampson • ELECTRONIC MATERIALS AND PROCESSES HANDBOOK
Hicks • STANDARD HANDBOOK OF ENGINEERING CALCULATIONS
Hodson • MAYNARD'S INDUSTRIAL ENGINEERING HANDBOOK
Johnson • ANTENNA ENGINEERING HANDBOOK
Juran and Gryna • JURAN'S QUALITY CONTROL HANDBOOK
Kaufman and Seidman • HANDBOOK OF ELECTRONICS CALCULATIONS
Lenk • McGRAW-HILL ELECTRONIC TESTING HANDBOOK
Lenk • LENK'S DIGITAL HANDBOOK
Mason • SWITCH ENGINEERING HANDBOOK
Schwartz • COMPOSITE MATERIALS HANDBOOK
Townsend • DUDLEY'S GEAR HANDBOOK
Tuma • ENGINEERING MATHEMATICS HANDBOOK
Waynant • ELECTRO-OPTICS HANDBOOK
Woodson • HUMAN FACTORS DESIGN HANDBOOK

Other

Boswell • SUBCONTRACTING ELECTRONICS
Gieck • ENGINEERING FORMULAS
Ginsberg • PRINTED CIRCUIT BOARD DESIGN
Johnson • ISO 9000
Lenk • McGRAW-HILL CIRCUIT ENCYCLOPEDIA AND TROUBLESHOOTING GUIDE, VOLS. 1 AND 2
Lubben • JUST-IN-TIME MANUFACTURING
Markus and Sclater • McGRAW-HILL ELECTRONICS DICTIONARY
Saylor • TQM FIELD MANUAL
Soin • TOTAL QUALITY CONTROL ESSENTIALS
Whitaker • ELECTRONIC DISPLAYS
Young • ROARK'S FORMULAS FOR STRESS AND STRAIN

To order or to receive additional information on these or any other McGraw-Hill titles, please call 1-800-822-8158 in the United States. In other countries, please contact your local McGraw-Hill office.

BC14BCZ

Library of Congress Cataloging-in-Publication Data

Automotive electronics handbook / Ronald Jurgen, editor in chief.
p. cm.
Includes index.
ISBN 0-07-033189-8
1. Automobiles—Electronic equipment. I. Jurgen, Ronald K.
TL272.5.A982 1994
629.25'49—dc 94-39724
CIP

2 3 4 5 6 7 8 9 0 AGM/AGM 9 0 9 8 7 6 5

ISBN 0-07-033189-8

The sponsoring editor for this book was Stephen S. Chapman, the editing supervisor was Virginia Carroll, and the production supervisor was Suzanne W. B. Rapcavage. It was set in Times Roman by North Market Street Graphics.

Printed and bound by Arcata Graphics/Martinsburg.

Information contained in this work has been obtained by McGraw-Hill, Inc. from sources believed to be reliable. However, neither McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that McGraw-Hill and its authors are supplying information, but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

This book is printed on acid-free paper.

This book is dedicated to Robert H. Lewis and to the memories of Douglas R. Jurgen and Marion Schappel.

CONTENTS

Part 4 Displays and Information Systems

Chapter 20. Instrument Panel Displays *Ronald K. Jurgen* 20.3

Chapter 21. Trip Computers *Ronald K. Jurgen* 21.1

Chapter 22. On- and Off-Board Diagnostics *Wolfgang Bremer, Frieder Heintz, and Robert Hugel* 22.1

Part 5 Safety, Convenience, Entertainment, and Other Systems

Chapter 23. Passenger Safety and Convenience *Bernhard K. Mattes* 23.3

Chapter 24. Antitheft Systems *Shinichi Kato* 24.1

Part 6 Electromagnetic Interference and Compatibility

Part 7 Emerging Technologies

CONTRIBUTORS

Robert E. Bicking *Honeywell, Micro Switch Division* (CHAP. 4)

Tracy Blake *Arizona State University* (CHAP. 30)

David S. Boehmer *Intel Corporation* (CHAP. 11)

Werner Brehm *Robert Bosch GmbH* (CHAP. 13)

Wolfgang Bremer *Robert Bosch GmbH* (CHAP. 22)

Wolfgang Bullmer *Robert Bosch GmbH* (CHAP. 13)

Jerry L. Cage *Allied Signal, Inc.* (CHAP. 15)

Tom Chrapkiewicz *Philips Semiconductor* (CHAP. 25)

Armin Czinczel *Robert Bosch GmbH* (CHAP. 16)

Jeffrey N. Denenberg *Noise Cancellation Technologies, Inc.* (CHAP. 31)

William C. Dunn *Motorola Semiconductor Products* (CHAP. 7)

Randy Frank *Motorola Semiconductor Products* (CHAPS. 2, 5, 32)

Robert L. French *R. L. French & Associates* (CHAP. 29)

Frieder Heintz *Robert Bosch GmbH* (CHAP. 22)

Gary C. Hirschlieb *Robert Bosch GmbH* (CHAP. 12)

Raymond S. Hobbs *Arizona Public Service Company* (CHAP. 30)

Gerhard Hötzel *Robert Bosch GmbH* (CHAP. 6)

Robert Hugel *Robert Bosch GmbH* (CHAP. 22)

Ronald K. Jurgen *Editor* (CHAPS. 1, 20, 21)

George G. Karady *Arizona State Univeristy* (CHAP. 30)

Donald B. Karner *Electric Transportation Application* (CHAP. 30)

Shinichi Kato *Nissan Motor Co., Ltd.* (CHAP. 24)

Bernhard K. Mattes *Robert Bosch GmbH* (CHAP. 23)

Fred Miesterfeld *Chrysler Corporation* (CHAP. 26)

Salim Momin *Motorola Semiconductor Products* (CHAP. 32)

James P. Muccioli *JASTECH* (CHAPS. 27, 28)

Klaus Müller *Robert Bosch GmbH* (CHAP. 10)

Kurt Neuffer *Robert Bosch GmbH* (CHAP. 13)

Harald Neumann *Robert Bosch GmbH* (CHAP. 6)

Paul Nickson *Analog Devices, Inc.* (CHAP. 3)

Johann Riegel *Robert Bosch GmbH* (CHAP. 6)

Makoto Sato *Honda R&D Co., Ltd.* (CHAP. 18)
Gottfried Schiller *Robert Bosch GmbH* (CHAP. 12)
Shari Stottler *Robert Bosch GmbH* (CHAP. 12)
Richard Valentine *Motorola Inc.* (CHAPS. 14, 19)
Helmut Weyl *Robert Bosch GmbH* (CHAP. 6)
Hans-Martin Wiedenmann *Robert Bosch GmbH* (CHAP. 6)
William G. Wolber *Cummins Electronics Co., Inc.* (CHAPS. 8, 9)
Akatsu Yohsuke *Nissan Motor Co., Ltd.* (CHAP. 17)

PREFACE

Automotive electronics as we know it today encompasses a wide variety of devices and systems. Key to them all, and those yet to come, is the ability to sense and measure accurately automotive parameters. Equally important at the output is the ability to initiate control actions accurately in response to commands. In other words, sensors and actuators are the heart of any automotive electronics application. That is why they have been placed first in this handbook where they are described in technical depth. In other chapters, application-specific discussions of sensors and actuators can be found.

The importance of sensors and actuators cannot be overemphasized. The future growth of automotive electronics is arguably more dependent on sufficiently accurate and low-cost sensors and actuators than on computers, controls, displays, and other technologies. Yet it is those nonsensor, nonactuator technologies that are to many engineers the more "glamorous" and exciting areas of automotive electronics.

In the section on control systems, a key in-depth chapter deals with automotive microcontrollers. Without them, all of the controls described in the chapters that follow in that section—engine, transmission, cruise, braking, traction, suspension, steering, lighting, windshield wipers, air conditioner/heater—would not be possible. Those controls, of course, are key to car operation and they have made cars over the years more drivable, safe, and reliable.

Displays, trip computers, and on- and off-board diagnostics are described in another section, as are systems for passenger safety and convenience, antitheft, entertainment, and multiplex wiring. Displays and trip computers enable the driver to readily obtain valuable information about the car's operation and anticipated trip time. On- and off-board diagnostics have of necessity become highly sophisticated to keep up with highly sophisticated electronic controls. Passenger safety and convenience items and antitheft devices add much to the feeling of security and pleasure in owning an automobile. Entertainment products are what got automotive electronics started and they continue to be in high demand by car buyers. And multiplex wiring, off to a modest start in production cars, holds great promise for the future in reducing the cumbersome wiring harnesses presently used.

The section on electromagnetic interference and compatibility emphasizes that interference from a variety of sources, if not carefully taken into account early on, can raise havoc with what otherwise would be elegant automotive electronic designs. And automotive systems themselves, if not properly designed, can cause interference both inside and outside the automobile.

In the final section on emerging technologies, some key newer areas are presented:

- Navigation aids and intelligent vehicle-highway systems are of high interest worldwide since they hold promise to alleviate many of vehicle-caused problems and frustrations in our society.
- While it may be argued that electric vehicles are not an emerging technology, since they have been around for many years, it certainly is true that they have yet to come into their own in any really meaningful way.
- Electronic noise cancellation is getting increasing attention from automobile designers seeking an edge over their competitors.

The final chapter on future vehicle electronics is an umbrella discussion that runs the gamut of trends in future automotive electronics hardware and software. It identifies potential technology developments and trends for future systems.

Nearly every chapter contains its own glossary of terms. This approach, rather than one overall unified glossary, has the advantage of allowing terms to be defined in a more application-specific manner—in the context of the subject of each chapter. It should also be noted that there has been no attempt in this handbook to cover, except peripherally, purely mechanical and electrical devices and systems. To do so would have restricted the number of pages available for automotive electronics discussions.

Finally, the editor would like to thank all contributors to the handbook and particularly two individuals: Otto Holzinger of Robert Bosch GmbH in Stuttgart, Germany and Randy Frank of Motorola Semiconductor Products in Phoenix, Arizona. Holzinger organized the many contributions to this handbook from his company. Frank, in addition to contributing two chapters himself and cocontributing a third, organized the other contributions from Motorola. Without their help, this handbook would not have been possible.

Ronald K. Jurgen

CHAPTER 2
PRESSURE SENSORS

Randy Frank
Technical Marketing Manager
Motorola Semiconductor Products

2.1 AUTOMOTIVE PRESSURE MEASUREMENTS

Various pressure measurements are required in both the development and usage of vehicles to optimize performance, determine safe operation, assure conformance to government regulations, and advise the driver. These sensors monitor vehicle functions, provide information to control systems, and measure parameters for indication to the driver. The sensors can also provide data log information for diagnostic purposes.

Depending on the parameter being measured, different units for indicating pressure will be used. Since pressure is force per unit area, basic units are pounds per square inch (psi) or kilograms per square centimeter. For example, tire pressure is usually indicated in psi. Manifold pressure is typically specified in kiloPascals (kPa). A Pascal, which is the international unit (SI or Systems Internationale) for pressure, is equal to 1 Newton per meter^2 or 1 kg $\cdot$ m^{-1} $\cdot$ s^{-2}. Other common units of pressure measurement include: inches, feet, or centimeters of water; millibars or bars, inches, or millimeters of mercury (Hg), and torr. The conversion constants as defined per international convention are indicated in Table 2.1.

Pressure can be measured by a number of devices that provide a predictable variation when pressure is applied. Sensors used on vehicles range from mechanical devices—with position movement when pressure is applied—to a rubber or elastomer diaphragm, to semiconductor-based silicon pressure sensors. Various pressure-sensing techniques are explained in Sec. 2.3.

The type of pressure measurement that is made can be divided into five basic areas which are independent of the technology used for the measurement: gage, absolute, differential, liquid level, and pressure switch.

2.1.1 Gage Pressure Measurements

The silicon pressure sensor technology explained in Sec. 2.3.5 is used to visualize the difference between gage, absolute, and differential pressure (refer to Fig. 2.1). For gage pressure measurements, the pressure is applied to the top of a (silicon) diaphragm (Fig. 2.1), creating a positive output. The opposite (back) side of the diaphragm is exposed to atmospheric pressure. Gage vacuum is a special case of gage pressure. For gage vacuum measurements, vacuum is applied to the back side of the diaphragm resulting in a positive output signal. Gage and gage vacuum are single-sided pressure measurements. Gage pressure is frequently indicated by psig.

TABLE 2.1 Pressure Unit Conversion Constants

(Most commonly used per international conventions)

	psi*	In H_2O†	In Hg‡	K Pascal	Millibar	cm H_2O§	mm Hg¶
psi*	1.000	27.680	2.036	6.8947	68.947	70.308	51.715
In H_2O†	3.6127×10^{-2}	1.000	7.3554×10^{-2}	0.2491	2.491	2.5400	1.8683
In Hg‡	0.4912	13.596	1.000	3.3864	33.864	34.532	25.400
K Pascal	0.14504	4.0147	0.2953	1.000	10.000	10.1973	7.5006
Millibar	0/01450	0.40147	0.02953	0.100	1.000	1.01973	0.75006
cm H_2O§	1.4223×10^{-2}	0.3937	2.8958×10^{-2}	0.09806	0.9806	1.000	0.7355
mm Hg¶	1.9337×10^{-2}	0.53525	3.9370×10^{-2}	0.13332	1.3332	1.3595	1.000

* PSI = pounds per square inch
† At 39 °F
‡ At 32 °F
§ At 4 °C
¶ At 0 °C

2.1.2 Absolute Pressure Measurements

An absolute pressure measurement is made with respect to a fixed (usually a vacuum) reference sealed within the sensor (Fig. 2.1). For a 100-kPa-rated absolute unit, the diaphragm is fully deflected with standard atmospheric pressure. Application of a vacuum restores the diaphragm to its undeflected (flat) position. The result is a high-level output with no vacuum applied and a low-level signal at full vacuum unless the zero is established at the full-scale deflection of the diaphragm. Pressure can also be applied to absolute units with appropriately designed diaphragms to withstand the additional applied stress. Absolute pressure is frequently indicated by psia.

2.1.3 Differential Pressure Measurements

Differential or Delta-P measurements are also shown in Fig. 2.1. The higher pressure is applied to the top of the diaphragm and the lower pressure, possibly a reference pressure, is applied to the opposite side. The diaphragm's deflection is a result of the pressure difference.

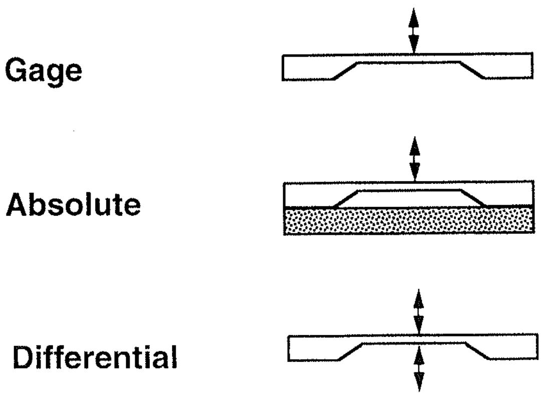


FIGURE 2.1 Types of pressure measurements: (*a*) gage, (*b*) absolute, and (*c*) differential.

Typically, the pressure differential is only a small percentage of the total line pressure and a system fault can expose one side of the sensor to the full line pressure. This must be taken into account when choosing the sensor and determining the rated pressure range that will be required. Differential pressure is frequently indicated by psid.

2.1.4 Liquid Level Measurements

The height of a column of liquid can be measured by a pressure sensor. The term *head* is frequently used in hydraulics to denote pressure. Measurements of inches or feet of water and centimeters of mercury are direct indications of the effect of pressure on liquid level. Other liquid levels are dependent on their specific weight and can be calculated by $h = (P_L - P_H)/w$, where $(P_L - P_L)$ is the pressure differential caused by the height of the fluid column and w is the specific weight of the liquid. The vapor pressure in a sealed enclosure will have an effect on the measurement of liquid height. Returning the reference side of a differential pressure sensor to the top of the enclosure will compensate for vapor pressure.

2.1.5 Pressure Switch

A pressure switch is typically achieved by mounting an electric contact on a diaphragm (rubber or any elastic material). The application of sufficient pressure (or vacuum) on one side of the diaphragm causes the movable contact to meet a stationary contact and close the circuit.

A pressure switch can also be achieved by any of the previously described techniques merely by establishing a reference threshold voltage that is calibrated to indicate the point that the pressure changes from an acceptable to unacceptable (or low to high) level. Once the threshold voltage is achieved, additional electronic circuits can be used to produce an electronic switch that can control loads such as an indicator lamp.

2.2 AUTOMOTIVE APPLICATIONS FOR PRESSURE SENSORS

Automotive requirements for pressure measurements range from the basic—oil pressure—to the sophisticated—air pressure differential from one side of the vehicle to the other. This section elaborates on the various possibilities for pressure measurements that exist either in the development, laboratory, or pilot phases of the vehicle, to actual volume production. Table 2.2 lists a number of potential pressure measurements versus vehicle systems and provides an indication of the pressure range and type of measurement.

Automotive specification and testing guidelines have been developed and published by the Society of Automotive Engineers (SAE) specifically for manifold absolute pressure (MAP) sensors. These documents are intended to assist in establishing test methods and specifications for other sensors. Other SAE documents that may apply to sensors are summarized in Table 2.3.

The packaging and testing requirements for automotive sensors can represent 50 to 80 percent of the sensor cost and over 90 percent of the warranty and in-service problems. The pressure-sensing applications that are presented in the following sections will include packaging requirements that are of particular concern.

2.2.1 Existing Applications for Pressure Sensors

A late twentieth century production vehicle is likely to have a number of pressure sensors for measurements such as manifold pressure and engine oil pressure, and has the potential for

TABLE 2.2 Pressure Sensing Requirements for Various Vehicle Systems

System	Parameter	Pressure range	Type
Engine control	Manifold absolute pressure	100 kPa	Absolute
	Turbo boost pressure	200 kPa	Absolute
	Barometric pressure (altitude)	100 kPa	Absolute
	EGR pressure	7.5 psi	Gage
	Fuel pressure	15 psi—450 kPa	Gage
	Fuel vapor pressure	15 in H_2O	Gage
	Mass air flow		Differential
	Combustion pressure	100 Bar, 16.7 Mpa	Differential
	Exhaust gas pressure	100 kPa	Gage
	Secondary air pressure	100 kPa	Gage
Elect transmission (continuously variable transmission)	Transmission oil pressure	80 psi	Gage
	Vacuum modulation	100 kPa	Absolute
Idle speed control	AC clutch sensor/switch	300–500 psi	Absolute*
	Power steering pressure	500 psi	Absolute*
Elect power steering (also elect assisted)	Hydraulic pressure	500 psi	Absolute*
Antiskid brakes/ traction control	Brake pressure	500 psi	Absolute*
	Fluid level	12 in H_2O	Gage
Air bags	Bag pressure	7.5 psi	Gage
Suspension	Pneumatic spring pressure	1 MPa	Absolute*
Security/keyless entry	Passenger compartment pressure	100 kPa	Absolute
HVAC (climate control)	Air flow (PC) Compressor pressure	300–500 psi	Absolute*
Driver information	Oil pressure	80 psi	Gage
	Fuel level	15 in H_2O	Gage
	Oil level	15 in H_2O	Gage
	Coolant pressure	200 kPa	Gage
	Coolant level	24 in H_2O	Gage
	Windshield washer level	12 in H_2O	Gage
	Transmission oil level	12 in H_2O	Gage
	Tire pressure	50 psi	Gage/absolute
	Battery fluid level	1–2 in below	Optical
Memory seat	Lumbar pressure	7.5 psi	Gage
Multiplex/diagnostics	Multiple usage of sensors		

* Gage measurement but absolute sensors used for failsafe

TABLE 2.3 SAE Specifications That Effect Pressure Sensors

Recommended environmental practices for electronic equipment design	SAE J1211
Performance levels and methods of measurement of electromagnetic radiation from vehicles and devices	SAEJ551
Performance levels and methods of measurement of EMR from vehicles and devices (narrowband RF)	SAEJ1816
Electromagnetic susceptibility procedures for vehicle components (except aircraft)	SAE J1113
Vehicle electromagnetic radiated susceptibility testing using a large TEM cell	SAE J1407
Open-field whole-vehicle radiated susceptibility 10 kHz–18 GHz, electric field	SAE J1338
Class B data communication network interface	SAE J1850
Diagnostic acronyms, terms, and definitions for electrical/electronic components	SAE J1930
Failure mode severity classification	SAE J1812
Guide to manifold absolute pressure transducer representative test method	SAE J1346
Guide to manifold absolute pressure transducer representative specification	SAE J1347

several other pressure measurements. Tighter emissions control and improved efficiency may necessitate further sensor use in future systems.

Manifold, Barometric and Turbo Boost Pressure. Manifold absolute pressure (MAP) is used as an input to fuel and ignition control in internal combustion engine control systems. The speed-density system that uses the MAP sensor has been preferred over mass air flow (MAF) control because it's less expensive, but stricter emission standards are causing more manufacturers to use mass air flow for future models.

Higher resolution from 32-bit engine controllers, with greater analog-to-digital (A/D) conversion capability and higher operating frequencies, will provide greater accuracy for a given MAP sensor during the critical transitions of the engine cycle. As shown in Fig. 2.2, previous changes from 8-bit to 16-bit controllers have resulted in a two-time improvement in resolution in the digital conversion for the intake manifold pressure. The 8-bit control unit performed the A/D conversion on a 4-ms timer interrupt in order to maintain a balance with other controls, with the resulting 1.1-ms lag time (worst case) during periods of overlapping interrupts. The 16-bit microcontroller performs the A/D conversion every 2 ms, which reduces the lag time to 0.3 ms. The actual system improvements that can result from using the higher performing microcontrol units is a result of other factors such as more precise and faster control of fuel injectors and sparkplugs, and additional and/or more accurate sensors and control algorithms.

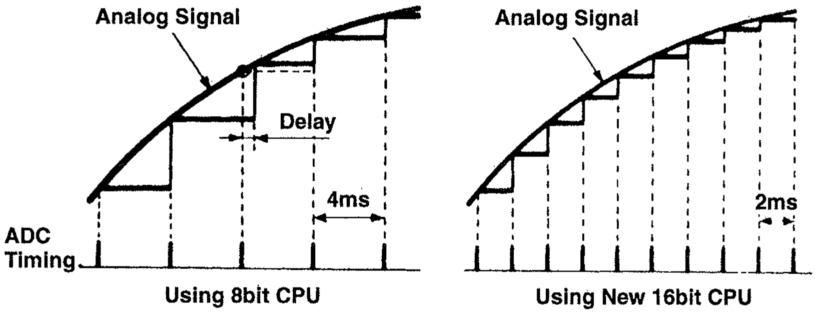


FIGURE 2.2 Effect of A-D on pressure measurements.

The MAP error band is also being tightened with a goal of 1 percent accuracy over the entire automotive temperature range. Existing specifications allow tolerances to increase as shown by the bowtie specification in Fig. 2.3 with associated multiplier(s).

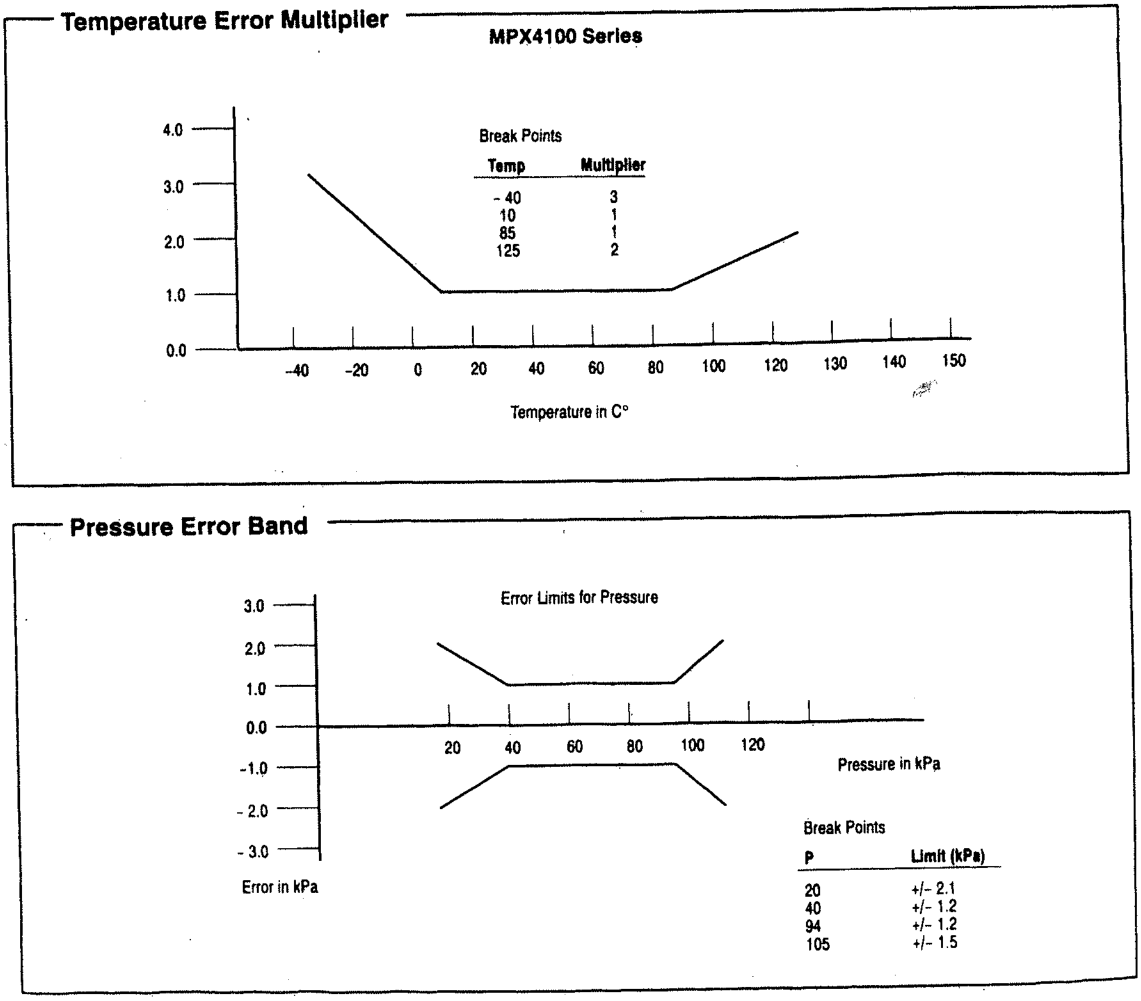
The need for barometric pressure is often desirable in MAF systems to provide altitude information to the engine control computer. The barometric pressure range is typically from 60 to 115 kpA with accuracy on the order of 1.5 percent over the operating pressure range. The error band tolerance increases by a temperature multiplier of up to 2× outside the 0 to 85 °C range. MAP and barometric pressure sensors are frequently mounted inside control modules making the mounting technique a key consideration for manufacturability. The increased usage of surface mount technology, and the need to reduce space so that additional features can be included in control modules are factors that will affect next-generation sensor designs.

A typical turbocharger can provide boost pressure in the range of 80 kPa over the naturally aspirated internal combustion engine. This increases the maximum rating for the sensor to 200 kPa absolute, but other requirements are scaled appropriately.

MPX4100 • MPX4101 SERIES

Transfer Function

Nominal Transfer Value: $V_{out} = V_S\ (0.01059 \times P - 0.1518)$
$+/-$ (Pressure Error x Temp. Mult. x $0.01059 \times V_S$)
$V_S = 5.1\ V \pm 5\%\ P_{in}$ kPa



Break Points

Temp	Multiplier
- 40	3
10	1
85	1
125	2

Break Points

P	Limit (kPa)
20	+/- 2.1
40	+/- 1.2
94	+/- 1.2
105	+/- 1.5

FIGURE 2.3 Error band for MAP sensor.

Oil Pressure. Oil pressure on automobiles has traditionally been measured by the deflection of a rubber diaphragm which closes a set of contacts (switch) providing a lamp indication with low oil pressure or moves a potentiometer to provide an analog signal for a gage.

A replacement for the conventional electromechanical oil pressure sending unit is an electronic device such as the one designed by Chrysler's Acustar Division. In addition to a silicon piezoresistive pressure sensor, the unit contains transient protection circuitry, signal amplification for the sensor output, and output drivers for both an electromagnetic gauge and a fuel pump. The FET output drivers are capable of handling 10 A based on the heat dissipated though a heatsink that is integral to the sensor package.

The unit utilizes a supply voltage from 9 to 16 V and operates over a media temperature range of –40 to +150 °C. The overall accuracy is ±3.25 percent with linearity ≤ ±0.25 percent over the entire operating pressure range of 200 psi. The switch point for the low pressure indication is 4 psi ± 1.5 psi.

The sensor package was specially designed for easy assembly. The housing interfaces to the sensor with an extremely reliable O-ring seal that can withstand a burst pressure over 400 psi. Special materials were used for both the package and the protective gel that covers the sensor, which allow it to survive qualification tests with over 1 million pressure cycles, including portions conducted at high temperatures. This exceeds the number of cycles that can be achieved with traditional diaphragm-driven potentiometers that have been used for providing the indication of oil pressure. The sensor has been designed for a 10-year/100,000-mile life that could be required for future vehicle warranties.

Media Compatibility in Automotive Measurements. Pressure sensors frequently have to interface with an environment that is more demanding than other electronic components. For example, the measurement of engine oil pressure, transmission oil pressure, fuel pressure, and so on, or fluid level (oil, gas, coolant, etc.) requires the sensor package to be exposed to one or more fluids that are detrimental to the operation of semiconductor circuitry. Each of these media interface problems is addressed separately, depending on the application. Automotive cost requirements usually limit the usage of stainless steel as the isolation technique. Instead, more cost-effective protective polymers and chemically tolerant plastic and rubber materials have been developed for sensor packages.

2.2.2 New Applications for Pressure Sensing

The list of potential applications for sensing pressure in the automobile includes several new applications. These measurements are frequently made during the development of new vehicle systems. Their actual use on the vehicle is determined by factors such as cost, legislated requirements, need for diagnostics, and value added to the system. Applications in this section will identify areas of concern, range of pressure measurement, and factors that affect the usage of a pressure sensor.

Transmission Oil Pressure and Brake Pressure. Transmission pressure is required as an input in computer-controlled transmission shift points. This pressure can be measured with sensors similar to those developed for engine oil pressure.

Pressure in a hydraulic system, such as the master cylinder of an antilock brake system (ABS), is much higher than transmission oil pressure typically requiring a sensor with minimum rating of 500 psig. Pressures in other locations in the ABS system can be lower. The dynamic pressures in brake tubing can be of interest during the development phase of passenger vehicles and may be of interest in heavy duty commercial vehicles. These pressures can be below 150 psig.

The ABS system controls front and rear tire slip. Tradeoffs that exist in developing an ABS system for a particular vehicle include stopping deceleration to achieve the shortest possible stopping distance versus more steering control. Increased yaw stability can be obtained by reducing the deceleration rate of the rear wheels. The addition of traction control to the system improves stability during acceleration and provides independent control of each wheel during a variety of driving maneuvers for improved vehicle performance.

Passenger vehicles may have a single pressure sensor to monitor the pressure in the hydraulic system. One system, General Motors' ABS-VI, provides information on the brake pressure by detecting the current going to motors in the system. For the ABS-VI system, a pressure sensor is not required to provide optimum brake pressure at each wheel. However, other systems rely on the rate of brake application and release to control lockup. Commercial vehicles may have several sensors for sensing brake pressures. Sensors close to brake cylinders report the actual pressure, which is compared to the reference value stored in the control unit.

Tire Pressure. The continuous monitoring of tire pressure offers increased fuel economy and safety to passenger cars or commercial vehicles. Underinflated tires create excessive rolling friction and therefore decrease fuel economy. Overinflated tires have excessive stress that can result in failure during operation. Improperly inflated, either over- or underinflated, tires have uneven wear patterns which decrease tire life. Available tire pressure systems consist of a tire pressure sensor (or switch) at each wheel, wheel speed indicator, temperature sensor, a radio frequency transmitter, electronic receiver/controller, and a display unit. The dashboard display provides an indication to the driver that the tires are improperly inflated. Tire pressure increases with temperature approximately 1.5 psi for every 10 °C of tire air temperature rise, so the system must provide correction for this effect. Abrupt increases in temperature and pressure can be sensed by these systems and provide a warning of eminent tire failure providing an additional safety factor.

Another tire pressure system utilizes a separate hand-held reader to easily verify the proper tire inflation when the vehicle is stationary. Yet another commercial vehicle system for trucks with dual tires operates while the vehicle is stationary and employs a visual indication for the driver that adequate pressure exists. This system provides a single fill point for the dual tires, maintains equal pressure under normal conditions, and provides an isolation valve in the event that a blowout or slow leak develops in one tire.

EGR Pressure. EGR (exhaust gas recirculation) back pressure and a pressure differential exist across the EGR valve used to control NO_x emissions in the engine control system. The valve is modulated by a vacuum which lifts a pintle from its seated position to allow exhaust gas to be recirculated. A change in vacuum pressure from 50 to 90 mm Hg is sufficient to fully open the valve, and a pressure differential across the valve of 200 mm Hg is typical. Pressure measurements are made during the development phase to determine system operating characteristics. However, a position sensor is typically used to measure the EGR valve's position and control NO_x emissions during normal vehicle operation.

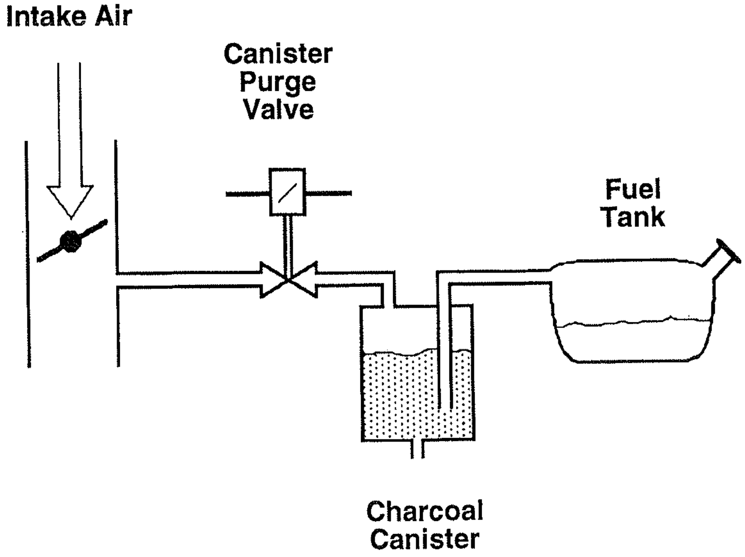


FIGURE 2.4 Canister purge system.

Fuel Rail and Vapor Pressure. Evaporative emissions that occur when the engine is off are currently stored in an activated charcoal canister of about 850 to 1500 cc until the engine is running, as shown in Fig. 2.4. The vapors are then consumed by the combustion chamber and catalytic converter. One implementation of on-board vapor containment of refueling hydrocarbons (on-board refueling vapor recovery or ORVR) would require refueling canisters on the vehicle that could be three to four times the volume of existing canisters. If leaks need to be detected in this system, a diagnostic pressure sensor may be required.

One approach to fuel routing, employed in the 5.9-liter Dodge Magnum engine, is to mount the fuel filter and pressure regulator at the fuel tank. The fuel pump is mounted inside the tank. Therefore, since only a fuel supply line to the fuel rail and a line to the evaporative canister are necessary, the fuel return line is eliminated. This system maintains lower fuel tank temperatures, resulting in lower evaporative emissions.

Monitors required for on-board diagnostic (OBD) systems per California Air Resources Board's (CARB's) OBDII legislation were originally targeted to be phased in between model year 1994 and model year 1996. A Bosch fuel injection system with on-board diagnostics is shown in Fig. 2.5 that identifies a differential air pressure sensor for tank vapor pressure.

Overpressure Occurrences. Fuel supply pressures in automobile fuel injection systems normally operate at pressures below 75 psi; however, fuel pumps develop pressures up to 3200 psi to open injectors. Pressure spikes can be reflected back through the fuel supply system that measure up to 300 psi during each fuel injection pulse.

Overpressure created by backfire can also apply a positive pressure of 75 psi and higher to the intake manifold absolute pressure sensor. Techniques used to prevent failure from over-

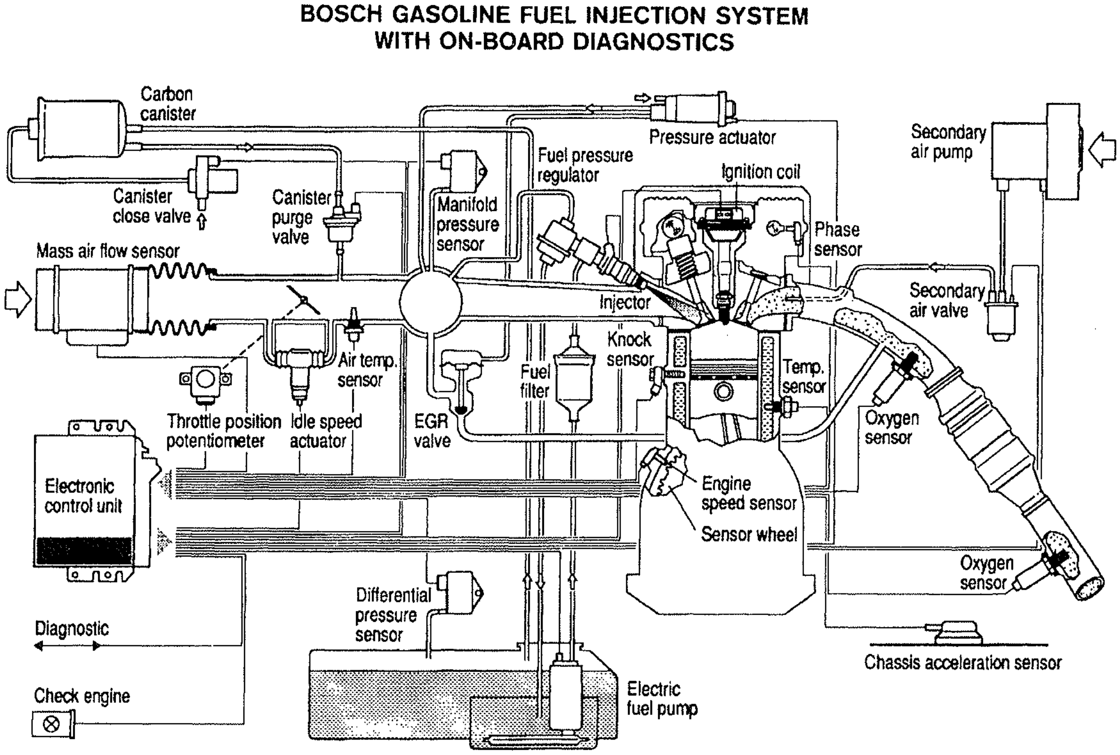


FIGURE 2.5 Fuel vapor control in electronic fuel injection system. *(Courtesy Robert Bosch, GMBh)*

pressure include overpressure stop built into the transducer, mechanical pulse filtering, and a sensor designed to operate within the overpressure range.

Mechanical stops have been a traditional protection technique for mechanical pressure sensors. This is possible where the amount of diaphragm deflection is large. Silicon pressure sensors have a modulus of elasticity that is the same as steel (30×10^6 psi) and a yield strength (180,000 to 300,000 psi) that is higher than steel, allowing high overpressure without diaphragm damage. However, the sensor package itself must be designed to handle the maximum pressure safely.

Snubbing, or mechanical filtering, is commonly used for static pressure measurements. A small diameter tube reduces the dynamic variation in applied pressure. If dynamic measurements are desired, the ac component of the desired signal may also be attenuated.

Increasing the diaphragm thickness of the sensor to safely handle the full range of pressure within normal operating range will also result in a lower sensitivity.

Alternate Fuel and Alternate Engine Implications to Pressure. Legislation that requires a percentage of the vehicles sold in California to be LEV (low emissions vehicles) and even ZEV (zero emissions vehicles) is increasing the demand for alternate fuel and electric vehicles. Among the alternate fuel vehicles, CNG (compressed natural gas) and hydrogen cells most likely would require sensors for pressure measurements. CNG is pressurized at 3000 psi and the distribution system includes pressure regulators, a transducer, valves, and idle air solenoids. Before the natural gas enters the engine, a regulator reduces the fuel pressure to near atmospheric pressure. Sensing may be required in both low- and high-pressure portions of the system. Development of alternative engines, such as the two-stroke engine for vehicle applications, will utilize electronics for control functions similar to four-cycle engines. However, the range and necessity for pressure measurements will differ from four-cycle engines. The pressure range for direct fuel injection is considerably higher for a two-stroke engine. The need to control the oil pump may necessitate pressure sensing in two-stroke systems as well.

Hydrogen fuel cells are another potential source of energy for use in electric vehicles. In one design, the proton-exchange membrane (PEM) design, a turbocompressor is used to pressurize the system and maintain hydration of the membrane. A pressure of at least three atmospheres (0.3 MPa) is required to remove the water. This pressure or the pressure drop across the membrane may need to be monitored during operation.

2.2.3 Other Applications for Pressure Sensors

Pressure sensors can be used on vehicles for measuring flow through pressure differential, or delta-P, measurements and for determining liquid level.

Delta-P (Flow-Sensing) Measurements. Applications on the vehicle for flow sensing include mass air flow; heating compartment flow; oil, fuel, and cooling liquid flow; and vehicle flow in an air stream. Mass air flow is typically accomplished by hot-wire anemometer or Karman vortex flow meters which do not use pressure-sensing techniques. Other vehicle flow requirements, including the pressure drop across the air filter, could be sensed and monitored by a differential pressure sensor. In addition to requirements such as media compatibility (Sec. 2.2.1), the lower-level signals require higher-sensitivity pressure sensors and/or additional amplification and must tolerate faults that could apply full line pressure to only one side of the sensor.

One of the more interesting applications of differential pressure measurements applied to flow analysis is the flow of the vehicle itself relative to crosswinds. A rear-wheel steering system developed by Daimler-Benz uses two pressure sensors to measure the pressure caused by wind on the vehicle's sides. An electronic control unit analyzes the pressure difference and inputs from other sensors, and alters the rear-wheel setting according to the wind strength. The system that measures the crosswinds directly is faster than yaw sensors, which are reactive and measure the change in attitude and direction of the vehicle.

Other laboratory measurements of airflow and crosswind force have also been made by Daimler-Benz that utilize 10 differential pressure gages with a range of ±3700 kPa. All pressure sensors were connected to a single pressure vessel to have a common reference. The reference pressure was measured by a 100-kPa absolute pressure sensor. The measurements were used to determine aerodynamic forces and moments and to compensate for wind effects in an active steering system.

Fluid Level Sensing. Various liquid levels can be measured in a vehicle, as shown in Table 2.4. All of these requirements, except fuel level, could be satisfied by a switch that simply detects that a predefined minimum level of liquid has been reached so that a driver indication can be provided. This can be accomplished by directly illuminating a dash lamp or through a microcontroller in a body computer which activates an output driver.

TABLE 2.4 Liquid Level Measurements

Level	Type	Range
Engine oil	Switch	38.1 cm of water
Transmission oil	Switch	30.5 cm of water
Coolant	Switch	61 cm of water
Windshield washer fluid	Switch	30.5 cm of water
Battery	Refraction switch	5.1 cm below reference
Power steering fluid	Switch	7.6 cm below reference
Brake fluid	Switch	30.5 cm of water
Fuel	Sensor and switch	38.1 cm of water

Sensing the fuel level has traditionally been performed by a float to sense the fluid level and a variable resistor with the wiper arm connected to the float. Configuration for the sensor depends upon the specific tank for which it was designed, necessitating a unique sensor for each vehicle. Manufacturers with several different vehicle models have the additional impetus to replace the existing techniques with a nonwearing, more accurate, self-calibrating electronic alternative. However, media compatibility for fuel level is among the harshest requirements for a pressure sensor. In addition to gasoline, oxygenated fuels containing ethanol, methanol, benzene, MTBE, engine additives, and even sour gas must be tolerated by the sensor. Nonintrusive differential sensors isolate the liquid from the sensor interface but must still tolerate fuel vapors. Also, the sloshing of the fuel in a vehicle's tank requires a time amplitude filter to smooth out the indication provided to the driver.

2.2.4 Combustion Pressure

The direct measurement of combustion pressure is being investigated for detecting misfire to meet CARB OBDII requirements. The high pressure (≥16 MPa) and temperature ranges combined with other environment factors make the design of a pressure sensor for this application extremely expensive. As a result, other techniques are being developed as alternatives to direct pressure measurement. These technologies include optical, fiberoptics measuring luminous emissions from combustion, and noncontact torque sensors. Section 2.3.7 explains a fiber optic technique.

The operation of the reciprocating-piston, internal combustion engine is represented by a constant volume process and the engine power cycle is analyzed by using pressure-versus-volume and pressure-versus-crank angle diagrams. To obtain these measurements in a laboratory environment a number of techniques have been developed. Direct (in-cylinder) pressure measurements have been performed with small diameter piezoresistive sensors placed in (or near) the sparkplug and piezoelectric washers placed under the spark plug. A high natural frequency is required for these sensors based on the dynamic measurement involved in the combustion process. Indirect measurements with shaft torque and optical phase shift are additional possibilities. The need to determine misfire due to component failure during vehicle operation is part of OBDII requirements. A sensor used on production vehicles will be required to survive the high pressure and temperature environment for the life of the vehicle, which could be 100,000 miles and 10 years. It must also have no need for periodic zeroing or calibration and be available at a low cost.

2.2.5 Other Pressure Measurements

An adaptive suspension system (see Chap. 17) can be accomplished with an air pressure controlled shock absorber damping, such as Mitsubishi's Active-ECS (Electronically Controlled Suspension). This system has two air pumps and nine solenoids that regulate air pressure based on inputs from sensors including an air pressure sensor in the rear of the vehicle that measures the passenger and cargo load. The driver can select soft, medium, or hard suspension. Another system utilizes an air reservoir charged to a pressure of 1 Mpa by a compressor. A pressure switch monitors when the pressure drops below 760 kPa to recharge the reservoir. Air springs operate at 300 kPa unloaded and at 600 kPa in the rear when fully loaded.

HVAC (heating, ventilating, and air conditioning) changes are occurring as manufacturers are required to convert from refrigerant CFC-12 to alternatives such as HFC-134a. The theoretical performance of these two refrigerants will mean about a 6 percent loss in efficiency, compressor discharge pressure that is 175 kPa higher, and a discharge temperature that is about 8 °C lower. Measuring the compressor discharge pressure (almost 1900 kPa for the HFC-134a system) is desirable for electric load control as vehicles add more requirements to the 12-V charging system. Also, the effect on engine performance and fuel economy when the A/C is used could make the refrigerant pressure sensor a standard vehicle sensor in the future. An absolute sensor used to measure gage pressure of the refrigerant provides a deadhead effect to prevent refrigerant loss in the event of a sensing diaphragm failure.

The measurement of the pressure developed when the air bag is inflated is part of the evaluation, qualification, and lot acceptance criteria of air bag inflating techniques. Time-to-peak pressure and peak tank pressure measurements require measurements in the tens of milliseconds range. Inflated bag pressures are in the range of 100 kPa or less. Hybrid inflators use a stored inert gas, such as argon, in place of sodium-azide propellant that requires a squib for ignition. The hybrid uses a pressure sensor to monitor the status of the stored gas.

Special heavy duty/commercial truck measurements require pressure measurements that are quite different from those made on passenger cars. Accumulator-type fuel injection systems for direct injection diesel engines have fuel pressurized to 20 to 100 MPa in the accumulator by a high-pressure pump. The accumulator pressure is monitored in this approach to reduce particulates. Another method to reduce diesel particulates utilizes a ceramic fiber as a filter in a canister. A pressure sensor monitors the backpressure and allows the full filter to be regenerated by burning off the accumulated particulates in the filter. A heater element in the trap has power supplied from the power-switching module. A temperature approaching 1300 °F (700 °C) is reached inside the filter cartridge to incinerate the particles.

Lumbar support systems utilize a pressurized system with a pressure sensor (≤7.5 psig) as the feedback element controlling the air pump to provide additional support to the driver's lower back in luxury vehicles. Pressure-sensitive grids have been used in the development process to automatically measure up to 3600 contact points for visual display and weight distribution analysis.

2.2.6 Partial Pressure Measurements

The oxygen (or lambda-) sensor in engine control systems is a chemical sensor that utilizes partial pressures to provide a feedback signal for the closed-loop control system. Lambda is defined as the actual air/fuel ratio divided by the stoichiometric (14.6) fuel ratio. The operation of the oxygen sensor is defined by the Nernst equation: $U_L = RT/4F \cdot \ln(P_{O2}''/P_{O2}')$, where U_L is the unloaded output voltage of the sensor, R is the universal gas constant, T is the absolute temperature, F is the Faraday constant, P_{O2}'' is the oxygen partial pressure of the air (about 2.9 psi), and P_{O2}' is the equilibrium partial pressure of the oxygen in the exhaust gases. Equilibrium occurs due to the catalytic activity of the platinum electrodes used to coat the inside and outside of the Y_2O_3 stabilized ZrO_2 ceramic electrode. The oxygen partial pressure changes by a factor of 10^7 at 900°C (or 10^{19} at 500°C) when the exhaust gas changes from a reducing environment (lambda = 0.999) to an oxidizing environment (lambda = 1.001).

2.3 TECHNOLOGIES FOR SENSING PRESSURE

A number of technologies have been used for on-vehicle measurements of static and dynamic pressure: diaphragm-potentiometer, linear variable differential transformer (LVDT), aneroid, silicon or ceramic capacitive, piezoresistive strain gage, piezoelectric ceramic or film, and optical phase shift (combustion pressure). Recent advances in sensing have focused on transducers that provide an electric signal easily interfaced to microcontrollers. Mechanical devices are frequently used in the laboratory for calibration and component development or on the vehicle during the development phase of the vehicle systems. Common mechanical devices include the Bourdon tube, bellows, diaphragms, deadweight gage, and manometer.

The Bourdon tube is a curved or twisted, flattened tube with one end closed that acts as a force collector. When pressure is applied at the open end, the tube tends to straighten and the resulting motion is used as an indication of the applied pressure.

The bellows, or pressure capsule, is a chamber that expands with applied pressure. Absolute pressure can be measured by sealing a reference pressure (e.g., vacuum) inside a closed unit and applying pressure to the outside. The movement of the chamber is proportional to the applied pressure.

Diaphragms are the most common force collector used in modern pressure sensors. The diaphragm material can be rubber, elastomer, stainless steel, silicon, ceramic, or even sapphire. Diaphragm shapes are circular or square and can be supported or clamped around their periphery.

A deadweight tester or piston gage can withstand extreme pressure changes and high overpressure occurrences. The piston is sealed in a cylinder using O-ring or Teflon seals. Pressure on the piston causes a deflection that can be measured by position-sensing techniques. Precision weights allow calibration for high-accuracy measurements. The deadweight tester is one of the few pressure-sensing techniques that measures pressure in terms of its fundamental units—force and area. Errors associated with the measurement are air-bouyancy corrections, gravity error, surface tension, fluidhead, and thermal expansivity. These errors are normally small but should be taken into account when high accuracy is required.

The manometer is used both as a pressure-measuring instrument and a standard for calibrating other instruments. Its simplicity and inherent accuracy result from it being the measurement of the height(s) of a column of liquid. Three basic types of manometers are the U-tube, well (cistern), and slant-tube.

Other measurement devices including McLeod, Pirani, Alphatron, and thermocouple gages which can measure vacuum in the range of 10^{-5} mm Hg.

Sensing techniques that provide a transducer for conversion from mechanical to electrical units include resistive, LVDT, capacitive, piezoresistive, and piezoelectric.

2.3.1 Potentiometric Pressure Sensors

Prior to electronic engine control systems, carburetor dashpots and distributor vacuum advance units used the distance that a rubber diaphragm traveled when pressure was applied as a mechanical indication of pressure. A diaphragm which moves a potentiometer (resistor with a sliding element or wiper) provides an electric signal that can be applied to a remote gage such as an oil pressure gage. Potentiometric sensors inherently have some level of noise and wear associated with their operation due to the contact of the wiper to the resistive element. Stiction or static friction is also a potential concern with these devices, especially if control of ≤0.5 percent of the total resistance is required. The finite resolution of wirewound potentiometers is overcome by the use of newer thin-film plastic resistor designs.

2.3.2 Linear Variable Differential Transformer

One of the earliest pressure inputs for engine control systems was provided by the LVDT pressure sensor. The principle of operation is demonstrated in Fig. 2.6. An LVDT pressure sensor consists of a primary winding and two secondary windings positioned on a movable cylindrical core. The core is attached to a force collector which provides differential coupling from the primary to the secondaries resulting in a position output that is proportional to pressure.

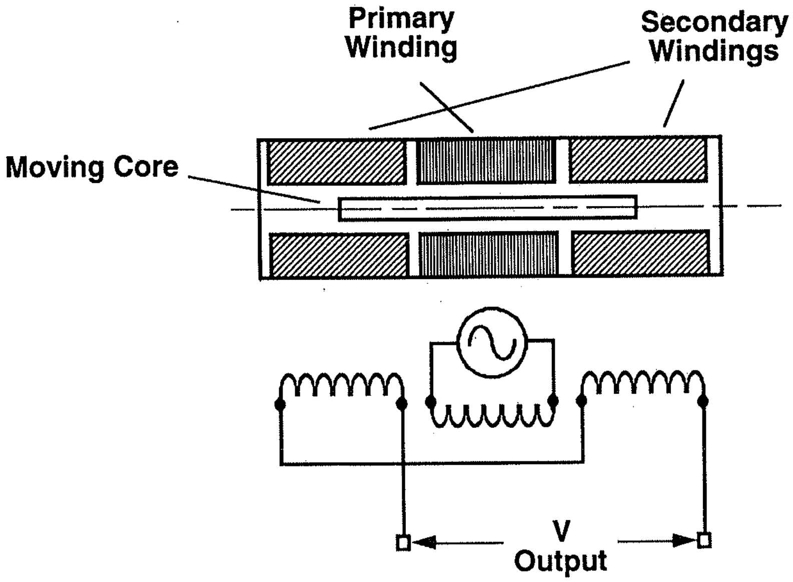


FIGURE 2.6 LVDT pressure sensor.

An alternating current is used to energize the primary winding, which results in a voltage being induced in each of the secondary windings. The windings are connected in series opposing, so the equal but opposite output of each winding tends to cancel (except for a small residual voltage called the null voltage). A pressure applied to a Bourdon tube or diaphragm causes the core to be displaced from its null position and the coupling between the primary and each of the secondaries is no longer equal. The resulting output varies linearly within the design range of the transducer and has a phase change of 180° from one side of the null position to the other. Since the core and coil structures are not in physical contact, essentially frictionless movement occurs.

Electronic devices necessary to signal condition the output of an LVDT consist of an oscillator for the supply voltage, circuitry to transform the constant voltage to a constant current, an amplifier with high input impedance for the output, a synchronous demodulator, and a filter with characteristics designed for quasistatic or dynamic measurements.

2.3.3 Aneroid Diaphragms

Another early design for sensing automotive manifold pressure consisted of dual sealed aneroid diaphragms. The diaphragms were bonded and sealed with a vacuum inside each unit to a metallized conductive ring on opposite sides of a ceramic substrate. The substrate served as the fixed plates of two separate capacitors. Manifold vacuum was applied to one chamber and the second served as a reference for compensating and signal conditioning that minimized common mode errors due to vibration and shock.

2.3.4 Capacitive Pressure Sensors

Capacitive pressure sensors have one plate that is connected to a force collector (usually a diaphragm), and the distance between the plates varies as a result of the applied pressure. The nominal capacitance is $C = Ae/d$, where A is the area of the plate, e is the permittivity, and d is the distance between the plates. Two common capacitive pressure sensors used in automotive applications are based on silicon and ceramic capacitors.

Silicon. A silicon capacitive absolute pressure (SCAP)-sensing element is shown in Fig. 2.7. The micromachined silicon diaphragm with controlled cavity depth is anodically bonded to a Pyrex® glass substrate. Feedthrough holes are drilled in the glass to provide a precise connection to the capacitor plates inside the unit. The glass substrate is metallized using thin-film deposition techniques. Photolithography is used to define the electrode configuration. After attaching the top silicon wafer to the glass substrate, the drilled holes are solder-sealed under vacuum to provide a capacitive-sensing element with an internal vacuum reference and solder bumps for direct mounting to a circuit board or ceramic substrate. The value of the capacitor changes linearly from approximately 32 to 39 pF with applied pressure from 17 to 105 kPa. The capacitive element is 6.7 mm × 6.7 mm and has a low-temperature coefficient of capacitance (–30 to 80 ppm/°C), good linearity (≈1.4 percent), fast response time (≈1 ms), and no

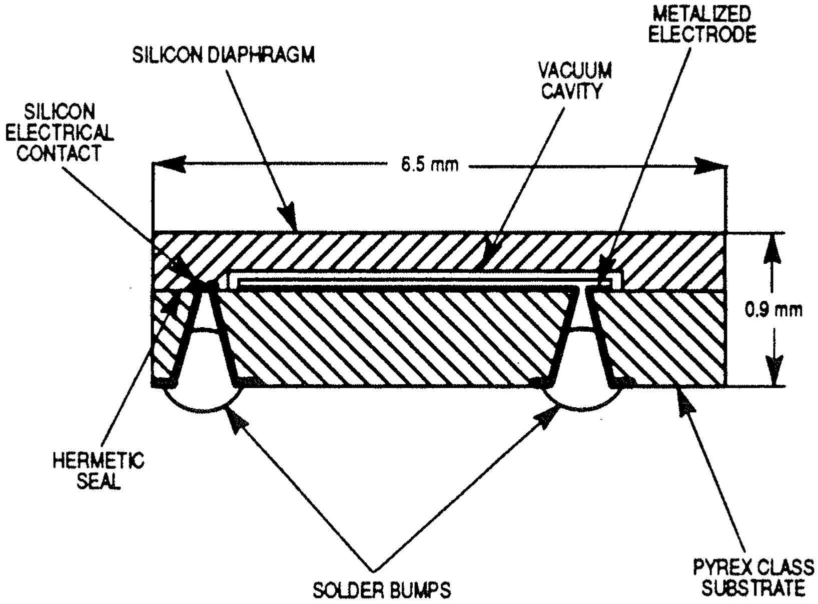


FIGURE 2.7 SCAP sensor.

exposed bond wires. The output of the sensor is typically signal-conditioned to provide a frequency variation with applied pressure for easy interface to microcontrollers.

Surface micromachining and bulk micromachined silicon-on-silicon techniques (see Sec. 2.3.5) have also been used to build silicon capacitive pressure sensors. These approaches also address the addition of signal-conditioning electronics on the same silicon structure.

Ceramic. The ability to make thin diaphragms from ceramic material combined with thin-film deposition to provide metal plates and connections has been used to manufacture ceramic pressure sensors. Their operation and signal-conditioning requirements are similar to the silicon capacitive sensor described in Sec. 2.3.4.

2.3.5 Piezoresistive Strain Gage

Strain-gage pressure sensors convert the change in resistance in four (sometimes in two) arms of a Wheatstone bridge. The change in the resistance of a material when it is mechanically stressed is called piezoresistivity.

The open-circuit voltage of an unbalanced Wheatstone bridge is given by $V_O = \mathbf{E}[(R1*R3 - R2*R4)/(R1 + R2)(R3 + R4)]$, where V_O is the output voltage, **E** is the applied voltage, and $R1$ through $R4$ are the resistive elements of the bridge. Additional variable resistive elements are typically added to adjust for zero-offset and sensitivity, and to provide temperature compensation.

Different approaches to piezoresistive strain gages range from traditional bonded and unbonded to the newest integrated silicon pressure sensors.

Bonded and Unbonded Strain Gages. The bonded resistance strain-gage pressure sensor consists of a filament-wire or foil, metallic or semiconductor, bulk material or deposited film bonded to the surface of a diaphragm. Pressure applied to the diaphragm produces a change in the dimension of the diaphragm and, consequently, in the length of the gage, and, therefore, a change in its resistance ($R = \rho\ L/A$). The change per unit length is called strain. The sensitivity of a strain gage is indicated by gage factor

$$GF = \Delta R/R \div \Delta L/L = 1 + 2\mu + \Delta\rho/\rho \div \Delta L/L \tag{2.1}$$

Foil strain gages have a negligible piezoresistive effect and their gage factor is usually between 2 and 3.

When a pressure sensor is used for measuring an applied force it is called a load cell.

Integrated Silicon Pressure Sensors. The GF for a strain gage is improved considerably (to about 150) by using a silicon strain gage. In addition to the conventional Wheatstone bridge, silicon processing techniques, and the relative size of piezoresistive elements in silicon enable the design of a unique piezoresistive sensor. The sensor signal can be provided from a single piezoresistive element located at a 45° angle in the center of the edge of a square diaphragm which provides an extremely linear measurement. The offset voltage and full scale span of the basic sensing element vary with temperature, but in a highly predictable manner. In addition to the basic sensing element, an interactively laser-trimmed four-stage network has also been integrated into a single monolithic structure (Fig. 2.8). The size of the silicon die, including the diaphragm, sensing element, and signal-conditioning electronics, is only 0.135 in by 0.145 in. The die is attached to the six-terminal package through the use of a stress-isolating layer of RTV (room temperature vulcanizing) silicone. This approach allows a minimum of external components for amplification to provide a usable output signal.

Two silicon wafers are used to produce the absolute piezoresistive silicon pressure sensor (Fig. 2.9). The top wafer is etched until a thin, square diaphragm, approximately one mil in thickness, is achieved. The square area is extremely reproducible as is the 54.7° angle of the cavity wall based on the characteristics of bulk micromachined silicon. The top wafer is

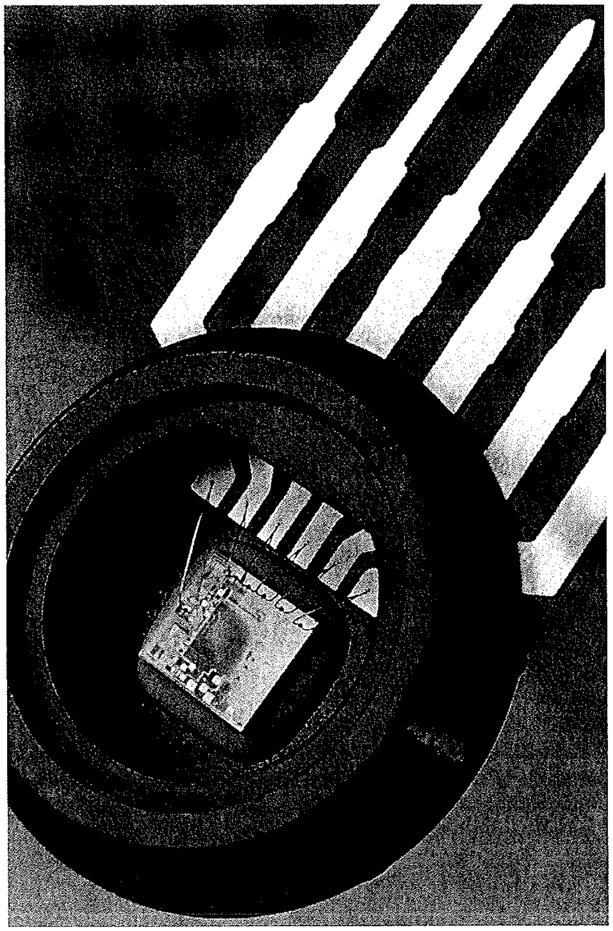


FIGURE 2.8 Silicon pressure sensor with integrated electronics. *(Courtesy Motorola, Inc.)*

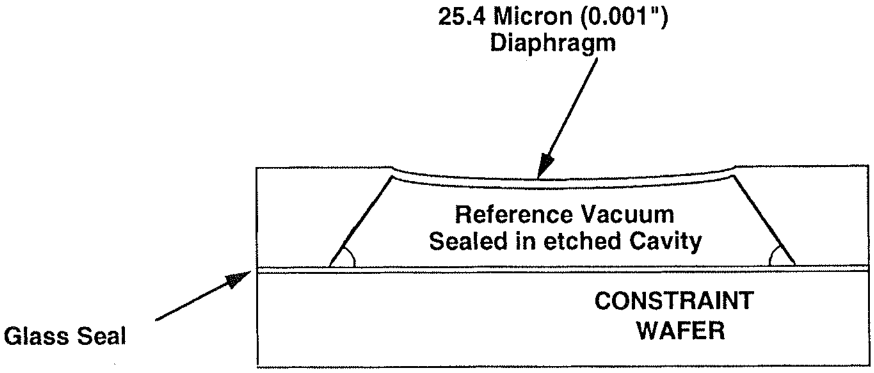


FIGURE 2.9 Cross section of piezoresistive silicon pressure sensor for measuring absolute pressure.

attached to a support wafer by a glass frit to provide a structure which is isolated from mounting stresses.

The bulk micromachining process used to form the diaphragm and the etched cavity in the majority of silicon pressure sensors is a chemical etching process that allows a thin (0.001-in) mechanical structure—the diaphragm—to be precisely etched from a silicon wafer that is approximately 0.015 in thick. Hundreds of sensors can be formed simultaneously on a 4-in (100-mm) diameter silicon wafer, and several wafers can be batch-processed to yield thousands of sensors in a single lot. Silicon pressure sensors can also be achieved by using surface micromachining techniques. In these sensors, a layer of sacrificial material (usually an oxide) is grown on top of a silicon wafer, and material such as polysilicon is then deposited on the sacrificial layer and patterned to achieve a particular structure. The sacrificial material is removed by a chemical etchant. Both bulk and surface micromachining techniques can be combined with semiconductor processing techniques to provide additional circuitry on the same monolithic structure. A number of new terms are used relative to silicon pressure sensors that are defined in the glossary to this section.

Both bulk and surface micromachining, discussed previously, are performed at the wafer level. A polysilicon thin-film sensor that consists of a thin film of silicon that is doped with boron and vapor-deposited over a stainless steel diaphragm is shown in Fig. 2.10. A thin deposited layer of silicon dioxide insulates the silicon from the stainless steel diaphragm. Silicon nitride is used to cover the strain-sensitive elements. Silicon-on-insulated-stainless-steel (SOISS) sensors are not formed using silicon wafer techniques, but they use batch-processing techniques and are inherently suited for harsh environments.

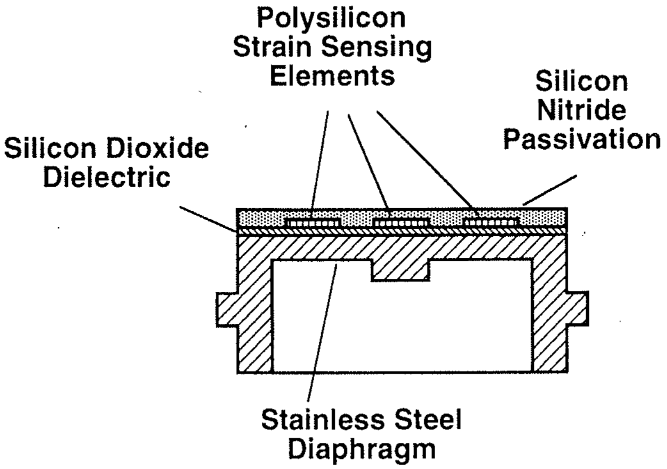

Polysilicon Strain Sensing Elements
Silicon Nitride Passivation
Silicon Dioxide Dielectric
Stainless Steel Diaphragm

FIGURE 2.10 Polysilicon pressure sensor on stainless steel diaphragm.

2.3.6 Piezoelectric Pressure Sensors

A piezoelectric sensor produces a change in electric charge when a force is applied across the face of a crystal or piezoelectric film. The inherent ability to sense vibration and the necessity for high-impedance circuitry are taken into account in the design of modern piezocrystal sensors. Transducers are constructed with rigid multiple plates and a cultured-quartz sensing element, which contains an integral accelerometer to minimize vibration sensitivity and suppress resonances. A typical unit also contains a built-in microelectronic amplifier to transform the high-impedance electrostatic charge output from the crystals into a low-impedance voltage signal. Units made in stainless steel housings have an invar diaphragm laser welded to seal the

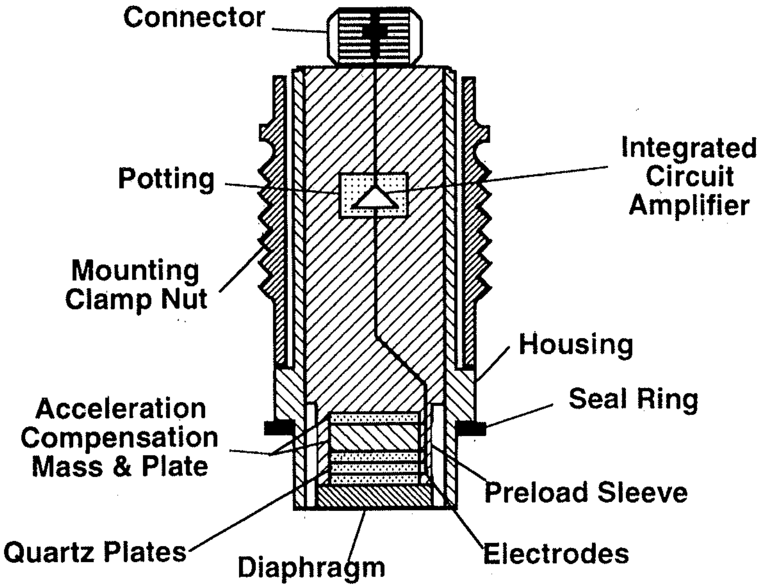


FIGURE 2.11 Piezoelectric pressure sensor.

sensing elements inside the package. Figure 2.11 shows the cross section of a piezoelectric pressure sensor.

More recently, piezo film sensors, which produce an output voltage when they are deflected, provide a very inexpensive method for pressure measurements. One approach has the piezo film cemented to a metallic dimple substrate with the dimple pointed toward the high-pressure source. As the pressure rises, a point is reached when the dimple snaps in the opposite direction and the movement is sufficient for the piezo film to generate a transient voltage.

Surface micromachining techniques have also been combined with piezoelectric thin-film materials, such as zinc oxide, to produce a semiconductor piezoelectric pressure sensor.

2.3.7 Fiber Optic Combustion Pressure Sensor

For extremely high pressure, or pressure measurements at high temperatures, different pressure measurement techniques are used. Figure 2.12 shows an alternative to traditional pressure-sensing techniques that is being developed to sense production vehicles' combustion

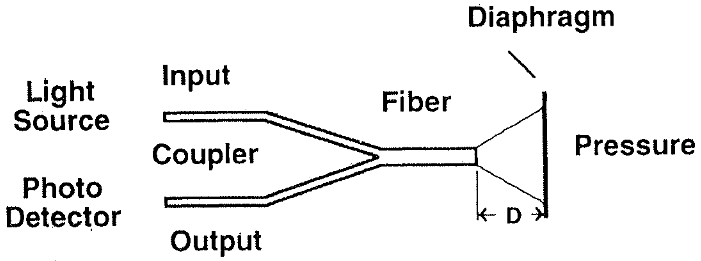


FIGURE 2.12 Fiber optic combustion pressure sensor.

pressure. The fiber optic pressure sensor can withstand temperatures (up to 550 °C) that are well above the normal range for piezoelectric sensors. Furthermore, the design has a normal pressure range of 0 to 1000 psi and overrange capability of 3000 psi.

The sensor's operation is the result of a light source input to an optical fiber coupler and a photodetector at the receiving end of the coupler. Light exits the optical fiber as a diverging cone which illuminates a diaphragm. The maximum angle is determined by an aperture. The amount of light that is returned to the sensor fiber after it is reflected from the diaphragm depends upon the gap D between the fiber and the diaphragm. The diaphragm can be sized to allow the sensor to be integrated into a spark plug for easy access to the combustion pressure of each cylinder. Accuracy within 5 percent has been demonstrated within the 550 °C operating temperature range.

2.3.8 Pressure Switch

A pressure switch can be simpler and more cost effective than a pressure sensor for an application that only requires detecting a single pressure level. Sufficient motion of the force-collecting diaphragm (e.g., elastomer, Kapton®, fluorosilicon) allows a spring to be compressed and a set of contacts to be closed in a traditional mechanical switch. The design of the contacts may allow several amperes to be switched.

Converting the output of any of the electric sensors in Sec. 2.3 to a threshold-sensing level requires additional circuits, including an electronic switch, such as a power FET, to conduct the current. The sensor can have multiple switch points depending on the amount of additional circuitry that is provided.

2.3.9 Pressure Valves/Regulators

Pressure is frequently controlled in automotive applications by pressure regulators or valves. The actuation of these mechanisms can be a result of applied pressure to a mechanical structure such as an integral piston or relief valve held closed by a spring force, or an electric signal generated from a sensor and subsequent activation of a solenoid, which opens a valve or moves a pintle in an orifice to change the pressure. The thermostat in the engine cooling system allows flow based on a minimum temperature being achieved. It must operate independently of the pressure variations in the cooling system. A common solution is an expansion-element thermostat which actuates a valve to redirect the flow of coolant into a radiator bypass line when the control temperature is reached.

Extremely small and precise silicon-based regulators, and even pumps, are possible using micromachining techniques. Figure 2.13 shows a silicon Fluistor™ (or fluidic transistor) microvalve that is approximately 5.5 mm by 6.5 mm by 2 mm. The valve is actually a thermopneumatic actuator which accepts an electric input. The cavity is etched in the middle silicon chip by bulk micromachining described in Sec. 2.3.4 and filled with a control liquid. When the liquid is heated, the silicon diaphragm moves outward over the valve seat. This approach has demonstrated a dynamic range of 100,000:1 controlling gas flows from 4 µlpm to 4 lpm at a pressure of 20 psid.

The microvalve combined with a pressure sensor and electronic feedback loop provides a solid state pressure regulator. It has potential for usage in both gas and liquid flow control applications on future vehicles. The integration of EGR and idle-air control (IAC) has already been accomplished in a somewhat conventional (patented) manner. A feedback-controlled valve with two inlets and a single outlet orifice eliminate stepper motor programming in the engine controller and only one calibration curve is needed for both EGR and IAC functions. The combination of this approach to control systems and newly developed technologies, such as the silicon microvalve, will allow additional advances in vehicle performance, efficiency, and control.

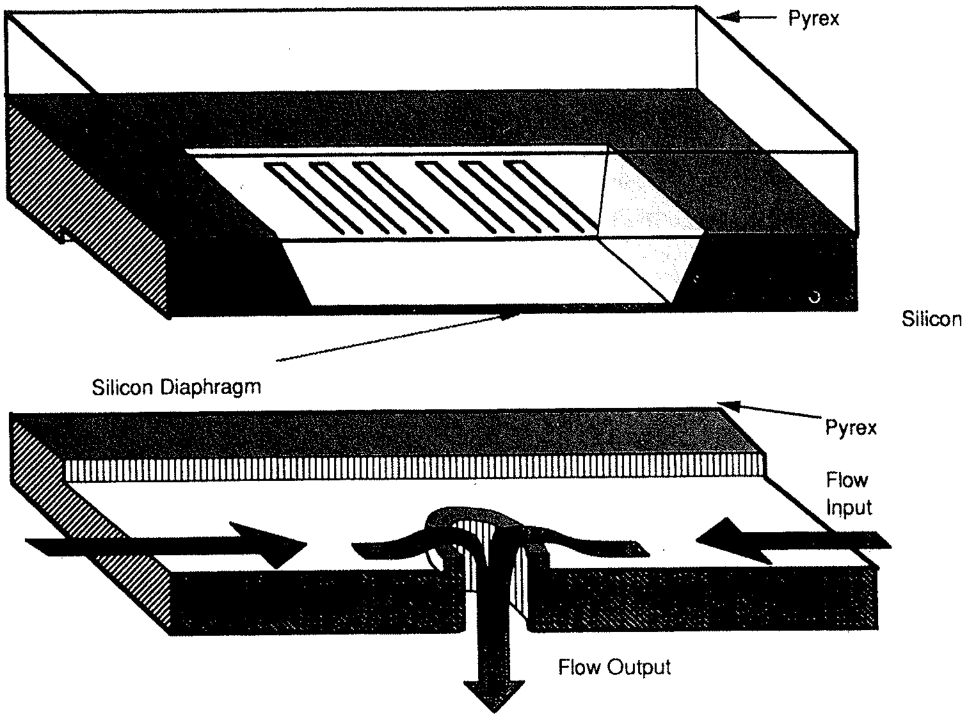


FIGURE 2.13 Silicon microvalve. *(Courtesy of Redwood MicroSystems, Inc.)*

2.4 FUTURE PRESSURE-SENSING DEVELOPMENTS

The different types of pressure measurements, different technologies for measuring pressure, and potential pressure measurements in automotive applications have been explained in this chapter. In addition, alternatives to making pressure measurements, such as indirect sensing and the use of pressure regulators, have been discussed. The use of semiconductor technology applied to sensing applications is producing sensors with inherently more decision and diagnostic capability that can communicate bidirectionally with host microcomputers in complex systems. The desire to directly produce a signal that is compatible with microcomputers rather than requiring analog signals to be converted to digital format through A/D converters is spurring development activity that could affect future automobile systems. Furthermore, recently developed fuzzy logic and neural network approaches to control systems and multiplexing of sensor outputs for use in several systems will make previously cost-prohibitive sensing applications a reality.

Increased and recently initiated sensing activity from industrial organizations such as SAE, the American National Standards Institute (ANSI), and the Institute of Electrical and Electronics Engineers (IEEE) should provide a greater level of understanding, common terminology, and improved specifications and test procedures for the numerous approaches that can be taken to sense pressure in automotive applications.

All trademarks are the property of their respective owners.

GLOSSARY

Altimetric pressure transducer A barometric pressure transducer used to determine altitude from the pressure-altitude profile.

Diaphragm The membrane of material that remains after etching a cavity into the silicon sensing chip. Changes in input pressure cause the diaphragm to deflect.

Error band The band of maximum deviations of the output values from a specified reference line or curve due to those causes attributable to the sensor. Usually expressed as "±% of full-scale output." The error band should be specified as applicable over at least two calibration cycles so as to include repeatability and verified accordingly.

Linearity error The maximum deviation of the output from a straight-line relationship with pressure over the operating pressure range. The type of straight-line relationship (end-point, least-square approximation, etc.) should be specified.

Operating pressure range The range of pressures between minimum and maximum pressures at which the output will meet the specified operating characteristics.

Overpressure The maximum specified pressure that may be applied to the sensing element of a sensor without causing a permanent change in the output characteristics.

Piezoresistance A resistive element that changes resistance relative to the applied stress it experiences (e.g., strain gauge).

Pressure error The maximum difference between the true pressure and the pressure inferred from the output for any pressure in the operating pressure range.

Pressure sensor A device that converts an input pressure into an electric output.

Ratiometric (ratiometricity error) At a given supply voltage, sensor output is a proportion of that supply voltage. Ratiometricity error is the change in this proportion resulting from any change to the supply voltage. Usually expressed as a percent of full-scale output.

Response time The time required for the incremental change in the output to go from 10 to 90 percent of its final value when subjected to a specified step change in pressure.

Temperature error The maximum change in output at any pressure in the operating pressure range when the temperature is changed over a specified temperature range.

BIBLIOGRAPHY

"Acustar Electronic Oil Pressure Sensor," *Automotive Industries,* March 1993, pp. 26–29.

Budd, John W., "A Look at Pressure Transducers," *Sensors,* July 1990, pp. 10–15.

Doeblin, Ernest O., *Measurement Systems Application and Design,* McGraw-Hill, New York, 1975.

He, Gang, and Marek T. Wlodarczyk, "Spark Plug-Integrated Fiber Optic Combustion Pressure Sensor," *Proceedings of Sensors Expo,* Chicago, Sept. 29–Oct. 1, 1992, pp. 211–216.

Holt, Daniel J., "ABS Testing," *Automotive Engineering,* March 1993, pp. 26–29.

Keebler, Jack, "Automakers, gas refiners debate vapor control units," *Automotive News,* Oct. 14, 1991, p. 39.

Lynch, Terrence, "Integrated Valve Meters EGR and Idle Air," *Design News,* Feb. 22, 1993, pp. 159–160.

Motorola Pressure Sensor Device Data Book, Q1/93, Phoenix, Ariz.

Norton, Harry N., "Transducers and Sensors," *Electronic Handbook,* McGraw-Hill, New York.

PCB Electronics, Inc., comments in "Piezoelectric Pressure Transducers," *Measurements & Control,* Oct. 1990, pp. 20–222.

Sawyer, Christopher A., "On-Board Diagnostics," *Automotive Industries,* May 1992, p. 57.

Siuru, Jr., William D., "Sensing Tire Pressure on the Move," *Sensors,* July 1990, pp. 16–19.

Tran, Van Truan, "Wind Forces and Moments," *Automotive Engineering,* April 1990, pp. 35–38.

ABOUT THE AUTHOR

Randy Frank is a Technical Marketing Manager for Motorola's Semiconductor Products Sector in Phoenix, Arizona. He has a BSEE, MSEE, and MBA from Wayne State University in Detroit, Michigan, and over 25 years' experience in automotive and control systems engineering. For the past 10 years he has been involved with semiconductor sensors, power transistors, and smart power ICs.

CHAPTER 3

LINEAR AND ANGLE POSITION SENSORS

Paul Nickson
Product Line Manager
Analog Devices, Inc.,
Transport & Industrial Products Division

3.1 INTRODUCTION

Position sensors of one form or another are an integral and necessary part of the modern automobile. Position sensors range in technology from the ubiquitous microswitch warning to the driver of a door ajar to linear variable differential transformers (LVDTs) used in sophisticated active suspension systems. Whether as monitors or as critical parts of safety systems, market and legislative pressures for longer warranties and lower emissions are opening avenues for a wide range of sensing technologies to find a place in the modern automobile.

The automotive systems designer must take into account many factors when choosing the appropriate technology for an application. Each sensor type has its own vocabulary, and it is important when making comparisons to understand how a figure of merit for one sensor relates to that of another. It is equally important to understand how the choice of output signal format, whether digital or analog, can affect the resolution of measurement and subsequently the performance or stability of an automotive system.

The purpose of this chapter is to give an overview of position sensor technologies currently used and available for use in automobiles and to compare their characteristics and suitability for particular applications. Consideration is given to the interfacing requirements of each type of sensor with an emphasis on the advantages and disadvantages of each method as they apply in the automotive environment. Where appropriate, descriptions of applications of the various sensor types in automobile applications are given. Other available technologies and technologies in development which have desirable characteristics for automotive applications are also discussed.

3.2 CLASSIFICATION OF SENSORS

Sensors may be classified in many different ways. From the perspective of a system designer, the basic questions are: What kind of information does the sensor provide and how is the sensor used? For the purposes of this discussion, a position sensor is defined as an electromechanical device that translates position information into electric signals. Sensors can be grouped into two basic categories.

3.2.1 Incremental or Absolute

Position information can be presented in two different ways. Incremental position sensors measure position as the distance from an arbitrary index or zero. Alternatively, position information may be provided that gives an unambiguous or absolute measure of the distance from a well-defined index.

Incremental sensors usually rely on some method of pulse counting. One pulse in the sequence is designed to be wider or of opposite polarity than the others so that it may be used as a nominal zero. A typical optical angle encoder consists of a glass disk marked with a number of equally spaced radial opaque lines and transparent gaps. The disk is illuminated on one side and a light sensor and associated electronics on the other side detect the passage of dark lines and gaps and generate corresponding electric pulses. Dedicated electronics built into the sensor or a remote microcontroller can be used to count the pulses. A zero is established by detecting the wider pulse, known as *North Marker* in optical encoder terminology, and then resetting the pulse counter. The advantage of this data format is that few wires are required to carry the information. Typically, four or five wires would be required depending on the exact details of the format (see Sec. 3.3.2). The biggest disadvantage of incremental sensors is that at power-up the system has no position information and requires a mechanical indexing cycle to find the marker pulse. Another disadvantage is that the system is prone to the effects of noise, which may lead to erroneous counts.

In contrast, absolute position sensors produce an unambiguous output at power-up. Each position or angle has a unique value. The output may be a voltage or frequency or other analog of the input position. Potentiometers are often used in applications requiring this characteristic. Many absolute position sensors have binary digital outputs. The digital formats vary depending on the construction of the sensor. Some optical encoders use Gray code to avoid ambiguities at code transitions. Other sensors, such as resolvers, do not directly produce a digital output but are almost always used with an analog-to-digital converter that may output in parallel or serial form one of the common formats—for example, two's complement or offset binary.

3.2.2 Contact or Proximity

Position sensors are designed to detect the position of components of mechanical systems by either being directly coupled by some shaft or linkage, as in the case of potentiometers or optical encoders, or by some noncontact or proximity means. Environmental issues are often a key influence in the choice of sensor for a given application. High levels of vibration, particularly in small engine applications, may cause rapid wear of the conductive track of a throttle-position-measuring potentiometer. Dirt and dust usually exclude optical sensors from underhood applications due to rapid degradation of the optical path.

The most common form of proximity sensors are based on various methods of magnetic field detection. A device based on magnetic field sensing principles may be more easily isolated from the destructive effects of the harsh environment encountered in many automotive applications.

3.3 POSITION SENSOR TECHNOLOGIES

3.3.1 Microswitches

The simplest form of contact sensor is a switch. Contact position sensors may be as simple as the microswitches that operate anything from brake lights to courtesy lights in the automobile. Many applications of microswitches in position sensing are as limit switches, usually

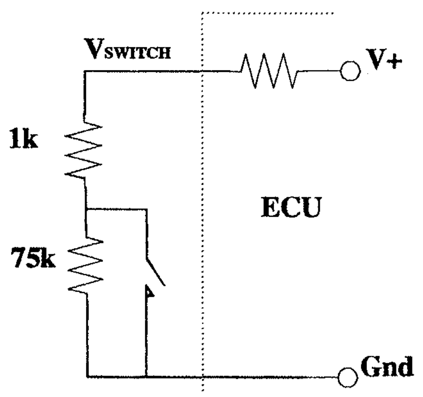


FIGURE 3.1 Diagnosable switch.

wired to limit or warn of the extent of travel of a mechanical component by disconnecting power to an electric motor or by operating an indicator lamp. In some cases, for safety reasons, it is desirable to be able to detect fault conditions that would make the switch inoperable. In some applications, it is possible to connect the switch as shown in Fig. 3.1. In this case, a diagnostic circuit measuring the voltage, V_{SWITCH}, can differentiate between the normal conditions of switch open or closed and can also determine if the switch is disconnected or if V_{SWITCH} is shorted to either power supply.

An undesirable characteristic of switches is that the contacts may bounce on closing. If it is important in the application that the first switching edge is detected, then a simple switch-debouncing logic that rejects noisy signals can be used. If a microcontroller is used to monitor the switch output, then debouncing can be accomplished by software means. This may be a better solution in applications subject to shock or heavy vibration, which may cause occasional false switching. In these cases, a microcontroller can be configured to poll the switch over some period of time and report switch closure only if a number of consecutive readings are the same.

3.3.2 Optical

Optical angle encoders for incremental shaft angular position measurement are constructed of a disk with a series of transparent and opaque equally spaced sectors. The disk can be made of glass for precision applications. Mylar film and metal disks offer high and medium resolutions, respectively, at low cost. The encoder disk is illuminated on one side and light sensors on the other side detect the passage of light and dark sectors as the disk is rotated. (Low-resolution versions such as the Hewlett-Packard HRPG series use an alternative reflective technology.) Most encoders have two sets of light sources and detectors offset by half the width of a sector. Figure 3.2 shows the relationship between the outputs of the light sensors as the encoder is rotated. This format is often referred to as "A quad B," since the signals are in quadrature. The passage of one pair of light and dark sectors over a detector is referred to variously as one cycle, one count, one line or 360 electrical degrees (°e). Encoder resolutions range from around 16 counts per revolution (CPR) for low-cost applications to over 6000 CPR for precision position control systems. Most encoders also include a third signal for use as an index or reference pulse. The index, or North marker as it is sometimes called, occurs once per revolution. The pulse width is usually equal to 90 °e.[1]

Four separate states, each of 90 °e, can be derived from the A and B outputs using integrated circuits available from a number of vendors. This allows a resolution of four times the number of lines on the encoder disk to be achieved. These ICs also determine the direction of rotation of the encoder by observing which output leads the other. By convention for clockwise rotation, the low-to-high transition of A leads the low-to-high transition of B. Control circuitry can be added to improve noise immunity by only allowing valid next states to be counted.

Incremental angle encoder accuracy specifications fall into two categories. The angular position accuracy is the difference between the actual shaft angle and the angle indicated by the encoder. This error is normally expressed in degrees or minutes of arc. The second category includes specifications for the symmetry and repeatability of individual cycles; these are usually

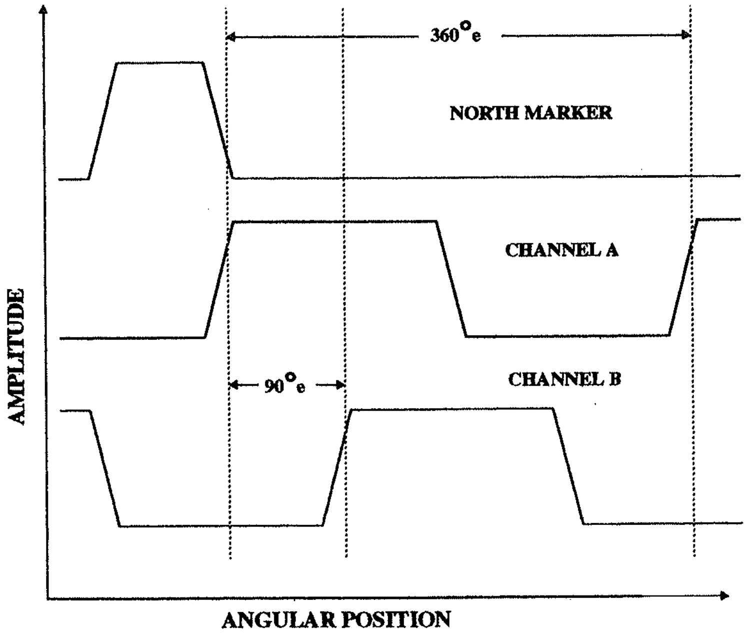


FIGURE 3.2 Encoder outputs.

expressed in electrical degrees. Typical characteristics are detailed in Fig. 3.3 and Table 3.1 Errors are caused by eccentricity and axial play of the code wheel and manufacturing defects in the lithography or etching of the code wheel. Modular encoders consisting of a light source and sensor head are available which use a collimated light source and an array of integrated photodiodes to minimize the effects of these error sources. Differential connection of the photodiodes ensures insensitivity to errors caused by light source variability due to environmental or other factors. A further source of error can occur if the encoder is rotated at high enough speeds that the rise and fall time of the digital outputs significantly affects the pulse width. The light sensor bandwidth usually determines the maximum rotational speed of the sensor. Typical bandwidths are below 100 kHz, which would limit the speed of a 100-CPR encoder to 1000 rpm.

Linear incremental optical encoders are available from many vendors. These allow direct measurement of linear motion. Modular sensor/emitter heads are available that can be used in these applications. The technology is basically the same as incremental angle encoders and the terminology used to describe specifications is the same as for angle encoders. Linear encoders are described in terms of their count density or resolution in counts per mm or mm

TABLE 3.1 Incremental Encoder Specifications

	Minimum	Typical	Maximum	Units
Position error		10	40	min of arc
Cycle error		3	10	°e
Pulse-width error		7	30	°e
Phase error		2	15	°e
State-width error		5	30	°e
Index pulse width	60	90	120	°e

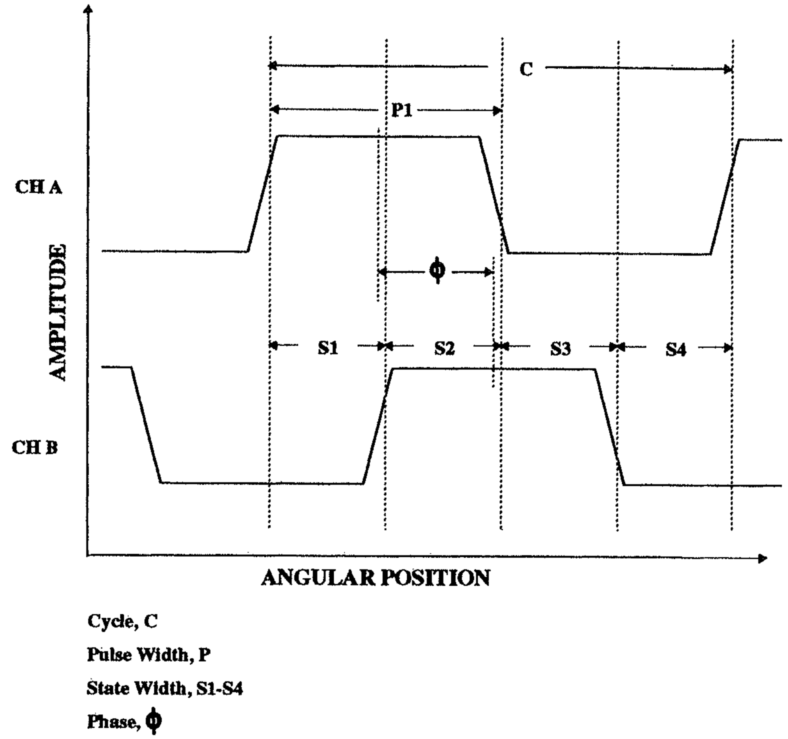


FIGURE 3.3 Encoder definitions.

per count. Line counts range up to approximately 8 per mm, allowing an ultimate linear resolution of around 30 μm.

If it is important to have an unambiguous measure of position as soon as power is applied to a system, then an absolute encoder must be used. Absolute optical encoders are manufactured with resolutions from one part in 2^6 to one part in 2^{16}. The data format can be binary, binary-coded decimal (BCD), or Gray code. An absolute angle encoder is divided into equal sectors which are arranged so that adjacent sectors contain consecutive digital words. The binary bits of each word form *N* concentric tracks on the encoder disk, where *N* is the digital word length. *N* sets of light sources and photodiodes detect the parallel word representing the input shaft angle.

Absolute optical encoders often use Gray code to eliminate code transition errors. In a natural binary sequence between zero and full scale on the disk, all the bits of the digital word change state together. Any misalignment of the code wheel and the light sensors, or sensor-to-sensor misalignment, will cause false codes to be generated. This could be disastrous for a position control system since a misread code could indicate an angle up to 180° away from the correct angle. Gray coding eliminates this problem. Gray code is a unit distance code; consecutive codes differ by only one binary bit. If a code transition is misread, the largest error will be one least significant bit of the digital word.

3.3.3 Potentiometric

Potentiometers are widely used as position sensors in automotive applications such as throttle and accelerator pedal position measurement. The automotive industry increasingly demands low-cost mechanically and electrically rugged sensors to provide control or measurement of position in the modern automobile. This has resulted in the development of potentiometers that are capable of operational life far in excess of the life of the average car, and in some cases capable of continuous rotational speeds of above 1000 rpm for more than 1000 hours.[2]

Potentiometers can be constructed using a wire-wound track. The resolution of these potentiometers is determined by the number of turns of wire used to wind the track. The resolution of rotary wire-wound potentiometers is often quoted as the number of turns per degree and can be anywhere between 1 (1° per turn) and 7 (8.5 arcmin per turn). The track resistance is proportional to the number of turns used and can be in the range of 10 ohms to 100 kilohms, with a tolerance of approximately 5 percent. Wire-wound potentiometers have advantages where low-value variable resistors are required but do not excel in linearity, resolution, or rotational life which can be as low as 10^5 revolutions. Potentiometers for position-sensing applications are constructed using a resistive track of conductive material, usually a graphite and carbon black doped plastic, and a collector track molded on some supporting substrate. A drive shaft or pushrod draws precious metal multifingered wipers along the tracks. Wiper damping is usually included to make the potentiometer insensitive to vibration. Potentiometers of this type are manufactured with a range of resistance from around 500 ohms to 20 kilohms with a tolerance of 10 to 20 percent (Table 3.2). Potentiometers of this type are capable of excellent linearity and very high resolution.

TABLE 3.2 Potentiometer Specifications

Parameter	Minimum	Maximum
Electrical travel	90°, 10 mm	360, 3000
Nominal resistance	500 ohms	20 kilohms
Resistance tolerance	10%	20%
Resistance temperature coefficient (TC)		500 ppm/°C
TC of V_{out} in voltage divider mode		5 ppm/°C
Linearity error	0.01%	1%

Potentiometric sensors are used as voltage dividers. A reference voltage is applied across the resistive element and the wiper voltage is used as an absolute measure of the position of the actuator. Linear potentiometers are available in a wide range of lengths from 10 mm to 300 cm. Rotary potentiometers are usually restricted to 355° of useful range due to the dead band created by the track-end contacts. Some versions are available with true 360° operation. These use multiple wipers and dedicated electronics to eliminate the dead band.

All potentiometers are ratiometric sensors. That is, the wiper voltage at a given position is some fraction of the reference voltage applied across the resistive track. If the reference voltage is varied, the potentiometer output remains in the same ratio to the reference voltage. Sensor potentiometers, when properly terminated, maintain ratiometric operation over a wide range of temperatures with temperature coefficients typically less than 5 parts per million (ppm) per degree centigrade. Without special signal processing, ratiometricity is compromised at the end of the electrical travel of a potentiometer by the change in resistivity of the track as it joins the end contact and by any parasitic external resistance in series with the track. Figure 3.4 shows the effective limitations on the potentiometer at its endpoints.

Ratiometricity is a very desirable characteristic for a sensor that is used with comparators or analog-to-digital converters. For example, if the same reference voltage that powers the sensor is used as the reference for an analog-to-digital converter, then the measurement system will be insensitive to the absolute value of the reference voltage. A given shaft angle will always be reported as the same digital code. In most automotive control systems, analog-to-digital converters (usually on board a microcontroller) use the regulated 5-V engine controller

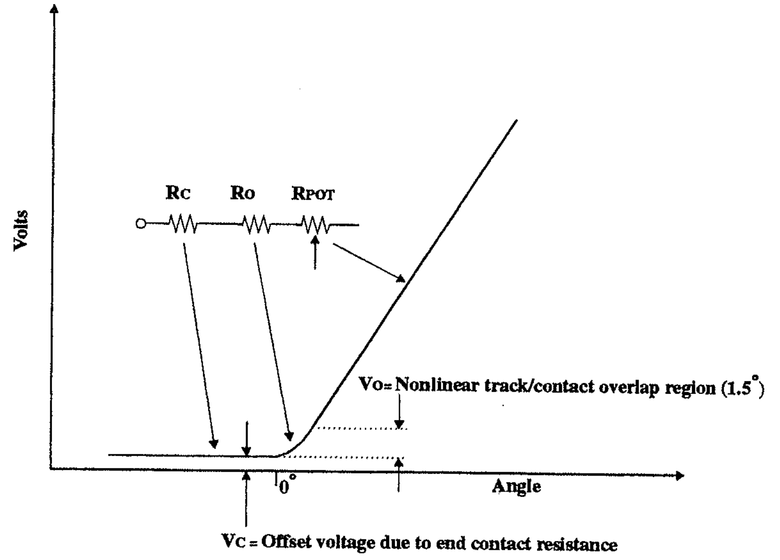

Volts
Rc
Ro
Rpot
Vo= Nonlinear track/contact overlap region (1.5°)
0°
Angle
Vc = Offset voltage due to end contact resistance

FIGURE 3.4 Potentiometer endpoint nonlinearity.

(ECU) power supply as reference to avoid the additional cost of a separate voltage reference chip. It is desirable that the same voltage, perhaps buffered to isolate the control module from accidental shorts, is used as reference by any ratiometric sensors in the automobile.

Potentiometers are subject to a number of sources of error, of which linearity is the most important. The linearity error is the difference between the actual transfer function of the potentiometer and the ideal transfer function (output voltage change versus mechanical travel) as a percentage of the applied reference voltage. Linearity specifications between 0.03 and 1 percent are available, with sensor cost inversely proportional to linearity. Microlinearity can be an important indicator of the suitability of the potentiometer for use in accurate control systems where large changes in the local gradient of the transfer function may cause instability. Ratiometricity, linearity, and offset errors can be caused by improperly loading the wiper of the potentiometer. The maximum error is at the center of travel of the wiper. If the load is a constant current, such as the input current of a buffer amplifier, a voltage will be developed across R_{EXT} that will cause an offset error at the extremes of travel of the wiper and an additional linearity error at the center of travel. Sensor potentiometers are usually specified with some maximum wiper current (100 nA, typically) to eliminate these errors. Plastic track potentiometers are capable of resolutions better than 0.001 percent of travel. This is primarily limited by the homogeneity of the resistive track material and hysteresis caused by limitations in the mechanical construction of the potentiometer related to bearings, wiper stiffness, and coefficient of friction of the track.

Plastic track sensor potentiometers are capable of operational life in excess of 10^7 revolutions for a rotary sensor or 10^7 strokes for a linear sensor. Unfortunately, no universal standards are available defining test conditions, and these may vary from vendor to vendor. Generally, two types of test are carried out. Dither testing simulates conditions which may exist in control applications or areas with high levels of vibration. The wiper is moved over a small proportion of the travel, say 1 or 2°, at a test frequency of 100 Hz. Information about contact resistance and local gradient changes in the potentiometer transfer function can be gathered rapidly using this technique. A change in gradient (ohms/percent of travel) relative to the mean gradient of the potentiometer can be equated to a change in loop gain in a control system. Sensitive systems may become unstable at worn points in the potentiometer track due to relative gradient errors.

A second method of potentiometer reliability testing is to repeatedly move the wiper at around 10 Hz over a large proportion of the available range. An excursion from 0 to 50 percent of the available travel will result in the maximum change in linearity error over a large number of cycles, since only one-half of the track is subject to wear. The criteria for failure of a potentiometer will be very application-specific and may, for example, be a doubling in linearity error. It is important to work with the potentiometer vendors to understand how their specifications can be extrapolated to make a prediction of operational life.

3.3.4 Magnetic

By far the largest category of position sensors relies on electromagnetic induction principles. This group of sensors can be broken into subgroups depending on the details of the employment of the phenomenon. Other sensors in this category rely on materials with magnetoresistive or magnetostrictive properties. Electromagnetic sensors have a number of advantages over other technologies. In general, this class of sensors measures or responds to changes in the relative position of components in a magnetic circuit. The components are always separated by an air gap and are not subject to friction wear. In many cases, it is possible to construct rugged sensors that are insensitive to the harshest automotive environments.

Variable Reluctance. The reluctance of a magnetic circuit determines the magnetomotive force (amp turns) required to produce a flux of a given value.[3] Variable reluctance devices operate by sensing changes in the reluctance within a magnetic circuit. In most cases, the reluctance change is caused by a change in the length of an air gap. The change in reluctance causes a change in the magnetic flux which induces a voltage in an output signal coil. The voltage induced is typically a bipolar pulse shape whose amplitude is proportional to the rate of change of flux (Faraday's law).

$$V = \left(\frac{d\Phi}{dt}\right) \tag{3.1}$$

This sensor technology cannot be used at zero speed since, if the rate of change of flux is zero, then the output will be zero.

In automotive applications, variable reluctance sensors are used to detect the position and speed of rotating toothed or slotted wheels in crank-, cam-, and wheel-monitoring applications. An easily magnetized or "soft" magnetic core or bobbin wound with a sense coil is magnetized by a strong permanent magnet such as samarium cobalt (Sm_2Co_{17}). The sense end of the core is placed in close proximity to a toothed gear wheel. The flux change that occurs when a tooth edge passes the sensor causes a voltage to be induced in the coil. Remote signal-conditioning electronics associated with the ECU are used to amplify the signal and produce a signal that a microcontroller can interpret as a position increment. An alternative construction[4] uses a coaxial pole piece to improve the magnetic circuit. This construction is particularly suited to sensing holes or apertures in a sense wheel. Sensors that detect slots and are positioned in close proximity to the target with a small air gap work at low reluctance and are less likely to be disturbed by interfering fields than sensors configured to detect teeth at a low mark space ratio.

Variable reluctance sensors are prone to a number of sources of error. Vibrations or resonance sometimes exacerbated by the attractive forces between the sensor and the target can seriously degrade the signal-to-noise ratio of the device. The sensors' target is usually a rotating ferromagnetic wheel or gear. Eddy currents will be generated by the movement of the wheel in the magnetic field of the sensor. This may lead to false readings from the sensor. In some refinements of variable reluctance sensors, the holes or apertures in the wheel are filled with an electrically conductive nonmagnetic material to homogenize eddy currents.[4]

Significant advantages of variable reluctance sensors in the automotive environment are their simple rugged construction and low cost. Additionally, they have a wide operating temperature range and require only two wires for operation. Variable reluctance sensors can also be used as variable inductance sensors by exciting the sense coil with alternating current and employing inductance-measuring means in the signal-conditioning electronics.

Hall Effect. Electric current is carried through the motion of electric charge. If a conductor is moved through a magnetic field with velocity v, the charges in the conductor will experience a force (Lorentz force) in a direction perpendicular to both the direction of motion and the magnetic field. This gives rise to an electric field of strength:

$$\mathbf{E} = v\mathbf{B} \tag{3.2}$$

The charges will move and a surface charge will develop on the conductor until an electrostatic field forms which counterbalances the electric field due to motion, $v\mathbf{B}$. A voltage due to the movement of charge can be detected with a voltmeter. The voltage is proportional to the field **B** and the velocity and length of the conductor. The result of this effect in thin films of material was first described by Hall over 100 years ago. When he passed current through a rectangle of gold foil in the presence of a magnetic field perpendicular to the plane of the foil, a voltage could be measured across the other axis of the foil.

Devices can be constructed using semiconductor materials which can utilize this effect to detect the strength of magnetic fields. Figure 3.5 illustrates the construction and operation of a silicon Hall effect device. A voltage is amplified across one axis of a thin block of high-resistivity n-type epitaxial material. In the presence of the field **B**, charges move in the direction of the arrow. A voltage directly proportional to **B**, the current density in the silicon and the Hall coefficient (scattering factor) can be measured at the point shown. The sensitivity is low and amplification is required to render a useful signal. For example, with a current of 10 mA flowing in an n-type silicon epitaxial layer 1 μm thick with a doping level of $10^{15}/\text{cm}^3$ and a field of 100 mT, a voltage difference of approximately 30 mV will be measured at 25 °C. Offsets caused by resistivity gradients, piezoelectric effects from packaging stress, and contact misalignment can amount to 10 mT or more. Layout techniques, such as cross-coupled structures to minimize the effects of resistivity gradients, can significantly improve offsets. Careful alignment of the Hall cell layout with crystal axes can mitigate piezoelectric effects. Silicon Hall effect devices are insensitive to magnetoresistive effects as the field strengths encountered in most applications have good linearity with errors of <0.1 percent for fields from 0 to >100 mT.[5]

The Hall voltage is strongly temperature dependent and, with constant current bias, is proportional to the magnetic field, the bias current, the carrier concentration, the scattering factor, and a constant G, which is a function of the geometry of the device.

$$VH = GI\mathbf{B}\left(\frac{r}{\text{qnt}}\right) \tag{3.3}$$

A typical silicon Hall device will exhibit a temperature coefficient of the Hall voltage of approximately 1000 ppm/°C under these bias conditions. The temperature dependence can be reduced with a current source, the temperature coefficient of which is designed to compensate for the Hall voltage TC. In this way, the sensitivity of a Hall effect device can be controlled within 1 or 2 percent over the range of temperatures normally encountered in automotive wheel position and speed applications. The temperature coefficient of the compensated device can be matched to the magnetic circuit if necessary; a typical requirement might be to provide a residual TC of 200 ppm/°C to compensate for the TC of a $\text{Sm}_2\text{Co}_{17}$ magnet.

Hall effect devices can be constructed from semiconductor materials other than silicon—for example, gallium arsenide (GaAs). GaAs offers higher carrier mobility and some promise for higher temperature operation than silicon. Silicon has the advantage of low cost and the availability of integrated circuit processing techniques that can be used to integrate Hall effect devices with sophisticated signal conditioning. Dielectrically isolated silicon processes,

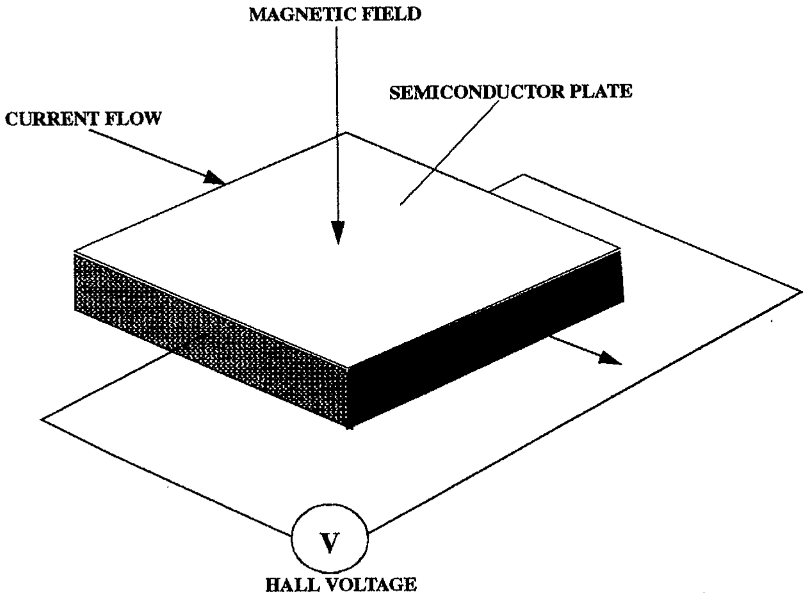


FIGURE 3.5 Hall effect device.

combining integrated circuit techniques with low leakage device isolation, can equal or better the high temperature performance of GaAs.

Hall effect integrated circuits are best categorized in terms of their output characteristics. Analog output devices are usually designed to provide a voltage output that is proportional to the applied magnetic field and also to the power supply voltage. An output ratiometric to the supply allows the device to be easily interfaced to analog-to-digital converters. Analog output devices can be used to construct noncontact absolute position transducers, where the Hall effect device measures a varying field which is designed to be proportional to an angle or linear position. Sensors such as these have no wearing parts other than bearings and can have significant reliability advantages over potentiometers in applications such as throttle-position measurement applications.

Digital output devices are used to construct limit switches or incremental position sensors. The Hall device can be designed to detect homopolar or bipolar fields. Important specifications for digital output devices are operate and release points and the differential between them. The operate point is the maximum field that must be applied to turn the output ON; where ON may be a current-sourcing or current-sinking function. The release point is the minimum field that will guarantee that the sensor is OFF. The differential is the difference between the actual operate and release points. The differential is built in to provide some hysteresis or noise margin to prevent false triggering, particularly at low rates of change of field. The differential may be considerably smaller than the difference between the specified operate and release points. Unipolar devices are specified with operate and release points of the same sign. Bipolar devices are specified with a positive operate point and a negative release point. A caution here is that some devices specified as bipolar do not always require a change of phase of field to operate and release; truly bipolar devices always do.

High-performance Hall effect ICs employ various circuit techniques to improve sensitivity.[6,7] Differential Hall sensors designed for use as gear wheel position sensors use two Hall cells ideally separated by half the gear tooth pitch. This kind of sensor is capable of detecting small changes in unipolar fields. Differential sensors produce an output pulse whose width depends on the rate at which the gear tooth passes the sensor. At very low and very high speeds a very small mark space ratio output results. At high speeds, the timing of the output will be delayed from the mechanical event by a significant proportion of the tooth pitch. A second method is to use a filter circuit to determine the average value of an alternating field and then detect variations from the average value. This method also eliminates any offset that the sensor may have. This method more accurately tracks the mechanical stimulus. A disadvantage is that the filtering function imposes a lower limit on the speed that can be tracked. The devices detailed in Refs. 6 and 7 have lower limits of around 4 Hz. An additional limitation is that the sensor may fail to operate with a high mark space ratio stimulus since the average value of the input will be close to the value of the longest part of the cycle.

A typical Hall sensor application is shown in Fig. 3.6. The Hall device is assembled into a probe with a biasing magnet. The Hall device orientation will depend on its mode of operation, whether unipolar or bipolar. In all cases, the device is sensitive to fields perpendicular to the plane of the silicon surface. Figure 3.6 also compares output waveforms of the Hall sensor configurations discussed earlier as they would appear in this application.

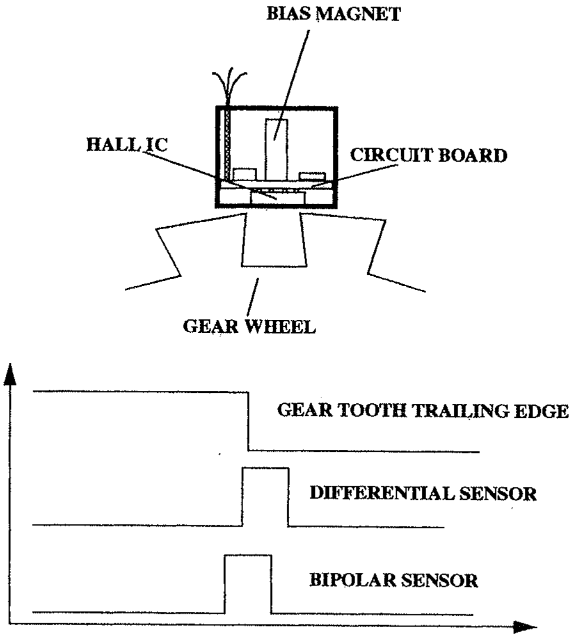


FIGURE 3.6 Hall probe gear position sensor.

Inductive Angle Transducers. Synchro resolvers, or simply resolvers, are absolute angle transducers. Due mainly to their construction, modern brushless resolvers offer the most rugged, reliable, and highest-resolution solution to angle sensing. Resolvers are often considered a high-cost transducer for automotive applications due to the high labor content in the production of most variants. Some designs[8] provide a cost-effective solution by employing sensing and output windings that can be produced on conventional armature-winding machines. Resolvers can be obtained either fully enclosed or as "pancake" devices, with the stator and rotor supplied separated to facilitate over shaft mounting. Resolvers are often referred to by their size. This is the diameter of the case of the device in inches, rounded up to the nearest 10th, and multiplied by 10. For example, a size 11 resolver will be a fraction less than 1.1 in in diameter. Resolver accuracies are specified in arc min. Typical values for accuracy lie in the range 7 arcmin, with more or less accurate versions available from some vendors.[9]

Resolvers are basically rotating transformers. The construction of a typical device and the output waveforms for a 360° rotation are shown in Fig. 3.7. An alternating voltage connected to the reference input provides primary excitation. The range of frequencies used can be 400 to 20 kHz depending on the construction of the resolver; most resolvers are optimized for the 2 to 5 kHz frequency range. The reference signal is coupled to the rotor via a transformer mounted at one end of the rotor shaft. A second rotor winding couples to two orthogonally oriented stator windings. The stator coils are wound so that as the rotor shaft rotates, the amplitudes of the outputs of the stator windings vary as sine and cosine of the shaft angle relative to some zero.

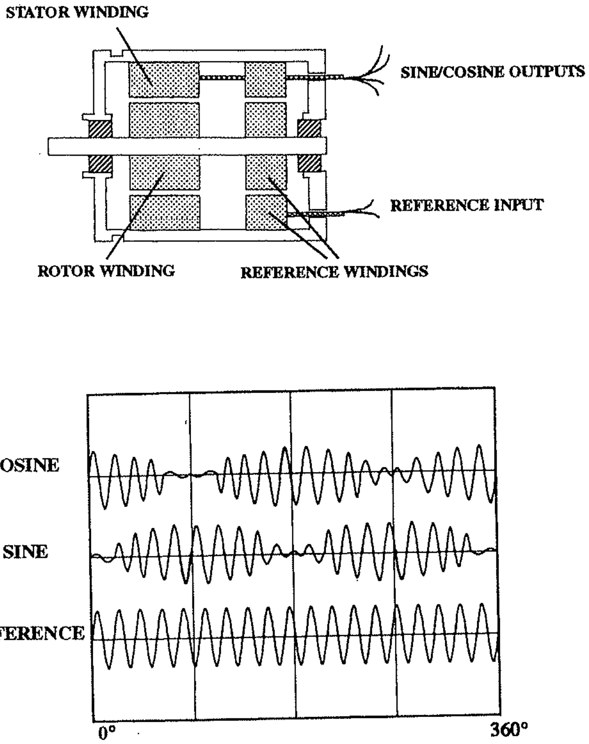


FIGURE 3.7 Resolver construction and output format.

By far, the most economical way of decoding the output of a resolver is to use an IC resolver-to-digital converter (R to D). A functional block diagram of a typical converter is shown in Fig. 3.8. The sine and cosine amplitude-modulated input signals from the resolver representing a shaft angle θ are multiplied by the cosine and sine, respectively, of the current value ϕ of the up/down counter. The resulting signals are subtracted giving:

$$VE = A \sin(\omega t) \sin(\theta - \phi) \tag{3.4}$$

where $A \sin(\omega t)$ represents the reference carrier.

This signal is synchronously demodulated and an integrator and voltage-controlled oscillator form a closed loop with the counter/multiplier, which seeks to null $\sin(\theta - \phi)$. When the

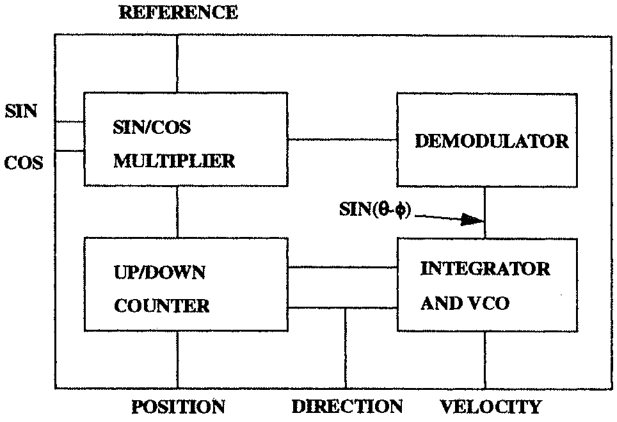


FIGURE 3.8 Tracking resolver-to-digital converter.

null is achieved, the counter value represents the resolver shaft angle within the rated accuracy of the converter. IC R-to-D converters are available that provide parallel or serial digital outputs with resolutions of from 10 to 16 bits and accuracies from 2 to 30 arcmin. Versions are available[10] that emulate standard optical encoder outputs for applications where absolute position measurement is not required but the environment is too harsh for an optical encoder. The system described is a type 2 servo loop which is characterized by zero position and velocity error (not including the limitations of the amplifiers in the IC). R-to-D converters of this type also provide a signal proportional to the angular speed of the resolver from zero to some upper limit, typically 1000 s of rpm, depending on the converter characteristics.

Inductive potential dividers are another class of transducers that are available in many variants.[11,12] A good example is the *rotary variable transformer* or ROVAT.[12] This device comprises a single coil wound on a circular ferromagnetic stator with a number of teeth wound as alternate polarity poles. The stator is excited with an AC signal of around 20 kHz. A rotor with a semicircular conductive screen on its inside surface encircles the stator. The screen reduces the flux linkage between the rotor and stator and the inductance of the screened portion of the stator is reduced, reducing the voltage drop across this portion of the stator. The voltage measured at a central tap on the stator is linearly proportional to the angle of the rotor. Further taps at 90° and 270° from the nominal zero allow a waveform to be measured with amplitude in quadrature with the signal measured at the center point. This allows a 360° absolute angle transducer to be realized, using decoding techniques similar to those which will be described later for the LVDT.

Inductive Linear Displacement Sensors. Shading ring or short-circuit ring sensors are absolute displacement sensors consisting of an E-shaped core with a winding on the central leg of the E. The winding is excited with high-frequency alternating current. An electrically conductive ring of Al or Cu is allowed to slide, maintaining an air gap along the central leg. The ring is attached to the mechanical component, whose position is to be measured. The ring is equivalent to a short-circuited secondary turn in a transformer. The ring has a shading effect, preventing any flux coupling between the legs of the core from its position along the central leg of the core to the open ends of the core (Fig. 3.9).[13] An inductance change can be measured at the terminals of the excitation coil. These sensors are usually used in a potential divider configuration with a reference inductance of similar construction to the main sensor connected in series with the main sensor. A reference alternating voltage is applied across the series-connected reference and sensor inductances, forming an inductive potential divider. The output is then proportional to the ratio of the inductances. This renders the sensor insensitive to tem-

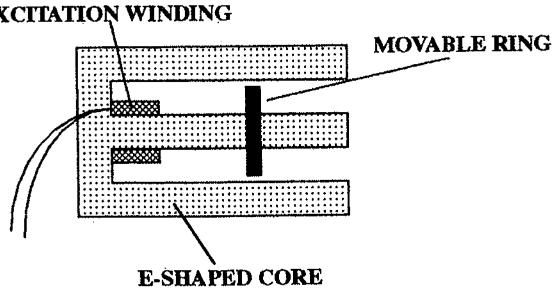


FIGURE 3.9 SCR sensor construction.

perature variations and allows easy adjustment of offsets. Signal-processing electronics can be used to rectify and filter the output and transmit the result to a remote control unit. An alternative construction is to replace the movable ring with an angled channel which can move, relative to the E-shaped core, in a plane perpendicular to the core. Again, the inductance of the sensor is proportional to the linear movement of the angled channel.

Another form of absolute linear displacement sensor is the linear variable differential transformer or LVDT.[14] LVDTs are rugged and reliable and capable of working in harsh environments. Suitable automotive applications include mounting inside hydraulic cylinders in suspension control systems.

LVDTs are constructed from a primary excitation coil positioned centrally on a cylindrical hollow former. Two identical secondary coils are positioned on either side of the primary. The coils have a common core which is free to move within the cylindrical former (Fig. 3.10). The secondaries are normally connected in series, with opposing phases such that with the core centrally positioned and coupling equally to each secondary, the voltage at the node common to both coils will be zero. With this connection, as the core is moved from one extreme of travel through the center to the other extreme, the output signal will vary from a maximum value in phase with the excitation through zero to a maximum value in antiphase with the excitation.

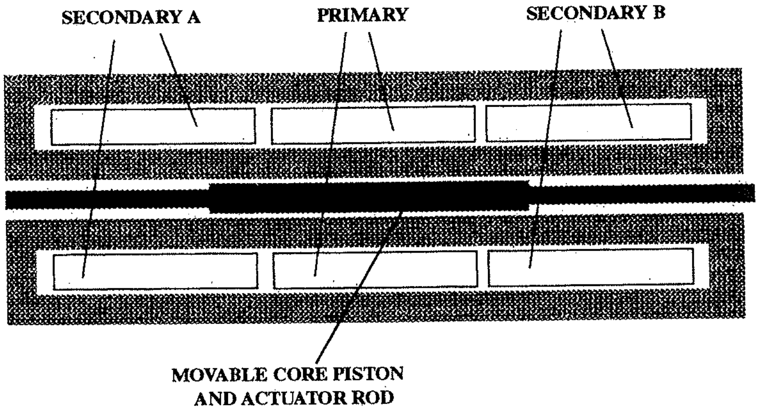


FIGURE 3.10 LVDT construction.

LVDTs are designed to give a linear output within some tolerance, typically ±0.25 percent, over a specified proportion of the available stroke length. The distribution of turns on the secondary coils is carefully arranged to maximize linearity over the widest possible range. LVDTs are available that maintain good linearity with stroke lengths from ±0.05 to ±10 in.

LVDTs operate with effective transformer ratios of between 10:1 to 2:1. The range of primary excitation frequencies can be from 20 Hz to 20 kHz, depending on the construction of the device. Most LVDTs are optimized for the 2- to 5-kHz frequency range. The output signal from an LVDT can be decoded in several different ways and a number of analog and digital integrated circuit solutions exist for this purpose.[15] An example of a typical connection scheme using an LVDT-to-digital converter is shown in Fig. 3.11. In this example, it is assumed that the sum of the voltages across the series-connected secondaries $V_A + V_B$ is a constant over the range of displacements of interest. The majority of LVDTs in production meet this criterion; for those that do not, an additional nonlinearity will result. The IC decodes the function $(V_A - V_B)/(V_A + V_B)$ over the range $[(V_A - V_B)] \le (V_A + V_B)/2$ into a 13-bit digital word that can be accessed via a three-wire serial interface. Additional bits indicate null and over or under range for signals outside the linear range. The ratiometric decoding scheme described here is insensitive to primary-to-secondary phase shifts, temperature, and any residual null voltage that the transducer may have due to stray capacitive coupling.

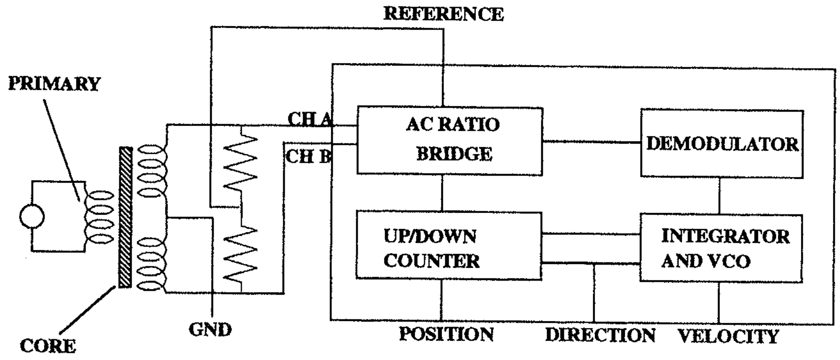


FIGURE 3.11 Tracking LVDT-to-digital converter.

Analog methods of decoding utilize the same basic algorithm as the converter already described. An analog decoder senses the secondary output voltages and evaluates the ratiometric function introduced earlier. The decoder output is filtered and amplified to produce an output voltage proportional to the position of the movable core.

Magnetoresistive. An interesting group of sensors utilize the property of some FeNi alloys such that their resistivity is strongly affected by the presence of a magnetic field. The magnetoresistive effects of one of the useful alloys, Permalloy, which is 81 percent nickel and 19 percent iron, enables sensitive magnetic field sensors with full-scale fields of 5 mT to be built. The variation in resistance is around 2.5 percent for a field of this magnitude. The resistance decreases with increasing field strength; the relationship between field strength and resistance is very nonlinear, approximating a cosine-squared function. Using thin films of Permalloy deposited on a silicon substrate allows signal-conditioning electronics to be integrated with the sensor. Despite the nonlinearity of the phenomenon, accurate linear sensors can be constructed by using the sensor in a bridge arrangement together with flux nulling means, such as a servo-driven coil surrounding the sensor, to effectively operate the sensor at constant resistance. An alternative construction uses opposing "barber pole" thin-film elements.[16] An element is composed of a rectangular thin film of Permalloy overlaid with a series of shorting stripes of aluminum at 45° to the long axis. Two series-connected elements, which are mirror images of each other (reflected about the long axis), form a potential divider. Magnetoresistive sensors generally exhibit high sensitivity, but this leaves them prone to interference from unwanted fields and, therefore, they are unsuitable for some applications.

Magnetostrictive. Magnetostriction is a property of materials that respond to a change of magnetic flux by developing an elastic deformation of their crystal structure. Magnetostrictive linear displacement sensors utilize this phenomenon by launching a compression wave down a cylindrical waveguide using electromagnetic means, usually a current pulse. The waveguide passes through a movable permanent magnet ring at some distance from a receiver site. The compression wave generated at the magnet position travels to the receiver site at approximately 2800 m/s where it causes a change of flux and generates a voltage pulse in a sense coil. The time of flight of the pulse can be measured to determine the distance of the movable ring magnet down the wire. Transducers with stroke lengths in excess of 7.5 m are available that use this technique.

3.3.5 Other Technologies

A number of other technologies can be applied to position-sensing problems, limited only by engineering ingenuity. Some, like capacitive-sensing techniques, are not tolerant of the automotive environment due to sensitivity to humidity, vibration, or temperature or pressure extremes. Others, like resolvers of traditional construction, are sufficiently rugged but prohibitively expensive unless low-cost manufacturing techniques can be found. Occasionally, a nonobvious method finds a niche. An example is a fuel-level sensor disclosed by workers at Bosch,[17] which, in principle, could be applied as a position sensor. The device operates by exciting a metal rod with acoustic waves such that it resonates. One end of the rod is immersed in the fuel. The resonant frequency is a function of the depth of immersion and the fuel level can be determined by suitable electronics.

A significant influence in the selection of technologies for automotive use is the mandatory inclusion of safety systems. Microwave or laser-ranging techniques can be applied to anticollision systems to anticipate obstacles at nighttime or in poor-visibility driving conditions. These will be a ubiquitous component of automobiles in years to come.

3.4 INTERFACING SENSORS TO CONTROL SYSTEMS

All sensors, whether they have digital or analog outputs, provide measurement of real-world phenomena that are then interpreted by another system to either indicate a value, a warning, or close a control loop. It is vitally important that the integrity of the data is maintained by the proper choice of interface.

Cost and reliability in automotive systems are primary drivers in the choice of interface. For a given performance, the sensor that requires the fewer connections will always be selected. In some cases, such as LVDTs where the basic sensor requires 5 or 6 wires, it may be advantageous to locate the signal-conditioning electronics with the sensor and communicate the processed data to a remote controller via a simple serial interface. Interfacing incremental or serial binary data to a microcontroller is straightforward. In the case of incremental data, typically the mechanical system being monitored is moved until some limit or index is detected. Knowledge of the absolute position of the mechanical component is then known. A counter or register can be set to an initial nonzero value or reset to zero. As the motion is detected, pulses from the incremental sensor can be counted and stored as a measure of position. The indexing cycle must be performed each time power is applied. Binary serial data can be read into a register and used directly with no further processing.

Analog-to-digital converters (ADCs) require some special considerations to optimize performance. Not least of these is grounding. In most applications, chassis ground returns cannot be used. Modern automobiles may have voltage drops of 1 V or more between the chassis at the ECU and the sensor site, due to return currents from electric equipment. This voltage is likely to be noisy with many transients and will certainly upset all but the crudest sensors.

Previous sections have discussed the advantages of ratiometric sensors that can use the same reference as the converter. The advantage is that a least-significant bit of the ADC is always a fixed percentage of the sensor span. This eliminates gain, offset, and temperature errors that may occur if separate references are used. Resolution and gain accuracy of a system can be further optimized by making certain that the span of the sensor output uses all of the available input span of the converter. Many ADCs include on board a microcontroller and use switch capacitor techniques to acquire the analog input values. These present a transient load to the sensor once or twice per conversion cycle. Some sensor outputs, particularly sensors with buffer amplifiers, require isolation from this transient to achieve rated accuracy. A simple technique is to use a simple RC filter on the output of the sensor. This limits the transient current that the sensor output sees and shunts the ADC input with a capacitor.

GLOSSARY

Absolute output sensor The sensor output is an unambiguous measure of position and is valid when power is applied.

Arcminute An angular measure. There are 60 arcminutes in 1 degree of arc.

Incremental sensor The sensor indicates changes in position. An additional position reference, such as a limit switch, is often used with this type of sensor.

Linearity error The amount by which the sensor output differs from an ideal characteristic. Usually expressed in percent.

Ratiometric output sensor An input stimulus causes the output to be a fraction of a reference voltage.

REFERENCES

1. Hewlett-Packard, *Optoelectronics Designers Guide,* San Jose, Calif., 1991–1992.
2. Novatechnik Position Sensor Data, Ostfildern, Germany, 1992.
3. P. Hammond, *Electromagnetism for Engineers,* Pergamon Press Ltd., Oxford, England, 1965.
4. Roland K. Kolter, European Patent Application EP 0 019 530 A1, Applicant Bendix, 1980.
5. Henry P. Baltes, and Popovic, Radivoje S., "Integrated magnetic field sensors," *Proceedings of the IEEE* vol. 74, no. 8, Aug. 1986, pp. 1107–1132.
6. Hartmut Jasberg, "Differential Hall IC for gear-tooth sensing," *Sensors and Actuators,* A21–A23, 1990, pp. 737–742.
7. AD22150 Data Sheet, Analog Devices, Norwood, Mass.
8. Charles S. Smith, U.S. Patent 4 962 331, Assigned to Servo-Tek Products Co. Inc., Hawthorne, N.J., 1990.
9. Clifton Precision, Analog Components Data, Clifton Heights, Pa.
10. AD2S90 Data Sheet, Analog Devices, Norwood, Mass.
11. Novatechnik Angle Sensor Data, Ostfildern, Germany, 1993.

12. Donald L. Hore, and Flowerdew, Peter M., "Developments in inductive analog transducers for 360° rotation or tilt and for linear displacement," *IEE International Conference,* No. 285, 1988.
13. E. Zabler, and Heintz, F., "Shading-ring sensors as versatile position and angle sensors in motor vehicles," *Sensors and Actuators* **3**, 1982/3, pp. 315–326.
14. *Schaevitz Linear and Angular Displacement Transducers Catalog,* Pennsauken, N.J.
15. AD2S93 and AD598 Data Sheets, Analog Devices, Norwood, Mass.
16. F. Heintz and Zabler, E., "Application possibilities and future chances of 'smart' sensors in the motor vehicle," *SAE Technical Paper Series* #890304.
17. E. Zabler, "Universal low-cost fuel-level sensor," Robert Bosch GmbH, Ettingen, Germany.

ABOUT THE AUTHOR

Paul Nickson is product line manager for Analog Devices' Sensor and Automotive Group in Wilmington, Mass. He has BSc Honours in Electronic and Electrical Engineering from the University of Birmingham, England and 16 years' experience in integrated circuit design. For the past five years, he has focused on silicon sensors and sensor signal conditioning.

CHAPTER 7
SPEED AND ACCELERATION SENSORS

William C. Dunn
Motorola Inc.
Semiconductor Products Sector

7.1 INTRODUCTION

In the automotive arena, speed and acceleration sensors are used in a wide variety of applications, from improving engine performance through safety to helping to provide creature comforts.

Speed sensing can be divided into rotational and linear applications. Rotational speed sensing has two major application areas: engine speed monitoring to enhance engine control, and performance and antilock breaking and traction control systems for improved road handling and safety. Linear sensing can be used for ground-speed monitoring for vehicle control, obstacle detection, and crash avoidance. Acceleration sensors are used in air bag deployment, ride control, antilock brake, traction, and inertial navigation systems.

In most cases, there are a number of different sensor types available for a specific monitoring function. However, the choice of sensor for a specific application can be difficult to make. The selection may be determined by the familiarity of the system's designer with the sensor. On the other hand, the output from one sensor can be used for several applications, and the individual requirements of each application may eventually determine the sensor to be used.

Electronics and electronic sensors are making rapid inroads into the automotive market. In order to analyze the large amounts of sensor data needed for low emissions and efficient engine control, it is necessary to process the information using microcontrollers (MCUs), which can operate at high speeds and in real time. Sensors that can convert information directly into a digital format for MCU compatibility have a distinct advantage over an analog output format. Digital signals are also supply-line voltage-insensitive, virtually unaffected by noise, and have better resolution than can be obtained with analog signals. If the addition of an analog-to-digital converter (on or off the MCU) is required for compatibility, the system cost is increased. The accuracy of the control system is only as good as the integrity of the sensor data supplied to the MCU. Hence, the importance of the performance of the sensor.

This chapter concentrates on speed and acceleration, and therefore does not go into all the other different types of applications for which many of these sensors can be used in the automobile.

7.2 SPEED-SENSING DEVICES

In automotive applications, the environment must be taken into consideration. The sensing must be accurate, the devices must be rugged and reliable, and they must function in the presence of oil, grease, dirt, and inclement weather conditions. These requirements have severely limited the use of a number of otherwise practical alternatives, such as optical sensors and contact sensing.

In the area of rotational speed monitoring, the most practical devices use magnetic field sensing. These sensors are Hall effect devices, variable reluctance (VR), and magnetoresistive or magnetic resistance element (MRE) devices. Both the Hall effect and VR devices have been widely used and have a proven track record. The MRE device has only recently come into its own with improved technology and provides a viable alternative to the Hall effect device. The MRE device has a higher sensitivity and a wider operating temperature range than the Hall effect device.

For the measurement of ground speed and object detection, optical, radar, laser, infrared, and ultrasonics have been explored. Linear sensing devices typically use the Doppler effect for speed sensing and pulse modulation for distance measuring. These devices are used for object detection in blind spots when reversing or changing lanes, and in such applications as collision avoidance systems.

7.2.1 Variable Reluctance Devices

Variable reluctance devices are in effect small ac generators with an output voltage proportional to speed, but they are limited in applications where zero speed sensing is required. The operating frequency range of the VR device is from about 10 Hz to 50 kHz. It is insensitive to mechanical stress and has a wide temperature operating range from –40 to 190 °C. The supply voltage and offset drift will depend upon the control electronics. The VR device, originally designed around existing automotive electromechanical systems, was adapted for electronic control. The ferrous metal in the VR system is designed for maximum output voltage at low rpm (revolutions per minute), to get as close as possible to zero speed sensing without generating excessive voltages at maximum rpm (up to 150 V). This device gives a linear output voltage with frequency. Most systems use MCUs for data processing, so the VR device needs an analog-to-digital (A/D) converter to generate a digital signal for compatibility with the MCU. Although the VR device itself is inexpensive, the extra costs for data conversion may eventually lead to its demise in many automotive applications.

7.2.2 Hall Effect Devices

The Hall effect exists when a current flowing in a carrier experiencing a magnetic field perpendicular to the direction of current flow results in the current being deflected perpendicular to the field and to the direction of the current. The Hall effect is shown in Fig. 7.1. A current I_c flowing through the device between terminals 1 and 2 will produce a potential VH between terminals 3 and 4, when a magnetic field **B** is applied perpendicular to the device. The potential VH is determined by the strength of the magnetic field and the current flowing. Hall effect devices can be manufactured with indium, gallium arsenide, or silicon. A comparison of their properties is given in Table 7.1.

As can be seen, silicon is the most sensitive material. It is compatible with ICs and has a wide operating temperature range. The Hall effect device is well known both in industry and in the automotive arena for rotational and position-sensing applications. However, recent developments in Hall sensing devices, such as differential sensing and integration, have given improved sensor characteristics, which may result in greater potential in automotive applications.

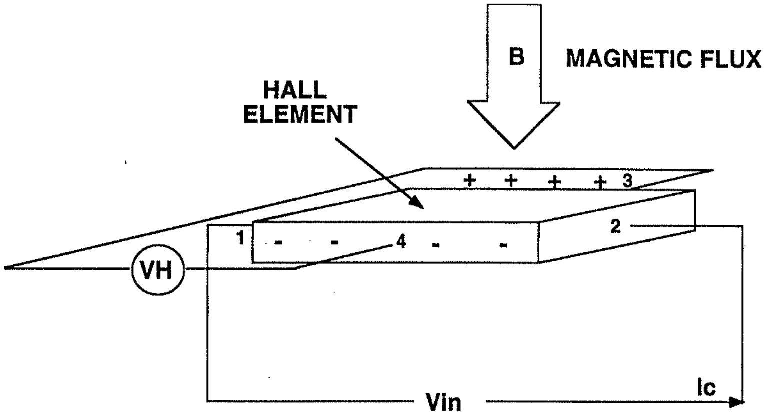


FIGURE 7.1 Hall effect.

The Hall effect device is very versatile, flexible in use, easy to package, and can be used for zero speed sensing (it can give an output when there is no rotation). Hall devices give a frequency output that is proportional to speed, making them compatible with MCUs. The Hall device is normally configured as a bridge to minimize temperature effects and to increase the sensitivity of the sensor.

A typical Hall effect sensor configuration with waveforms is shown in Fig. 7.2. The teeth of the ferrous wheel concentrate the magnetic flux when the teeth come into close proximity to the Hall sensor and magnet. The output from the sensor is a sinusoidal waveform, whose frequency is the rpm of the ferrous wheel multiplied by the number of teeth on the wheel. The resolution of the system depends on the number of teeth in the wheel (typically 20).

7.2.3 Magnetoresistive Devices

The magnetoresistive effect is the property of a current-carrying ferromagnetic material to change its resistivity in the presence of an external magnetic field. For example, a ferromagnetic (permalloy) element (an alloy containing 20 percent iron and 80 percent nickel) will change its resistivity by 2 to 3 percent when it experiences a 90° rotation in a magnetic field.[1] The resistivity value rises to a maximum when the direction of current and magnetic field are coincident, and is a minimum when the fields are perpendicular to each other. This relationship is shown in Fig. 7.3. This attribute is known as the Anisotropic Magneto Resistive Effect. The resistance *R* of an element is related to the angle q between the current and the magnetic field directions by the expression:

$$R = R_{\mathrm{ll}} \cos^2 q + R_l \sin^2 q \tag{7.1}$$

TABLE 7.1 Comparison of Properties of Hall Effect Materials

Material	Operating temp. °C	Supply voltage	Sensitivity @ 1 kA/m	Frequency range
Indium	–40 to 100	1V	7 mV	0 to 1 MHz
GaAs	–40 to 150	5 V	1.2 mV	0 to 1 MHz
Si (with conditioning)	–40 to 150	12 V	94 mV	0 to 100 kHz

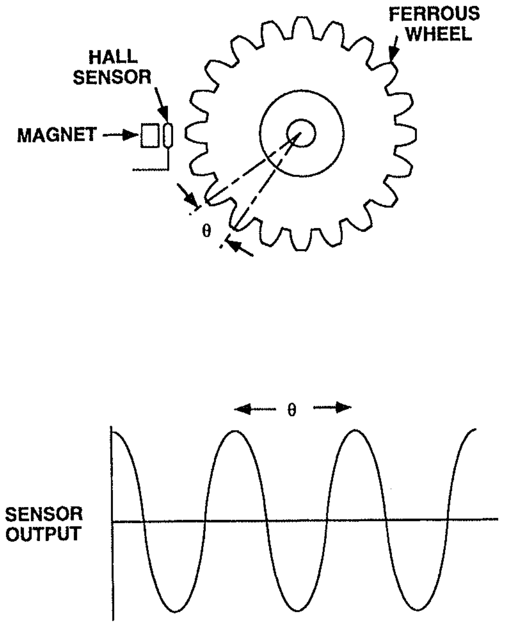


FIGURE 7.2 Hall sensor and output waveforms.

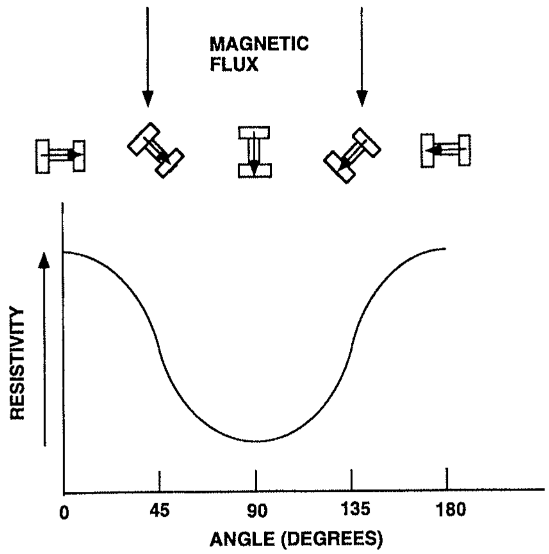


FIGURE 7.3 Relationship between magnetic field and resistivity change in MRE devices.

where R_{II} is the resistance when the current and the magnetic field directions are parallel and R_l is the resistance when the current and the magnetic field directions are perpendicular.

MRE devices give an output when stationary, which make them suitable for zero speed sensing. MRE devices also give an output frequency that is proportional to speed, making for ease of interfacing with an MCU. For good sensitivity and to minimize temperature effects, a bridge configuration is normally used. In an MRE sensor, aluminum strips can be put across the permalloy element to linearize the device. This configuration is shown in Fig. 7.4 together with a typical MRE characteristic. The low-resistance aluminum stripes cause the current to flow at 45° in the permalloy element, which biases the element into a linear operating region.

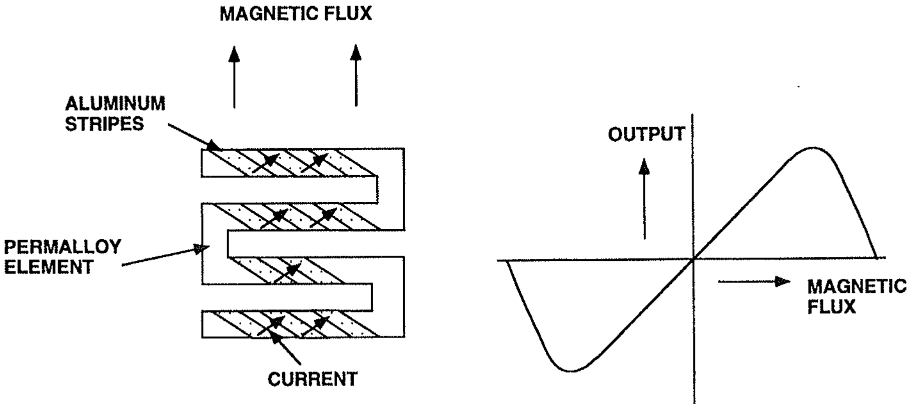


FIGURE 7.4 Use of permalloy strip for linearization in MRE devices.

Integrated MRE devices can typically operate from –40 to 150 °C, over a supply voltage range of 8 to 16 V, and at frequencies from 0 to 1 MHz. In comparing the MRE sensor to a Hall effect device, the MRE has a higher sensitivity, is less prone to mechanical stress, has a wider operating frequency range, has the potential of being more cost effective, gives better linearity, and is more reproducible. However, it is more sensitive to external magnetic fields. Table 7.2 shows a comparison of rotational sensing devices.

TABLE 7.2 Comparison of Sensing Devices

Sensor type	Operating temperature (°C)	Sensitivity 1 kA/m (mV)	Frequency range	Mechanical stress
Hall effect	–40 to 150	90	0 to 100 kHz	High
MRE	–40 to 150	140	0 to 1 MHz	Low
VR	–40 to 190		1 to 50 kHz	None
Magnetic transistor	–40 to 150	250	0 to 500 kHz	Low

It should be noted that Hall effect and MRE devices have many applications in the automobile outside of rotational speed sensing, such as position sensing, fuel-level sensing, and active suspension. The magnetic transistor is showing potential in rotational speed sensing and position-sensing applications, and may eventually be another viable contender to the Hall effect device in the automotive market.

7.2.4 Ultrasonic Devices

Ultrasonic devices can be used to measure distance, ground speed, and as a proximity detector. To give direction and beam shape, the signals are transmitted and received via specially configured horns and orifices. The transmitter and receiver horns are similar in shape, but are normally separate to accommodate different characteristics. The ultrasonic devices are made from PZT crystal-oriented piezoelectric material ($PbZrO_3$—$PbTiO_3$).

For the measurement of distance or object detection, a pulse of ultrasonic energy is transmitted and the time is measured for the reflected pulse to return to the receiver. The frequency of the transmitted ultrasonic waves are typically about 40 kHz and travel with a velocity of 340 m/s at 15 °C. This velocity changes with temperature and pressure. However, these parameters can be measured and corrections made if high accuracy is required. The repetition frequency and power requirements depend on the distance to be measured. To measure speed, the distance variation with time can be measured and the velocity calculated. A more common method is to use the Doppler effect, which is a change in the transmitted frequency as detected by the receiver due to motion of the target (or, in this case, the motion of the transmitter and receiver).

7.2.5 Optical and Radio Frequency Devices

Optical devices are still being used for rotational speed sensing. They are normally light-emitting diodes (LEDs) with optical sensors. Figure 7.5 shows a typical optical sensor system. An optical sensor detects light from an LED through a series of slits cut into the rotating disc, so that the output from the sensor is a pulse train whose frequency is equal to the rpm of the disc multiplied by the number of slits. The higher the number of slits in the disc, the smaller the angle of rotation that can be measured. The optical sensor can be a single photodiode or a photodiode array as shown. This array gives a more accurate determination of the position of the slit, resulting in higher resolution of the position of the disc.

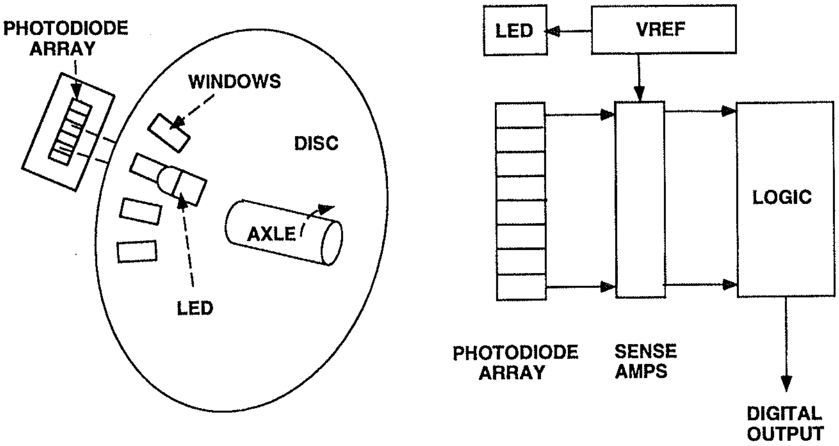


FIGURE 7.5 Optical sensor.

Optical and radio frequency (RF) devices are used for object detection, linear approach speed, and distance measurements in crash avoidance systems where distances greater than about 10 m are involved. These devices use the same principles as the ultrasonic devices. Optical devices normally use lasers or infrared devices for the transmitting source and optical sensors for the receivers. RF devices use gallium arsenide or Gunn devices to obtain the power and high frequency (about 100 GHz) required in the transmitter. The high operating frequency is set to a large extent by the need for a small antenna. These applications are under development and are discussed in Sec. 7.7.2.

7.3 AUTOMOTIVE APPLICATIONS FOR SPEED SENSING

There are several applications for rotational speed sensing. First it is necessary to monitor engine speed. This information is used for transmission control, engine control, cruise control, and possibly for a tachometer. Electronics and electronic sensing in the automobile were brought about by the need for higher-efficiency engines, better fuel economy, increased power and performance, and lower emissions. Second, wheel speed sensing is required for use in transmissions, cruise control, speedometers, antilock brake systems (ABS), traction control (ASR), variable ratio power steering assist, four-wheel steering, and possibly in inertial navigation and air bag deployment applications.

Linear speed sensing can be used to measure the ground speed. This measurement also has the possibly of use in ABS, ASR, and inertial navigation. Similar types of sensors can be used in crash avoidance, proximity, and obstacle detection applications.

7.3.1 Rotational Applications

The high timing accuracy that can be obtained with fuel injection systems and replacement of points by sensors have made cost-effective engine control and low maintenance a reality. Adjustment of the stoichiometric ratio of air to fuel, accurate ignition timing, and oxygen sensors in the exhaust system, have vastly improved engine performance and greatly reduced emissions over widely varying operating conditions. The two important factors in engine control are the engine speed in rpm and the crank angle. These signals are used by the engine control MCU for determination of fuel injection and ignition timing. The engine rpm measurement range is from 50 to (say) 8000 rpm. A resolution of about 10 rpm is required for an accuracy of about 0.2 percent. For injection and ignition control in a six-cylinder engine, the interval between combustion at maximum rpm is 2.5 ms, so that this time sets the injection period. In practice, a crank angle accuracy of between 1 and 2 degrees per revolution is required. Newer systems with sequential fuel injection, may also require information on TDC (top dead center) for each cylinder to determine the timing. With the low frequencies involved in this application, either Hall effect or MRE devices can be used for monitoring both the engine rpm and crank angle.

Vehicle speed measurements are in the range 0 to 180 km/h (120 mph) and digital displays must have an accuracy of 1 km/h. Some systems have a mechanical pickoff from the drive shaft, which can then use optical sensors for the measurement of road speed. However, newer systems have a pickoff located directly on the drive shaft, which makes optical devices less practical. It is preferred to eliminate the remote sensing via mechanical coupling to save the cost of the associated mechanical components, seals, maintenance, and so on. One method of pickoff is a ring magnet with between 4 and 20 magnetic poles (depending on the required resolution). Figure 7.6 shows such a system using an MRE sensing device. The magnetic flux changes are sensed by an MRE bridge sensor when the magnet disc is rotating. The bridge is supplied from a voltage reference circuit, and its output is amplified and shaped to give a frequency output that is proportional to shaft rotation speed. A ferrous toothed wheel pickoff with magnet and flux concentrator can also be used (see Fig. 7.2). Vehicle speed sensing can be performed with Hall effect, MRE, or VR devices. The number of pulses P per second from the detector are counted to measure speed S, from the following relationship:

$$P = N \times S \times K \tag{7.2}$$

where N = the number of magnetic poles on ring magnet or wheel teeth
K = a constant determined by axle ratio and wheel size

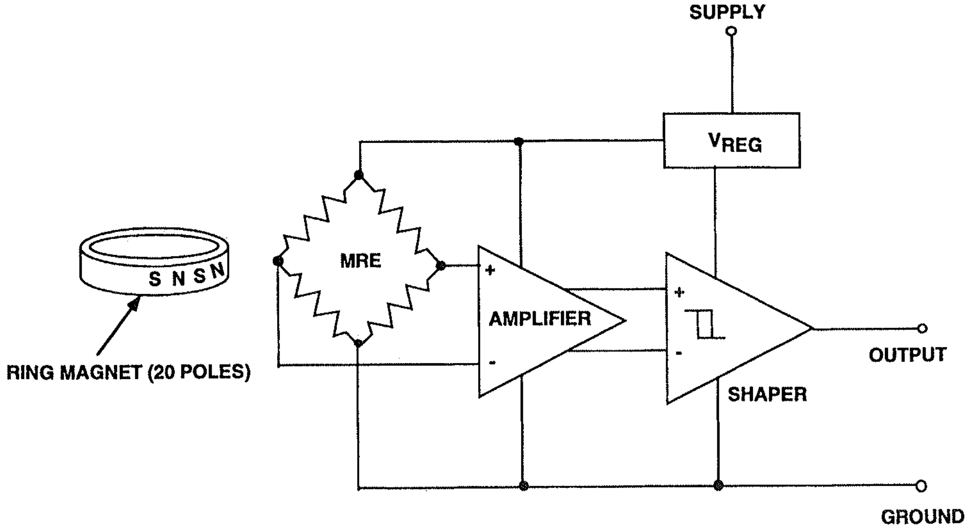


FIGURE 7.6 MRE speed-sensing module.

The resolution in vehicle speed is then:

$$\frac{P}{S} = N \times K \tag{7.3}$$

The typical system requirements are an operating temperature of –40 to 120 °C, rotational speed detection of 5 to 3000 rpm (1000 p/s), and a duty cycle ratio of 50 ± 10%.

In applications such as ABS, ASR, and four-wheel steering, additional speed sensors are attached to all four wheels so that the slip differential between the wheels can be measured. VR devices have been used and are very cost effective in this application. But the cost of other devices is dropping and as they become cost effective, they are being designed into new systems. In electronic transmission applications, information from the road and engine speed sensors, as well as torque data and throttle position are required for the MCU to select the optimum gear ratio. Electronic control can ensure smooth transition between gear ratios. Transmissions using electronic control are also smaller than conventional automatic transmissions, thus enabling more gear ratios for better performance, higher torque, efficiency, and acceleration.

Cruise control systems require information from the road and engine speed sensors to control the throttle position, and possibly the optimum selection of transmission ratios. Variable ratio assisted power steering also requires information from the wheel speed sensors for adjustment of the steering ratios for ease of turning at low speeds and good road control at high speeds. If automatic tire pressure adjustment becomes a reality, this system may also require information from the wheel speed sensors.

Another application for rotational speed sensing is to control the speed of the radiator cooling fan. The speed of the fan is determined by the coolant temperature. Hall effect devices (MRE can also be used) have been used to monitor the position of the armature and speed of the cooling fan motor. The motor controller uses this information to modulate the power to the motor through a three-phase bridge driving circuit for the control of the fan motor speed.

7.3.2 Linear Applications

Under linear applications are the detection of obstacles close to the vehicle, crash avoidance, distance of the chassis relative to the ground for ride control, measurement of ground speed for ABS, ASR, and inertial navigation. Ultrasonic devices are normally used for short distance measurements (<10 m) and RF devices for long distance measurements (see Sec. 7.6.2). For the measurement of objects from 0.5 to 2 m using ultrasonics, a pulse repetition rate of about 15 Hz is used. The reflected pulses take from 3 to 12 ms to return. The return time T is given by

$$L = C \times \frac{T}{2} \text{ (m)} \tag{7.4}$$

where L = distance to target
C = the transmission speed [given by $C = 331 + 0.6$ t (m/sec)]
T = temperature (at 15 °C), the speed of ultrasonic waves is 340 m/s.

In the case of chassis-to-ground measurements for ride control and ground speed measurements, the distance to be measured is from 15 to 50 cm and a higher pulse repetition rate can be used (up to 50 Hz). In this case, the reflected pulse takes from 0.9 to 3 ms to return. For ride control applications, an accelerometer has an advantage over distance measurement, in that it is unaffected by varying distance measurements over rough terrain.

7.4 ACCELERATION SENSING DEVICES

Acceleration sensors vary widely in their construction and operation. In applications such as crash sensors for air bag deployment, mechanical devices (simple mechanical switches) have been developed and are in use. Mechanical switches are normally located at the point of impact in the crash zone. With the development of micromachined devices, solid state analog accelerometers have been designed for air bag applications. The analog accelerometers are centrally placed on the automotive frame. These devices can be very cost effective when compared to mechanical switches and are rapidly replacing the electromechanical devices. Silicon micromachined sensors provide a higher degree of functionality, can be programmed, have high reliability, have excellent device-to-device uniformity, and can be integrated with memory circuits to create a more accurate sensor. Additional features such as self-test and diagnostics are also available.

7.4.1 Mechanical Sensing Devices

Mechanical switches are simple make-break devices. Figure 7.7 shows the cross section of a Breed type of switch or sensor. The device contains a spring, a metal ball, and electric contacts in a tube. On impact, the inertia in the ball causes it to move against the retaining force of the spring and closes the electric contacts at the end of travel. An alternative to this device is the one shown in Fig. 7.8. It consists of a cylindrical mass wound in a flat spring. The seismic mass

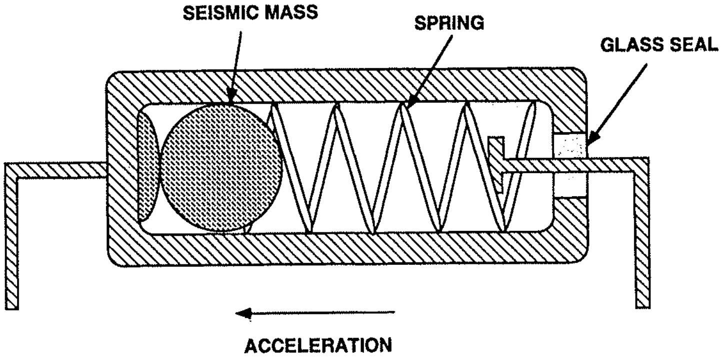


FIGURE 7.7 Cross section of mechanical sensor.

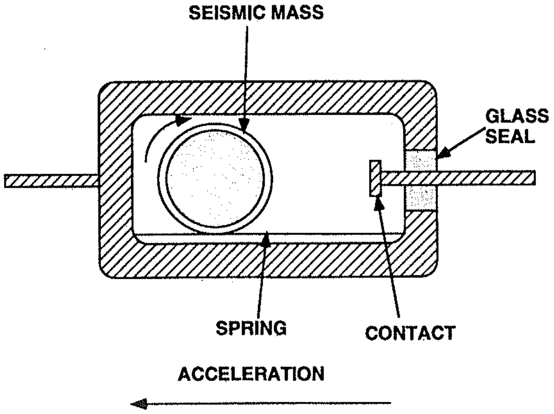


FIGURE 7.8 Cross section of mechanical switch.

rolls on impact against the spring tension, and again makes electrical contact at the end of travel. The machining tolerances on these devices are high and give wide variations in the acceleration trigger point.

7.4.2 Piezoelectric Sensing Devices

Piezoelectric devices consist of a layer of piezoelectric material (such as quartz) sandwiched between a mounting plate and a seismic mass. Electric connections are made to both sides of the piezoelectric material. The cross section of such a device is shown in Fig. 7.9. Piezoelectric material has the unique property that when a force or pressure is applied to opposite faces of the material, an electrical charge is produced. This charge can be amplified to give an output voltage that is proportional to the applied force. Piezoelectric devices can be effective in some applications, but are not suitable for sensing zero- or low-frequency acceleration, that is <5 Hz due to offset and temperature problems (pyroelectric effect). Piezoelectric sensors have a high Q, low damping, and a very high output impedance. Self-test features are also difficult to implement. The main advantages of piezoelectric devices are their wide operating temperature range (up to 300 °C), and high operating frequency (100 kHz).

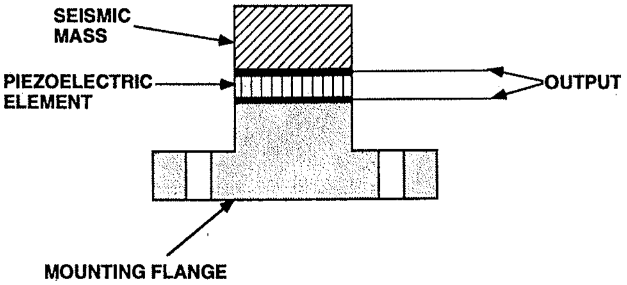


FIGURE 7.9 Cross section of a piezoelectric accelerometer.

Figure 7.10 shows a typical signal-conditioning circuit[2] and the trimming network used with piezoelectric sensors. The output from the sensor is fed to a charge amplifier, which converts the charge generated by the sensor into a voltage proportional to the charge. The circuit

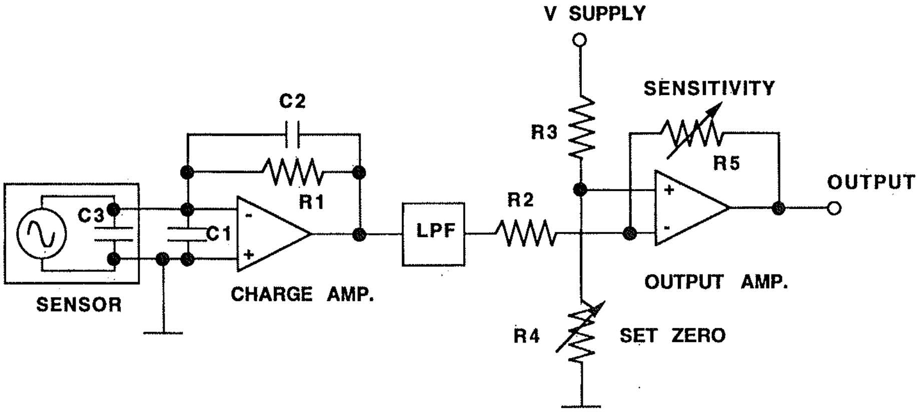


FIGURE 7.10 Piezoelectric signal-conditioning circuit.

is a modified virtual ground voltage amplifier. Feedback via capacitor C2 and resistor R1 is used to maintain the input at a virtual ground potential. This type of circuit minimizes the effect of stray or ground capacitance C1. The output voltage from the amplifier is fed via a low-pass filter (LPF) to an output amplifier, where it is trimmed for offset by R4 and sensitivity by R5. In system development, the sensitivity is set by the piezoelectric material used. Higher-sensitivity materials however, exhibit higher sensitivity to temperature variations.

7.4.3 Piezoresistive Sensing Devices

The property of some materials to change their resistivity when exposed to stress is called the piezoresistive effect. In silicon, the sensing resistors can be either P or N type doped regions, which can be very sensitive to strain. The resistors are also sensitive to temperature, so that the strain gauge is normally designed as a bridge configuration to minimize temperature effects and to obtain higher sensitivities (see also Chap. 2). In order to maintain good linearity, the operating temperature of piezoresistive devices is limited to about 100 °C. The nonlinearity is caused by excessive junction leakage current at high temperatures. Higher operating temperatures have been obtained using oxide-isolated strain gages (up to 175 °C). An uncompensated strain gauge has a typical error of 3 percent over the operating temperature range –20 to 80 °C. This error can be reduced with a compensating resistor, and still further reduced to about 0.5 percent by the use of thermistors, over an improved operating temperature range of –40 to 110 °C.

Piezoresistive sensing can be used with bulk micromachined accelerometers. Such a device is shown in Fig. 7.11. The strain-sensing elements are diffused into the suspension arms. These elements can then detect strain in the arms caused by acceleration forces on the seismic mass.

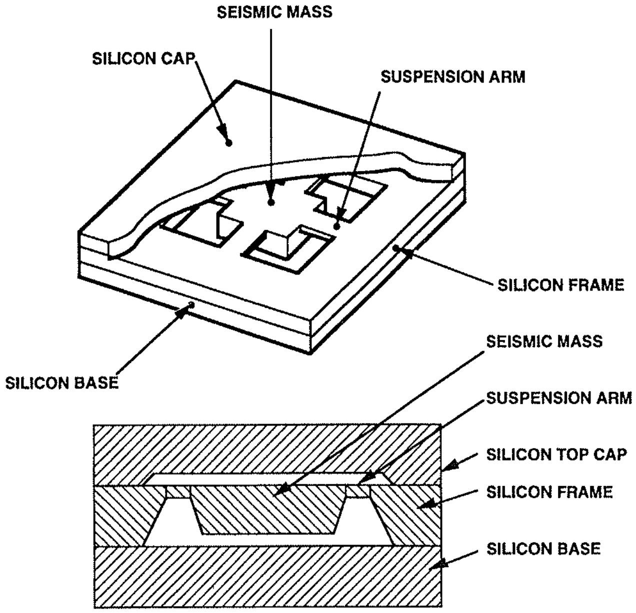


FIGURE 7.11 Bulk micromachined accelerometer.

7.4.4 Capacitive Sensing Devices

When used with micromachined structures as shown in Figs. 7.11 and 7.12, differential capacitive sensing has a number of attractive features when compared to other methods of sensing: easily implemented self-test, temperature insensitivity, and smaller size. In addition, comparing capacitive sensing to piezoelectric sensing reveals that capacitive sensing has the advantages of dc and low-frequency operation and well-controlled damping. When compared to piezoresistive sensing; differential capacitive sensing has the advantage of a wider operating temperature range and requires less complex trimming. Capacitive sensing has one other major advantage over other sensing methods in that it can be used in closed-loop servo systems. In these systems, voltages can be applied to the capacitive plates to produce electrostatic forces, which will balance the forces on the seismic mass due to acceleration. The main advantage of closed-loop operation is to make the sensor to a large extent independent of process variations. Signal-conditioning circuits can be designed to detect changes in capacitance of <0.1 fF, so that plate capacitances in the range of 200 to 400 fF can be used. The small spacing between capacitor plates in micromachining technology (2 μ) enables practical acceleration sensors as small as 500 μ × 500 μ. When designing capacitive sensors, care must be taken to ensure that the sensing voltages are properly balanced to minimize offsets due to electrostatic forces. These forces can also be produced by internal noise sources. The attributes of capacitive sensing are a linear response, operation over wide temperature ranges (–40 to 150 °C), and a frequency response from dc to about 2 kHz.

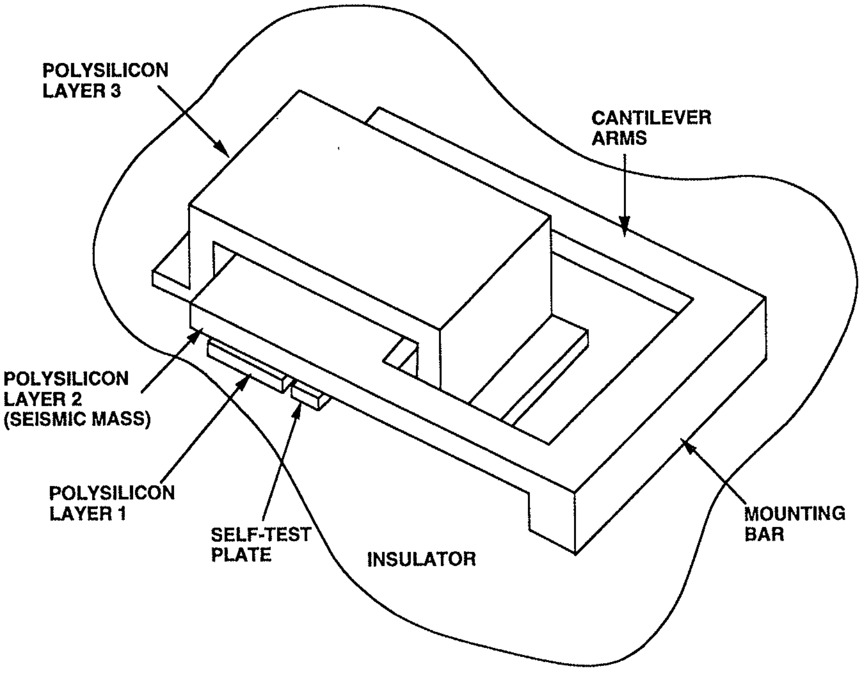


FIGURE 7.12 Surface micromachined accelerometer.

Micromachined Structures. There are a variety of types of micromachined structures that can be used in accelerometers. These structures fall into two technologies: bulk micromachined structures and surface micromachined structures. Bulk micromachined devices are structures etched out of silicon wafers. Figure 7.11 shows the cross section of a bulk micromachined device consisting of three layers of silicon bonded together. The center layer is shaped to form a seismic mass suspended by four arms[3] (a cantilever structure has also been designed with two suspension arms[4]). When acted upon by acceleration forces, the seismic mass moves between the top and bottom plates. In this case, the movement can be sensed using piezoresistive elements diffused into the suspension arms, or differential capacitive sensing can be used between the seismic mass and the upper and lower silicon plates. The top and bottom plates can also be made of glass with metalized areas to form the top and bottom capacitors. Such devices have been designed to operate from the high g range (>1000 g), down to sensors with resolution in the μg range. Closed-loop control techniques are normally used in these lower g ranges.

The surface micromachined device shown in Fig. 7.12 is built using layers of polysilicon and sacrificial glass, which are alternately deposited and shaped. In this case, three layers of polysilicon and two layers of sacrificial glass were used. After deposition of the third polysilicon layer, the sacrificial glass is etched away leaving the freestanding structure as shown. A number of etch holes are normally placed in the second and third layer of polysilicon to speed up the etch process. These etch holes are also used to control the squeeze film damping and bandwidth of the device. The seismic mass of the second-layer polysilicon in these devices is of the order of 5×10^{-10} kg. A second plate under the middle polysilicon can be used for self-test. This function is achieved by applying a voltage to the self-test plate, which in turn will produce an electrostatic force on the center polysilicon plate causing it to deflect. This deflection will simulate an external acceleration force. An alternative to the polysilicon and glass structure is nickel with sacrificial copper.[5]

Differential capacitive sensing is used with all of these structures. Both bulk and surface micromachined devices have a very rugged construction. These devices use squeeze film damping to control bandwidth and to ensure critical damping of the resonant frequency (about 2 kHz for bulk and 10 kHz for surface micromachined devices). Film damping also ensures high resistance to shock in the sensing direction. In the directions perpendicular to the sensitive axes, the devices are rugged by construction with low cross-axis sensitivity (<3 percent). An accelerometer designed to sense a few g will typically have a shock tolerance of well over 5000 g. As already noted, surface micromachined devices that have been developed for air bag deployment have analog outputs. These devices normally operate from a 5-V supply, have a bandwidth of about 1 kHz and a sensitivity of 40 mV per g, giving a full-scale output with ±50 g input. Both open- and closed-loop techniques have been used for sensing. In comparing bulk and surface micromachined devices, the bulk structure is larger, using crystal-oriented etching with end stops, which require extra diffusions; whereas, the surface micromachined device uses isotropic etching (masking) with different materials acting as end stops. Surface structures have the potential for easier integration and use a simpler less costly process, but do require annealing.

Open-Loop Sensing. Open-loop signal-conditioning circuits amplify and convert the capacitance changes into a voltage. Such a CMOS circuit using switched capacitor techniques is shown in Fig. 7.13. The circuit contains a virtual ground amplifier to minimize the effect of stray capacitance. The positive input of the amplifier is referenced to a voltage of $V_{\mathrm{REF}}/2$, when switch $S2$ is closed the amplifier has unity gain, and the voltage on the middle plate of the sensor is set to $V_{\mathrm{REF}}/2$. After $S2$ is opened, $S1$ is switched so that any differences between sensor capacitors C_1 and C_2 produces a charge at the negative input of the amplifier. This charge produces a voltage (V_{out}) at the output of the integrating amplifier. The output voltage of the amplifier is given by

$$V_{\mathrm{out}} = V_{\mathrm{REF}} \frac{C_1 - C_2}{C_3} \tag{7.5}$$

where C_1 and C_2 are the sensor capacitances
C_3 is the integrator capacitance

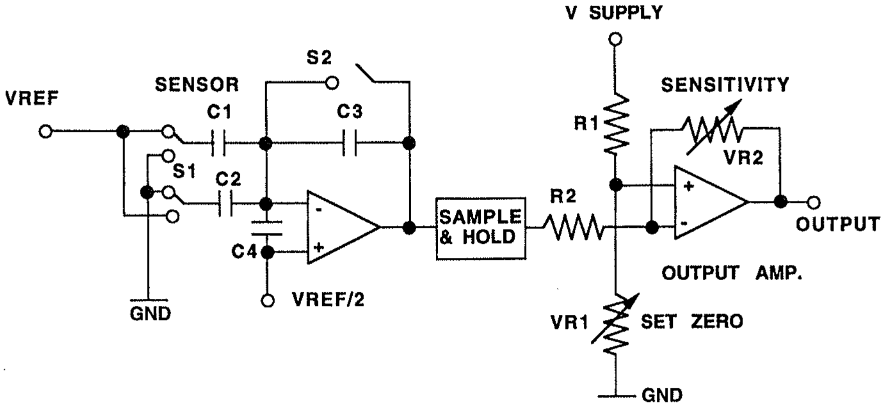


FIGURE 7.13 Capacitive sensing integrator circuit.

If the reference voltage is made proportional to the supply voltage, a ratiometric output is obtained. That is, the system gain is proportional to the supply voltage. This is a requirement in some systems to facilitate the design of the A/D converter.

A block diagram of the system is shown in Fig. 7.14. The system contains an internal oscillator, voltage reference, amplifier, sample and hold, switched capacitor filter, trim network, and output buffer. The output voltage in such a circuit is proportional to the capacitance change. This change is proportional to 1/displacement, or 1/acceleration, giving rise to some nonlinearities. However, the displacement is small compared to the spacing between the plates, so that the output voltage approximates to acceleration, giving less than 3 percent nonlinearity. The filter is used for noise reduction and to set the bandwidth for specific applications. Trimming is used to set the zero operating point and sensitivity of the system.

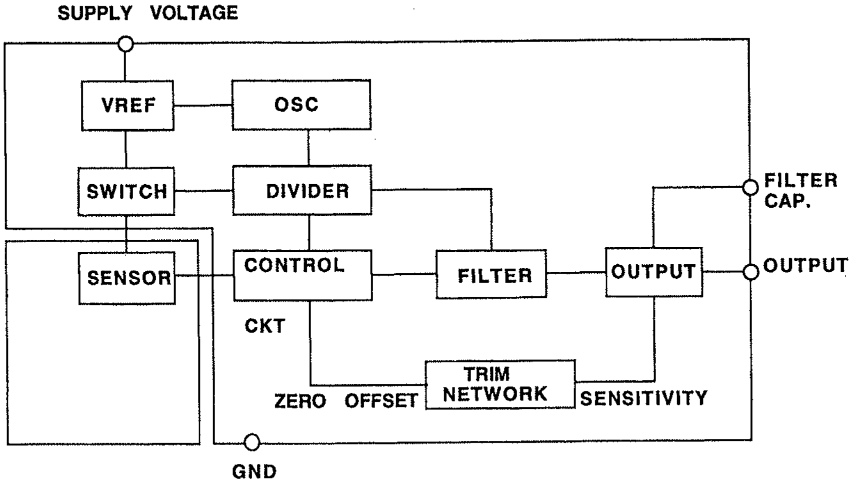


FIGURE 7.14 Signal-conditioning block diagram.

An alternate circuit with improved linearity is shown in Fig. 7.15. In this case, the output voltage is fed back to the input of the integrator, forming a bridge circuit. The feedback also sets the amplitude of the driving voltage across the sensing capacitors, so that it is proportional to their displacement. This also balances the electrostatic forces on the middle plate.

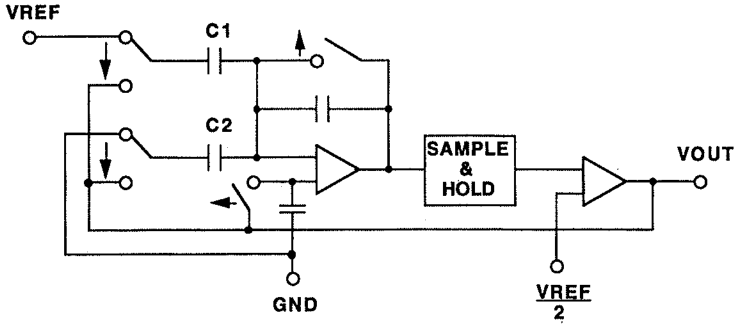


FIGURE 7.15 Linearized circuit schematic.

In this case

$$V_{\text{out}} = \frac{(C_1 - C_2)}{(C_1 + C_2)} \frac{V_{\text{REF}}}{2}, \tag{7.6}$$

where $C_1 \propto 1/d_1$
$C_2 \propto 1/d_2$
$d_1 + d_2 = K$ (constant)

so that

$$V_{\text{out}} = V_{\text{in}} \frac{(d_1 - d_2)}{K} \tag{7.7}$$

showing that, in this case, V_{out} is proportional to displacement and acceleration giving improved linearity (<1 percent nonlinearity).

Closed-Loop Sensing. An alternative to the open-loop sensing circuit is the closed-loop sensing circuit, which can be configured to give an analog or digital output. Figure 7.16*a* shows the balanced electrostatic forces exerted on a seismic mass by upper and lower capacitor plates, which are at voltages of $+V$ and $-V$. If the seismic mass experiences a force due to acceleration, and a voltage δV is applied to the middle plate to generate enough electrostatic force to counterbalance the acceleration force, then the forces are as shown in Fig. 7.16*b*.

That is

$$m \times a = \frac{C(V + \delta V)^2}{2d} - \frac{C(V - \delta V)^2}{2d} \tag{7.8}$$

from which

$$m \times a = 2\,C \times V \times \frac{\delta V}{d} \tag{7.9}$$

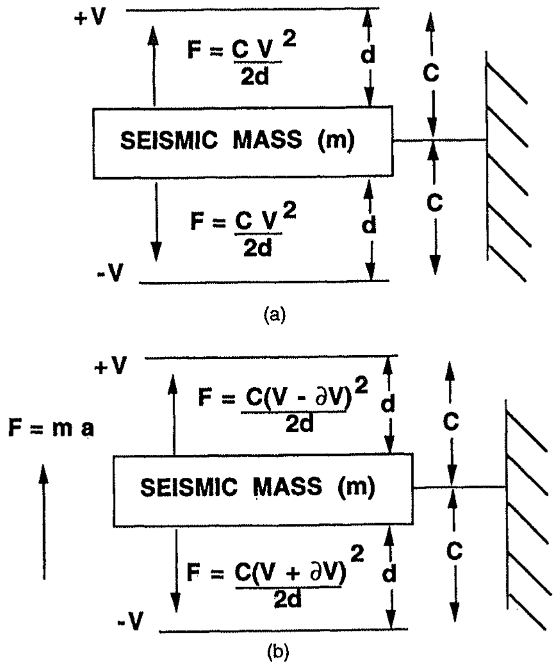


FIGURE 7.16 Electrostatic force diagram.

This shows that an acceleration force can be balanced by a linear voltage applied to the center plate. This voltage can be amplified to give a linear output voltage proportional to acceleration.[6]

In a micromachined structure, the electrostatic force produced by a δV of about 1 V can counterbalance the force produced by an acceleration of 50 g on the seismic mass. Figure 7.17 shows a block diagram of the analog closed-loop system. The top and bottom plates are dc-biased by the resistor divider network consisting of R1, R2, and R3. The ac antiphase signals used for sensing the position of the center plate are fed to the top and bottom sensor plates via the capacitors (C1, C2) from the control logic. The analog output voltage from the filter is feedback to the positive input of the integrator. When the integrator is clocked into the unity gain phase, the feedback voltage is applied to the center plate of the sensor. The electrostatic forces produced on the center plate by the differential voltage between the top and bottom plates, and the feedback voltage on the center plate, will force the center plate back to its normalized position as given in Eq. (7.9). Other methods that have been used for closed-loop operation are pulse width modulation (PWM), and delta sigma modulation (DSM).

The block diagram and waveforms of a PWM system are shown in Fig. 7.18. In this case, the center plate is held at a fixed voltage (0). The output of the integrator is amplified and converted into a PWM signal (VP), this signal and the inverted signal (VPN) are fed to the top and bottom plates of the sensor. The leading edges of the VP and VPN signals are used to sense the position of the middle plate. The electrostatic force generated by the voltage × time differential between the middle plate and the top and bottom plates, will act on the middle plate to counterbalance the forces on the plate due to acceleration. With zero acceleration, the PWM signal has a 50 percent duty cycle, so that the resulting electrostatic forces on the seismic mass are zero. The transfer function of the PWM system is given by

$$\text{Duty cycle} = \frac{\text{Acc} \times g \times d}{C \times V_{\text{REF}}{}^2} \tag{7.10}$$

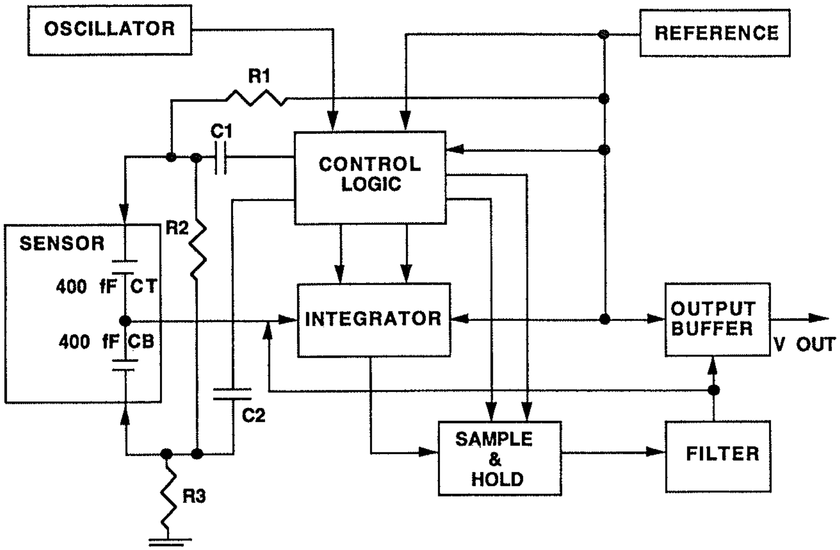


FIGURE 7.17 Block diagram of analog closed-loop system.

where g = gravitational constant
d = plate spacing
C = plate capacitance

The output can either be the PWM digital signal, or an analog output (obtained by feeding the PWM signal through a low pass filter).

Figure 7.19 shows the block diagram and waveforms of a DSM system. In this case, the middle plate is held at VF (volts). As can be seen, transient edges of the plate-drive waveforms are used for sensing the position of the center plate. The output from the integrator is fed to a comparator, whose output is then clocked into a latch where it sets up a "1" or a "0" (high or low) depending on the output from the integrator. The output from the latch is used to apply a voltage V_{REF} to the appropriate top or bottom capacitor plate, so that the electrostatic forces generated by the voltage V_{REF} – VF will maintain the center plate in its no-load position. The one-bit serial data stream from the latch can be fed directly to the MCU, or fed via a decimator circuit (which will convert the data into an 8-bit word) to the MCU. Alternatively, an LPF can be used to convert the data into an analog output. Bipolar or CMOS circuits can be used for signal conditioning. However, CMOS signal conditioning has the following advantages: a very high input impedance, good switching characteristics, low power requirements, small size, compatibility with MCU processes with the prospect of future integration, switched capacitor filters are available for noise reduction and bandwidth control, and EPROM technology is available for trimming (see Fig. 7.14). BiMOS circuits have also been used for signal conditioning.[6] In the BiMOS circuits, thin-film resistors are used to enable laser trimming of the zero offset and voltage reference; external capacitors and resistors are used for filtering and gain control.

Single vs. Multichip Control Circuits. The processing of sensors is not completely compatible with IC fabrication. Consequently, for the integration of sensors and ICs, a number of additional steps are required, which can have a detrimental effect on yields. The question then arises as to which is the most cost effective: a sensor die plus a control die with the additional cost of assembly, or a monolithic approach.

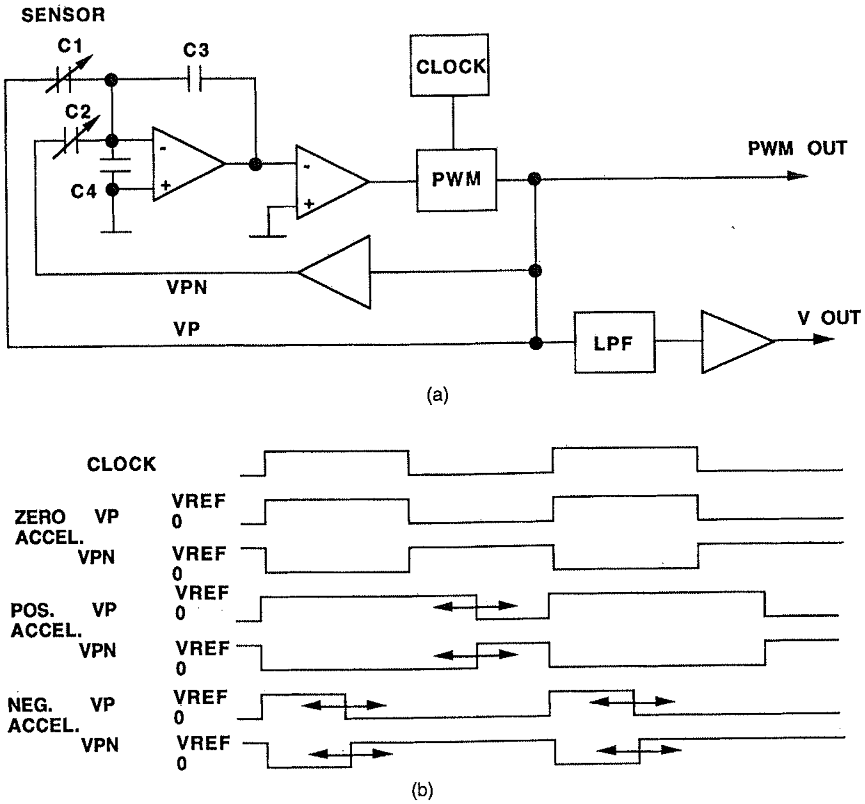


FIGURE 7.18 PWM block diagram.

In the case of micromachined devices, there are a number of advantages to using the dual-chip approach, such as flexibility, in that one type of sensor can be interfaced with a number of different types of control die, or several types of sensors can be interfaced with one type of control die. This provides a variety of input and output options. The control die and sensor can be developed simultaneously, minimizing development time. Problem solving is made easier, and the processing for both die can be optimized for performance. With the two-die approach, the sensor can also be capped and sealed during processing in a clean room atmosphere, thus eliminating contaminants and particles for good longevity. In the monolithic approach, this is not the case. In the monolithic approach, changes required to improve one section can affect the other section, which may then require additional changes in that section. The main disadvantage of the dual-chip approach is the introduction of parasitic capacitances. However, these can be addressed by existing control circuit design techniques.

7.5 AUTOMOTIVE APPLICATIONS FOR ACCELEROMETERS

Accelerometers have a wide variety of uses in the automobile. The initial application is as a crash sensor for air bag deployment. This application is normally associated with head-on collisions, but can also be applied to rear-end impact collisions to prevent rebound impact between the passengers and the windshield. An extension of this application is the use of

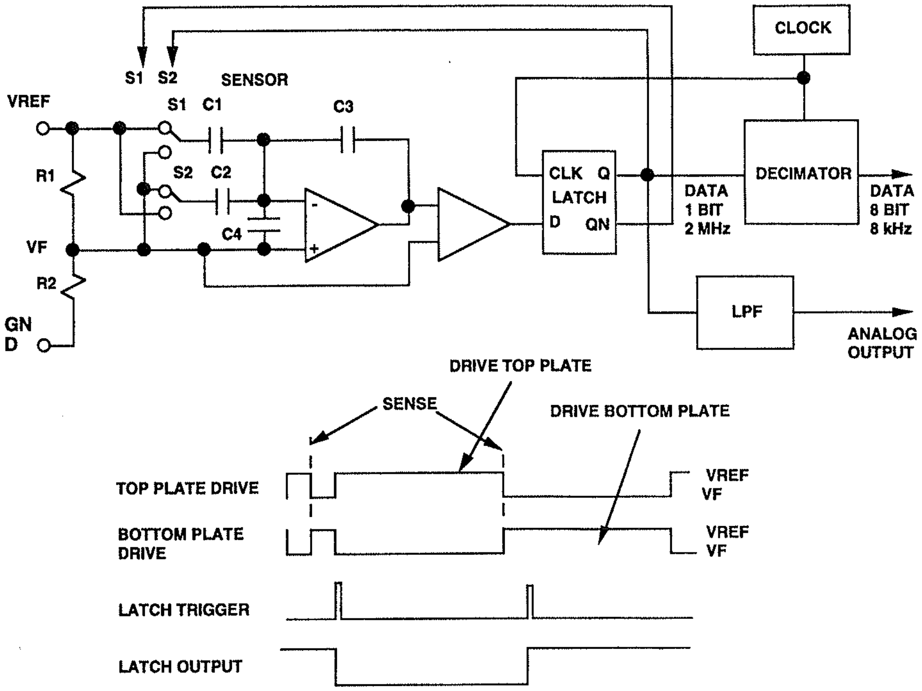


FIGURE 7.19 Block diagram of Delta Sigma Modulator.

accelerometers for the detection of side impact. This application will require additional air bags to the side of the occupants. Other low *g* linear accelerometers are being developed for ride control, ABS, traction, and inertial navigation applications.

Solid state acceleration sensors have special mounting requirements that are different from normal integrated circuits. These are to ensure that acceleration forces are transmitted to the sensor package. An advantage of the solid state device is self-test features for diagnostics. In acceleration applications, the devices are required to operate over the temperature range −40 to 85 °C (125 °C under the hood), and to withstand >2000 *g* shock. Other similar devices that have application in the automotive arena are vibration devices.

7.5.1 Air Bag Deployment Application

Crash sensors that use mechanical switches (sensors) are typically located some 40 cm from the point of impact, which necessitates the use of multiple sensors (normally 3 to 5 sensors are used in multipoint sensing) for crash sensing and air bag deployment. These devices are velocity change detectors, and are calibrated to make contact when the change of velocity in the passenger compartment is at least 20 km/h, this being the velocity change at which the front seat occupants will strike the windshield.

A centrally located analog sensor can be used as a crash sensor (single point). In the case of a centrally located accelerometer, the g level to be sensed is lower than that of a point-of-impact device. However, only one device is required to monitor the crash signature. This signature will vary with different types of chassis and different types of impacts. Consequently, an MCU is used to monitor the output of the accelerometer to determine if a crash has occurred. The typical output of a centrally located accelerometer during a 48 km/h crash is shown in Fig. 7.20. Deceleration of the vehicle and occupant displacement are also shown. At 48 km/h, the sensor has 20 ms to detect the crash and trigger the air bag. This results in inflation of the air bag 50 ms after impact, at which time the occupant has moved about 18 cm or approximately halfway to the windshield and at the contact point with the inflated air bag. During the initial 20 ms, deceleration can reach 20 *g*, but the average is about 5 *g* when the air bag is triggered. The centrally located accelerometer can take one of several forms: a piezoelectric sensor, a piezoresistive device, or a capacitive sensor.

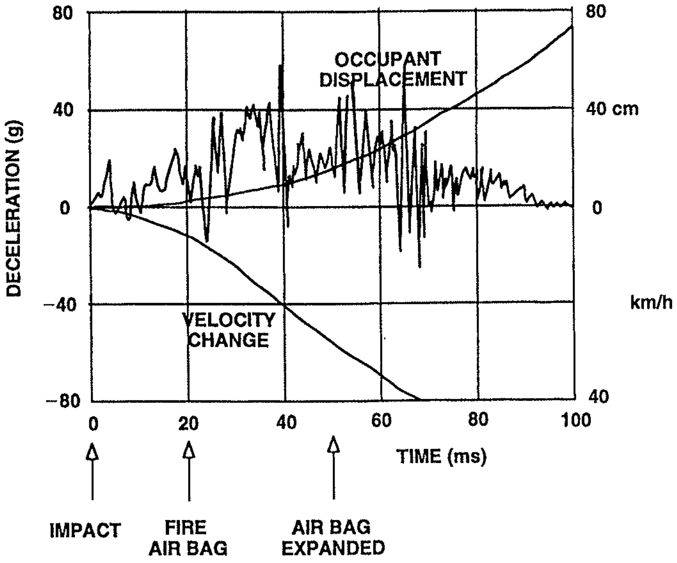


FIGURE 7.20 Typical 48 km/h crash waveform.

The centrally located accelerometer has a number of performance advantages over its mechanical counterpart. These are the reduction in the number of sensors and required bussing, which makes the centrally located system much more cost effective. There is an improvement in sensing and signal-processing accuracy with the single-point sensing accelerometer over the mechanical sensor. This gives a better-defined trigger point and overall improved performance across different chassis types. Capacitive sensors appear to have the edge in this application, because they have the potential of being cost effective, meet the requirements of the application, and have self-test features plus diagnostics available. In this application, a typical accelerometer specification is ±50 *g* full-scale output, accuracy ±5 percent over temperature, bandwidth dc to 750 Hz, and cross-axis sensitivity <3 percent. During impact, the crash sensor can also be used for seat belt locking.

7.5.2 Ride Control Application

In ride control systems, the leaf or coil springs located on the axles are replaced by four *wheel stations,* which form an active suspension. Each wheel station contains an oil-filled cylinder with a piston to set the distance of the frame above the axles and to isolate the frame from axle vibration. This is achieved using a servo feedback system. When a vehicle with conventional suspension encounters a foreign object on the highway, the load on the wheel increases as it moves up to negotiate the obstacle. This load increase makes the vehicle rise up. With a fully active suspension, the increase in load is detected and a servo valve is opened to transfer the necessary amount of oil from the appropriate cylinder to a storage container. Consequently, the load exerted on the chassis by each wheel is maintained at its specific level and the chassis remains at its static level. After the object has been traversed, oil is pumped back into the cylinder to reestablish the static load conditions.

An alternative to the active suspension is the adaptive suspension system. In this case, information from the front wheels is gathered and used to predict road conditions for the control of the rear wheels. The advantage over the fully active suspension is one of cost, as the number of acceleration sensors is halved. During cornering, oil is also pumped into the outside wheel cylinders to minimize roll angle.[7]

A combination of sensors is used for active suspension. These are accelerometers, wheel speed sensors, chassis-to-ground sensing, and piston-level sensing in the suspension system. The low g accelerometers used on the axles of the four wheels to detect the load changes on the wheels have the following specifications: ±2 *g* full-scale, accuracy ±5 percent over temperature, bandwidth dc to 10 Hz, and cross-axis sensitivity <3 percent. The acceleration information and data from the wheel speed sensors is used to provide the information necessary for the MCU to operate the servo control valves. Hall effect, MRE, and opto sensors have been used for monitoring the level of the pistons in the wheel stations cylinders.

7.5.3 Vibration Applications

Lean-burn engines are being developed for improved emission levels and for better fuel economy (10 to 15 percent improvement). NO_x emissions are greatly reduced to meet federal standards. Lean-burn engines use high stoichiometric ratios; 20:1 and higher are necessary. At these ratios, combustion becomes unstable and torque fluctuations large. Consequently, antiknock and vibration sensors are required to supply the information necessary to the MCU, so that it can adjust the injected fuel amount and ignition timing for stability over widely varying conditions.

There are two types of solid state sensors that can be used in this application: piezoelectric devices and capacitively coupled vibration sensors. A typical vibration sensor contains a number of fingers of varying length which vibrate at their resonant frequencies when those frequencies are encountered. The resonance is capacitively coupled to the sensing circuit, and the outputs as shown are obtained. Optical sensors have also been used as antiknock sensors. In this case, the ignition spectrum is monitored for the detection of misfiring or knocking. Vibration sensors can also be used for vibration monitoring in maintenance applications.

7.5.4 Antilock Brake System Applications

In antilock brake systems, speed sensors are attached to all wheels to determine wheel rotation speed and slip differential between wheels. VR devices, as well as Hall effect and MRE devices, can be used in this application, as zero speed sensing is not required. VR devices have been used and shown to be cost effective in this application, but Hall effect and MRE devices are now being designed into these systems. Pressure sensors are used to monitor brake fluid pressure, and an accelerometer or ground-speed sensor can be used to provide information on changes in the vehicular speed. Brake pedal position and brake fluid pressure information are also required for control. All of this information is fed to an MCU, which processes the data and adjusts the brake fluid pressure to each wheel for optimum braking. Many of the elements of the ABS system can be used for the detection of lateral slippage on high-speed cornering, and can be used for traction and the direction of power to the wheels. Traction control applies in particular to slippery surfaces and with four-wheel-drive vehicles. Additional information over that used in ABS systems is required by the MCU for ASR applications, such as engine speed and throttle angle. In this application, servo feedback to the throttle may also be necessary.

A more cost-effective, but less accurate, system for ABS and ASR is the adaptive control system in which accelerometers are normally used to measure deceleration when braking, and acceleration when the throttle is opened. If skidding occurs during braking, the brake pressure is reduced and adjusted for maximum deceleration, or the throttle adjusted for maximum traction. Typical specifications for the accelerometer required in this application are: ±1 g full-scale output, accuracy ±5 percent over temperature, bandwidth 0.5 to 50 Hz, and cross-axis sensitivity <3 percent.

7.6 NEW SENSING DEVICES

New cost-effective sensors are continually being developed. The technology and cost are often pushed by the application and volume requirements of the automotive industry and federal mandates. Today's silicon sensors and control electronics are limited in operating temperature to 150 °C to ensure long life of the devices. This operational temperature is adequate for most applications, but higher temperature operation may be required for sensors mounted in the engine compartment. The limit on the operating temperature of silicon devices can be extended to between 200 and 250 °C by the use of special isolation techniques such as dielectric isolation (this operating temperature applies to surface micromachined devices). For higher-temperature operation, alternative materials such as GaAs or SiC are being developed, but the cost of these devices limits their use at present.

A list of semiconductor conductor materials and maximum practical operating temperatures is given in Table 7.3. Higher operating temperatures have been reported but with poor longevity.

TABLE 7.3 Device Operating Temperatures

Material	Maximum practical operating temperature, °C
Si	150
Si (dielectric iso.)	250
GaAs	300
AlGaAs	350
GaP	400
SiC	500

7.6.1 New Rotational Speed-Sensing Devices

A number of new devices are being investigated to detect magnetic fields. These are flux gate, Weigand effect, magnetic transistor, and magnetic diode. The magnetic transistor at present is showing the most promise. The device operates on a similar principle to the Hall effect device. That is, the current division between split collectors (bipolar) or split drains (MOS) can be changed by a magnetic field. This current differential can then be detected and amplified to give an output voltage proportional to magnetic field strength. These devices can use either majority or minority carriers, and can be either vertical or lateral bipolar or MOS devices. The magnetic transistor has the potential of higher sensitivity than the Hall effect device.

Figure 7.21 shows the cross section of a lateral PNP magnetic transistor. The current from each collector is equal until a magnetic field is applied perpendicular to the surface of the device. The magnetic field causes an imbalance of current between the two collectors. Sensitivities with this type of structure have been reported as being an order of magnitude greater than in the Hall effect device.[8] Magnetic transistors and diodes can be directly integrated with the signal-conditioning circuits, which could make them very cost effective in future applications. A comparison of the magnetic transistor to other practical devices is given in Table 7.2.

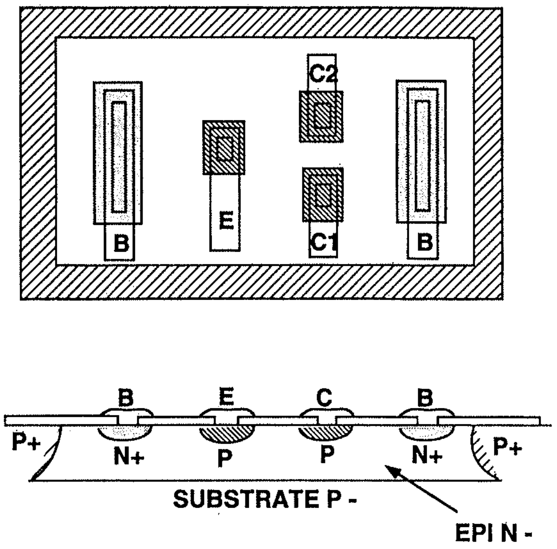


FIGURE 7.21 Cross section of a field-assisted PNP magnetic transistor.

7.6.2 New Linear Speed-Sensing Devices

A number of different sensing technologies can be used for distance, object detection, and approach speed measurements. Shown in Fig. 7.22 are the areas covered by blind-spot, rear, and forward-looking sensors. Ultrasonics, infrared, laser, and microwaves (radar) can be used in the detection of objects behind vehicles and in the blind areas. From a practical standpoint, no technology has come to the forefront. However, with new innovations in technology the situation may change very rapidly. Ultrasonics and infrared sensors are cost effective but degrade with inclement weather conditions such as ice, rain, snow, and the accumulation of road grime. Infrared devices are also color-sensitive, in that the sensitivity to shiny black objects is very low compared to other colors. Microwave devices appear to have the edge

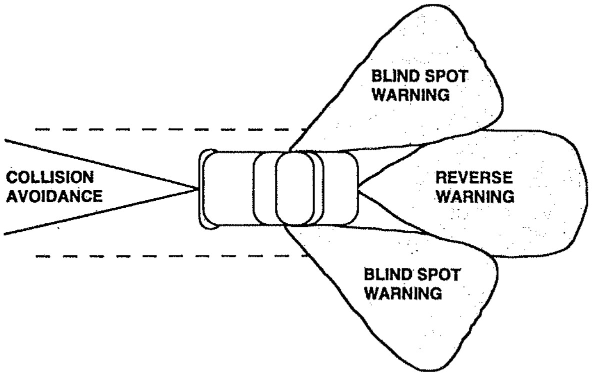


FIGURE 7.22 Collision-avoidance patterns.

when considering environmental conditions,[9] but are expensive, and radar can be affected by false return signals and clutter.

For the detection of obstacles and vehicles in front of a vehicle, the choice is between laser and microwaves due to the distances involved (up to 90 m). Microwaves have the disadvantage of high cost and large antenna size when considering available devices in the 60 GHz range. Frequencies greater than 100 GHz are preferred for acceptable antenna size. However, collision avoidance radar in the 76/77-GHz band has been developed in Europe. Collision avoidance radar in the 77, 94, and 144 GHz is being considered in the United States. A typical system uses a 38.5-GHz VCO (voltage controlled oscillator) with frequency-doubling to obtain about 40 mW of power at 77 GHz. A frequency-modulated continuous wave or pulse-modulated system can be used. The system uses GaAs devices to meet the frequency and power requirements. Lasers can be cost effective in this application, but also have their drawbacks: degradation of performance by fog, reflections from other light sources (sun, etc.), build-up of road grime on sensor surfaces, and poor reflecting surfaces at laser frequencies, such as grimy and shiny black surfaces.

7.6.3 New Inertial and Acceleration-Sensing Devices

Recent developments in solid state technology have made possible very small cost-effective devices to sense angular rotation. The implementation of one such gyroscopic device is shown in Fig. 7.23. This device is fabricated on a silicon substrate using surface micromachining techniques. In this case, three layers of polysilicon are used, with the first and third layers being fixed and the second layer free to vibrate about its center. The center is held in position by four spring arms attached to four mounting posts as shown. This device can sense rotation about the *X* and *Y* axes and sense acceleration in the direction of the *Z* axis. The center layer of polysilicon, driven by electrostatic forces, vibrates about the *Z* axis. These forces are produced by voltages applied between the fixed comb fingers and the comb fingers of the second polysilicon. Capacitor plates as shown are formed between the first and third polysilicon on the *X* and *Y* axes, and the second layer of polysilicon. Differential capacitive sensing techniques can then be used to sense any displacement of the vibrating disc caused by angular rotation. For example, if angular rotation takes place about the *X* axis, Coriolis forces produce a deflection of the disc about the *Y* axis. This deflection can be detected by the capacitor plates on the *X* axis. The sensing of the three functions is achieved by using a common sensing circuit that alternatively senses the *X* rotation, *Y* rotation, and acceleration. The gyroscope is designed to have a resolution of <10 degrees per hour for angular rate measurements, and an acceleration resolution of 0.5 mg.

7.7 *FUTURE APPLICATIONS*

New applications to increase creature comfort and safety are constantly being developed, but their rate of introduction will depend on the cost effectiveness of the technology, demand, and government mandates. Other concerns of automotive manufacturers are size, weight, power requirements, and adverse effects on styling and appearance. Many of the new sensor technologies are in their infancy, and thus are not yet cost effective on medium- and low-priced automobiles, but are being made available as options on luxury cars.

7.7.1 Future Rotational Speed-Sensing Applications

A future application for speed-sensing devices will be in continuously variable transmissions. In this application, engine and wheel speed, as well as torque, will be measured, and the infor-

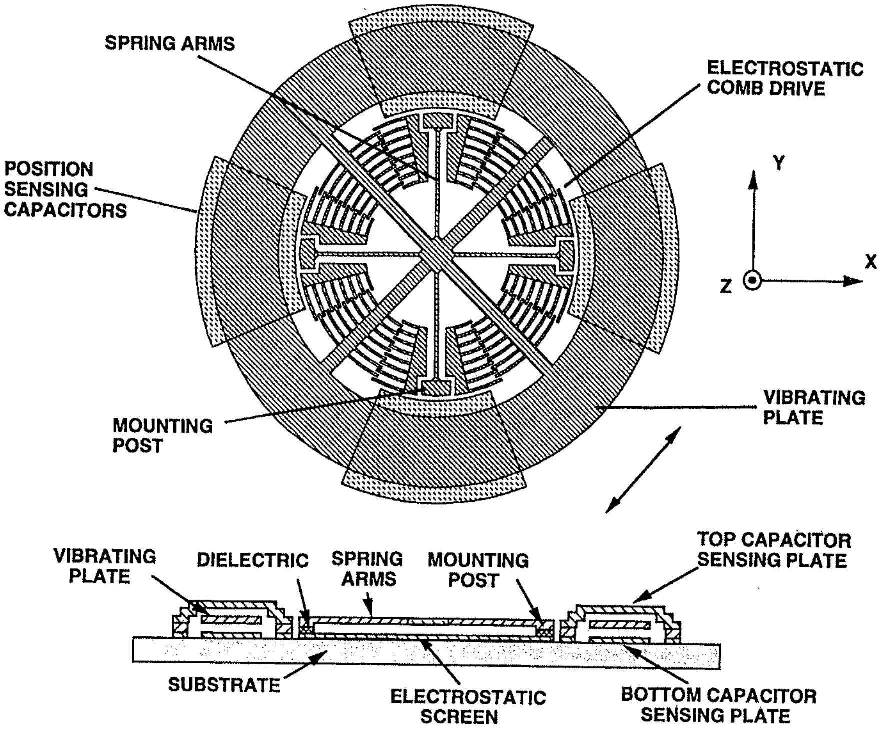


FIGURE 7.23 Solid state gyroscope.

mation processed by an MCU to optimize transmission ratios for engine performance and efficiency. All-wheel steering is also under development, and requires speed-sensing information, in addition to steering, front-, and rear-wheel angle position data for processing and control.

7.7.2 Future Linear Speed-Sensing Applications

Another application that has been developed is the use of speed- and distance-measuring devices for collision avoidance. These devices fall into three categories: near-obstacle detection (rear), blind-spot detection, and semiautomatic frontal object detection and control[9] (see Fig. 7.22).

Near-obstacle detection is used to prevent accidents during reversing. Blind-spot detection is used to prevent accidents due to careless lane changing, and when backing out of a driveway, garage, or alley into traffic. The semiautomatic frontal detection is a long-range system. The distance and closing speed between vehicles, or between a vehicle and a fixed object, can be measured and the speed adjusted as necessary to avoid a collision,[10] or the driver can be warned of impending danger. An addition to this is to monitor road surface conditions for friction—for example, dry roads compared to wet or icy roads—and also to use this information to adjust approach speeds and distance. Without collision avoidance, road condition monitoring can be used to caution vehicle operators. Collision avoidance systems can be used to minimize collisions, or can be used to operate protection systems before an unavoidable collision happens to protect the automobile passengers. In this case, vehicles closing at or approaching an obstacle at 80 km/h will be less than 7 m apart before a collision-is-imminent determination can be made. This gives 200 ms decision-making time for the system MCU. This is, however, long compared to today's air bag deployment systems, which have 20 ms decision-making time after the event. In the future, an idealistic system may be a combination of the two systems.

7.7.3 Future Acceleration Applications

One of the future applications being considered is the expansion of the air bag system to include side impact protection. The sensor used for crash sensing is unidirectional, so that it can only detect forward impact. A similar sensor, mounted perpendicular to the air bag sensor, can be added to the system to detect side impact and to deploy protection for the passengers. This device will typically require a 250-*g* accelerometer. Another application for accelerometers is to detect slippage during cornering in advanced steering systems. These systems will employ a low-*g* accelerometer (1–2 *g*).

7.7.4 Inertial Navigation Applications

A number of inertial navigation systems are being developed for short- and long-range travel. Long-range inertial navigation systems normally obtain their location by using a triangulation method. This method references three navigation satellites with known locations in fixed orbits. However, there are certain conditions under which contact with all three satellites is lost. This occurs when the vehicle is in the shadow of tall buildings or high hills, and triangulation is not possible. Under these conditions, the guidance system has to rely on such devices as gyroscopes, which sense angular rotation or change in direction, and/or monitor vehicular motion relative to the road.

Short-range inertial navigation systems or inertial measurement units (IMU) rely to a large extent on high-accuracy accelerometers and gyroscopes. A typical accelerometer specification for this application is: ±2-*g* full-scale output, accuracy 0.5 percent over temperature, bandwidth dc to 20 Hz, and a cross-axis sensitivity <0.5 percent. A centrally located IMU can be expanded to cover other applications such as suspension, ABS, ASR, and working with crash avoidance sensors. This may be the way to handle cost-effective system design in the future. The IMU can also be designed to provide location data for intelligent vehicle highway systems. These systems (Prometheus,[11] Amtics[12]) improve travel efficiency and reduce fuel consumption and pollution by selecting the optimum route to a given destination. The route is chosen to avoid traffic congestion, road construction, and accidents (see Chap. 29).

7.8 SUMMARY

In this chapter, a number of speed-sensing devices, both rotary and linear, have been described, together with potential applications. VR, MRE, Hall effect, and opto devices (possibly magnetic transistor in future applications) can be used in rotational applications for engine control, transmission, and wheel speed sensing. Of these devices, Hall effect, VR, and opto have been widely used. With the tendency for direct pickoff, optical devices may become impractical. MRE devices are being designed in and will become a serious contender to the Hall effect device. In linear applications for crash avoidance, microwave devices have the edge over performance and optical devices in terms of cost. However, as the cost of microwave devices declines, they could become cost effective. For blind-area alert and reversing obstacle detection, ultrasonics and infrared devices are cost effective, but performance degrades during inclement weather.

The accelerometer has possibly the greatest potential for applications in the automobile. These applications range from crash sensing, ride control, ABS, and ASR to IMU systems. Accelerometers needed will range in sensitivities from 50 *g* in crash sensing to 1 *g* in the IMU. Advances in technology are providing a number of new sensors that are showing potential, such as the magneto transistor and the micromachined gyroscope. To summarize, Figs. 7.24 and 7.25 show the types of sensors used in specific applications, and the technologies used for specific sensors. As can be seen, one type of sensor can be used in a number of applications. In applications where a sensor output is shared, care must be taken to ensure that a failure in one system does not disable the sensor or other systems. Because of the similarities in several of the systems and the use of shared sensors, the greatest potential for cost-effective system design is a single control system. The IMU shows great potential to be the controller for ride control, ABS, ASR, four-wheel-drive, and steering applications. The rate of introduction of new sensors and systems will depend on federal mandates, customer demand, and the need to improve engine fuel efficiency and to reduce emissions.

GLOSSARY

Adaptive suspension A suspension system that monitors motion of the front wheels and adjusts the suspension of the rear axle accordingly.

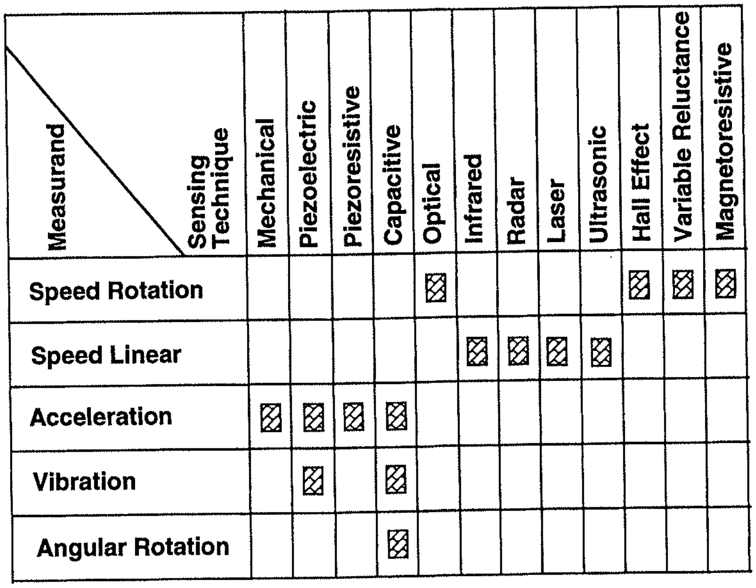
Arntics Acronym for Advanced Mobile Traffic Information and Communication System.

ASR (traction) A system to prevent wheel spin on slippery surfaces, to give maximum traction and acceleration.

Sensor \ Application	Air Bag Deployment	Ride Control	ABS	Engine Control	Transmission	4 Wheel Drive	All Wheel Stearing	Engine Vibrarion	Cruise Control	Seat Belts	Power Steering Assist	Traction (ASR)	Collision Avoidance	Obstacle Detection	Inertial Navigation
Speed Rotation	▨	▨	▨	▨	▨	▨	▨		▨		▨	▨			
Speed Linear			□		□				□		□	□	□	□	□
Acceleration	▨	▨	▨	▨						▨		▨			▨
Vibration								▨							
Angular Rotation															▨

▨ USED □ MAY BE USED

FIGURE 7.24 Sensor applications.



Measurand \ Sensing Technique	Mechanical	Piezoelectric	Piezoresistive	Capacitive	Optical	Infrared	Radar	Laser	Ultrasonic	Hall Effect	Variable Reluctance	Magnetoresistive
Speed Rotation					▨					▨	▨	▨
Speed Linear						▨	▨	▨	▨			
Acceleration	▨	▨	▨	▨								
Vibration		▨		▨								
Angular Rotation				▨								

FIGURE 7.25 Sensor technologies.

Coriolis forces Forces exerted by a spinning body to oppose any motion at right angles to the axle.

Crash signature Shock waveform projected through a chassis during a collision.

Inertial navigation Guidance system giving accurate location.

IMU (inertial measurement unit) System used for guidance between two locations indicates road hazards and delays.

Lean burn Engine with high compression ratios and high air-to-fuel ratios for increased efficiency and low emissions.

Micromachining Manufacturing technology for micromechanical structures using chemical etching techniques.

NO_x Chemical symbol for oxides of nitrogen.

Prometheus Acronym for PROgraM for a European Traffic with Highest Efficiency and Unprecedented Safety.

Switched capacitor filter Technique for switching capacitors to simulate high-value resistors for low-frequency filters to minimize size.

REFERENCES

1. Osamu Ina, Yoshimi Yoshino, and Makio Lida, "Recent Intelligent Sensor Technology in Japan," S.A.E. paper 891709, 1989.
2. D. E. Bergfried, Mattes, B., and Rutz, R., "Electronic Crash Sensors for Restraint Systems," *Proceedings of the International Congress on Transportation Electronics,* Detroit, Oct. 1990, pp. 169–177.

3. E. Peeters, Vergote, S., Puers, B., and Sansen, W., "A Highly Symmetrical Capacitive Microaccelerometer with Single Degree of Freedom Response," *Tranducers 91,* 1991, pp. 97–103.
4. J. T. Suminto, "A Simple High Performance Piezoresistive Accelerometer," *Transducers 91,* 1991, pp. 104–107.
5. J. C. Cole, "A New Capacitive Technology for Low-Cost Accelerometer Applications," *Sensors Expo. International,* 1989.
6. T. A. Core, Tsang, W. K., and Sherman, S. J., "Fabrication Technology for an Integrated Surface-Micromachined Sensor," *Solid State Technology,* Oct. 1993, pp. 39–47.
7. H. Wallentowitz, "Scope for the Integration of Powertrain and Chassis Control Systems: Traction Control—All Wheel Drive—Active Suspension," *Proceedings of the International Conference on Transportation Electronics,* S.A.E. paper 901168, 1990.
8. H. Kaneko, Muro, H., and French, P. J., "Optimization of Bipolar Magneto-Transistors," *Micro Systems Technologies 90,* 1990, pp. 599–604.
9. L. Raffaelli, Stewart, E., Borelli, J., and Quimby, R., "Monolithic Components for 77 GHz Automotive Collision Avoidance Radars," *Proceedings Sensors Expo,* 1993, pp. 261–268.
10. S. Aono, "Electronic Applications for Enhancing Automotive Safety," *Proceedings of the International Conference on Transportation Electronics,* SAE 901137, 1990, pp. 179–186.
11. J. Hellaker, "Prometheus-Strategy," *Proceedings of the International Congress Transportation Electronics,* SAE 901139, 1990, pp. 195–200.
12. H. Okamoto, and Hase, M., "The Progress of Amtics-Advanced Mobile Traffic Information and Communication System," *Proceedings of the International Congress Transportation Electronics,* SAE 901142, 1990, pp. 217–224.

ABOUT THE AUTHOR

William C. Dunn is a member of the technical staff in the Advanced Custom Technologies group at Motorola's Semiconductor Product Sector in Phoenix, Arizona. He has over 30 years' experience in circuit design and systems engineering. For the past 15 years he has been involved in the development of semiconductor sensors, smart power devices, and control systems for the automotive market. Prior to joining Motorola, he worked for several large corporations in the United Kingdom and United States, has written over 30 papers, and has over 30 patents issued and pending on mechanical sensor structures, semiconductor technology, and circuit design.

CHAPTER 11
AUTOMOTIVE MICROCONTROLLERS

David S. Boehmer
Senior Applications Engineer
Intel Corporation

A microcontroller can be found at the heart of almost any automotive electronic control module or ECU in production today. Automotive systems such as antilock braking control (ABS), engine control, navigation, and vehicle dynamics all incorporate at least one microcontroller within their ECU to perform necessary control functions. Understanding the various features and offerings of microcontrollers that are available on the market today is important when making a selection for an application. This chapter is intended to provide a look at various microcontroller features and provide some insight into their characteristics from an automotive application point of view.

11.1 MICROCONTROLLER ARCHITECTURE AND PERFORMANCE CHARACTERISTICS

A microcontroller can essentially be thought of as a single-chip computer system and is often referred to as a single-chip microcomputer. It detects and processes input signals, and responds by asserting output signals to the rest of the ECU. Fabricated upon this highly integrated, single piece of silicon are all of the features necessary to perform embedded control functions. Microcontrollers are fabricated by many manufacturers and are offered in just about any imaginable mix of memory, I/O, and peripheral sets. The user customizes the operation of the microcontroller by programming it with his or her own unique program. The program configures the microcontroller to detect external events, manipulate the collected data, and respond with appropriate output. The user's program is commonly referred to as code and typically resides on-chip in either ROM or EPROM. In some cases where an excessive amount of code space is required, memory may exist off-chip on a separate piece of silicon. After power-up, a microcontroller executes the user's code and performs the desired embedded control function.

Microcontrollers differ from microprocessors in several ways. Microcontrollers can be thought of as a complete microcomputer on a chip that integrates a CPU with memory and various peripherals such as analog-to-digital converters (A/D), serial communication units (SIO, SSIO), high-speed input and output units (HSIO, EPA, PWM), timer/counter units, and standard low-speed input/output ports (LSIO). Microcontrollers are designed to be embedded within event-driven control applications and generally have all necessary peripherals integrated onto the same piece of silicon. Microcontrollers are utilized in applications ranging from automotive ABS to household appliances in which the microcontroller's function is predefined and limited user interface is required.

Microprocessors, on the other hand, typically require external peripheral devices to perform their intended function and are not suited to be utilized in single-chip designs. Microprocessors basically consist of a CPU with register arrays and interrupt handlers. Peripherals such as A/D and HSIO are rarely integrated onto microprocessor silicon. Microprocessors are designed to process large quantities of data and have the capability to handle large amounts of external memory. Although microprocessors are typically utilized in applications which are much more human-interface and I/O intensive such as personal computers and office workstations, they are beginning to find their way into embedded applications.

Choosing a microcontroller for an application is a process that takes careful investigation and thought. Items such as memory size, frequency, bus size, I/O requirements, and temperature range are all basic requirements that must be considered when choosing a microcontroller. The microcontroller family must possess the performance capability necessary to successfully accomplish the intended task. The family should also provide a memory, I/O, and frequency growth path that allows easy upgradability to meet market demands. Additionally, the microcontroller must meet the application's thermal requirements in order to guarantee functionality over the intended operating temperature range. Items such as these must all be considered when choosing a microcontroller for an automotive application.

11.1.1 Block Diagram

Usually the first item a designer will see when opening a microcontroller data book or data sheet is a block diagram. A block diagram provides a high-level pictorial representation of a microcontroller and depicts the various peripherals, I/O, and memory functions the microcontroller has to offer. The block diagram gives the designer a quick indication if the particular microcontroller will meet the basic memory, I/O, and peripheral needs of their application. Figure 11.1 shows a block diagram for a state-of-the-art microcontroller. It depicts 32 Kbytes

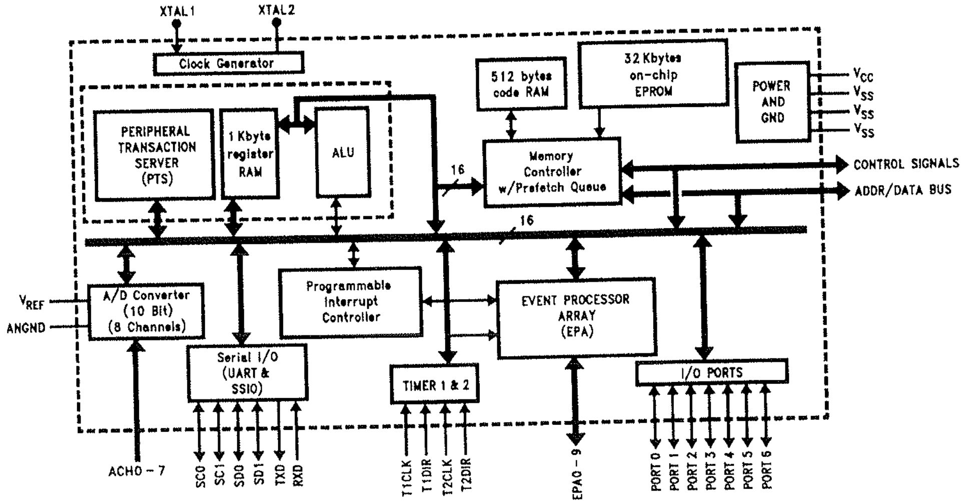


FIGURE 11.1 Microcontroller block diagram.

of EPROM, 1 Kbyte of register RAM, 6 I/O ports, an A-to-D converter, 2 timers, high-speed input/output (I/O) channels, as well as many other peripherals. These features may be "excessive" to a designer looking for a microcontroller to implement in an automotive trip-computer application but would be excellently suited for automotive ABS/traction control or engine control.

11.1.2 Pin-Out Diagram

A microcontroller's pin-out diagram is used to specify the functions assigned to pins relative to their position on a given package. An example pin-out diagram is shown in Fig. 11.2. Note that most pins have multiple functions assigned to them. Pins that can support more than one function are referred to as multifunction pins. The default function for multifunction pins is normally that of low-speed input and output (discussed later in this chapter). If the user should wish to select the secondary or special function associated with the pin, he or she can do so by writing to the appropriate special function register. There are some exceptions. A good example is pins used for interfacing to external memory. If the device is instructed to power-up executing from external memory as opposed to on-chip memory, the address data bus and associated control pins will revert to their special function as opposed to low-speed I/O.

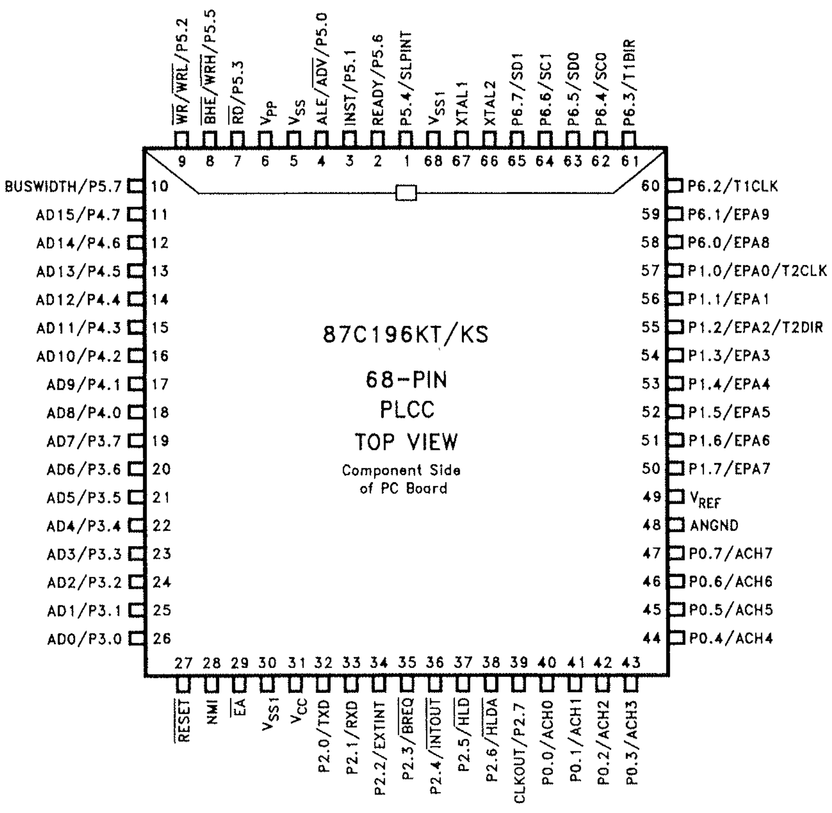


FIGURE 11.2 Microcontroller pin-out diagram.

11.1.3 Central Processing Unit

The central processing unit or CPU can be thought of as the brain of a microcontroller. The CPU is the circuitry within a microcontroller where instructions are executed and decisions are made. Mathematical calculations, data processing, and control signal generation all take place within the CPU. Major components of the CPU include the arithmetic logic unit (ALU), register file, instruction register, and a microcode engine. The CPU is connected to the bus controller and other peripherals via a bidirectional data bus.

Microcontrollers are, for the most part, digital devices. As digital devices, microcontrollers utilize a binary numbering system with a base of 2. Binary data digits or *bits* are expressed as either a logic "1" (boolean value of true) or a logic "0" (boolean value of false). In a 5-V system, a logic "1" may be simply defined as a +5-V state and a logic "0" may be defined as a 0-V state. A bit is a single memory or register location that can contain either a logic "1" or a logic "0" state. Bits of data can be arranged as a *nibble* (4 bits of data), a *byte* (8 bits of data), or as a *word* (16 bits of data). It should also be noted that, in some instances, a word may be defined as the data width that a given microcontroller can recognize at a time, be it 8 bits or 16 bits. For purposes of this chapter, we will refer to a word as being 16 bits. Data can also be expressed as a double word which is an unsigned 32-bit variable with a value between 0 and 4,294,967,295. Most architectures support this data only for shifts, dividends of a 32-by-16 divide, or for the product of a 16-by-16 multiply.

The most common way of referring to a microcontroller is by the width of its CPU. This indicates the width of data that the CPU can process at a time. A microcontroller with a CPU that can process 8 bits of data at a time is referred to as an 8-bit microcontroller. A microcontroller with a CPU that can process 16 bits of data at a time is referred to as a 16-bit microcontroller. With this in mind, it is easy to see why 16-bit microcontrollers offer higher performance than their 8-bit counterparts. Figure 11.3 illustrates a typical 16-bit CPU dia-

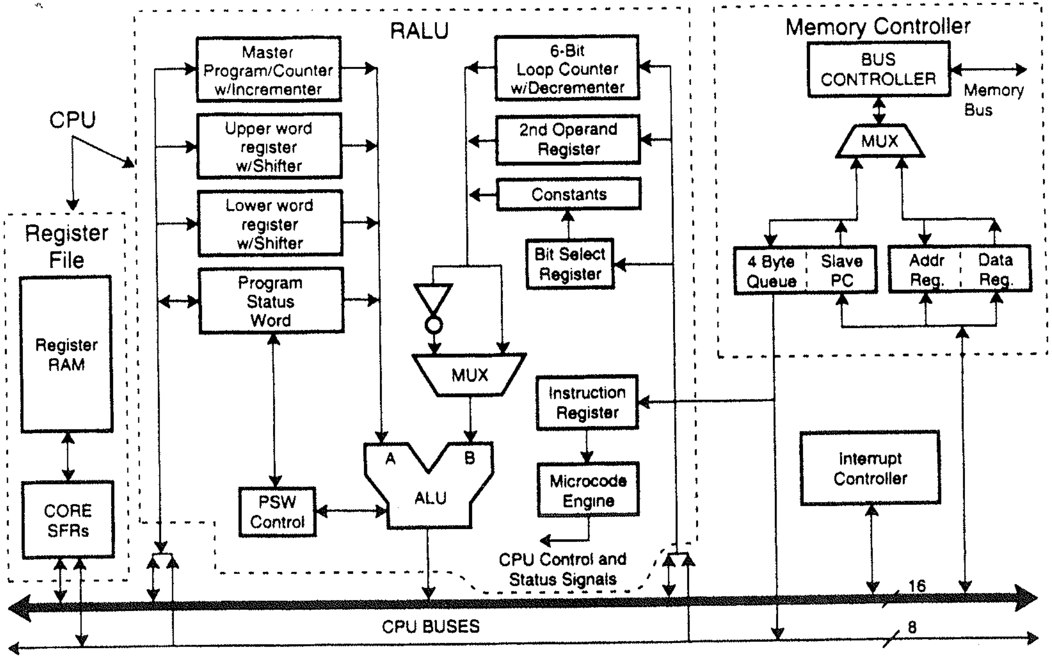


FIGURE 11.3 16-bit CPU.

gram. The microcode engine controls the CPU. Instructions to the CPU are taken from the instruction queue and temporarily stored in the instruction register. This queue is often referred to as a *prefetch queue* and it decreases execution time by staging instructions to be executed. The microcode engine then decodes the instructions and generates the correct sequence of events to have the ALU perform the desired function(s).

Arithmetic Logic Unit. The ALU is the portion of the CPU that performs most mathematical and logic operations. After an instruction is decoded by the microcode engine, the data specified by the instruction is loaded into the ALU for processing. The ALU then processes the data as specified by the instruction.

Register File. The register file consists of memory locations that are used as temporary storage locations while the user's code is executing. The register file is implemented as RAM and consists of both RAM memory locations and special function registers (SFRs). RAM memory locations are used as temporary data storage during execution of the user's code. After power-up, RAM memory locations default to a logic "0" and data in SFR locations contain default values as specified by the microcontroller manufacturer.

Special Function Registers. SFRs allow the user to configure and monitor various peripherals and functions of the microcontroller. By writing specific data to an SFR, the users can configure the microcontroller to meet the exact needs of their application. Figure 11.4 shows an example of a serial port SFR used for configuration. Note that each bit location within the SFR determines a specific function and can be programmed to either a logic "1" or "0". If more than two configuration choices are possible, two or more bits will be combined to produce the multiple choices. An example of this would be the mode bits (M1 and M2) in the example SFR (Fig. 11.4). Bit locations marked "RSV" are reserved and should be written to with a value as indicated by the manufacturer.

SP_CON (1FBBH)

7	6	5	4	3	2	1	0
0	0	PAR	TB8	REN	PEN	M2	M1

M2, M1	Mode	Function
	00	Mode 0: Synchronous
	01	Mode 1: Standard asynchronous
	10	Mode 2: Asynchronous (receiver interrupt on 9th bit = 1)*
	11	Mode 3: Asynchronous (9th bit = parity or data)**

PEN Parity Enable. Enables the Parity function for Mode 1 or Mode 3; cannot be enabled for Mode 2.

REN Receiver Enable. Enables the receiver to write to SBUF_RX.

TB8 Transmission Bit 8. Set the ninth data bit for transmission (Modes 2 and 3). Cleared after each transmission; not valid if parity is enabled.

PAR*** 0 = even parity
1 = odd parity

Bits 6, 7 Reserved; write as zeros for future product compatibility.

* Mode 2: Asynchronous (receiver: interrupt on 9th bit = 1; transmitter: 9th bit = TB8)
** Mode 3: Asynchronous (receiver: always interrupt on 9th bit; transmitter: 9th bit = parity for PEN = 1
*** Par bit only available on 8XC196KT and KS devices. 9th bit = TB8 for PEN = 0
For 8XC196KR, JR, KQ, JQ devices, this bit should be written as a zero
to maintain compatibility with future devices.

FIGURE 11.4 Special function control register example.

SP_STAT (1FB9H)

7	6	5	4	3	2	1	0
RB8/RPE	RI	TI	FE	TXE	OE	X	X

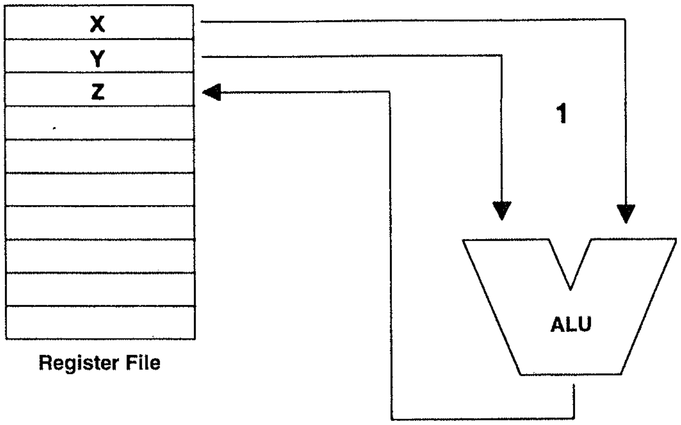
Bits 0, 1 Reserved; ignore data.
OE Set on buffer overrun error.
TXE Set on transmitter empty. When set may write 2 bytes to transmit buffer.
FE Framing error; set it no STOP bit is found at the end of a reception. When set may write 1 byte to transmit buffer.
TI Transmit interrupt; set at the beginning of the STOP bit transmission.
RI Receive interrupt; set after the last data bit is received.
RPE (Parity enabled) Receive parity error (Modes 1 and 3 only); set if parity is enabled and a parity error occurred.
RB8 (Parity disabled) Received Bit 8 (Modes 2 and 3 only); set if the 9th bit is high on reception.

FIGURE 11.5 Special function status register example.

Some SFRs can be read by the user to determine the current status of a given peripheral. Figure 11.5 shows an example of a serial port status register that, when read, indicates the current status of the microcontroller's serial port. Note that each bit location corresponds to a particular state of the serial port. Bit locations marked "RSV" are reserved and should be ignored when read.

Register Direct vs. Accumulator-Based Architectures. Microcontroller architectures can be classified as either the register-direct or accumulator-based type. These terms refer to the means by which the CPU must handle data when performing mathematical, logical, or storage operations.

Register-direct architectures allow the programmer to essentially use most, if not all, of the microcontroller's entire RAM array as individual accumulators. That is, the programmer can perform mathematical or storage operations directly upon any of the RAM locations. This simplifies task switching because program variables may be left in their assigned registers while servicing interrupts. Figure 11.6 illustrates a register-to-register type architecture (such



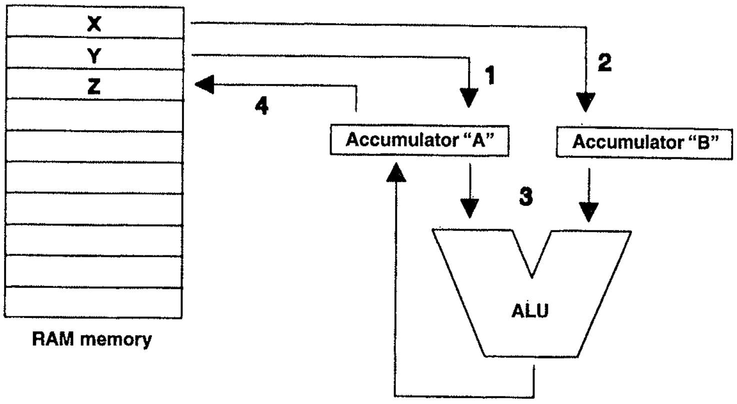
Z=X+Y

1. MUL Z,X,Y

FIGURE 11.6 Register-to-register architecture example.

as Intel's MCS®-96). This architecture essentially has 232 "accumulators" (more are available through a windowing mechanism) of which any can be operated on directly by the RALU. The true advantage of this type of architecture is that it reduces accumulator bottleneck and speeds throughput during program execution.

Accumulator-based architectures require the user to first store the data to be manipulated into a temporary storage location, referred to as an "accumulator," prior to performing any type of data operation. After the operation is completed, the user program must then store the result to the desired destination location. Figure 11.7 depicts an example of an accumulator-based architecture.



```
Z=X+Y
1. LD A,Y
2. LD B,X
3. MUL A,B
4. ST A,Z
```

FIGURE 11.7 Accumulator-based architecture.

Program Counter. The Program Counter (PC) controls the sequencing of instructions to be executed. The PC is a 16-bit register located within the CPU which holds the address of the next instruction to be executed. After an instruction is fetched, the PC is automatically incremented to point to the next instruction.

Stack and Stack Pointer. The stack is an area of memory (typically user-assigned) that is used to store data temporarily in a FILO (first-in, last-out) fashion. The stack is primarily used for storing program information (such as the program counter or interrupt mask registers) when an interrupt service routine is invoked. It is also sometimes used to pass variables between subroutines. The stack is typically accessed through PUSH and POP instructions. Execution of the PUSH instruction "pushes" the contents of the spec-

SP starting address (SP+12):	
(SP+10):	80FFh
(SP+8):	A5A5h
(SP+6):	6E20h
(SP+4):	5555h
(SP+2):	0000h
SP ending address (SP):	8000h

FIGURE 11.8 Stack pointer example.

POP instruction "pops" the contents of the specified operand off of the stack.The stack pointer (SP) is a register which points to the next available word location on the stack. Consider the example shown in Fig. 11.8 which shows the contents of the stack after the following code sequence is executed:

PUSH #80FFh	pushes immediate data 80FFh onto stack
PUSH #0A5A5h	pushes immediate data A5A5h onto stack
PUSH 82h	pushes data @ 82h (assume it's 6E20h) onto stack
PUSH #5555h	pushes immediate data 5555h onto stack
PUSH 4Eh	pushes data @ 4Eh (assume it's 0000h) onto stack
PUSH #8000h	pushes immediate data 8000h onto stack

Continuing with the preceding example, if a POP instruction were executed, the data at the current SP address (SP) would be "popped" off the stack and stored to the address specified by the instruction's operand. Executing the POP instruction results in the SP being incremented by 2.

Program Status Word and Flags. The program status word (PSW) is a collection of boolean flags which retain information concerning the state of the user's program. These flags are set or cleared depending upon the result obtained after executing certain instructions as specified by the microcontroller manufacturer. PSW flags are not directly accessible by the user's program; access is typically through instructions which test one or more of the flags to determine proper program flow. Following is a summary of common PSW flags as supported by Intel's MCS-96® architecture:

Z: The *Zero* flag is set when an operation generates a result equal to zero. The Z flag is never set by the add-with-carry (ADDC/ADDCB) or subtract-with-carry (SUBC/SUBCB) instructions, but is cleared if the result is nonzero. These two instructions are normally used in conjunction with ADD/ADDB and SUB/SUBB instructions to perform multiple-precision arithmetic. The operation of the Z flag for these instructions leaves it indicating the proper result for the entire multiple-precision calculation.

N: The *Negative* flag is set when an operation generates a negative result. Note that the N flag will be in the algebraically correct state even if overflow occurs. For shift operations, the N flag is set to the same value as the most significant bit of the result.

V: The *oVerflow* flag is set when an operation generates a result that is outside the range for the destination data type. For shift-left instructions, the V flag is set if the most significant bit of the operand changes at any time during the shift. For an unsigned word divide, the V flag is set if the quotient is greater than 65,535. For a signed word divide, the V flag is set if the quotient is less than –32,768 or greater than 32,767.

VT: The *oVerflow Trap* flag is set when the V flag is set, but it is only cleared by instructions which are specially designated to clear the VT flag (such as CLRVT, JVT, and JNVT). The VT flag allows for testing possible overflow conditions at the end of a sequence of related arithmetic operations. This is normally more efficient than testing the V flag after each instruction.

C: The *Carry* flag is set to indicate either (1) the state of the arithmetic carry from the most significant bit of the ALU for an arithmetic operation or (2) the state of the last bit shifted out of an operand for a shift. Arithmetic borrow after a subtract operation is the complement of the C flag (i.e., if the operation generated a borrow, then C = 0).

ST: The *STicky* bit flag is set to indicate that, during a right shift, a 1 has been shifted first into the C flag and then shifted out. The ST flag can be used along with the C flag to control rounding after a right shift. Imprecise rounding can be a major source of error in a numerical calculation; use of both the C and ST flags can increase accuracy as described in the following paragraphs

Consider multiplying two 8-bit quantities and then scaling the result down to 12 bits:

```
MULUB    AX, CL, DL    (CL * DL = AX)
SHR      AX, #4        (AX is shifted right by 4 bits)
```

If the C flag is set after the shift, it indicates that the bits shifted off the end of the operand were greater than or equal to one-half the least significant bit of the 12-bit result. If the C flag is cleared after the shift, it indicates that the bits shifted off the end of the operand were less than half the LSB of the 12-bit result. Without the ST flag, the rounding decision must be made on the basis of the C flag alone. (Normally the result would be rounded up if the C flag is set.) The ST flag allows a finer resolution in the rounding decision as shown here:

C	ST	Bits shifted off
0	0	Value = 0
0	1	0 < Value < ½ LSB
1	0	Value = ½ LSB
1	1	Value > ½ LSB

Jump instructions are the most common instructions to utilize PSW flags for determining the operation to perform. Instructions that test PSW flags are very useful when program flow needs to be altered dependent upon the outcome of an arithmetic operation. The most common example of this would be for program loops that are to be executed a certain number of times. Following are examples of several MCS-96 instructions whose operation is dependent upon the state of one or more program status word flags:

JC (Jump if C flag is set.) If the C (carry) bit is set, the program will jump to the address location specified by the operand. If the C flag is cleared, control will pass to the next sequential instruction.

JGT (Jump if signed greater than.) If both the N (negative) and the Z (zero) flags are clear, the program will jump to the address location specified by the operand. If either of the flags is set, control will pass to the next sequential instruction.

JLE (Jump if signed less than or equal.) If either the N or Z flags is set, the program will jump to the address location specified by the operand. If both the N and Z flags are cleared, control will pass to the next sequential instruction.

11.1.4 Bus Controller

The bus controller serves as the interface between the CPU and the internal program memory and the external memory spaces. The bus controller maintains a queue (commonly called the prefetch queue) of prefetched instruction bytes and responds to CPU requests for data memory references. The prefetch queue decreases execution time by staging instructions to be executed. The capacities of prefetch queues vary but for the MCS-96 architecture, it is 4 bytes deep.

When using a logic analyzer to debug code it is important to consider the effects of the prefetch queue. It is not possible to accurately determine when an instruction will execute by simply watching when it is fetched from external memory. This is because the prefetch queue is filled in advance of instruction execution. It is also important to consider the effects when a jump or branch occurs during program execution. When the program sequence changes because of a jump, interrupt, call, or return, the PC is loaded with the new address, the queue is flushed and processing continues. Consider the situation in which the external address/data bus is being monitored when a program branch occurs. Because of the prefetch queue, it will appear as if instructions past the branch point were executed, when in fact they were only loaded into the prefetch queue.

11.1.5 Frequency of Operation

Microcontrollers are being offered in an ever-increasing range of operating frequencies. Most high-end automotive applications currently use microcontrollers operating in the 12- to 20-MHz range, with 24 MHz becoming not so uncommon. Microcontrollers with frequencies as high as 30 and 32 MHz are available as prototypes and will soon be available for production. Operating frequency becomes especially important when a microcontroller must perform high-speed event control such as required in ABS braking and engine control. Applications such as these typically have to detect, calculate, and respond to external events within a given amount of time. In ABS applications, this time is commonly referred to as loop time and defines the amount of time that the microcontroller has to execute the main loop of the software algorithm to achieve optimal performance.

Operating frequency can be directly related to the speed at which a microcontroller will execute the user's code. For instance, let's look at how long it takes for a particular microcontroller to execute the following generic subroutine. For this example, consider the execution times rather than the operations each instruction is performing.

```
 {6}  PUSHF
 {3}  NOTB   PTS_COUNT_EPA1
 {5}  ADDB   NUM_OF_PULSES_1, PTS_COUNT_EPA1, #00h
 {5}  SUB    INV_SPEED_1, FTIME_1, ITIME_1
{27}  DIV    INV_SPEED_1, NUM_OF_PULSES_1
 {5}  LD     Templ+2, #Speed_high_constant
 {5}  LD     Templ, #Speed_low_constant
{27}  DIV    Templ, INV_SPEED_1
 {4}  ST     Templ, EPA1_FREQ
{11}  RET
```

In this example, the numbers in brackets {} denote how many state times it will take the microcontroller to execute the given line of code. A state time is the basic time measurement for all microcontroller operations. For this MCS-96 family microcontroller, a state time is based on the crystal frequency divided by two. A state time for other microcontrollers may be based upon the crystal frequency divided by three. For this particular microcontroller, a state time can be calculated by the following formula (other microcontroller families use similar formulas):

$$1 \text{ state time} = 1[(\text{frequency of operation})/2]$$

Applying this formula, 1 state time = 125 ns when operating at 16 MHz, and 167 ns when operating at 12 MHz. The example code sequence takes the microcontroller 98 state times to execute. This equates to 16.37 μs to execute at an operating frequency of 12 MHz. At 16 MHz, it takes only 12.25 μs for the microcontroller to execute the subroutine. An operating frequency of 16 MHz results in the microcontroller executing approximately 34 percent more instructions in a given time than at a frequency of 12 MHz.

Another consideration when choosing an operating frequency is the clocking resolution of on-chip timer/counters. The maximum clocking rate of on-chip timer/counters is limited by the frequency the microcontroller is being clocked at. As an example, if an on-chip timer/counter is set up to increment/decrement at a rate determined by CLOCK/4, this would result in 333 ns resolution at 12 MHz. However, if the clock speed were increased to 16 MHz, a higher and more desirable resolution of 250 ns is achieved.

11.1.6 Instruction Set

An often overlooked feature that gives a microcontroller the capability to perform desired operations and manipulate data is its instruction set. A microcontroller's instruction set consists of a set of unique commands which the programmer uses to instruct the microcontroller on what operation to perform.

An *instruction* is a binary command which is recognized by the CPU as a request to perform a certain operation. Examples of typically supported operations are loads, moves, and stores which transfer data from one memory location to another. There are also jumps and branches which are used to alter program flow. Arithmetic instructions include various multiples, divides, subtracts, additions, increments, and decrements. Instructions such as ANDs, ORs, XORs, shifts, and so forth, allow the user to perform logical operations upon data. In addition to these basic instructions, microcontrollers often support specialized instructions unique to their architecture or intended application.

Instructions can be divided into two parts, the *opcode* and *operand.* The opcode (sometimes referred to as the machine instruction) specifies the operation to take place and the operand specifies the data to be operated upon. Instructions typically consist of either 0, 1, 2, or 3 operands to support various operations. As an example, consider the following MCS-96 architecture instructions:

PUSHF (0 operands) is an instruction that pushes the program status word (PSW) onto the stack. Since this instruction operates on a predefined location, no operand is necessary.

Format: PUSHF

PUSH (1 operand) is an instruction that pushes the specified word operand onto the stack.

Format: PUSH (SRC)

ADD (2 operands) adds two words together and places the result in the destination (leftmost) operand location.

Format: ADD (DST),(SRC)

ADD (3 operands) adds two words together as the 2-operand ADD instruction, but in this case, a third operand is specified as the destination.

Format: ADD (DEST),(SRC1),(SRC2)

Instructions support one or more of six basic addressing types to access operands within the address space of the microcontroller. If programmers wish to take full advantage of a microcontroller architecture, it is important that they fully understand the details of the supported addressing types. The six basic types of addressing modes are termed register-direct, indirect, indirect with autoincrement, immediate, short-indexed, and long-indexed. The following descriptions describe these modes as they are handled by hardware in register-to-register architectures.

The *register-direct* addressing mode is used to directly access registers within the lower 256 bytes of the on-chip register file. The register is selected by an 8-bit field within the instruction and the register address must conform to the operand type's alignment rules. Depending upon the instruction, typically up to three registers can take part in the calculation.

Examples:

ADD AX,BX,CX	AX = BX + CX
MUL AX,BX	AX = AX*BX
INCB CL	CL = CL + 1

The *indirect addressing* mode accesses a word in the lower register file containing the 16-bit operand address. The indirect address can refer to an operand anywhere within the address space of the microcontroller. The register containing the indirect address is selected by an 8-bit field within the instruction. An instruction may contain only one indirect reference; the remaining operands (if any) must be register-direct references.

Examples:

LD BX,[AX]	BX = mem_word(AX)

In this example, assume that before execution:

contents of AX = 2FC2h
contents of 2FC2h = 3F26h

Then after execution,

contents of BX = 3F26h

ADDB AL,BL,[CX]	AL = BL + mem_byte(CX)

The *indirect with autoincrement* addressing mode is the same as the indirect mode except that the variable that contains the indirect address is autoincremented after it is used to address the operand. If the instruction operates on bytes or short integers, the indirect address variable is incremented by one; if it operates on words or integers, the indirect address will be incremented by two.

Examples:

LD BX,[AX]+	BX = mem_word(AX)
	AX = AX + 2

ADDB AL,BL,[CX]+	AL = BL + mem_byte(CX)
	CX = CX + 1

For the *immediate addressing* mode, an operand itself is in a field in the instruction. An instruction may contain only one immediate reference; the remaining operand(s) must be register-direct references.

Example:

ADD AX,#340	AX = AX + 340 (decimal)

For the *short-indexed addressing* mode, an 8-bit field in the instruction selects a word variable (which is contained in square brackets) in the lower register file that contains an address. A second 8-bit field in the instruction stream is sign-extended and summed with the word variable to form an operand address.

Since the 8-bit field is sign-extended, the effective address can be up to 128 bytes before the address in the word variable and up to 127 bytes after it. An instruction may contain only one short-indexed reference; the remaining operand(s) must be register-direct references.

Example:

LD AX,4[BX]	AX = mem_word(BX + 4)

In this example, assume that before execution:

contents of BX = A152h

The operand address is then A152h + 04h = A156h

The *long-indexed addressing* mode is like the short-indexed mode except that a 16-bit field is taken from the instruction and added to the word variable to form the operand. No sign extension is necessary. An instruction may contain only one long-indexed reference and the remaining operand(s) must be register-direct references.

Examples:

ST AX,TABLE[BX]	mem_word(TABLE + BX) = AX
AND AX,BX,TABLE[CX]	AX = BX and mem_word(TABLE + CX)

11.1.7 Programming Languages

The two most common types of programming languages in use today for automotive microcontrollers are *assembly languages* and *high-level languages* (HLLs). Program development begins with the user writing code in either an assembly language or an HLL. This code is written as a text file and is referred to as a source file. The source file is then assembled or compiled using the appropriate assembler/compiler program. The assembler translates the source code into object code and creates what is referred to as an object file. The object file contains machine language instructions and data that can be loaded into an evaluation tool for debugging and validation. The object can also be converted into a hex file for EPROM programming or ROM mask generation as discussed later in this chapter. The program development process is illustrated in Fig. 11.9.

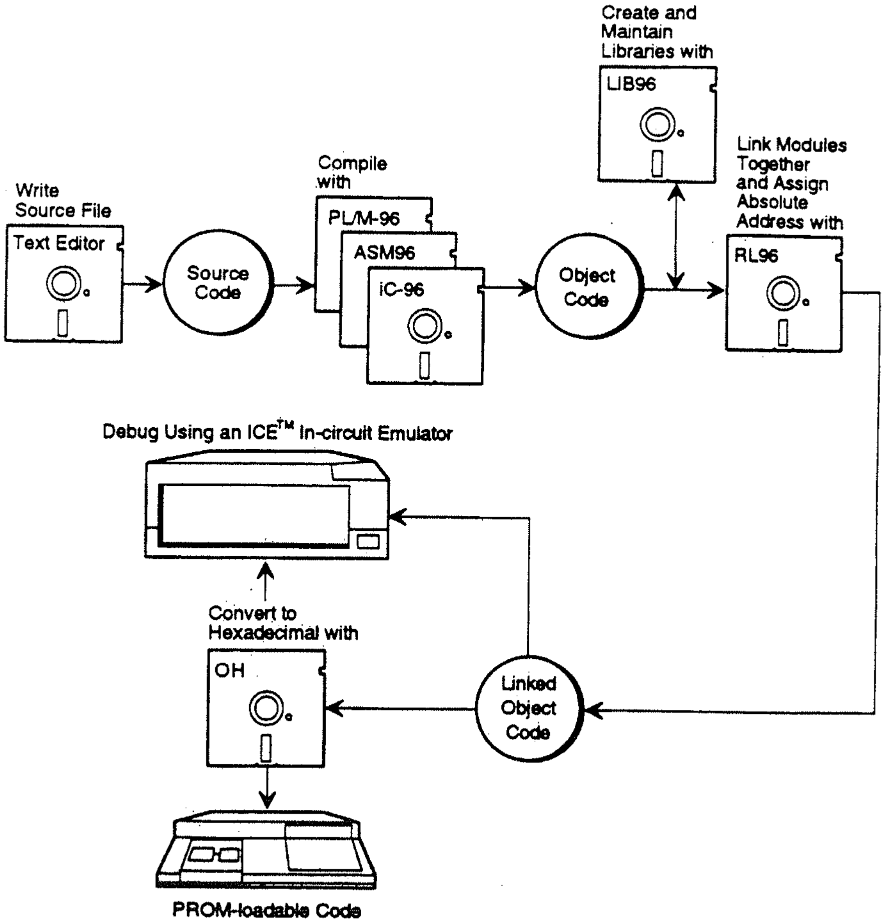
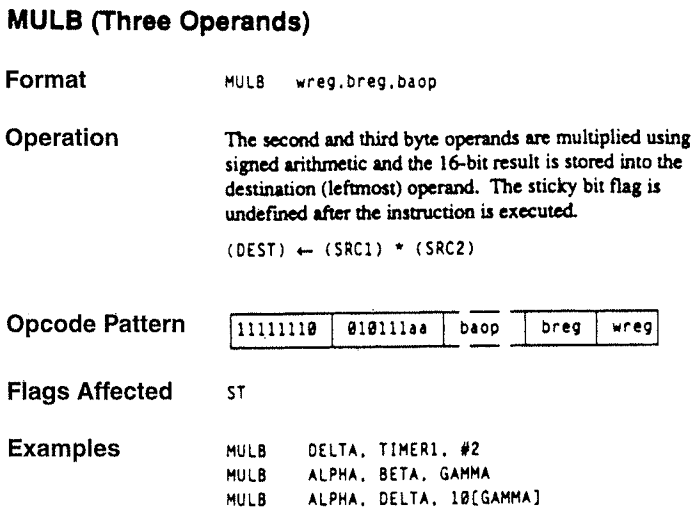

Create and Maintain Libraries with
LIB96
Link Modules Together and Assign Absolute Address with
RL96
Compile with
PL/M-96
ASM96
iC-96
Write Source File
Text Editor
Source Code
Object Code
Debug Using an ICE™ In-circuit Emulator
Convert to Hexadecimal with
OH
Linked Object Code
PROM-loadable Code

FIGURE 11.9 The program development process.

Assembly Language Programming. An assembly language is a low-level programming language that is specific to a given microcontroller family. Assemblers translate language operation codes (mnemonics) directly into machine instructions that instruct the microcontroller on what operation to perform. Because the programmer is essentially using the microcontroller's machine code to write assembly language programs, more precise control of the device can be achieved through the direct manipulation of individual bits within registers. Because of their efficiency, assembly language programs require less code space than high-level languages. Assembly language programs consist of three parts: machine instructions, assembler directives, and assembler controls.

A *machine instruction* is a machine code that can be executed by the microcontroller's CPU. The collection of machine instructions that a particular microcontroller can execute is referred to as its instruction set. An example of a machine instruction is the opcode for the MULB instruction (Fig. 11.10) from Intel's MCS-96 assembly language. MULB is the mnemonic that represents the machine instruction which performs the specified multiplication operation. When executed by the microcontroller, the MULB opcode results in the multiplication of the two byte operands with the result being placed in a word destination location.



MULB (Three Operands)

Format MULB wreg,breg,baop

Operation The second and third byte operands are multiplied using signed arithmetic and the 16-bit result is stored into the destination (leftmost) operand. The sticky bit flag is undefined after the instruction is executed.

(DEST) ← (SRC1) * (SRC2)

Opcode Pattern

11111110	010111aa	baop	breg	wreg

Flags Affected ST

Examples

```
MULB    DELTA, TIMER1, #2
MULB    ALPHA, BETA, GAMMA
MULB    ALPHA, DELTA, 10[GAMMA]
```

FIGURE 11.10 Machine instruction example: MULB.

Assembler directives allow the user to specify auxiliary information (such as storage reservation, location counter control, definition of nonexecutable code, object code relocation, and flow of assembler processing) that determines the manner in which the assembler generates object code from the user's source file input.

Assembler controls set the mode of operation for the assembler and direct the flow of the assembly process. Assembler controls can be classified into primary controls and general controls. Primary controls are set at the beginning of the assembly process and cannot be changed during the assembly. Primary controls allow the user to specify items such as print options, page lengths and widths, error messages, and cross-referencing. General controls can be specified in the invocation line or on control lines anywhere in the source file and can appear any number of times in the program. General controls either cause an immediate action or an immediate change of conditions in which the condition specified remains in effect until another general control causes it to change.

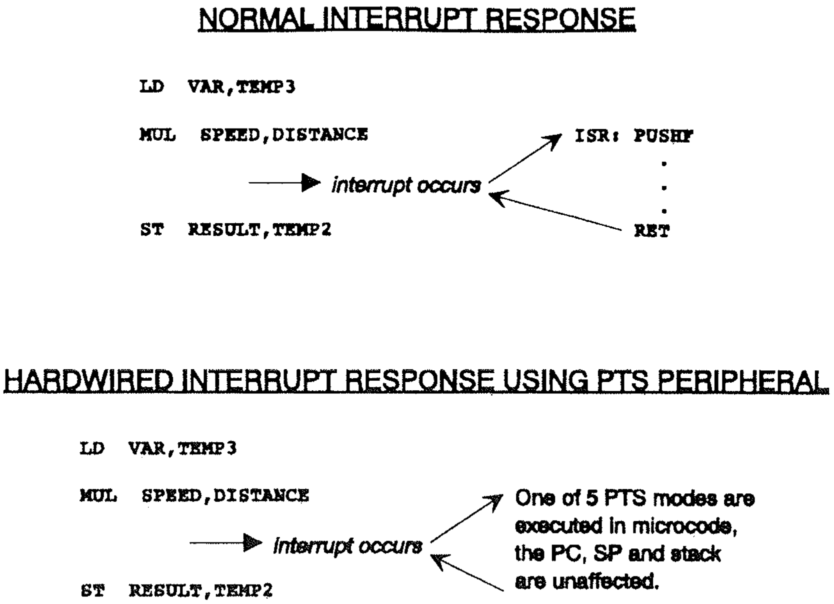
High-level Language Programming. Unlike low-level languages (such as assembly languages), a high-level language is a general purpose language that can support numerous microcontroller architectures. The most common high-level language used for automotive applications is C. C programs are written with statements rather than specific instructions from a microcontroller's instruction set. High-level languages utilize a software program known as a *compiler* to translate the user's source code into the specific microcontroller's machine language. Each microcontroller family has its own unique compiler to support selected high-level languages. Although high-level languages tend to be less efficient than assembly languages, their advantage lies in ease of writing code and better debugging capability. The use of statements as opposed to specific instructions better suits high-level languages toward control of procedures (to implement complex software algorithms) as opposed to the microcontroller itself.

11.1.8 Interrupt Structure

The interrupt structure is one of the more important features of an automotive microcontroller. Applications such as automotive ABS and engine control can be referred to as event-driven control systems. Event-driven control systems require that normal code execution be halted to allow a higher-priority task or event to take place. These higher-priority tasks are known as interrupts and can initiate a change in the program flow to execute a specialized routine. When an interrupt occurs, instead of executing the next instruction, the CPU branches to an interrupt service routine (ISR). The branch can occur in response to a request from an on-chip peripheral, an external signal, or an instruction. In the simplest case, the microcontroller receives the request, performs the desired operation and returns to the task that was interrupted.

ISRs are typically serviced via software but it is becoming common for microcontroller manufacturers to implement special on-chip hardware ISR functions for commonly performed operations. These ISRs are typically microcoded or *hardwired* into the microcontroller as described later in this section.

Software, or Normal, Servicing of Interrupts. The software servicing of interrupts is fairly straightforward as shown in Fig. 11.11. When an interrupt source is enabled by the user and a


NORMAL INTERRUPT RESPONSE
interrupt occurs
HARDWIRED INTERRUPT RESPONSE USING PTS PERIPHERAL
interrupt occurs
One of 5 PTS modes are executed in microcode, the PC, SP and stack are unaffected.

```
LD   VAR,TEMP3
MUL  SPEED,DISTANCE        ISR: PUSHF
                                 .
                                 .
                                 .
ST   RESULT,TEMP2               RET
```

```
LD   VAR,TEMP3
MUL  SPEED,DISTANCE
ST   RESULT,TEMP2
```

FIGURE 11.11 Comparison of normal interrupts and hardwired interrupts.

valid interrupt event occurs, the CPU will fetch the starting address of the ISR from the interrupt vector table. The interrupt vector table is a dedicated section of memory that contains the user-programmed start address of the various ISRs. After fetching the ISR address, the CPU automatically pushes the current program counter (PC) onto the stack and loads the PC with the ISR beginning address. This results in the program flow vectoring to the ISR address. The user-programmed ISR is then executed. The last instruction within the ISR is a return instruction that pops the old PC off the stack. This results in program flow continuing from where it was interrupted.

Interrupt mask registers allow the user to prevent or *mask* undesirable interrupts from occurring during various sections of the program. This is a very desirable feature and allows for custom tailoring of the interrupt structure to meet the needs of a particular application. Enabling or disabling of all interrupts (known as globally enabling/disabling) is typically supported with a software instruction such as DI (globally disable all interrupts) or EI (globally enable all interrupts).

Hardware, or Microcoded, Interrupt Structures. Hardware interrupt structures differ from software interrupts in that the user doesn't have to provide the ISR to be executed when the interrupt occurs. With a hardware interrupt structure, the ISR is predefined by being hardwired or *microcoded* into the microcontroller. This is advantageous because it requires less code space and requires less CPU overhead. Stack operations are not necessary since interrupt vectors do not have to be fetched. Most microcontroller manufacturers have their own proprietary solution for hardware ISR's, which are all somewhat similar to one another. For purposes of this section, we will briefly describe the peripheral transaction server as implemented on members of Intel's MCS-96 family of microcontrollers.

The PTS provides a microcoded hardware interrupt handler which can be used in place of a normal ISR. The PTS requires much less overhead than a normal ISR since it operates without modification of the stack. Any interrupt source can be selected by the user to trigger a PTS interrupt in place of a normal ISR. The PTS is similar to a direct memory access (DMA) controller in that when a PTS interrupt, or *cycle,* occurs, data is automatically moved from one location of memory to another as specified by the user. Figure 11.11 compares a regular ISR to a PTS interrupt cycle.

The PTS allows for five modes of operation; single-byte transfer, multiple-byte transfer, PWM, PWM toggle, and A/D scan mode. Each mode is configurable through an 8-byte, user-defined PTS control block (PTSCB) located in RAM. The user may enable virtually any normal interrupt source to be serviced by a PTS interrupt by simply writing to the appropriate bit in an SFR known as the PTS_SELECT register. When a PTS interrupt is enabled and the event occurs, a microcoded interrupt service routine executes in which the contents of the PTSCB are read to determine the specific operation to be performed. More details on the PTSCB can be found in the application example found in this section.

The major advantage of the PTS for automotive applications is its fast response time. The PTS is ideally suited for transferring single or multiple bytes/words of data in response to an interrupt. An example of this is the serial port example which will be described shortly. Another example of the usefulness of the PTS (using A/D scan mode) would be if the user wanted to automatically store A/D conversion results every time a conversion completed within a user-defined scan of A/D channels. The PTS could also be configured to automatically transfer a block of data between memory locations every time an interrupt occurs.

Application Example of PTS Single-Byte Transfer Mode. This example shows how the PTS can be used to automatically transmit and receive 8-byte messages over the serial port. Data to be transmitted and received data are stored in separate tables. The use of the PTS for this purpose greatly reduces CPU overhead and code-space requirements. The layout of the user-defined PTSCB for single-byte transfer mode is shown in Fig. 11.12. PTS_DEST within the PTSCB contains the destination address for the data transfer and PTS_SOURCE contains the source address for the transfer.

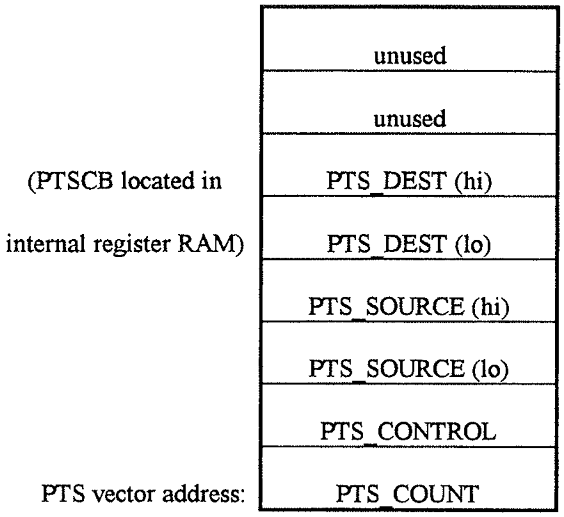


FIGURE 11.12 PTS control block for single-byte transfer mode.

Two PTSCBs are set up for this example, one in response to receive (RX) interrupts and one in response to transmit (TX) interrupts. The RX PTSCB's PTS_DEST is initialized with the start address of the receive data table and the TX PTSCB's PTS_DEST is initialized with the address of the serial port's transmit buffer.

PTS_CONTROL is a byte that specifies the PTS operation to be performed. Its layout is shown in Fig. 11.13.

PTS_COUNT is a down counter that is used to keep track of how many PTS interrupts or cycles have occurred since the last initialization. PTS_COUNT is initialized by the user to any value below 256 and is decremented everytime the corresponding PTS cycle occurs. It is often used to keep track of how many pieces of data have been transferred. In this example, PTS_COUNT is used to determine when a complete 8-byte message has been transmitted or received. After PTS_COUNT expires, an "end-of-PTS" or "normal" ISR occurs, in which the user utilizes the data as required by the application. When an interrupt source is enabled by the user to be a PTS interrupt, the following sequence of events occurs every time the corresponding interrupt occurs:

1. Instead of a normal interrupt, the user has selected it to do a PTS cycle.
2. The microcoded PTS routine fetches the PTS_CONTROL byte from the PTSCB whose start address is specified by the user in the PTS interrupt vector table. The microcoded PTS routine then:

   reads data to be transferred from address specified by PTS_SOURCE

   writes the data to address specified by PTS_DEST

   optionally increments/updates PTS_SOURCE and PTS_DEST addresses

   decrements PTS_COUNT
3. When PTS_COUNT reaches "0", an end of PTS interrupt occurs and the normal ISR is executed in which the user utilizes the received data as necessary (for RX interrupts) or reloads the transmit table with new data (for TX interrupts).

Interrupt Latency. Interrupt latency is defined as the time from when the interrupt event occurs (not when it is acknowledged) to when the microcontroller begins executing the first

PTS_CONTROL BYTE (for single and multiple-byte transfers)

7	6	5	4	3	2	1	0
M2	M1	M0	B/W	SU	DU	SI	DI

M2, M1, M0:	Mode	Function
	000	PTS Block Transfer
	100	PTS Single Transfer

B/W:	Byte/Word:	"0" = Word; "1" = Byte
SU:	Source Update:	"1" = update source address
DU:	Destination Update:	"1" = update destination address
SI:	Increment Source:	"1" = Increment Source address
DI:	Increment Destination:	"1" = Increment Destination address

FIGURE 11.13 PTS control byte for single- and multiple-byte transfer modes.

instruction of the interrupt service routine. Interrupt latency must be carefully considered in timing-critical code as is found in many automotive applications.

There is a delay between an interrupt's triggering and its acknowledgment. An interrupt is not acknowledged until the currently executing instruction is finished. Further, if the interrupt signal does not occur at least some specified (assume four for this discussion) state times before the end of the current instruction, the interrupt may not be acknowledged until after the next instruction has been executed. This is because an instruction is fetched and prepared for execution a few state times before it is actually executed. Thus, the maximum delay between interrupt generation and its acknowledgment is approximately four state times plus the execution time of the next instruction.

It should also be noted that most microcontrollers have protected instructions (such as RETURN, PUSH, POP) which inhibit interrupt acknowledgment until after the following instruction is executed. These instructions can increase interrupt-to-acknowledgment delay.

When an interrupt is acknowledged, the interrupt pending bit is cleared and a call is forced to the location indicated by the corresponding interrupt vector. This call occurs after the completion of the current instruction, except as noted previously. For the MCS-96 architecture, the procedure of fetching the interrupt vector and forcing the call requires 16 state times. The stack being located in external memory will add an additional two state times to this number.

Latency is the time from when an interrupt is generated (not acknowledged) until the microcontroller begins executing interrupt code. The maximum latency occurs when an interrupt occurs too late for acknowledgment following the current instruction. The worst case is calculated assuming that the current instruction is not a protected one. The worst-case latency is the sum of three terms:

1. The time for the current instruction to finish (assume four state times).
2. The state times required for the next instruction. This time is basically the time it takes to execute the longest instruction used in the user's code (assume it's a 16-state DIV instruction).
3. The response time (assume 16 states, 18 for an externally located stack).

Thus, for this scenario, the maximum delay would be 4 + 16 + 16 = 36 state times. This equates to approximately 4.5 μs for a MCS-96 microcontroller operating at 16 MHz. This latency can increase or decrease depending upon the longest execution-time instruction used. Figure 11.14 illustrates an example of this worst-case scenario.

Interrupt latency can be reduced by carefully selecting instructions in areas of code where interrupts are expected. Using a protected instruction followed immediately by a long instruction increases the maximum latency because an interrupt cannot occur after the protected instruction.

11.1.9 Fabrication Processes

The basic fabrication processes that are widely used for automotive microcontrollers today are NMOS (N-channel metal-oxide semiconductor) and CMOS (complementary MOS). The scope of this chapter does not allow for an in-depth discussion of these processes, although a brief description of the structures used to build on-chip circuitry will be discussed. These terms refer to the components used in the construction of MOSFET (MOS field effect transistor) inverters which are the basis of logic on digital devices. NMOS inverters are constructed of N-channel transistors only, whereas CMOS inverters are constructed of both N-channel and P-channel transistors. This section will describe the basic operation of each inverter along with its pros and cons.

Simply stated, a P-channel transistor conducts when a logic "0" is applied to its gate. Conversely, N-channel transistors conduct when a logic "1" is applied to their gate. Figures 11.15 and 11.16 show a simplified cross-sectional view and the electrical symbol for N- and P-channel devices, respectively.

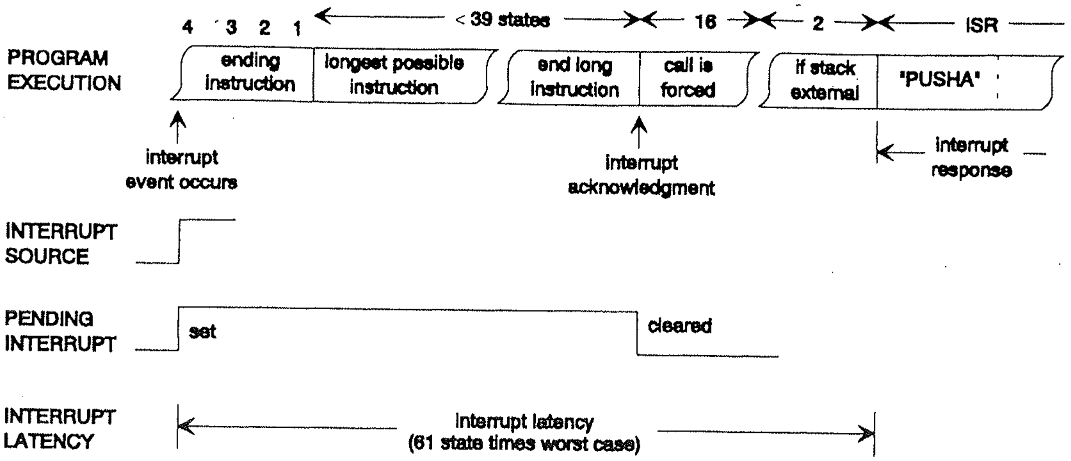


FIGURE 11.14 Worst case interrupt latency example.

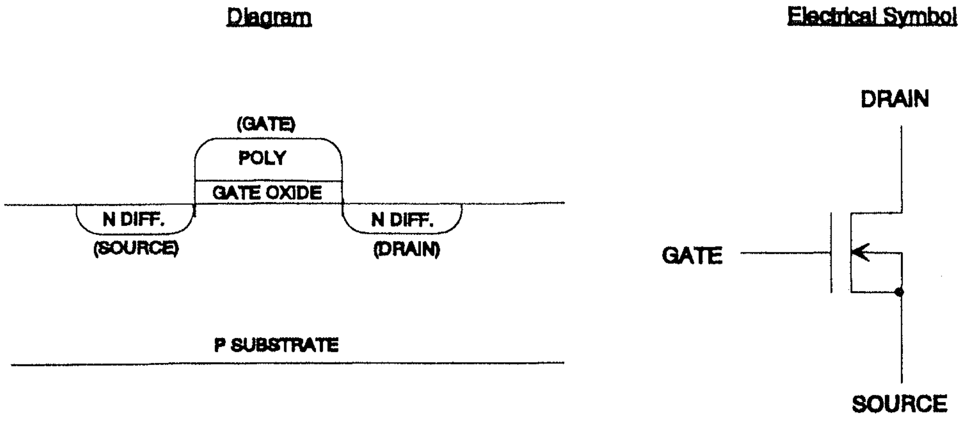


FIGURE 11.15 N-channel transistor.

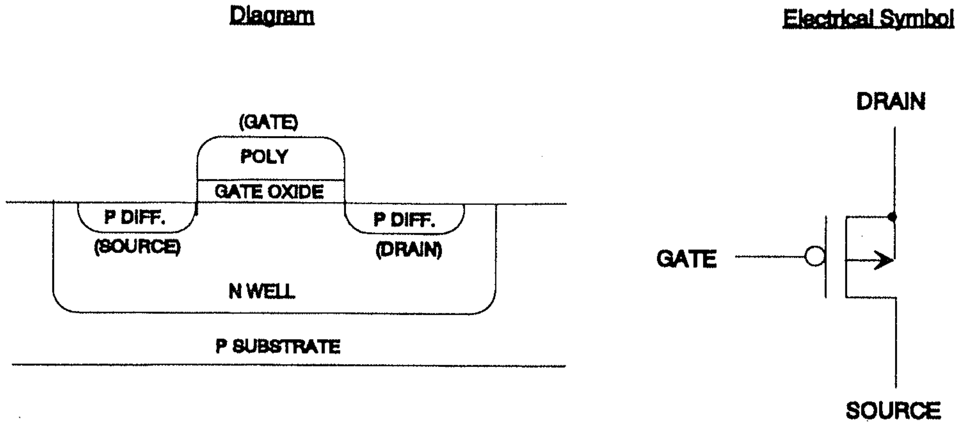


FIGURE 11.16 P-channel transistor.

NMOS Inverters. NMOS inverters are constructed of two NMOS transistors in which one is utilized as a resistance (Q2) and the other is utilized as a switch (Q1). A depletion-mode NMOS transistor is commonly utilized for the resistance device. A basic NMOS inverter is shown in Fig. 11.17. Note that Q2 is always on and acts as a resistor.

When a logic "0" is applied to the inverter's input, Q1 is turned off, which results in Q2 driving a logic 1 at the output. When a logic "1" is applied to the inverter's input, Q1 is turned on and overcomes Q2. This results in a logic "0" at the output.

NMOS microcontrollers are still produced in large quantities today. An advantage of NMOS processes is the simplistic circuit configuration which results in higher chip densities. NMOS devices are also less sensitive to electrostatic discharge (ESD) than CMOS devices. An inherent disadvantage of NMOS design is the slower switching speeds and higher power dissipation due to the dc current path from power to ground through Q1 and Q2 when the inverter is driving a logic "0".

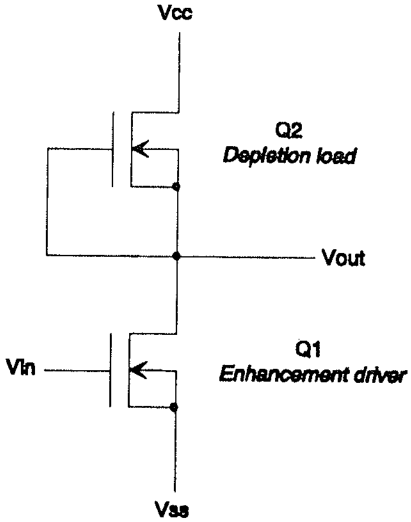


FIGURE 11.17 NMOS inverter.

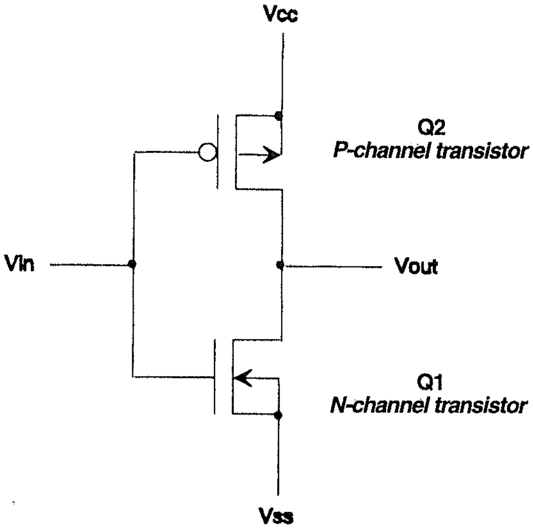


FIGURE 11.18 CMOS inverter.

CMOS Inverters. The CMOS is the most widely used process for automotive microcontrollers today. CMOS inverters are constructed of both P-channel and N-channel transistors that have their inputs tied together as shown in Fig. 11.18. When a logic "0" is applied to the inverter's input, Q1 is turned off and Q2 is turned on, which results in Q2 driving a logic "1" at the output. When a logic "1" is applied to the inverter's input, Q2 is turned off and Q1 is turned on, which results in Q1 driving a logic "0" at the output. Note that only one of these two devices will conduct at a time when the input is "1" or "0". While the input switches, both Q1 and Q2 may conduct for a short time resulting in a small amount of power dissipation.

The main advantages of CMOS logic are greatly improved switching times and lower power consumption, which is due to the complementary design of the inverter. A disadvantage of CMOS logic is that it is more expensive due to its increased complexity and more demanding fabrication process. CMOS logic is more susceptible to ESD damage, although microcontroller manufacturers have countered this by incorporating very effective ESD protection devices onto the silicon.

11.1.10 Temperature Range

Another important factor that must be considered when choosing a microcontroller is the temperature range in which it will be required to operate. The two most common temperature specifications specified by microcontroller manufacturers are *ambient temperature under bias* (TA) and *storage temperature.* These specifications are based upon package thermal characteristics as determined through device and package testing. Storage temperature refers to the temperature range that a microcontroller can be subjected to during periods of nonoperation. Storage temperature specifications are more extreme than ambient temperature under bias temperatures and are usually all the same regardless of the specified ambient temperature range. The common storage temperature range in industry is –60 to +150 °C. While powered-down, a given microcontroller must not be subjected to temperatures that exceed its specified storage temperature range.

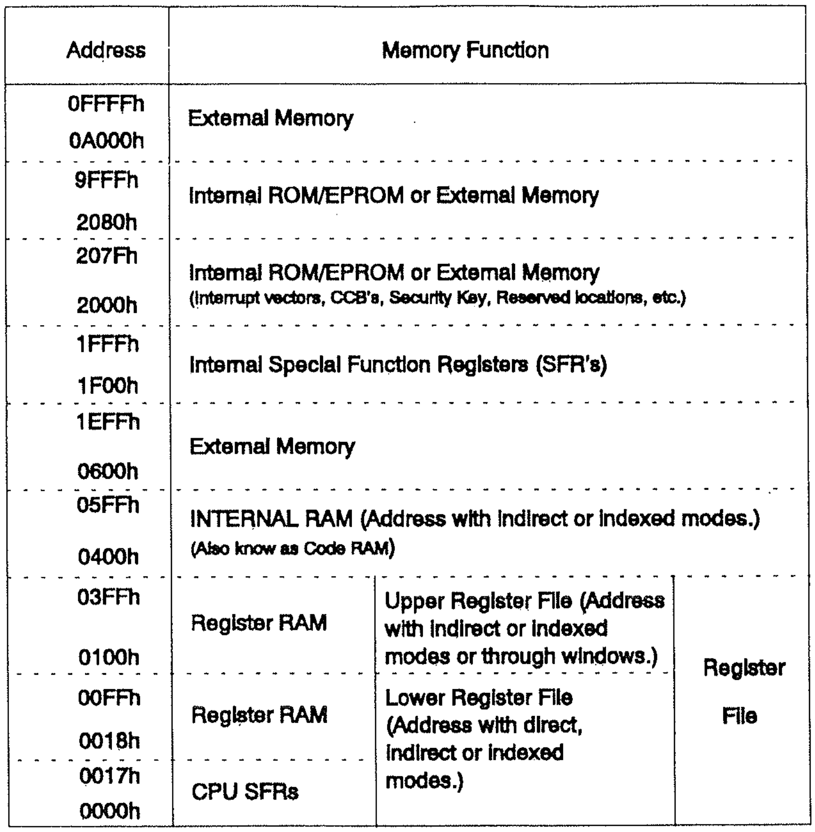
Ambient temperature under bias (TA) refers to the temperature range that the microcontroller is guaranteed to operate at within a given application. While powered-up or operating, a microcontroller must not be subjected to temperatures that exceed its specified ambient temperature range. The most common ambient temperature ranges in industry are:

Commercial	0 to +70 °C
Extended	–40 to +85 °C
Automotive	–40 to +125 °C

11.2 MEMORY

Microcontrollers execute customized programs that are written by the user. These programs are stored in either on-chip or off-chip memory and are often referred to as the *user's code.* On-chip memory is actually integrated onto the same piece of silicon as the microcontroller and is accessed over the internal data bus. Off-chip memory exists on a separately packaged piece of silicon and is typically accessed by the microcontroller over an external address/data bus.

A memory map shows how memory addresses are arranged in a particular microcontroller. Figure 11.19 shows a typical microcontroller memory map.



Address	Memory Function		
0FFFFh 0A000h	External Memory		
9FFFh 2080h	Internal ROM/EPROM or External Memory		
207Fh 2000h	Internal ROM/EPROM or External Memory (Interrupt vectors, CCB's, Security Key, Reserved locations, etc.)		
1FFFh 1F00h	Internal Special Function Registers (SFR's)		
1EFFh 0600h	External Memory		
05FFh 0400h	INTERNAL RAM (Address with indirect or indexed modes.) (Also know as Code RAM)		
03FFh 0100h	Register RAM	Upper Register File (Address with indirect or indexed modes or through windows.)	Register File
00FFh 0018h	Register RAM	Lower Register File (Address with direct, indirect or indexed modes.)	
0017h 0000h	CPU SFRs		

FIGURE 11.19 Microcontroller memory map.

Memory is commonly referred to in terms of Kbytes of memory. One Kbyte is defined as 1024 bytes of data. Memory is most commonly arranged in bytes which consist of 8 bits of data. For instance, a common automotive EPROM is referred to as a "256k × 8 EPROM". This EPROM contains 256-Kbytes 8-bit memory locations or 2,097,152 bits of information.

11.2.1 On-Chip Memory

On-chip microcontroller memory consists of some mix of five basic types: random access memory (RAM), read-only memory (ROM), erasable ROM (EPROM), electrically erasable ROM (EPROM), and flash memory. RAM is typically utilized for run-time variable storage and SFRs. The various types of ROM are generally used for code storage and fixed data tables.

The advantages of on-chip memory are numerous, especially for automotive applications, which are very size and cost conscious. Utilizing on-chip memory eliminates the need for external memory and the "glue" logic necessary to implement an address/data bus system. External memory systems are also notorious generators of switching noise and RFI due to their high clock rates and fast switching times. Providing sufficient on-chip memory helps to greatly reduce these concerns.

RAM. RAM may be defined as memory that has both read and write capabilities so that the stored information can be retrieved (read) and changed by applying new information to the cell (write). RAM found on microcontrollers is that of the static type that uses transistor cells connected as flip-flops. A typical six-transistor CMOS RAM cell is shown in Fig. 11.20. It consists of two cross-coupled CMOS inverters to store the data and two transmission gates, which provide the data path into or out of the cell. The most significant characteristic of static memory is that it loses its memory contents once power is removed. After power is removed, and once it is reapplied, static microcontroller RAM locations will revert to their default state of a logic "0". Because of the number of transistors used to construct a single cell, RAM memory is typically larger per bit than EPROM or ROM memory.

Although code typically cannot be executed from register RAM, a special type of RAM often referred to as *code RAM* is useful for downloading small segments of executable code. The difference between code and register RAM is that code RAM can be accessed via the

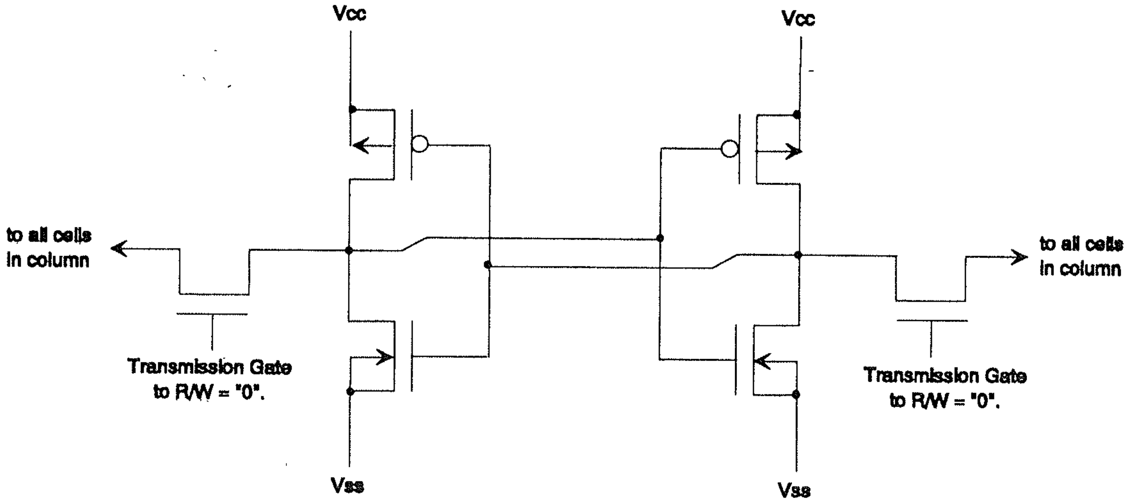


FIGURE 11.20 CMOS RAM memory cell.

memory controller, thus allowing code to be executed from it. Code RAM is especially useful for end-of-line testing during ECU manufacturing by allowing test code to be downloaded via the serial port peripheral.

ROM. Read-only memory (ROM), as the name implies, is memory that can be read but not written to. ROM is used for storage of user code or data that does not change since it is a nonvolatile memory that retains its contents after power is removed. Code or data is either entered during the manufacturing process (masked ROM, or MROM) or by later programming (programmable ROM, or PROM); either way, once entered it is unalterable.

A ROM cell by itself (Fig. 11.21) is nothing more than a transistor. ROM cells must be used in a matrix of word and bit lines (as shown in Fig. 11.22) in order to store information. The word lines are connected to the address decoder and the bit lines are connected to output buffers. The user's code is permanently stored by including or omitting individual cells at word and bit line junctions within the ROM array. For MROMs, this is done during wafer fabrication. For PROMs, this is done by blowing a fuse in the source/drain connection of each cell. To read an address within the array, the address decoder applies the address to the memory matrix. For any given intersection of a word and bit line, the absence of a cell transistor allows no current to flow and causes the transistor to be off. This indicates an unprogrammed ROM cell. The presence of a complete cell conducts and is sensed as a logical "0", indicating a programmed cell. The stored data on the bit lines is then driven to the output buffers.

MROMs are typically used for applications whose code is stable and in volume production. After the development process is complete and the user's program has been verified, the user submits the ROM code to the microcontroller manufacturer. The microcontroller manufacturer then produces a mask that is used during manufacturing to permanently embed the program within the microcontroller. This mask layer either enables or disables individual ROM cells at the junctions of the word and bit lines. An advantage of MROM microcontrollers is that they come with user code embedded, which saves time and money since post-production programming is not necessary. A disadvantage of MROM devices is that, since the mask with the user code has to be supplied early in the manufacturing process, throughput time (TPT) is longer.

Some versions of ROM (such as Intel's Quick-ROM) are actually not ROMs, but rather EPROMs, which are programmed at the factory. These devices are packaged in plastic devices, which prevents them from being erased since ultraviolet light cannot be applied to the actual EPROM array. Throughput time for QROMs is faster since the user code isn't required until after the actual manufacturing of the microcontroller is complete. As with

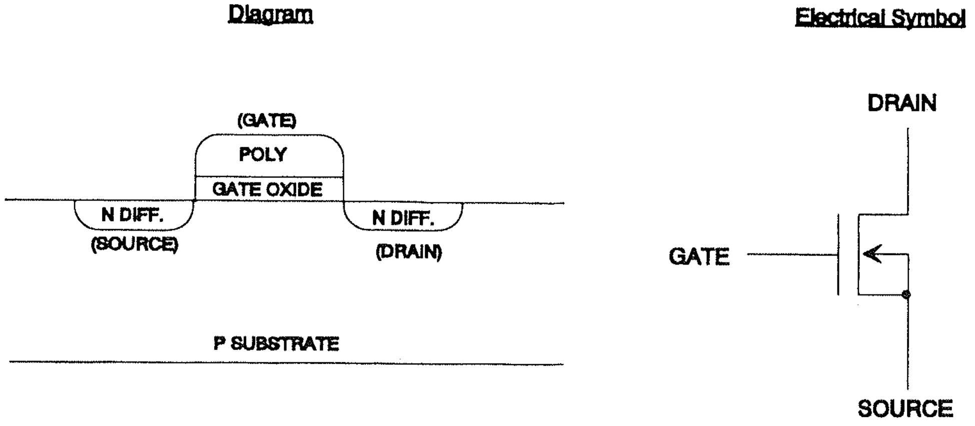


FIGURE 11.21 ROM memory cell.

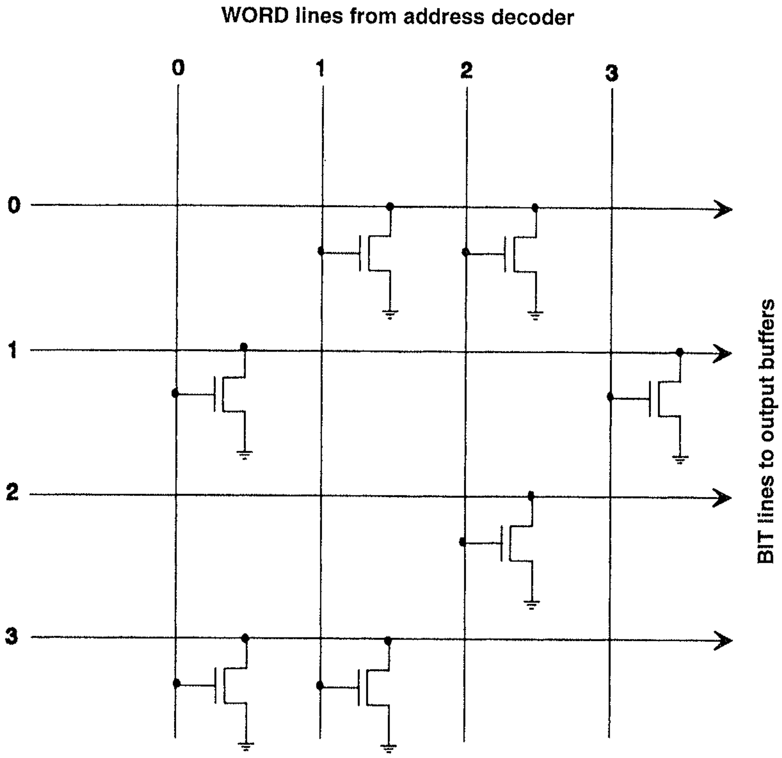


FIGURE 11.22 Simplified ROM memory matrix.

MROMs, the user supplies the ROM code to the microcontroller manufacturer. Instead of creating a mask with the ROM code, the manufacturer programs it into the device just prior to final test.

EPROM. EPROM devices are typically used during application development since this is when user code is changed often. EPROMs are delivered to the user unprogrammed. This allows the user to program the code into memory just prior to installation into an ECU module. Many EPROM microcontrollers actually provide a mechanism for in-module programming. This feature allows the user to program the device via the serial port while it is installed in the module. EPROM devices come assembled in packages either with or without a transparent window. Windowed devices are true EPROM devices that allow the user to erase the memory contents by exposing the EPROM array to ultraviolet light. These devices may be reprogrammed over and over again and thus are ideally suited for system development and debug during which code is changed often. EPROM devices assembled in a package without a window are commonly referred to as *one-time programmable devices* or OTPs. OTPs may only be programmed once, since the absence of a transparent window prevents UV erasure. OTPs are suited for limited production validation (PV) builds in which the code will not be erased.

A typical EPROM cell is shown in Fig. 11.23. It is basically an N-channel transistor that has an added poly1 floating gate to store charge. This floating gate is not connected and is surrounded by insulating oxide that prevents electron flow. The mechanism used to program an EPROM cell is known as *hot electron injection.* Hot electron injection occurs when very high drain (9-V) and select gate (12-V) voltages are applied. This gives the negatively charged electrons enough energy to surmount the oxide barrier and allows them to be stored on the gate.

This has the same effect as a negative applied gate voltage and turns the transistor off. When the cell is unprogrammed, it can be turned on like a normal transistor by applying 5 V to the poly2 select gate. When it is programmed, the 5 V will not turn on the cell. The state of the cell is determined by attempting to turn on the cell and detecting if it turns on. Erasure is performed through the application of ultraviolet (UV) light, which gives just the right amount of energy necessary for negatively charged electrons to surmount the oxide barrier and leave the floating gate.

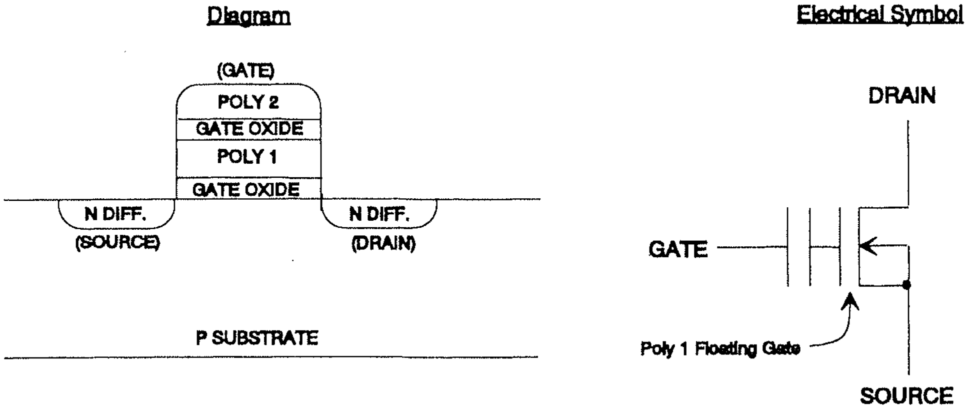


FIGURE 11.23 EPROM memory cell.

Flash. Flash memory is the newest nonvolatile memory technology and is very similar to EPROM. The key difference is that flash memory can be electrically erased. Once programmed, flash memory contents remain intact until an erase cycle is initiated via software. Like EEPROM, flash memory requires a programming and erase voltage of approximately 12.0 V. Since a clean, regulated 12-V reference is not readily available in automotive environments, this need is often provided for through the incorporation of an on-chip charge pump. The charge pump produces the voltage and current necessary for programming and erasure from the standard 5-V supply voltage. The advantage of flash is in its capability to be programmed *and* erased in-module without having to be removed. In-module reprogrammability is desirable since in-vehicle validation testing doesn't always allow for easy access to the microcontroller. Flash also allows for last-minute code changes, data table upgrades, and general code customization during ECU assembly. Since a flash cell is nearly identical in size to that of an EPROM cell, the high reliability and high device density capable with EPROM is retained. The main disadvantage of flash is the need for an on-chip charge pump and special program and erase circuitry, which adds cost.

A flash memory cell is essentially the same as an EPROM cell, with the exception of the floating gate. The difference is a thin oxide layer which allows the cell to be electrically erased. The mechanism used to erase data is known as *Fowler-Nordheim tunneling,* which allows the charge to be transferred from the floating gate when a large enough field is created. Hot electron injection is the mechanism used to program a cell, exactly as is done with EPROM cells. When the floating gate is positively charged, the cell will read a "1", when negatively charged, the cell will read a "0".

EEPROM. EEPROM (electrically erasable and programmable ROM, commonly referred to as E^{2}ROM) is a ROM that can be electrically erased and programmed. Once programmed, EEPROM contents remain intact until an erase cycle is initiated via software. Like flash, programming and erase voltages of approximately 12 V are required. Since a clean, regulated 12-V reference is not readily available in automotive environments, this requirement is satisfied using an on-chip charge pump as is done for flash memory arrays. Like flash, the advantage of EEPROM is its capability to be programmed and erased in-module. This allows the user to erase and program the device in the module without having to remove it. EEPROM's most significant disadvantage is the need for an on-chip charge pump. Special program and erase circuitry also adds cost.

An EEPROM cell is essentially the same as an EPROM cell with the exception of the floating gate being isolated by a thin oxide layer. The main difference from flash is that Fowler-Nordheim electron tunneling is used for *both* programming and erasure. This mechanism allows charge to be transferred to or from the floating gate (depending upon the polarity of the field) when a large enough field is created. When the floating gate is positively charged, the cell will read a "1"; when negatively charged, the cell will read a "0".

11.2.2 Off-Chip Memory

Off-chip memory offers the most flexibility to the system designer, but at a price; it takes up additional PCB real estate as well as additional I/O pins. In cost- and size-conscience applications, such as automotive ABS, system designers almost exclusively use on-chip memory. However, when memory requirements grow to sizes in excess of what is offered on-chip (such as is common in electronic engine control), the system designer must implement an off-chip memory system. Off-chip memory is flexible because the user can implement various memory devices in the configuration of his choice. Most microcontrollers on the market today offer a wide variety of control pins and timing modes to allow the system designer flexibility when interfacing to a wide range of external memory systems.

Accessing External Memory. If circuit designers must use external memory in their applications, the type of external address/data bus incorporated onto the microcontroller should be considered. If external memory is not used, this will have, if any, impact upon the application. There are two basic types of interfaces used in external memory systems. Both of these are parallel interfaces in which bits of data are moved in a parallel fashion and are referred to as *multiplexed* and *demultiplexed* address/data buses.

Multiplexed Address/Data Buses. As the name implies, multiplexed address/data buses allow the address as well as the data to be passed over the same microcontroller pins by multiplexing the two in time. Figure 11.24 illustrates a typical multiplexed 16-bit address/data bus system as is implemented with Intel's 8XC196Kx family of microcontrollers.

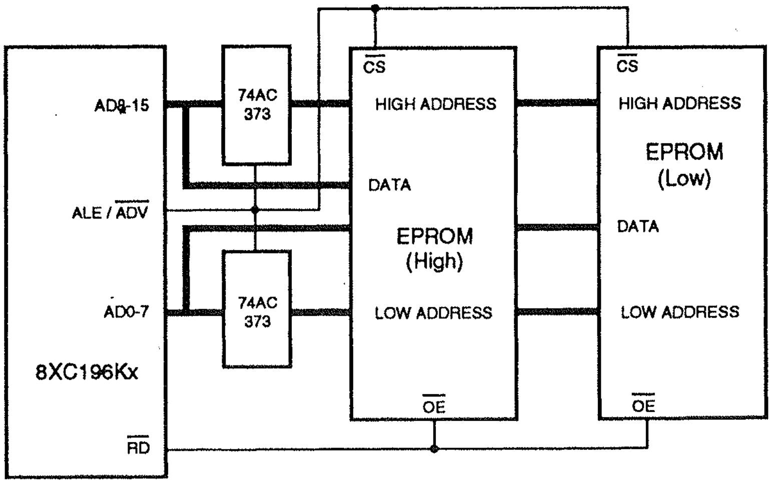


FIGURE 11.24 Multiplexed address/data bus system.

During a multiplexed bus cycle (refer to Fig. 11.25), the address is placed on the bus during the first half of the bus cycle and then latched by an external address data latch. The signal to latch the address comes from a signal generated by the microcontroller, called address latch enable (ALE). The address must be present on the bus for a specified amount of time prior to ALE being asserted. After the address is latched, the microcontroller asserts either a read (RD#) or a write (WR#) signal to the external memory device.

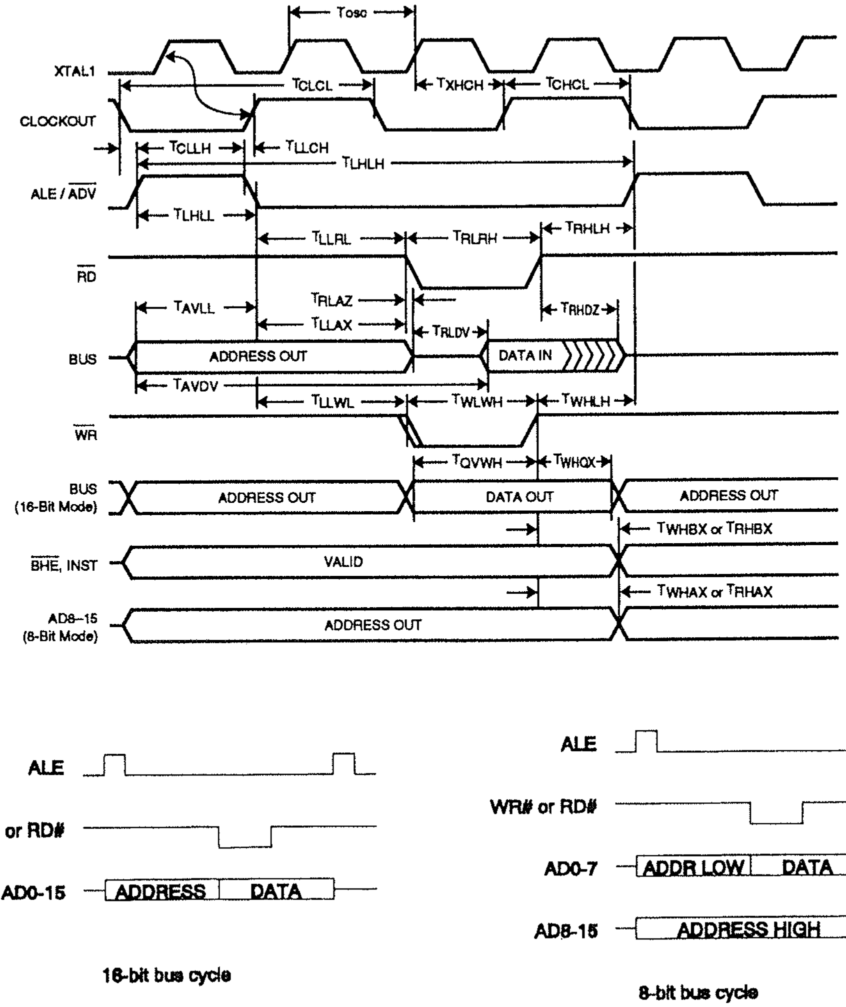


FIGURE 11.25 Multiplexed bus cycle and timing diagram.

For a read cycle, the microcontroller will pull its RD# output pin low and float the bus to allow the memory device to output the data located at the address latched on its address pins. The data returned from external memory must be on the bus and stable for a specified setup time before the rising edge or RD#, which is when the microcontroller latches the data.

For a write cycle, the microcontroller will pull its WR# pin low and then output data on the bus to be written to the external memory. After a specified setup time, the microcontroller will release its WR# signal, which signals to the memory device to latch the data on the bus into the address location present on its address pins.

Advantages of multiplexed address/data bus systems are that fewer microcontroller pins are required since address and data share the same pins. For a true 16-bit system, this translates into a multiplexed system requiring 16 fewer pins (for address and data) than would be required by a demultiplexed system. A disadvantage is that an external latch is required to hold the address during the second half of the bus cycle; this adds to the component count.

Demultiplexed Address/Data Buses. Microcontrollers with demultiplexed address/data buses implement separate, dedicated address and data buses as shown in Fig. 11.26.

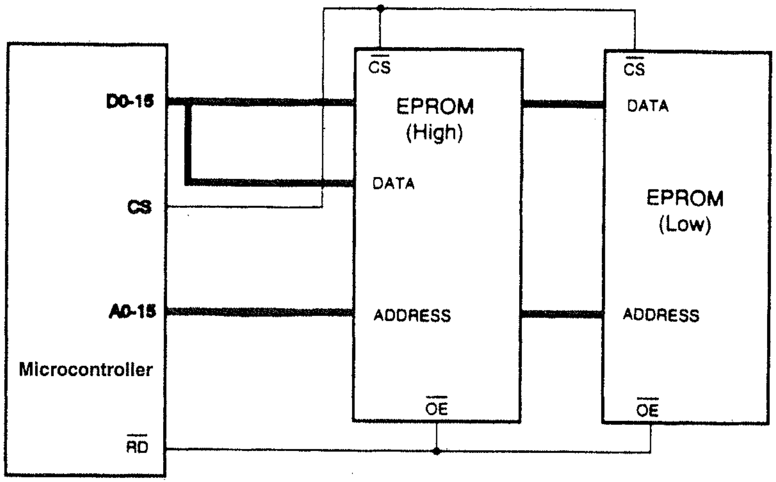


FIGURE 11.26 Typical demultiplexed address/data bus system.

The operation of a demultiplexed address/data bus is basically the same as the multiplexed type with the exception of not having an ALE signal to latch the address for the second half of the bus cycle. The operation of the RD#,WR#, address, and data lines is essentially the same as for that of a multiplexed system.

During a demultiplexed bus cycle, the microcontroller places the address on the address bus and holds it there for the entire bus cycle. For a read of external memory, the microcontroller asserts the RD# signal (or WR# for a write signal) just as would be done for a multiplexed bus cycle. The memory device will respond accordingly by either placing the data to be read on the data bus or by latching the data to be written off of the data bus. Figure 11.27 illustrates a simplified demultiplexed bus cycle.

An advantage of multiplexed address/data bus systems is that external data latches are not necessary, which saves on system component count. A disadvantage, as mentioned earlier, is that more microcontroller pins must be allocated for the interface, which leaves fewer pins for other I/O purposes.

11.3 LOW-SPEED INPUT/OUTPUT PORTS

Low-speed input/output (LSIO) ports allow the microcontroller to read input signals as well as provide output signals to and from other electronic components such as sensors, power drivers,

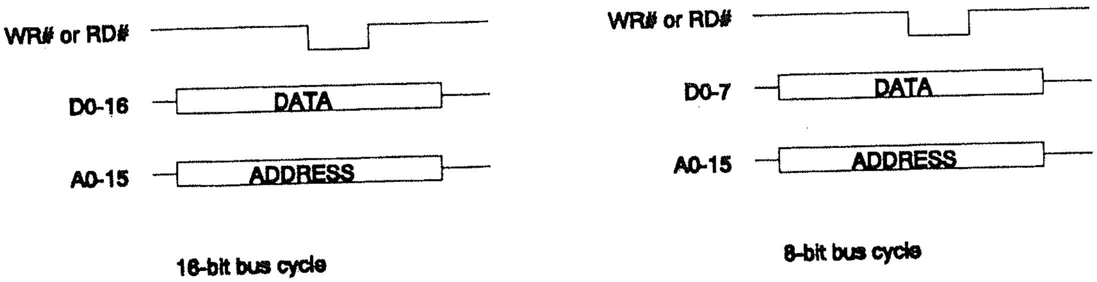


FIGURE 11.27 Demultiplexed bus cycle.

digital devices, actuators, and other microcontrollers. The term "low-speed" is used to describe these ports because unlike high-speed I/O (HSIO) ports which are interrupt driven, LSIO port data must be manually read and written by the user program. Interrupt-driven I/O is typically not possible on port pins configured for LSIO operation. It is common for modern high-performance microcontrollers to utilize multifunctional port pins which can be configured for a special function as well as LSIO. LSIO ports most commonly consist of eight port pins in parallel, which are supported by byte registers. For example, by writing to a single-byte special function register, an entire port can be configured, read, or written. Manipulating individual bits in the port register allows the user flexibility in accessing either single or multiple port pins.

11.3.1 Push-Pull Port Pin Configuration

The term *push-pull,* or *complementary,* output is commonly used to define a port pin that has the capability to output either a logic "1" or "0". Figure 11.28 shows a basic push-pull port pin configuration. Referring to Fig. 11.28, writing a "1" to the data output register enables the P-channel MOSFET and pulls the pin to +5 V, thus driving a logic "1" at the port pin. When a "0" is written

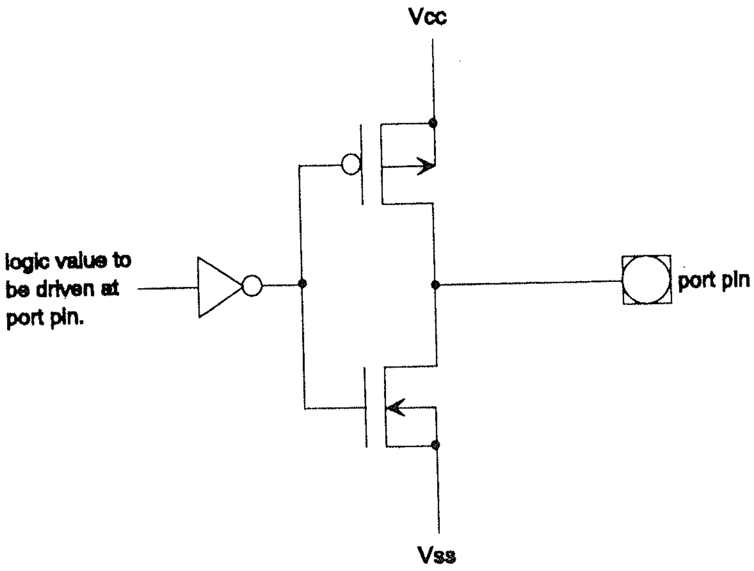


FIGURE 11.28 Push-pull port pin.

to the data register, the N-channel MOSFET is enabled and thus provides a current path to ground which results in a logic 0 at the port pin. Note that during this time the P-channel pull-up MOSFET is disabled to prevent contention at the port pin. Also note that the port logic design does not allow both the P-channel and the N-channel devices to be driving at the same time.

11.3.2 Open-Drain Port Pin Configuration

Open-drain port pins (Fig. 11.29) are useful for handshaking signals over which multiple devices will have control. The fact that the P-channel transistor is either omitted or disabled dictates the need for an external pull-up resistor. An example of an application for open-drain port pins would be for a bus contention line between two microcontrollers communicating on a common bus. During normal operation, the line is pulled high by the external pull-up resistor to signal to either microcontroller that no contention exists. If one of the microcontrollers should detect contention on the bus, it simply outputs a logic "0", which signals the contention to the other processor. To output the "0", the port only has to overcome the external pull-up which the user should appropriately size to match the port drive specifications.

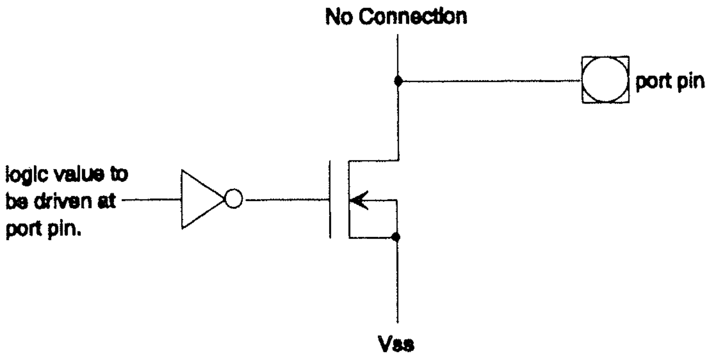


FIGURE 11.29 Open-drain port pin.

11.3.3 High-Impedance Input Port Pin Configuration

High impedance, or "Hi-z," port pins (Fig. 11.30) are used strictly as inputs since no drivers exist on these types of pins. Hi-z refers to the relatively high input impedance of the port pin. This high input impedance prevents the port pin circuitry from actively loading the input signal. Note that the pin is connected to the gates of a CMOS inverter, which drives internal circuitry. Usually a certain amount of hysteresis is built into these pins and is specified in the data sheet.

11.3.4 Quasi Bidirectional Port Pin Configuration

Quasi bidirectional (QBD) port pins are those that can be used as either input or output without the need for direction control logic. QBD port pins can output a strong low value or a weak high value. The weak high value can be externally overridden, providing an input function. Figure 11.31 shows a QBD port pin diagram and its transfer characteristic.

Writing a "1" to the port pin disables the strong low driver (Q2) and enables a very weak high driver (Q3). To get the pin to transition high quickly, a strong high driver (Q1) is enabled for one state time and then disabled (leaving only Q3 active).

It is important to keep in mind that since the port pin can be externally overridden with a logic "0", reading the port pin could falsely indicate that it was written as a logic "0".

The ability to overdrive the weak output driver is what gives the quasi bidirectional port pin its input capability. To reduce the amount of current that flows when the pin is externally pulled low, the weak output driver (Q4) is turned off when a valid logic "0" is detected. The input transfer characteristic of a quasi bidirectional port pin is shown in Fig. 11.31.

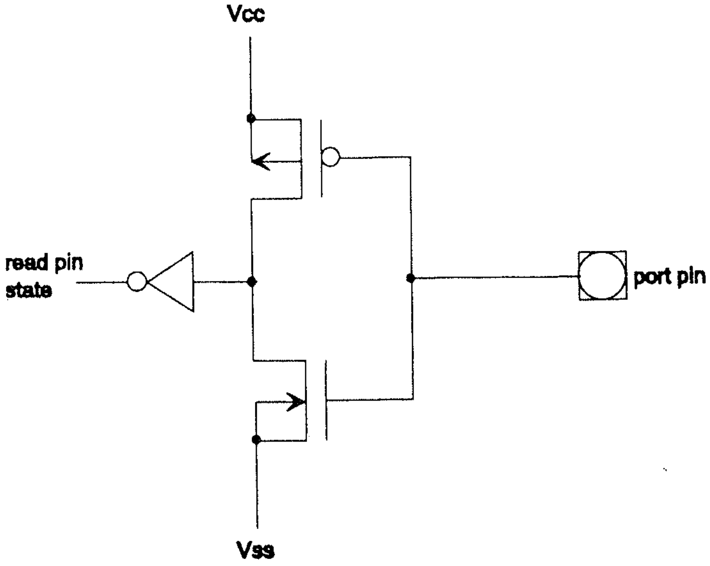


FIGURE 11.30 High-impedance input port pin.

11.3.5 Bidirectional Port Example

The following example describes the operation of a state-of-the-art bidirectional port structure. This particular structure is used upon newer members of Intel's MCS-96 automotive microcontroller family. A single port consists of eight multifunction, parallel port pins (see Fig. 11.32), which are controlled (on a by-pin basis) with four special function registers referred to as Px_PIN, Px_REG, Px_MODE, and Px_DIR. As is common with other high-performance microcontrollers, the pins of this port are shared with alternate special functions controlled by other on-chip peripherals. The Px_MODE register allows the programmer to choose either LSIO or the associated special function for any given port pin. Writing a "1" to the appropriate bit selects the corresponding pin as special function whereas a "0" selects LSIO. The function of the Px_PIN and Px_REG registers is fairly straightforward. In order to read the value on the pin, the user simply reads the Px_PIN register. To write a value to the Px_REG register, the user simply writes the desired output value to the Px_REG register. The Px_DIR register allows the user to configure the port pin as either input or output.

In order to prevent an undefined pin state during reset, port pins revert to a default state during reset. For the Intel Kx bidirectional port structure, this state is defined as a weak logic "1". The transistor that drives this state is labeled as WKPU in Fig. 11.32 and is asserted in reset until the user writes to the Px_MODE register to configure the port pin.

Ports such as this offer the user much flexibility in assigning their function within an application. Following are three examples that depict how these ports may be configured by the user by writing values to the appropriate bit within the port SFR. Also note that the eight pins of a port may be configured individually on a pin-by-pin basis.

To configure a given port pin as a high-impedance input pin, the user must write the fol-

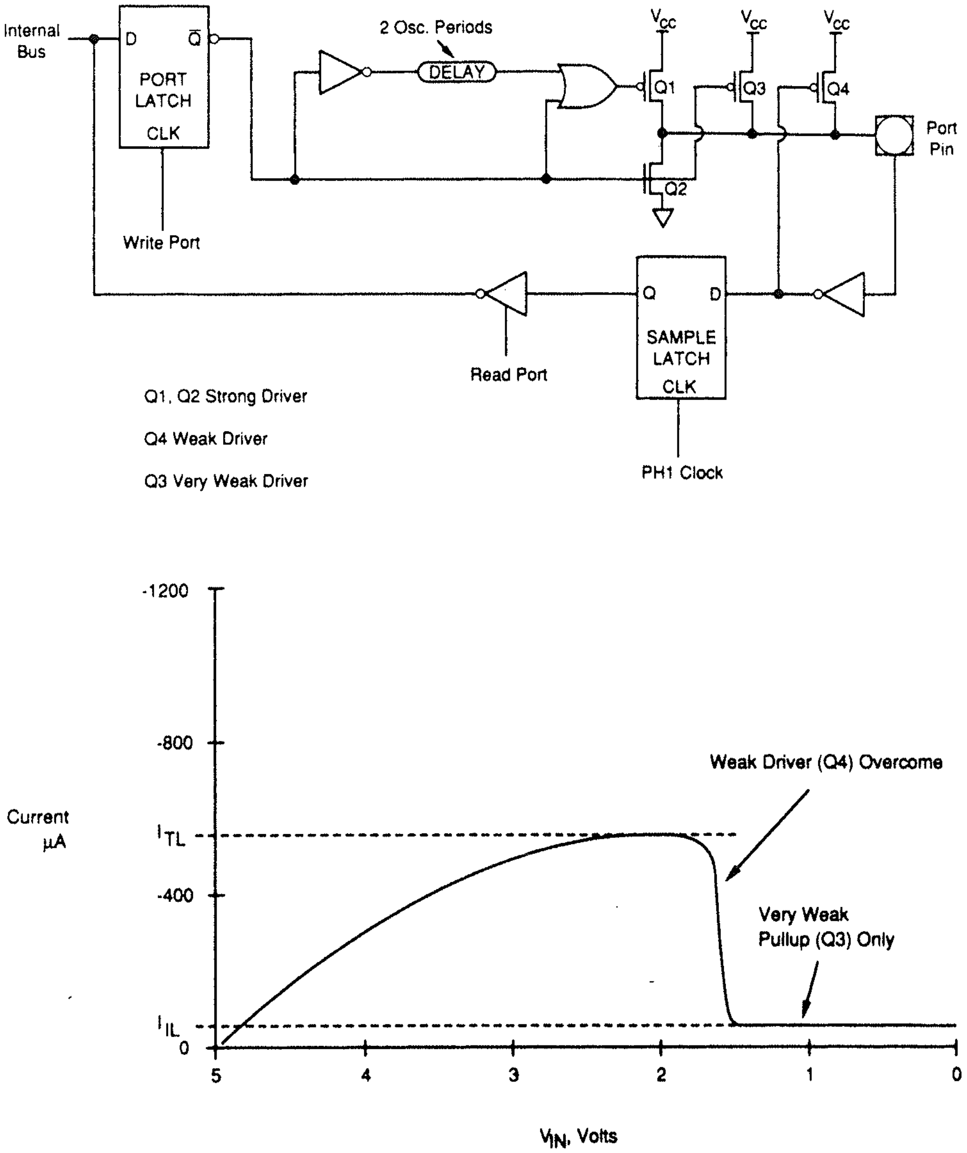


FIGURE 11.31 Quasi bidirectional port pin and transfer characteristic.

Px_MODE: "0" selects the pin as LSIO and disables weak pull-up.
Px_DIR: "1" disables operation of the N-channel transistor.
Px_REG: "1" disables the N-channel transistor.

To configure a given port pin for push-pull operation, the following values must be written to the corresponding bit within the port SFR.

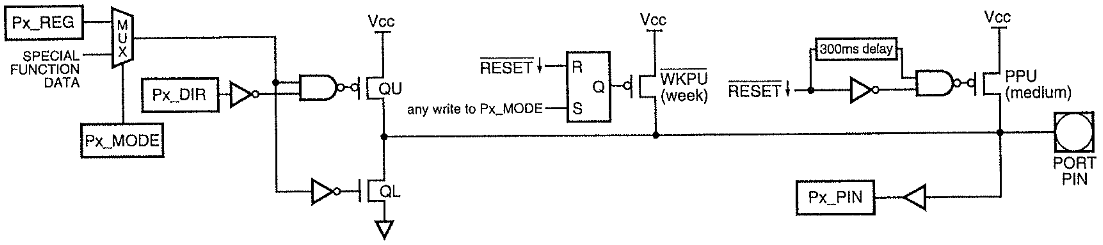


FIGURE 11.32 Bidirectional port structure example.

Px_MODE: "0" selects the pin as LSIO and disables weak pull-up transistor.
Px_DIR: "0" enables operation of both the N- and P-channel transistors.
Px_REG: "0" or "1" drives that value at the port pin.

To configure a port pin for open-drain operation, the user must write the following values to the corresponding bits within the port SFR.

Px_MODE: "0" selects the pin as LSIO and disables weak pull-up transistor.
Px_DIR: "1" disables operation of the N-channel transistor.
Px_REG: "1" disables the P-channel transistor / achieves Hi-Z state.
"0" enables the N-channel transistor / drives "0" at pin.

11.4 HIGH-SPEED I/O PORTS

Perhaps the most demanding of automotive microcontroller applications is electronic engine control and antilock braking/traction control. These applications both require the microcontroller to detect, process, and respond to external signals or "events" within relatively short periods of time. Sometimes referred to as a capture/compare module, a microcontroller's HSIO (high-speed input/output) peripheral allows the microcontroller to capture an event as it occurs. The term *capture* refers to a series of events that begins with the microcontroller detecting a rising or falling edge upon a high-speed input pin. At the precise moment this edge is detected, the value of a software timer is loaded into a time register and an interrupt is triggered. This gives the microcontroller the relative time at which the event occurred. An HSIO peripheral also provides compare functions by detecting an internal event, such as a timer reaching a particular count value. When the particular count value is detected, the HSIO unit will generate a specified event (rising or falling edge) on a port pin. This feature is ideal for generating PWM waveforms or synchronizing external events with internal events.

For example, consider a typical ABS microcontroller which must detect, capture, and calculate wheel speeds; respond with signals to hydraulic solenoids; and perform many other background tasks all within a loop time of about 5 ms. The wheel speed signals are input to the microcontroller as square waves with frequencies up to 7000 Hz (approximately one edge every 71 μs). The microcontroller must have the performance necessary to capture and process these edges on as many as four wheel speed inputs. HSIO peripherals, along with the interrupt structure, play a major role in the microcontroller's ability to perform this function.

Nearly every microcontroller manufacturer has its own proprietary HSIO peripheral. For purposes of this section, the event processor array (EPA) HSIO peripheral, which is used by Intel's 87C196KT automotive microcontroller, will be discussed.

11.4.1 High-Speed Input and Output Peripheral

High-speed input/output peripherals typically consist of a given number of capture/compare modules, a timer/counter structure, control and status SFRs, and an interrupt structure of some type. Figure 11,33 shows a block diagram of the EPA peripheral. The main components of the EPA are ten capture/compare channels, two compare only channels, and two timer/counters. The capture/compare channels are configured independently of each other. The two timer/counters are shared between the various capture/compare channels. Each capture/compare channel has its own dedicated SFR's: EPAx_TIME and EPAx_CON (x designates the channel number).

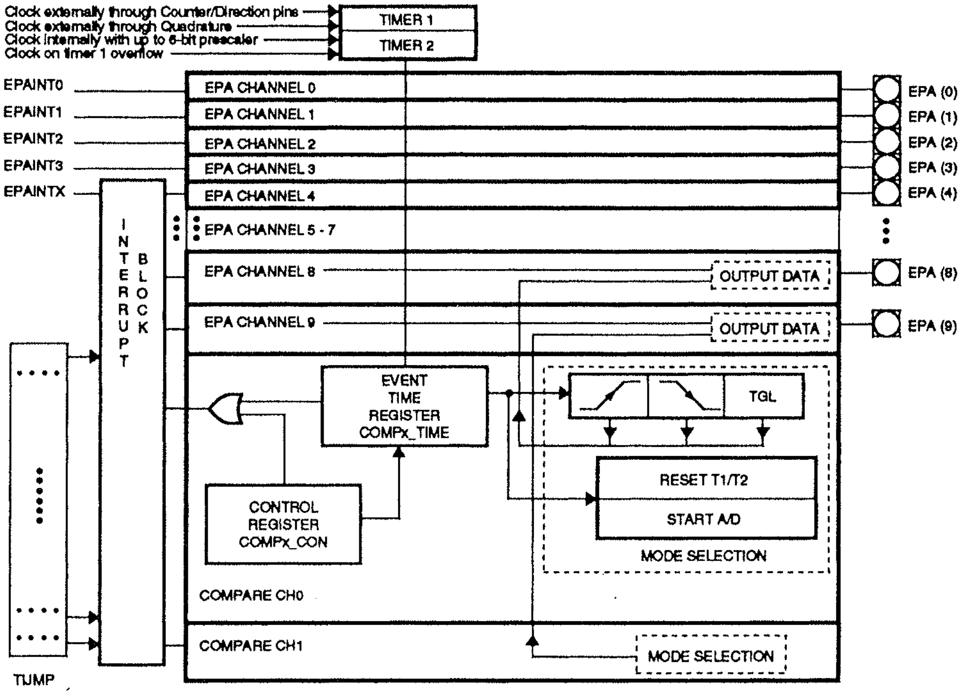


FIGURE 11.33 Example HSIO peripheral: Intel's EPA peripheral.

11.4.2 Timer/Counter Structures

High-performance microcontrollers typically integrate one or more timer/counters onto their silicon. A microcontroller's timer/counter structure provides a time base to which all HSIO events are referenced. Timers are clocked internally, whereas counters are clocked from an external clock source. Timers are often very flexible structures, in which programmers have the capability to configure the timer/counters to meet their application's particular needs. The 87C196KT has two 16-bit timer/counters referred to as TIMER1 and TIMER2. As 16-bit timer/counters, each timer has the capability of counting to 2^{16} or 65,536 before overflowing. The user has the option of triggering an interrupt upon overflow of a timer/counter. Each of these two timers can be independently configured using the TxCONTROL SFR as shown in Fig. 11.34, where x specifies either 1 or 2 for Timer1 or Timer2, respectively.

Bits number 3, 4, and 5 are the mode bits that allow the user to configure the clocking source and direction of each timer/counter. The clock rate can be based either upon the fre-

TxCONTROL SFR

7	6	5	4	3	2	1	0
CE	UD	M2	M1	M0	P2	P1	P0

CE: Count Enable: "0" = disable timer, "1" = enable timer

UD: Up/Down: "0" = count up, "1" = count down

MODE:	M2, M1, M0	Clock source	Direction determined by:
	0 0 0	XTAL/4	state of UD bit
	0 0 1	TxCLK pin	state of UD bit
	0 1 0	XTAL/4	state of TxDIR pin
	0 1 1	TxCLK pin	state of TxDIR pin
	1 0 0	Timer1 overflow	state of UD bit
	1 1 0	Timer1 overflow	same as Timer1
	1 1 1	Quadrature clocking using TxCLK and TxDIR pins	

Prescale:	P2, P1, P0	Clock prescale values
	0 0 0	÷ by 1 (250 ns @ 16 MHz xtal frequency)
	0 0 1	÷ by 2 (500 ns @ 16 MHz xtal frequency)
	0 1 0	÷ by 4 (1 µs @ 16 MHz xtal frequency)
	0 1 1	÷ by 8 (2 µs @ 16 MHz xtal frequency)
	1 0 0	÷ by 16 (4 µs @ 16 MHz xtal frequency)
	1 0 1	÷ by 32 (8 µs @ 16 MHz xtal frequency)
	1 1 0	÷ by 64 (16 µs @ 16 MHz xtal frequency)
	1 1 1	reserved

FIGURE 11.34 Timer control SFR example.

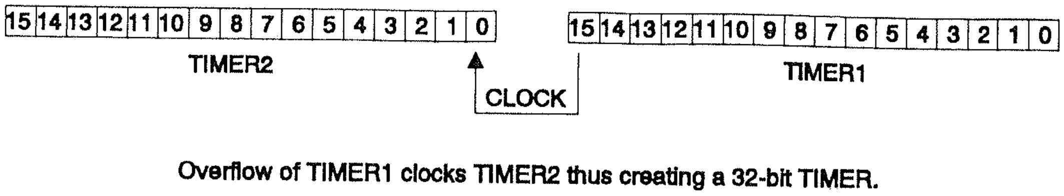


FIGURE 11.35 Cascading of timer/counters.

quency that the microcontroller is being clocked at the XTAL pins or upon the input frequency on another pin referred to as TxCLK. The user also has the option of either having the logic level of another pin (TxDIR) or the UD bits in TxCONTROL determine the direction (up/down) that the timer/counter is clocked.

For those applications that require a 32-bit timer/counter, the user has the option (using the mode bits) to direct the overflow of TIMER1 to clock TIMER2. This is known as cascading and essentially creates a 32-bit timer/counter as shown in Fig. 11.35.

11.4.3 Input Capture

Input capture refers to the process of capturing a current timer value when a specific type of event occurs. An excellent example of high-speed input capture can be illustrated with a basic automotive ABS input capture algorithm that calculates the frequency of a wheel speed input. The signals from the wheel speed sensors are input into the microcontroller's EPA pins as square waves. Consider the generic wheel speed input capture example shown in Fig. 11.36.

Two timers (1 and 2) are used in this example. Timer1 is used in conjunction with an EPA channel to provide a 5-ms software timer (this is a compare function that will be discussed in the next section). The 5 ms is the main loop time used in generic ABS algorithms. Timer2 is used in conjunction with one or more EPA channels to capture the relative times at which edges occur on wheel speed inputs. The EPA is configured to capture falling edges and initiate an interrupt, which stores the event time and increments an edge count. To simplify this example, we will consider only a single input channel.

The process starts by EPA interrupts being enabled after Timer1 starts a new 5-ms timer count. The first falling edge causes an interrupt that stores the event time (T2) into a variable *initial time* and increments an edge count. The next edge causes an interrupt in which the event time (T2+x) is stored into a variable called *final time* and increments the edge count.

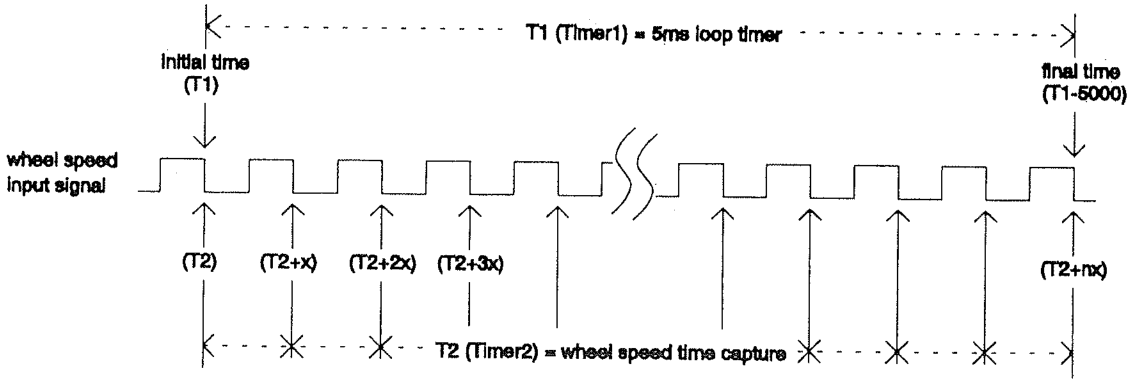


FIGURE 11.36 Input capture example using EPA peripheral.

Subsequent edges' event times are also stored into *final time* until Timer1's 5-ms count expires. At this point, final time contains the time at which the last edge to occur was captured. The average period of the input waveform can then be calculated with the following equation:

input period = (*final time* – *initial time*) / edge count

11.4.4 Output Compare

Output compare refers to the process of generating an event when a timer value matches a predetermined time value. The event may be to generate an interrupt, toggle an output pin, perform an A/D conversion, and so forth. Following is an example that shows the steps necessary to generate an event every 50 μs:

1. Enable the output compare channel's interrupt.
2. Initialize the timer to count up at 1 μs per timer tick.
3. Initialize the output compare channel to re-enable and reset the timer (to zero) when a timer match occurs.
4. Initialize the output compare channel to produce the desired event when a timer match occurs.
5. Write 32h (50 decimal) to the appropriate output compare channel's time register.
6. Enable the timer to start the process.
7. A compare channel interrupt will be generated every 50 μs.

Since the example re-enables and zeros the timer, the event will occur continuously until the user's program halts the process.

Software Timers. Software timers such as the 5-ms timer used in the ABS wheel speed capture example can be set up easily using a compare channel and a timer. The following software timer procedure is very similar to that used in the previous output compare example:

1. Enable the compare channel's interrupt.
2. Initialize the timer to count up at 1 us per timer tick.
3. Initialize the output compare channel to re-enable and reset the timer (to zero) when a timer match occurs.
4. Initialize the output compare channel to produce an interrupt (5-ms ISR) when a timer match occurs.
5. Write 1388h (5000 decimal) to the appropriate output compare channel's time register.
6. Enable the timer to start the process.
7. An compare channel interrupt will be generated every 5 ms.

11.4.5 Pulse-Width Modulation (PWM)

Pulse-width modulation (PWM) peripherals provide the user with the ability to generate waveforms that have specified frequencies and duty cycles. PWM waveforms are typically used to generate pulsed waveforms used for motor control or they may be filtered to produce a smooth analog signal. HSIO peripherals typically provide for PWM waveform generation, although the methods are not usually as efficient as dedicated PWM peripherals. A basic example of creating a PWM waveform using an HSIO peripheral's output compare function is described in Sec. 11.4.4.

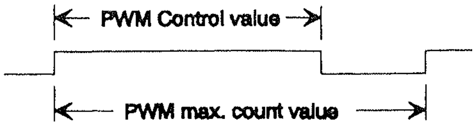
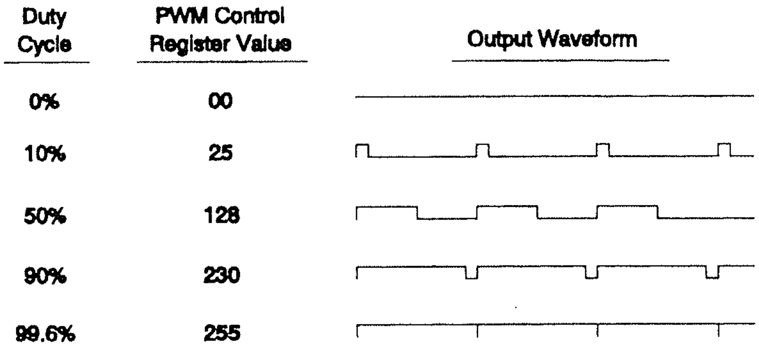


FIGURE 11.37 PWM waveform time values.

PWM Peripheral. The components of a basic automotive microcontroller's PWM peripheral include a counter (typically 8-bit), a comparator, a holding register, and a control register. The counter typically has a prescaler that allows the user to select the clock rate of the counter, which allows for selectable PWM frequencies. Without prescaling capability, an 8-bit counter would only allow for a period of 256 state times. The PWM control register determines how long the PWM output is held high during the pulse, effectively controlling the duty cycle as shown in Fig. 11.37. For an 8-bit PWM counter, the value written to the PWM control register can be from 0 to 255 (equating to 255 state times with no prescaling). Note that PWM peripherals do not typically allow for a 100 percent duty cycle because the output must be reset when the counter reaches zero.

The operation of a PWM peripheral is rather simple. The PWM control register's value (assume 8-bit for this example) is loaded into a holding register when the 8-bit counter overflows. The comparator compares the contents of the holding register to the counter value. When the counter value is equal to zero, the PWM output is driven high. It remains high until the counter value matches the value in the holding register, at which time the output is pulled low. When the counter overflows, the output is again switched high. Figure 11.38 shows typical PWM output waveforms.



Duty Cycle	PWM Control Register Value	Output Waveform
0%	00	
10%	25	
50%	128	
90%	230	
99.6%	255	

FIGURE 11.38 PWM output waveforms.

11.5 SERIAL COMMUNICATIONS

It is often necessary for automotive microcontrollers to have the capability to communicate with other devices both internal and external to the ECU. Within an ECU a microcontroller may have to communicate with other devices such as backup processors, shift registers, watchdog timers, and so forth. It is not uncommon for automotive microcontrollers to communicate with devices external to the ECU, such as other modules within the vehicle and even diagnostic computers at a service station. All of these communication examples require a large quantity of data to be transmitted/received in a short period of time. Also consider that this communication must utilize as few pins of the microcontroller as possible in order to save valuable PCB board space. These requirements all support the need for serial communications.

Serial communications provides for efficient transfer of data while utilizing a minimum number of pins. Serial communications is performed by transferring a group of data bits, one at a time, sequentially over a single data line. Each transmission of a group of bits (typically a

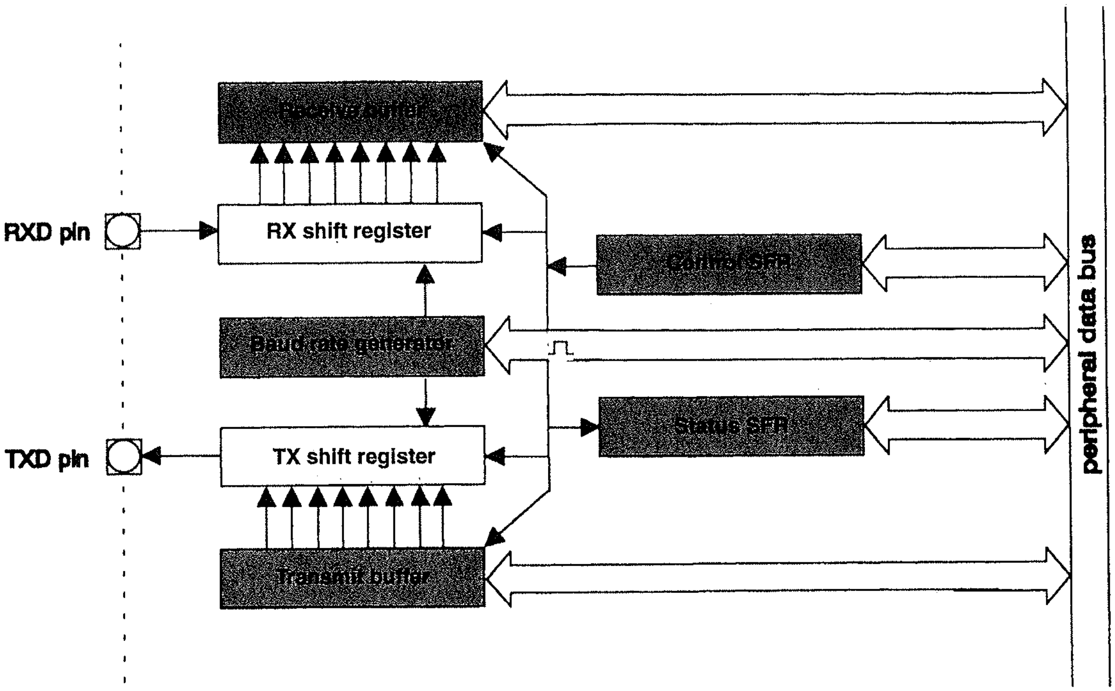


FIGURE 11.39 Serial port block diagram.

byte of data) is known as a data frame. This transfer of data takes place at a given speed, which is referred to as the baud rate and is typically specified in bits/second.

A typical microcontroller serial port consists of data buffers, data registers, and a baud rate generator. Interface to the outside world takes place via the transmit (TXD) and receive (RXD) pins. A block diagram for a typical serial port peripheral is shown in Fig. 11.39. By writing to the serial port control register, users are able to customize the operation of the serial port to their particular application's requirements.

The baud rate generator is used to provide the timing necessary for serial communications and determines the rate at which the bits are transmitted. In synchronous modes, the baud rate generator provides the timing reference used to create clock edges on the clock output pin. In asynchronous modes, the baud rate generator provides the timing reference used to latch data into the RX pin and clock it out of the TX pin.

11.5.1 Synchronous Serial Communications

Sometimes an application does not allow asynchronous serial communications to take place due to variations in clock frequency, which results in unacceptable baud rate error. Some applications simply require some sort of shift register I/O. Synchronous communication involves an additional clock pin, which is used to signal the other device that data being transferred are valid and ready to be read. Often when the user configures the serial port to work in a synchronous mode, the TXD pin automatically reverts to supplying the clock and the RXD pin automatically becomes the data pin. This configuration prevents an additional pin from having to be reserved for use as a serial clock pin. When a synchronous data transfer is initiated, a series of eight clock pulses is emitted from the clock pin at a predetermined baud rate as shown in Fig. 11.40.

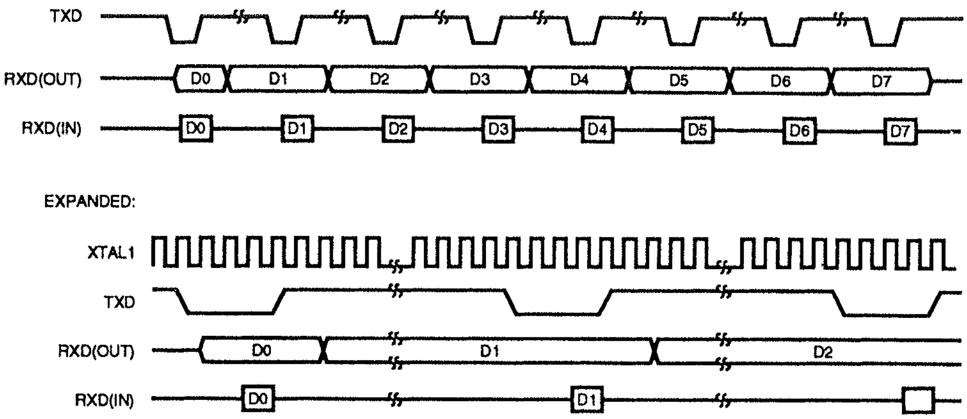


FIGURE 11.40 Synchronous serial mode data frame.

An example of synchronous serial communications is shown in Fig. 11.41. Assume that processor A is to transfer a byte of data to processor B. The program executing in processor A initiates a serial transmission by writing the data byte to be transmitted into the transmit buffer. Assuming microcontroller A's serial port is enabled for transmission, writing to the transmit buffer results in a series of eight clock pulses to be emitted from microcontroller A's clock pin. The first falling edge of the clock will signal to processor B that bit 0 (LSB) is ready to be read into its receive buffer. Microcontroller A will place the next data bit on the TXD pin with each rising clock edge. With B's serial port enabled for reception, each falling edge will result in another data bit being shifted into B's receive buffer. When B's receive buffer is full, the received data byte will be loaded into its receive register and will signal its CPU that the reception has been completed and the data is ready for use.

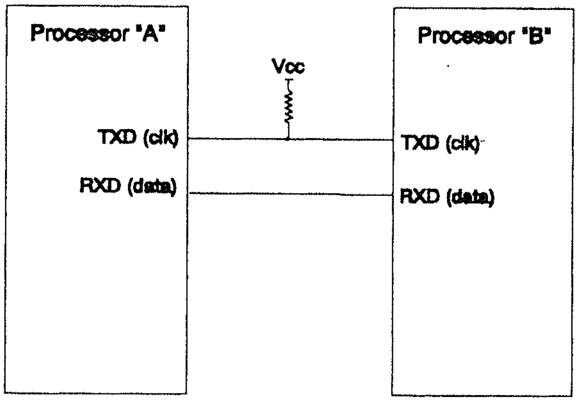


FIGURE 11.41 Synchronous serial communications example.

Shift Register Based I/O Expansion. A common application for synchronous serial transmission is shift register based I/O expansion as shown in Fig. 11.42. In this circuit, a 74HC164 8-bit serial-in/parallel-out shift register is used to provide eight parallel outputs with a single serial input. The 74HC165 8-bit parallel-in/serial out shift register shown provides a single serial input resulting from eight parallel input signals. This allows the system designer to

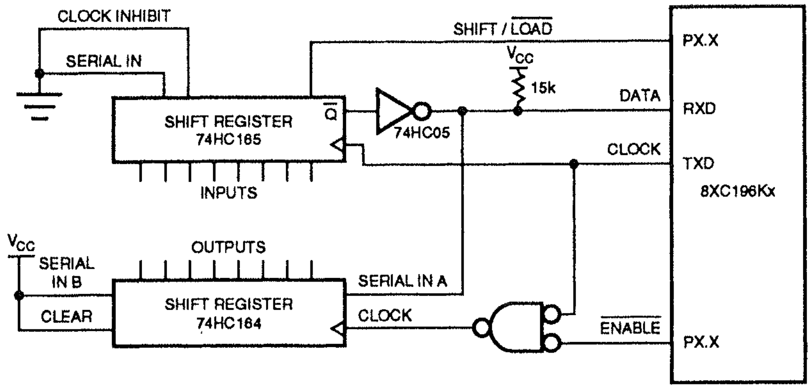


FIGURE 11.42 Shift register based I/O expansion example.

implement an additional 8-bit output port and additional 8-bit input port (16 signals total) using only four pins on the microcontroller. This expansion scheme allows a designer to achieve a greater number of I/O pins without having to upgrade to a microcontroller with a higher pin count.

To output data using this I/O expansion method, the user code simply writes a byte to the serial port transmit register to initiate data transfer. This causes the written byte to be shifted out of the microcontroller's RXD pin and into the 74HC164 one bit at a time. The data is reflected at the output pins of the 74HC164 as each bit is shifted in. For address/data bus emulation, another microcontroller pin may be utilized to indicate valid data to the intended receiving device.

To receive eight bits of data in parallel using this method, the user's code must latch the data on the 74HC165's input pins into its shift register by asserting the *shift/load* signal. After this is accomplished, the user's code simply needs to enable the serial port receive circuitry to receive the data one bit at a time into its receive buffer.

11.5.2 Asynchronous Serial Communications

The most common type of serial communications is asynchronous. As its name implies, asynchronous communication takes place between two devices without use of a clock line. Data is transmitted out the transmit buffer and received into the receive buffer independently at a speed determined by the baud rate generator. Most microcontrollers offer several modes of asynchronous serial communication.

Standard Asynchronous Mode. The standard asynchronous mode consists of 10 bits: a start bit, eight data bits (LSB first), and a stop bit, as shown in Fig. 11.43. After the user initiates a transmission, data is automatically transmitted from the TX pin at the specified baud rate.

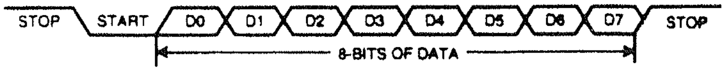


FIGURE 11.43 Standard asynchronous mode data frame.

A parity function is also implemented, which provides for a simple method of error-detection. Data transmitted will consist of either an odd or even number of logical "1"s. If even parity is enabled, the parity bit will either be set to a "1" or a "0" to make the number of "1"s in the data byte even. If odd parity is enabled, the parity bit will be set to the appropriate value to make the number of "1"s in the data byte odd. For instance, consider the data byte 11010010b. If even parity is enabled, the parity bit will be set to a "0" since there is already an even number of "1"s. If odd parity were enabled, the parity bit would be set to a "1" since another "1" would be needed to provide an odd number of "1"s. If the parity function is enabled (usually through a serial port control register), the parity bit is sent instead of the eighth data bit and parity is checked on reception. The occurrence of parity errors is typically flagged in a serial port status register to alert the microcontroller to corrupted data in the receive register.

Multiprocessor Asynchronous Serial Communications Modes. Two other common serial communications modes which are used on automotive microcontrollers are the asynchronous 9th-bit recognition mode and the asynchronous 9th-bit mode. These two modes are commonly used together for multiprocessor communications where selective selection on a data link is required. Both modes are similar to the standard asynchronous mode with the exception of an additional ninth data bit in the data frame as shown in Fig. 11.44.

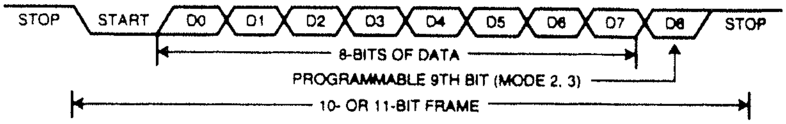


FIGURE 11.44 Asynchronous 9th-bit data frame.

The 9th-bit recognition mode consists of a start bit, nine data bits (LSB first), and a stop bit. For transmission, the ninth bit can be set to "1" by setting a corresponding bit in the serial port control register before writing to the transmit buffer. During reception, the receive interrupt bit is *not* set unless the ninth data bit being received is set to a logic "1".

The 9th-bit mode uses a data frame identical to that of the 9th-bit recognition mode. In this mode, a reception will always cause a receive interrupt, regardless of the state of the ninth data bit.

A multiprocessor data link is fairly simple to implement using these two modes. Microcontrollers within the system are connected as shown in Fig. 11.45. The master microcontroller is set to the 9th-bit recognition mode so that it is always interrupted by serial receptions. The slave microcontrollers are set to operate in the 9th-bit recognition mode so that they are interrupted on receptions only if the ninth data bit is set. Two types of data frames are used: address frames, which have the ninth bit set, and data frames, which have the ninth bit cleared. When the master processor wants to transmit a block of data to one of several slaves, it first sends out an address frame which identifies the target slave. Slaves in the 9th-bit recognition mode are not interrupted by a data frame, but an address frame interrupts all slaves. Each slave can examine the received byte and see if it is being addressed. The addressed slave then switches to the 9th-bit mode to receive data frames, while the slaves that were not addressed stay in the 9th-bit recognition mode and continue without interruption.

11.6 ANALOG-TO-DIGITAL CONVERTER

Analog-to-digital converter (A/D) peripherals allow automotive microcontrollers to sense and assign digital values to analog input voltages with considerable accuracy. An analog input

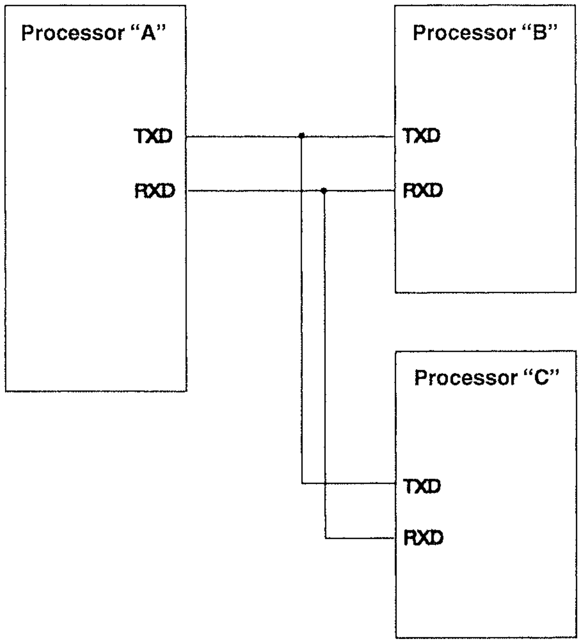


FIGURE 11.45 Asynchronous 9th-bit data frame.

may be defined as a voltage level that varies over a continuous range of values as opposed to the discrete values of digital signals.

11.6.1 Types of A/D Converters

The vast majority of A/D converters available on microcontrollers are of the successive approximation (S/A) type. Other types include flash A/D converters, in which conversions are completed in a parallel fashion and are performed at speeds measuring tens-of-nanoseconds. The drawback is that flash A/D converters require a great deal of die space when integrated on a microcontroller. It is because of their relatively large size that flash A/D converters are seldom offered on microcontrollers. Dual-slope A/D converters offer excellent A/D accuracy but typically take a relatively long period of time to complete a conversion. S/A A/D converters are very popular because they offer a compromise among accuracy, speed, and die-size requirements. The main drawback to successive approximation converters is that implementing the capacitor and resistor ladders takes a considerable amount of die space, although somewhat less than flash A/Ds. These converters are also somewhat susceptible to noise, although there are proven ways to reduce the effects of noise within a given application. The advantage of S/A converters is that they combine the best of other types of converters. They are relatively fast and do not take up excessive die space.

S/A converters typically consist of a resistor ladder, a sample capacitor, an input multiplexer, and a voltage comparator. A typical S/A converter is shown in Fig. 11.46. The resistor ladder is used to produce reference voltages for the input voltage comparison. A sample capacitor is utilized to capture the input voltage during a given period of time known as the sample time. Sample time can be defined as the amount of time that an A/D input voltage is applied to the sample capacitor.

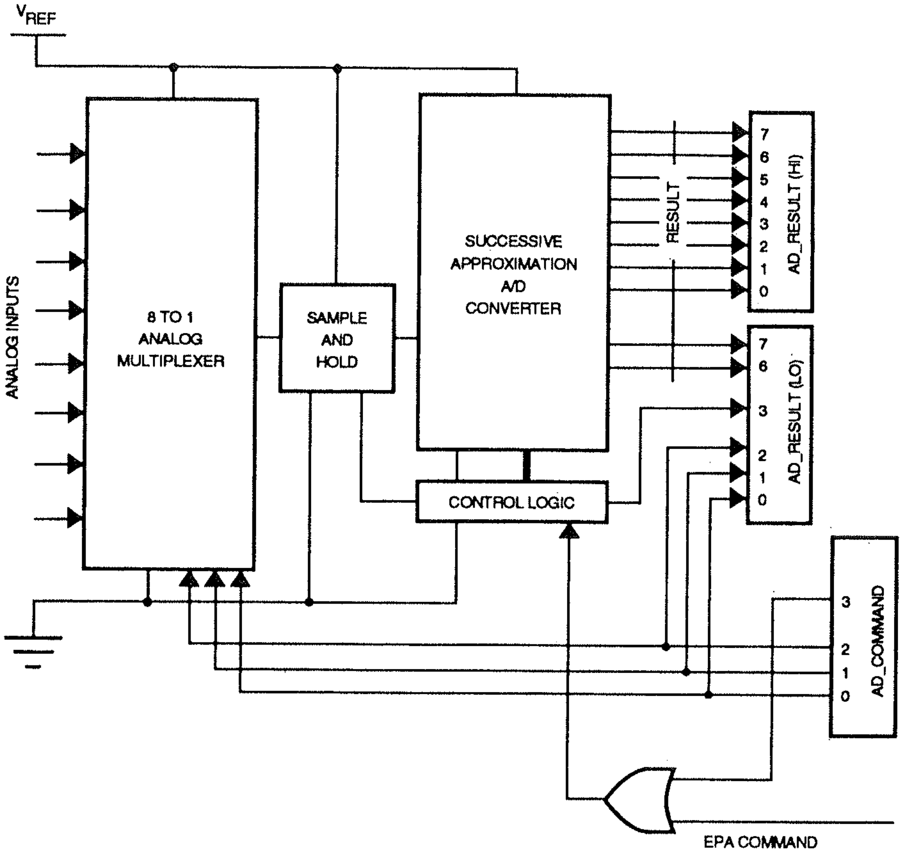


FIGURE 11.46 Typical successive approximation converter.

A successive approximation algorithm is used to perform the A/D conversion. A typical S/A converter consists of a 256-resistor ladder, a comparator, coupling capacitors, and a 10-bit successive approximation register (SAR), along with SFRs and logic to control the process. The resistor ladder provides 20-mV steps (with $V_{ref} = 5.12$ V), while capacitive coupling creates 5-mV steps within the 20-mV ladder voltages. Therefore, 1024 internal reference voltage levels are available for comparison against the analog input to generate a 10-bit conversion result. Eight-bit conversions use only the resistor ladder, providing 256 levels.

11.6.2 The A/D Conversion Process

The successive approximation conversion compares a reference voltage to the analog input voltage stored in the sampling capacitor. A binary search is performed for the reference voltage that most closely matches the input. The ½ full-scale reference voltage is the first tested. This corresponds to a 10-bit result in which the most significant bit is zero and all other bits are one (0111 1111 11b). If the analog input is less than the test voltage, bit 10 is left at zero and a new test voltage of ¼ full scale (0011 1111 11b) is tested. If this test voltage is less than the analog input voltage, bit 9 of the SAR is set and bit 8 is cleared for the next test (0101 1111

11b). This binary search continues until 8 or 10 tests have occurred, at which time the valid 8-bit or 10-bit result resides in the SAR where it can be read by software.

11.6.3 A/D Interfacing

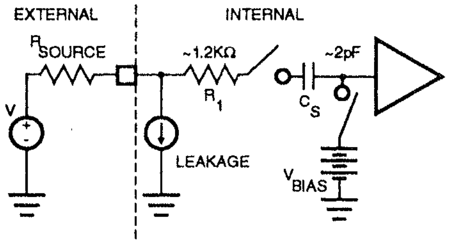


FIGURE 11.47 Idealized interface circuitry.

The external interface circuitry to an analog input is highly dependent upon the application and can impact converter characteristics. Several important factors must be considered in the external interface design: input pin leakage, sample capacitor size, and multiplexer series resistance from the input pin to the sample capacitor. These factors are idealized in Fig. 11.47.

The following example is for a 1-μs sample time and a 10-bit conversion. The external input circuit must be able to charge a sample capacitor (C_S) through a series resistance (R_1) to an accurate voltage, given a dc leakage (I_L). For purposes of this example, assume C_S of 2 pf, R_1 of 1.2 kΩ, and I_L of 1 μA.

External circuits with source impedances of 1 kΩ or less can maintain an input voltage within a tolerance of about 0.2 LSB (1.0 kΩ × 1.0 μA = 1.0 mV) given the dc leakage. Source impedances above 5 kΩ can result in an external error of at least one LSB due to the voltage drop caused by the 1-μA leakage. In addition, source impedances above 25 kΩ may degrade converter accuracy because the internal sample capacitor will not charge completely during the sample time.

Typically, leakage is much lower than the maximum specification specified by the microcontroller manufacturer. Given typical leakage, source impedance may be increased substantially before a one-LSB error is apparent. However, a high source impedance may prevent the internal sample capacitor from fully charging during the sample window. This error can be calculated using the following formula:

$$\text{Error (LSBs)} = \left(e^{\frac{-T_{\text{SAM}}}{RC}} \right) \times 1024$$

where T_{SAM} = sample time, μs
$R = R_{\text{SOURCE}} + R_1, \Omega$
$C = C_S$, μf

The effects of this error can be minimized by connecting an external capacitor C_{EXT} from the input pin to ANGND. The external signal will charge C_{EXT} to the source voltage. When the channel is sampled, a small portion of the charge stored in C_{EXT} will be transferred to the internal sample capacitor. The ratio of C_S to C_{EXT} causes the loss in accuracy. If C_{EXT} is .005 μf or greater, the maximum error will be –0.6 LSB.

Placing an external capacitor on each analog input also reduces the sensitivity to noise because the capacitor combines with series resistance in the external circuit to form a low-pass filter. In practice, one should include a small series resistance prior to the external capacitor on the analog input pin and choose the largest capacitor value practical, given the frequency of the signal being converted. This provides a low-pass filter on the input, while the resistor also limits input current during overvoltage conditions.

11.6.4 Analog References

To achieve maximum noise isolation, on-chip A/D converters typically separate the internal A/D power supply from the rest of the microcontroller's power supply lines. Separate supply pins, V_{ref} and An_{gnd}, usually supply both the reference and digital voltages for the A/D converter. Keep in mind that V_{ref} and An_{gnd} are the reference for a large resistor ladder on successive approximation converters. Any variation in these supplies will directly affect the reference voltage taps within the ladder, which in turn directly affect A/D conversion accuracy.

If the on-chip A/D converter is not being used, or if accuracy is not a concern, the Vref and Angnd pins can simply be connected to V_{cc} and V_{ss}, respectively. However, since the reference supply levels strongly influence the absolute accuracy of the A/D converter, a precision, well-regulated reference should be used to supply V_{ref} to achieve the highest performance levels. It is also important to use bypass capacitors between V_{ref} and An_{gnd} to minimize any noise that may be present on these supplies. In noise-sensitive applications running at higher frequencies, the use of separate ground planes within the PCB (circuit board) should be considered, possibly as shown in Fig. 11.48. This will help minimize ground loops and provide for a stable A/D reference.

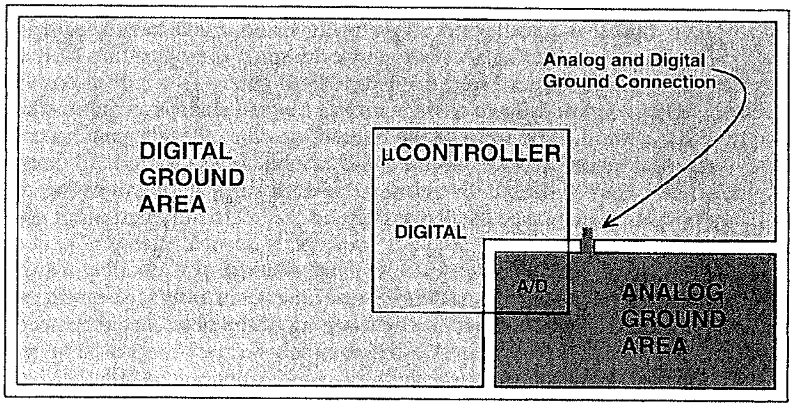


FIGURE 11.48 Example of separate analog and digital ground planes.

11.7 FAILSAFE METHODOLOGIES

The amount and complexity of automotive electronics incorporated into automobiles has increased at an incredible rate over the last decade. This trend has contributed significantly towards the impressive safety record of modern automobiles. Although microcontrollers are extremely reliable electronic devices, it is possible for failures to occur, either elsewhere in the module or within the microcontroller itself. It is critical that these failures be detected and responded to as quickly as possible in safety-related applications such as automotive antilock braking. If proper failsafe methodologies and good programming practices are followed, the chances of a failure going undetected are drastically reduced. The application of *failure mode and effect analysis* (FMEA) is an excellent tool for identifying potential failure modes, detection strategies, and containment methods. Used properly, FMEA will assist the designer in providing a high-quality, reliable automotive module. Although the scope of this chapter does not provide for a discussion on this topic, the author highly encourages the use of FMEA.

11.7.1 Hardware Failsafe Methods

Sometimes a hardware solution is required for detection of and response to certain failure modes. It is difficult for software alone to detect failures external to the device. As an example, consider a case in which electrical overstress (EOS) [illegible] read or drive an incorrect value. In this case, it can be difficult for software to detect because it would base its response on an incorrect value read from a pin.

Watchdog Timers (WDTs). An on-chip hardware watchdog is an excellent method of detecting failures which otherwise may go undetected. An example of this would be a microcontroller fetching either erroneous address or data (due to noise, etc.) and becoming "lost." WDTs commonly utilize a dedicated 16-bit counter, which provides for a count of 2^{16}(65,536) clocked at a rate of one tick per state time. If users wish to take advantage of this feature, they simply write to a register to enable the count. Once enabled, the user program must periodically clear the watchdog by writing a specific bit pattern to the Watchdog SFR. Clearing the WDT at least every 4.1 ms (65,535 * 1 state time at 16 MHz) will prevent the device from being reset. The strategy is that if the WDT initiates a reset, the assumption can be made that a failure has occurred and the microcontroller has became lost.

External Failsafe Devices. It is common for systems to incorporate an external failsafe device, such as another microcontroller or an *application-specific integrated circuit* (ASIC). The function of a failsafe device is to monitor the operation of the primary microcontroller and determine if it is operating properly.

The simplest failsafe devices output a signal such as a square wave for the microcontroller to detect and respond to. If the microcontroller doesn't respond correctly, a reset is typically asserted by the failsafe and the ECU reverts to a safe mode of operation. More complex failsafe devices will actually monitor several critical functions for failures such as low Vcc, stopped or decreased oscillator frequency, shorted/opened input signals, and so forth.

Oscillator Failure Detection. It is possible for the clocking source (typically an oscillator) to fail for various reasons. Since most microcontrollers are static devices, a particularly difficult failure mode to detect is the clocking of the device at a reduced frequency. To detect this failure, an *oscillator failure detection* circuit is often integrated upon the microcontroller. This circuitry will detect if the oscillator clock input signal falls below a specified frequency, in which case an interrupt will be generated or the device will reset itself.

Redundancy/Cross-checking. A common failsafe methodology is achieved by designing a redundant, or backup, processor into the module. In this case, the secondary microcontroller usually executes a subset of the main microcontroller's code. The secondary microcontroller typically processes critical input data and performs cross-checks periodically with the main microcontroller to insure proper operation. A failsafe routine is initiated if data exchanged between the two devices did not correlate.

11.7.2 Software Failsafe Techniques

Failsafe methodologies implemented in software are ideal for detecting failure modes that can interfere with proper program flow. Examples of these types of failures include noise glitches, which are notorious for causing external memory systems to fetch invalid addresses. ROM/EPROM memory corruption could cause an ISR start address to be fetched from an invalid interrupt vector location. Interrupts occurring at a rate faster than anticipated can cause problems such as an overflowing stack. Fortunately, failure modes such as these can be dealt with by implementing software failsafe methods. It is simply good programming practice to anticipate these types of failure modes and provide a failsafe strategy to deal with them. Following are several software strategies commonly used to deal with specific types of failure modes:

Checksum. One possible error that must be accounted for is ROM/EPROM memory corruption. An effective method of detecting these types of failures is through the calculation of a checksum during the initialization phase of a user's program. A checksum is the final value obtained as the result of performing some arithmetic operation upon every ROM/EPROM memory location. The obtained checksum is then compared against a stored checksum. If the two match, the ROM/EPROM contents are intact. An error routine is called if the two checksums do not match. The most common arithmetic operation used to perform a checksum is addition. The checksum is calculated by adding the contents of all memory locations. When the addition is performed, the carry is ignored which provides for a byte or word checksum. The final result is then used as the checksum.

Unused Interrupt Vectors. It is a rare occasion when all interrupt sources are enabled within an application. If, for some unforeseen reason, the program should vector to an unused interrupt source, some sort of failsafe routine should be implemented to respond to the failure. The failsafe routine could be as simple as vectoring to a reset instruction or it can be as complicated as the programmer wishes.

Unused Memory Locations. A strategy should be in place to detect if, for some unforeseen reason, the program sequence should begin to execute in an unused area of ROM/EPROM. It is uncommon for the user's code to fill the entire ROM/EPROM array of a microcontroller. It is good programming practice to fill any unused locations with the opcode of an instruction such as *Reset.* On the MCS-96 family, executing the opcode FFh (which happens to be the blank state of EPROM) will initiate a reset sequence. Other microcontroller families have similar instructions.

Unimplemented Opcode Interrupt Vectors. Microcontrollers often dedicate one or more interrupt vectors for failsafe purposes. An *unimplemented opcode* interrupt is designed to detect corrupted instruction fetches. The corresponding interrupt service routine is executed whenever an unsupported opcode is fetched for execution. The interrupt service routine contains the user's failsafe routine, which is tailored to address this failure for the specific application.

11.8 FUTURE TRENDS

There are several significant trends developing in automotive electronics as ECU manufacturers strive to meet the challenges of a demanding automotive electronics market. The challenges that are bringing about these trends are: decreasing cost targets, decreasing form-factor goals, increasing performance requirements, and increasing system-to-system communication requirements. As the most significant component of an ECU, microcontrollers are bearing the brunt of these demands. This section will discuss these challenges and provide some insight into some of the ways microcontroller manufacturers are addressing these trends.

11.8.1 Decreasing Cost Targets

Microcontroller manufacturers are aproaching cost reduction in two ways: indirectly and directly. *Indirect* cost reductions are achieved by integrating features onto the microcontroller which allow the system designer to reduce cost elsewhere in the system. The key to this approach being successful is in the microcontroller manufacturer's ability to integrate the feature cheaper than the cost of providing an external solution. Integration is not always the cheaper solution, therefore each feature must be evaluated individually to determine the feasability of integration. An example of an indirect cost reduction would be the integration of watchdog and failsafe functions onto the microcontroller. This would eliminate the need for external watchdog components and thus reduce cost.

Another example would be through the integration of communications protocols such as CAN (Controller Area Network) or J1850 onto the same piece of silicon as the microcontroller. This will reduce the system chip count (and thus cost) by at least one integrated circuit device (the CAN chip) and several interfacing components. In most cases, a reduced chip count will translate into a PCB size decrease and a cost savings.

By *directly* addressing decreasing cost targets, microcontroller manufacturers actually reduce the manufacturing cost of the microcontroller itself. An example of this would be utilizing smaller geometry processes for manufacturing. Process geometry refers to the transistor channel width that is implanted onto a piece of silicon for a given fabrication process. Smaller processes allow for a higher transistor density on an integrated circuit. Higher densities allow for smaller die sizes which relate to lower costs. Most automotive microcontrollers manufactured today are fabricated with a 1.0-micron, or larger, process. As technology advances, future automotive microcontrollers will be manufactured upon submicron processes, such as 0.6 micron.

11.8.2 Increasing Performance Requirements

Automotive applications, such as ABS and engine control, require the processing of a substantial amount of data within a limited period of time. Higher-performance microcontrollers are required as system complexity increases and new features, such as traction control and vehicle dynamics, are incorporated into the ECU.

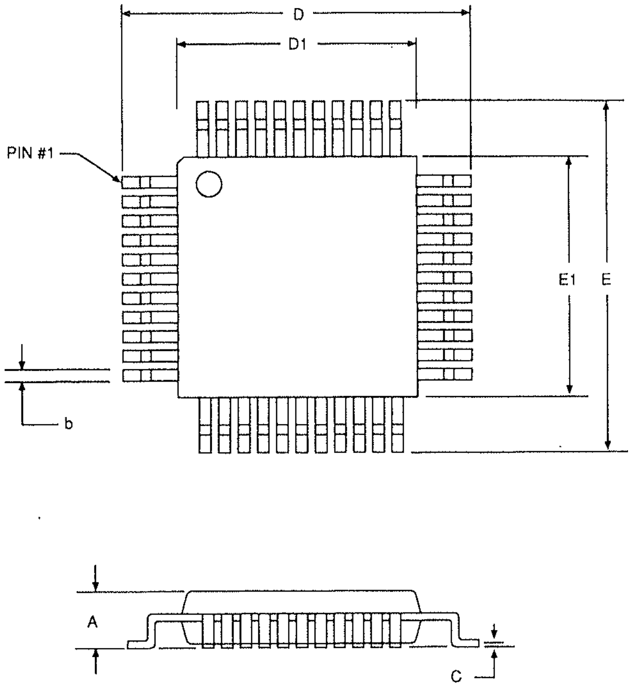
Microcontroller performance can be directly related to speed. Therefore, a rather straightforward approach to increased performance is through increasing clock speed. Today, most automotive microcontrollers have the capability to operate at frequencies of 16 MHz with speeds up to 20 MHz becoming common. Future microcontrollers will have the ability to be operated at frequencies of 24 or even 32 MHz. This allows more code to be executed in the same amount of time, and thus improves performance.

The method of increasing performance is not limited to just increasing the clock frequency. Microcontrollers can also achieve higher performance by enhancing existing peripherals for more efficient operation. This may be in the form of improved data handling or new features which suit the needs of a specific automotive application.

11.8.3 Increasing System-to-System Communication Requirements

The increasing complexity of automotive electronics requires that an increasing amount of information (diagnostics, etc.) be shared between various ECUs within an automobile. To fulfill this need, high-speed data links are utilized to transfer messages between multiple ECUs utilizing protocols such as Bosch's Controller Area Network (CAN) and SAE's J1850. To provide further size and cost savings, it is becoming more and more common to see these protocols supported or integrated onto automotive microcontrollers as opposed to separate integrated circuits.

The theory of centralized body computing is also receiving a closer look due to increased government regulations concerning fuel economy and diagnostics. A centralized body computer would link all ECUs (ABS and traction control, engine, transmission, suspension, instrumentation, etc.) together over a high-speed, in-vehicle serial network. One common scenario would have the central computer (possibly a microprocessor as opposed to a microcontroller) performing the more intense data-crunching tasks, while peripheral microcontrollers located in each individual ECU would perform system I/O functions. These communication protocols provide for efficient two-wire, high-speed serial communications between multiple ECUs utilizing protocols such as CAN and J1850. Supporting these protocols places additional loading upon the microcontroller. Increased microcontroller performance is necessary to manage this loading.



SHRINK QUAD FLATPACK				
SYMBOL	DESCRIPTION	MIN.	NOM.	MAX.
N	Lead Count	80		
A	Overall Height			1.66
A1	Stand Off	0.00	-	
b	Lead Width	0.14	0.20	0.26
c	Lead Thickness	0.117	0.127	0.177
D	Terminal Dimension	13.70	14.00	14.30
D1	Package Body		12.0	
E	Terminal Dimension	13.70	14.00	14.30
E1	Package Body		12.0	
e1	Lead Pitch	0.40	0.50	0.60
L1	Foot Length	0.35	0.50	0.70
T	Lead Angle	0.0°		10.0°
Y	Coplanarity			0.10

FIGURE 11.49 Shrink quad flat pack (SQFP) package.

11.8.4 Decreasing Form Factor Goals

Automobile manufacturers striving to build compact, more fuel efficient automobiles are putting pressure upon ECU suppliers to build smaller, lighter modules.

ECU size is directly affected by PCB size. The easiest way to achieve a smaller PCB is through integration and utilization of smaller integrated circuit packages. To support this demand, automotive microcontroller manufacturers are beginning to offer smaller, fine-pitch packages. A package commonly used today is the 68-lead plastic leaded chip carrier (PLCC) which has its pins placed on 1.27-mm centers and a body that is 24.3 mm^2. An example of a possible automotive package solution for the future would be the 80-lead shrink quad flat pack (SQFP, Fig. 11.49) which has pins on 0.50-mm centers and a body that is 12.0 mm^2. It is relatively easy to see that the SQFP package offers 12 additional pins in a package that is half the size of the PLCC. This high pin density, fine-pitch packaging allows for a smaller package to be utilized for the same size microcontroller die.

Another technology that is quickly becoming popular for automotive applications is referred to as *multichip modules* (MCMs). An MCM is a collection of unpackaged integrated circuit die (from various manufacturers) which are mounted upon a common substrate and packaged together. The advantage of MCMs is that they require much less PCB space than if the ICs were packaged separately.

GLOSSARY

Accumulator A register within a microcontroller that holds data, particularly data on which arithmetic or logic operations are to be performed.

Arithmetic logic unit (ALU) The part of a microcontroller that performs arithmetic and logic operations.

Analog-to-digital converter An electronic device that produces a digital result that is proportional to the analog input voltage.

Assembly language A low-level symbolic programming language closely resembling machine language.

Central processing unit (CPU) The portion of a computer system or microcontroller that controls the interpretation and execution of instructions and includes arithmetic capability.

EPROM Erasable and programmable read-only memory.

High-speed input/output unit (HSIO) A microcontroller peripheral which has the capability to either capture the time at which a certain input event occurs or create an output event at a predetermined time, both relative to a common clock. HSIO events are configured by the programmer to occur automatically.

Interrupt service routine (ISR) A predefined portion of a computer program which is executed in response to a specific event.

Low-speed input/output The input/output of a digital signal by "manually" reading or writing a register location in software.

Machine language A set of symbols, characters, or signs used to communicate with a computer in a form directly usable by the computer without translation.

Program counter (PC) A microcontroller register which holds the address of the next instruction to be executed.

Program status word (PSW) A microcontroller register that contains a set of boolean flags which are used to retain information regarding the state of the user's program.

Pulse-width modulation (PWM) The precise and timely creation of negative and positive waveform edges to achieve a waveform with a specific frequency and duty cycle.

Random access memory (RAM) A memory device which has both read and write capabilities so that the stored information (write) can be retrieved (reread) and be changed by applying new information to the inputs.

Read-only memory (ROM) A memory that can only be read and not written to. Data is either entered during the manufacturing process or by later programming; once entered, it is unalterable.

Register/arithmetic logic unit (RALU) A component of register-direct microcontroller architectures that allows the ALU to operate directly upon the entire register file.

Serial input/output (SIO) A method of digital communication in which a group of data bits is transferred one at a time, sequentially over a single data line.

Special function register (SFR) A microcontroller RAM register which has a specific, dedicated function assigned to it.

BIBLIOGRAPHY

ASM96 Assembler User's Manual, Intel Corp., 1992.

Automotive Electrics/Electronics, Robert Bosch GmbH, 1988.

Automotive Handbook, Intel Corporation, 1994.

Automotive Handbook, 2d ed., Robert Bosch GmbH, 1986.

Corell, Roger J., "How are semiconductor suppliers responding to the growing demand for automotive safety features?," *Intel Corp.,* 1993.

Davidson, Lee S., and Robert M. Kowalczyk, "Microcontroller technology enhancements to meet ever-increasing engine control requirements," Intel Corp., 1992.

Fink, Donald G., and Donald Christiansen, *Electronics Engineers' Handbook,* 3d ed. McGraw-Hill, 1989.

iC-96 Compiler User's Manual, Intel Corp., 1992.

Introduction to MOSFETS and EPROM Memories, Intel Corp., 1990.

MCS®-51 Microcontroller Family User's Manual, Intel Corp., 1993.

Millman, Jacob, and Arvin Grabel, *Microelectronics,* McGraw-Hill, 1987.

Packaging Handbook, Intel Corporation, 1994.

Ribbens, William B., *Understanding Automotive Electronics,* Howard Sams Company, Carmel, Ind. 1992.

8XC196Kx User's Manual, Intel Corporation, 1992.

8XC196KC/8XC196KD User's Manual, Intel Corp., 1992.

ABOUT THE AUTHOR

David S. Boehmer is currently a senior technical marketing engineer for the Automotive Operation of Intel's Embedded Microprocessor Division located in Chandler, Ariz. He is a member of SAE.

CHAPTER 12
ENGINE CONTROL

Gary C. Hirschlieb, Gottfried Schiller, and Shari Stottler
Robert Bosch GmbH

12.1 OBJECTIVES OF ELECTRONIC ENGINE CONTROL SYSTEMS

The electronic engine control system consists of sensing devices which continuously measure the operating conditions of the engine, an electronic control unit (ECU) which evaluates the sensor inputs using data tables and calculations and determines the output to the actuating devices, and actuating devices which are commanded by the ECU to perform an action in response to the sensor inputs.

The motive for using an electronic engine control system is to provide the needed accuracy and adaptability in order to minimize exhaust emissions and fuel consumption, provide optimal driveability for all operating conditions, minimize evaporative emissions, and provide system diagnosis when malfunctions occur.

In order for the control system to meet these objectives, considerable development time is required for each engine and vehicle application. A substantial amount of development must occur with an engine installed on an engine dynamometer under controlled conditions. Information gathered is used to develop the ECU data tables. A considerable amount of development effort is also required with the engine installed in the vehicle. Final determination of the data tables occurs during vehicle testing.

12.1.1 Exhaust Emissions

Exhaust Components. The engine exhaust consists of products from the combustion of the air and fuel mixture. Fuel is a mixture of chemical compounds, termed hydrocarbons (HC). The various fuel compounds are a combination of hydrogen and carbon. Under perfect combustion conditions, the hydrocarbons would combine in a thermal reaction with the oxygen in the air to form carbon dioxide (CO_2) and water (H_2O). Unfortunately, perfect combustion does not occur and in addition to CO_2 and H_2O, carbon monoxide (CO), oxides of nitrogen (NO_x), and hydrocarbons (HC) occur in the exhaust as a result of the combustion reaction. Additives and impurities in the fuel also contribute minute quantities of pollutants such as lead oxides, lead halogenides, and sulfur oxides. In compression ignition (diesel) engines, there is also an appreciable amount of soot (particulates) created. Federal statues regulate the allowable amount of HC, NO_x, and CO emitted in a vehicle's exhaust. On diesel engines, the amount of particulates emitted is also regulated.

Spark Ignition Engines

Air/fuel Ratio. The greatest effect on the combustion process, and therefore on the exhaust emissions, is the mass ratio of air to fuel. The air/fuel mixture ratio must lie within a certain range for optimal ignition and combustion. For a spark ignition engine, the mass ratio for complete fuel combustion is 14.7:1; i.e., 14.7 kg of air to 1 kg of fuel. This ratio is known as the stoichiometric ratio. In terms of volume, approximately 10,000 liters of air would be required for 1 liter of fuel. The air/fuel ratio is often described in terms of the excess-air factor known as lambda (λ). Lambda indicates the deviation of the actual air/fuel ratio from the theoretically required ratio:

$$\lambda = \frac{\text{quantity of air supplied}}{\text{theoretical requirement (14.7 for gasoline)}}$$

At stoichiometry: $\lambda = 1$
For a mixture with excess air (lean): $\lambda > 1$
For a mixture with deficient air (rich): $\lambda < 1$

Effect of Air/Fuel Ratio on Emissions

CO emissions. In the rich operating range ($\lambda < 1$), CO emissions increase almost linearly with an increasing amount of fuel. In the lean range ($\lambda > 1$), CO emissions are at their lowest. With an engine operating at ($\lambda = 1$), the CO emissions can be influenced by the cylinder distribution. If some cylinders are operating rich and others lean with the summation achieving $\lambda = 1$, the average CO emissions will be higher than if all cylinders were operating at $\lambda = 1$.

HC emissions. As with CO emissions, HC emissions increase with an increasing amount of fuel. The minimum HC emissions occur at $\lambda = 1.1 \ldots 1.2$. At very lean air/fuel ratios, the HC emissions again increase due to less than optimal combustion conditions resulting in unburned fuel.

NO_x emissions. The effect of the air/fuel ratio on NO_x emissions is the opposite of HC and CO on the rich side of stoichiometry. As the air content increases, the oxygen content increases and the result is more NO_x. On the lean side of stoichiometry, NO_x emissions decrease with increasing air because the decreasing density lowers the combustion chamber temperature. The maximum NO_x emissions occur at $\lambda = 1.05 \ldots 1.1$.

Catalytic Converters. To reduce the exhaust gas emission concentration, a catalytic converter is installed in the exhaust system. Chemical reactions occur in the converter that transform the exhaust emissions to less harmful chemical compounds. The most commonly used converter for a spark ignition engine is the three-way converter (TWC). As the name implies, it simultaneously reduces the concentration of all three regulated exhaust gases: HC, CO, and NO_x. The catalyst promotes reactions that oxidize HC and CO, converting them into CO_2 and H_2O, while reducing NO_x emissions into N_2. The actual chemical reactions that occur are:

$$2CO + O_2 \rightarrow 2CO_2$$

$$2C_2H_6 + 7O_2 \rightarrow 4CO_2 + 6H_20$$

$$2NO + 2CO \rightarrow N_2 + 2CO_2$$

In order for the catalytic converter to operate at the highest efficiency for conversion for all three gases (HC, CO, NO_x), the average air/fuel ratio must be maintained within less than 1 percent of stoichiometry. This small operating range is known as the *lambda window* or *catalytic converter window.* Figure 12.1 is a graph of lambda (λ) versus the exhaust emissions both before and after the catalytic converter. Up to 90 percent of the exhaust gases are converted to less harmful compounds by the catalytic converter.

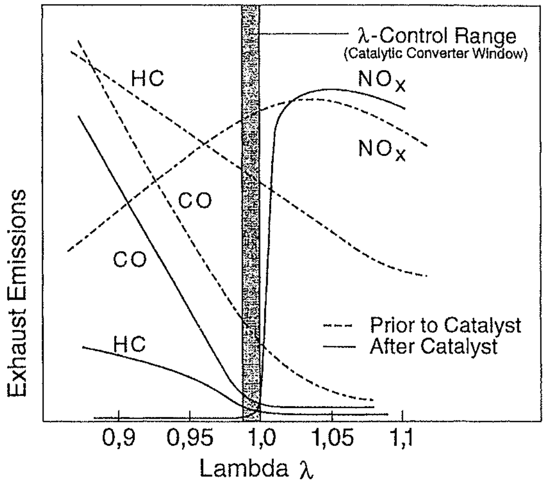


FIGURE 12.1 Lambda effect on exhaust emissions prior to and after catalyst treatment.

To remain within the catalytic converter window, the air/fuel ratio is controlled by the lambda closed-loop fuel control system, which is part of the electronic engine control system. The key component in this system is the lambda sensor. This sensor is installed in the exhaust system upstream of the catalytic converter and responds to the oxygen content in the exhaust gas. The oxygen content is a measure of the excess air (or deficiency of air) in the exhaust gases. A detailed discussion of the lambda closed-loop control system occurs in Sec. 12.2.1.

Ignition Timing. The ignition timing is defined as the crankshaft angle before top dead center (TDC) at which the ignition spark occurs. The ignition timing of the air/fuel mixture has a decisive influence on the exhaust emissions.

Effect of ignition timing on exhaust emissions.

- CO emissions are almost completely independent of the ignition timing and are primarily a function of the air/fuel ratio.
- In general, the more the ignition is advanced, the higher the emissions of HCs. Reactions initiated in the combustion chamber continue to occur after the exhaust valve opens, which depletes the remaining HCs. With advanced timing due to lower exhaust temperatures, these postreactions do not readily occur.
- With increased timing advance, the combustion chamber temperatures increase. The temperature increase causes an increase in NO$_x$ emissions regardless of air/fuel ratio.

To provide the optimal ignition timing for exhaust emissions, precise control of the ignition timing is required. It is imperative that the ignition timing be coordinated with the air/fuel ratio since they have a combined effect on exhaust emissions as well as fuel consumption and driveability. Ignition timing is generally controlled by the ECU. Ignition timing control is discussed in detail in Sec. 12.2.1.

Exhaust Gas Recirculation (EGR). Exhaust gas recirculation (EGR) is a method of reducing emissions of oxides of nitrogen. A portion of the exhaust gas is recirculated back to the combustion chamber. Exhaust gas is an inert gas and, in the combustion chamber, it lowers the peak combustion temperature. Depending on the amount of EGR, NO$_x$ emissions can be reduced by up to 60 percent, although an increase in HC emissions would occur at such high levels of EGR.

Some internal EGR occurs due to the overlap of the exhaust and intake valves. Additional quantities are supplied by a separate system linking the exhaust manifold to the intake manifold. The quantity of EGR flow to the intake system is metered by a pneumatic or electronic valve. The EGR valve is controlled by the ECU. The maximum flow of EGR is limited by an increase in HC emissions, fuel consumption, and engine roughness. EGR control is discussed in detail in Sec. 12.2.1:

Compression Ignition (Diesel) Engines. There are some key distinctions between an SI engine and a CI engine. The CI engine uses high pressure and temperature instead of a spark to ignite the combustible air/fuel mixture. To achieve this, the CI engine compression ratio is in the range of 21:1, as opposed to roughly 10:1 for an SI engine. In a CI engine, the fuel is injected directly into the cylinder near the top of the compression stroke. Mixing of the fuel and air, therefore, occurs directly in the cylinder.

Air/fuel ratio. Diesel engines always operate with excess air ($\lambda > 1$). Where:

$$\lambda = \frac{\text{quantity of air supplied}}{\text{theoretical requirement}}$$

The excess air ($\lambda = 1.1 \ldots 1.2$) reduces the amount of soot (particulates), HC, and CO emissions.

Catalytic Converters. An oxidizing catalyst is used that converts CO and HC to CO_2 and H_2O. The NO_x reduction that occurs for an SI engine three-way catalyst (TWC) is not possible with a diesel because the diesel operates with excess air. The optimal conversion of NO_x requires a stoichiometric ratio ($\lambda = 1$) or a deficiency of air ($\lambda < 1$).

Injection Timing. In a compression ignition engine, the start of combustion is determined by the start of fuel injection. In general, retarding the injection timing decreases NO_x emissions, while overretarding results in an increase in HC emissions. A 1° (crankshaft angle) deviation in injection timing can increase NO_x emissions by 5 percent and HC emissions by as much as 15 percent. Precise control of injection timing is critical. Injection timing on some systems is controlled by the ECU. Feedback on injection timing can be provided by a sensor installed on the injector nozzle. Further discussion on injection timing occurs in Sec. 12.3.1.

Exhaust Gas Recirculation (EGR). As with an SI engine, exhaust gas can be recirculated to the combustion chamber to significantly reduce NO_x emissions. The quantity of EGR allowed to enter the intake is metered by the EGR valve. If the quantity is too high, HC emissions, CO emissions, and soot (particulates) increase as a result of an insufficient quantity of air. The EGR valve is controlled by the ECU, which determines how much EGR is tolerable under the current engine operating conditions.

12.1.2 Fuel Consumption

Federal statutes are currently in effect that require each automobile manufacturer to achieve a certain average fuel economy for all their models produced in one model year. The requirement is known as *corporate average fuel economy* or CAFE. The fuel economy for each vehicle type is determined during the federal test procedure, the same as for exhaust emissions determination, conducted on a chassis dynamometer. Because of the CAFE requirement, it is critical that fuel consumption be minimized for every vehicle type produced.

The electronic engine control system provides the fuel metering and ignition timing precision required to minimize fuel consumption. Optimum fuel economy occurs near $\lambda = 1.1$. However, as discussed previously, lean engine operation affects exhaust emissions and NO_x is at its maximum at $\lambda = 1.1$.

During coasting and braking, fuel consumption can be further reduced by shutting off the fuel until the engine speed decreases to slightly higher than the set idle speed. The ECU determines when fuel shutoff can occur by evaluating the throttle position, engine RPM, and vehicle speed.

The influence of ignition timing on fuel consumption is the opposite of its influence on exhaust emissions. As the air/fuel mixture becomes leaner, the ignition timing must be advanced to compensate for a slower combustion speed. However, as discussed previously, advancing the ignition timing increases the emissions of HC and NO_x. A sophisticated ignition control strategy permitting optimization of the ignition at each operating point is necessary to reach the compromise between fuel consumption and exhaust emissions. The electronic engine control system can provide this sophisticated strategy.

12.1.3 Driveability

Another requirement of the electronic engine control system is to provide acceptable driveability under all operating conditions. No stalls, hesitations, or other objectionable roughness should occur during vehicle operation. Driveability is influenced by almost every operation of the engine control system and, unlike exhaust emissions or fuel economy, is not easily measured. A significant contribution to driveability is determined by the fuel metering and ignition timing. When determining the best fuel and ignition compromises for fuel consumption and exhaust emissions, it is important to evaluate the driveability. Other factors that influence driveability are the idle speed control, EGR control, and evaporative emissions control.

12.1.4 Evaporative Emissions

Hydrocarbon (HC) emissions in the form of fuel vapors escaping from the vehicle are closely regulated by federal statutes. The prime source of these emissions is the fuel tank. Due to ambient heating of the fuel and the return of unused hot fuel from the engine, fuel vapor is generated in the tank. The evaporative emissions control system (EECS) is used to control the evaporative HC emissions. The fuel vapors are routed to the intake manifold via the EECS and they are burned in the combustion process. The quantity of fuel vapors delivered to the intake manifold must be metered such that exhaust emissions and driveability are not adversely affected. The metering is provided by a purge control valve whose function is controlled by the ECU. Further discussion on the operation of the evaporative emissions control system occurs in Sec. 12.2.1.

12.1.5 System Diagnostics

The purpose of system diagnostics is to provide a warning to the driver when the control system determines a malfunction of a component or system and to assist the service technician in identifying and correcting the failure (see Chap. 22). To the driver, the engine may appear to be operating correctly, but excessive amounts of pollutants may be emitted. The ECU determines a malfunction has occurred when a sensor signal received during normal engine operation or during a system test indicates there is a problem. For critical operations such as fuel metering and ignition control, if a required sensor input is faulty, a substitute value may be used by the ECU so that the engine will continue to operate.

When a failure occurs, the malfunction indicator light (MIL), visible to the driver, is illuminated. Information on the failure is stored in the ECU. A service technician can retrieve the information on the failure from the ECU and correct the problem. Detailed examples of system diagnostics are discussed in Sec. 12.2.3.

12.2 SPARK IGNITION ENGINES

12.2.1 Engine Control Functions

Fuel Control. For the purpose of discussing fuel control strategies, a multipoint pulsed fuel injection system is assumed. Additional discussions of fuel control for different types of fuel systems such as carbureters, single-point injection, and multipoint continuous injection appear in Sec. 12.2.4 (Fuel Delivery Systems).

In order for the fuel metering system to provide the appropriate amount of fuel for the engine operating conditions, the mass flow rate of incoming air, known as the air charge, must be determined.

$$F_m = \frac{A_m}{\text{requested air-fuel ratio}}$$

where F_m = fuel mass flow rate
A_m = air mass flow rate

The air mass flow rate can be calculated from:

$$A_m = A_v A_d$$

where A_v = volume flow rate of intake air
A_d = air density

There are three methods commonly used for determining the air charge: speed density, air flow measurement, and air mass measurement. In the speed density method, the air charge is calculated by the engine electronic control unit based on the measurement of air inlet temperature, intake manifold pressure, and engine RPM. The temperature and pressure are used to determine the air density and the RPM is used to determine the volume flow rate. The engine acts as an air pump during the intake stroke. The calculated volume flow rate can be determined as follows:

$$A_{\text{RPM}} = \frac{\text{RPM}}{60} \times \frac{D}{2} \times V_{\text{E}}$$

where RPM = engine speed
D = engine displacement
V_{E} = volumetric efficiency

In an engine using exhaust gas recirculation (EGR), the volume flow rate of EGR must be subtracted from the calculated volume flow rate.

$$A_v = A_{\text{RPM}} - A_{\text{EGR}}$$

The volume flow rate of EGR can be determined empirically based on the EGR valve flow rate and the EGR control strategy being used.

In the air flow measurement method, the air flow is measured using a vane type meter and air density changes are compensated for by an air inlet temperature sensor. The vane meter uses the force of the incoming air to move a flap through a defined angle. This angular movement is converted by a potentiometer to a voltage ratio. Because only the fresh air charge is measured, no compensation is required for EGR.

In the air mass measurement method, the air charge is measured directly using a hot-wire or hot-film air mass flow sensor. The inlet air passes a heated element, either wire or film. The element is part of a bridge circuit that keeps the element at a constant temperature above the inlet air temperature. By measuring the heating current required by the bridge circuit and converting this to a voltage via a resistor, the air mass flow passing the element can be determined. Again, because only the fresh air charge is measured, no compensation for EGR is required. However, sensing errors may occur due to strong intake manifold reversion pulses, which occur under certain operating conditions. In such cases, a correction factor must be determined and applied.

Calculation of Injector Pulse Width. The base pulse width is determined from the required fuel mass flow rate (F_m) and an empirical injector constant. The injector constant is determined by the design of the injector and is a function of the energized time versus the flow volume. This constant is normally determined with a constant differential pressure across the injector (from fuel rail to intake manifold). When the pressure across the injector does not remain constant (i.e., there is no pressure regulator intake manifold vacuum reference), an entire map of injector constants for different manifold pressures may be required.

The effective injector pulse width is a modification of the base pulse width. The base pulse width is adjusted by a number of correction factors depending on operating conditions. For example, a battery voltage correction is required to compensate for the electromechanical characteristics of the fuel injectors. Injector opening and closing rates differ depending on the voltage applied to the injector, which affects the amount of fuel injected for a given pulse width. Other common correction factors may include hot restart, cold operation, and transient operation corrections. Figure 12.2 is a flowchart of a typical injector effective pulse-width calculation method.

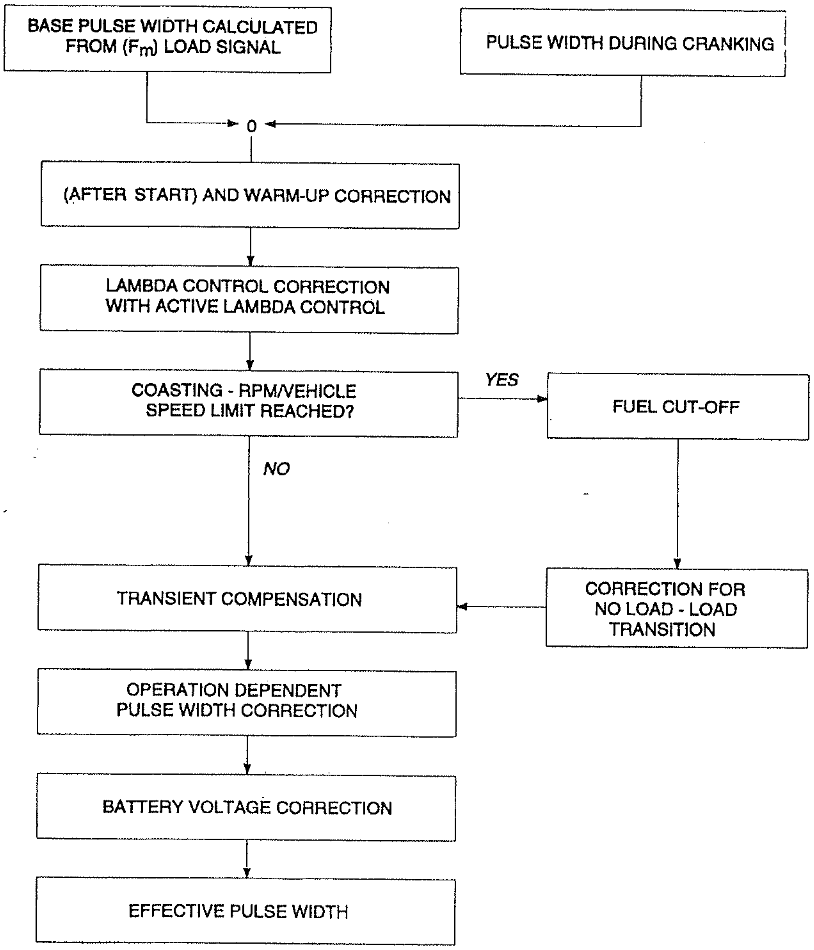

BASE PULSE WIDTH CALCULATED FROM (F_m) LOAD SIGNAL
PULSE WIDTH DURING CRANKING
0
(AFTER START) AND WARM-UP CORRECTION
LAMBDA CONTROL CORRECTION WITH ACTIVE LAMBDA CONTROL
COASTING - RPM/VEHICLE SPEED LIMIT REACHED?
YES
FUEL CUT-OFF
NO
TRANSIENT COMPENSATION
CORRECTION FOR NO LOAD - LOAD TRANSITION
OPERATION DEPENDENT PULSE WIDTH CORRECTION
BATTERY VOLTAGE CORRECTION
EFFECTIVE PULSE WIDTH

FIGURE 12.2 Determination of effective injector pulse width.

Injection Strategies. There are three commonly used fuel injection strategies for multipoint fuel metering systems: simultaneous injection, group injection, and sequential injection. Figure 12.3 is a diagram of the different strategies. Some engines use simultaneous injection during crank and switch over to sequential after the engine is running. This allows for shorter starting times since no synchronization with the camshaft is necessary before fuel injection begins. A description of each strategy follows.

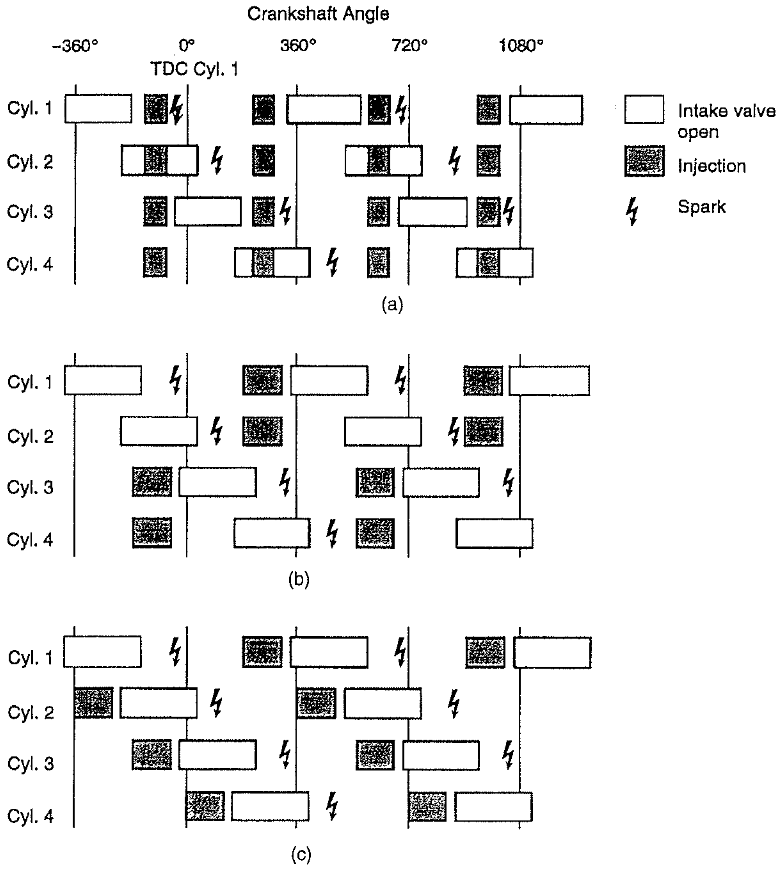


FIGURE 12.3 Fuel injection strategies: (*a*) simultaneous injection, (*b*) group injection, and (*c*) sequential injection.

Simultaneous injection. Injection of fuel occurs at the same time for all cylinders every revolution of the crankshaft. Therefore, fuel is injected twice within each four-stroke cycle. The injection timing is fixed with respect to crank/camshaft position.

Group injection. The injectors are divided into two groups that are controlled separately. Each group injects once per four-stroke cycle. The offset between the groups is one crankshaft revolution. This arrangement allows for injection timing selection that eliminates spraying fuel into an open intake valve.

Sequential injection. Each injector is controlled separately. Injection timing, both with reference to crank/camshaft position and pulse width, can be optimized for each individual cylinder.

Lambda Control. A subsystem of the fuel control system is lambda closed-loop control. Lambda (λ) is defined as the excess-air factor that indicates the deviation of the actual air/fuel ratio from the theoretically required ratio:

$$\lambda = \frac{\text{quantity of air supplied}}{\text{theoretical requirement (14.7 for gasoline)}}$$

The lambda sensor, or exhaust gas oxygen sensor, is installed in the engine exhaust stream upstream of the catalytic converter. The sensor responds to the oxygen content of the exhaust gas. The signal from the lambda sensor serves as feedback to the fuel control system. This provides the fine-tuning needed to remain within the limited catalytic converter window for optimal catalyst performance. (See Sec. 12.1.1 for more discussion on the catalytic converter window.) For a lean mixture ($\lambda > 1$), sensor voltage is approximately 100 mV. For a rich mixture ($\lambda < 1$), the sensor voltage is approximately 800 mV. At roughly $\lambda = 1$ (a stoichiometric mixture), the sensor switches rapidly between the two voltages. The input from the lambda sensor is used to modify the base pulse width to achieve $\lambda = 1$.

Lambda closed-loop control requires an operationally ready lambda sensor, typically one which has reached an operating temperature threshold. Sensor output is monitored by the ECU to determine when the sensor is supplying usable information. An active sensor signal, along with other requirements, such as engine temperature, must be achieved before lambda closed-loop control will be activated.

Under steady state conditions, the lambda control system oscillates between rich and lean around the lambda window. As the lambda sensor switches, the injector pulse width is adjusted by the amount determined by a control factor until the lambda sensor switches again to the opposite condition. The control factor can be defined as the allowable increase or decrease in the commanded fuel injector pulse width. The frequency of oscillation is determined by the gas transport time and the magnitude of the control factor. The gas transport time is defined as the time from air/fuel mixture formation to lambda sensor measurement.

Under transient conditions, the gas transport time results in a delay before the lambda sensor can indicate that the operating conditions have changed. Using only the lambda sensor for closed-loop fuel control would result in poor driveability and exhaust emissions because of this delay. Therefore, the engine control unit uses an anticipatory control strategy that uses engine load and RPM to determine the approximate fuel requirement. The engine load information is provided by the manifold pressure sensor for speed density systems and by the air meter for air flow and air mass measurement systems and by the throttle valve position sensor. The engine control unit contains data tables for combinations of load and RPM. This allows for rapid response to changes in operating conditions. The lambda sensor still provides the feedback correction for each load/RPM point. The data used for these data tables are largely developed from system modeling and engine development testing.

Due to production variations in engines, variations in fuel and changes due to wear and aging, the control system must be able to adapt to function properly for every engine over the engine's life. Therefore, the electronic control unit has a feature for adapting changes in the fuel required for the load/RPM points. At each load/RPM point, the lambda sensor continuously provides information that allows the system to adjust the fuel to the commanded A/F ratio. The corrected information is stored in RAM (random access memory) so that the next time the engine reaches that operating point (load/RPM), the anticipatory value will require less correction. These values remain stored in the electronic control unit even after the engine is shut off. Only if power to the electronic control unit is disrupted (i.e., due to a dead battery), will the correction be lost. In that case, the electronic control unit will revert back to the original production values that are written in ROM (read-only memory).

Lambda sensors do not switch symmetrically from lean to rich and rich to lean. Because of this, the control strategy is modified to account for the asymmetry. This can be accomplished either by delaying the modification by the control factor after the sensor switches or by using control factors of different magnitudes for rich-to-lean and lean-to-rich switching.

Ignition Timing Control. The goal of the engine control system for ignition timing is to provide spark advance which optimizes engine torque, exhaust emissions, fuel economy, and driveability, and which minimizes engine knock. Data tables with the base ignition timing, depending on engine load and RPM, are stored in ROM in the electronic control unit. The values in these tables are optimized for fuel economy, exhaust emissions, and engine torque. They are developed through engine experimentation, usually with an engine dynamometer. Corrections to the base timing values are needed for temperature effects, EGR, hot restart, barometric pressure, and engine knock. In addition, some systems use ignition timing to vary the engine torque for improvement in automatic transmission shift quality or for idle speed control. Figure 12.4 is a flowchart of a typical ignition timing calculation method.

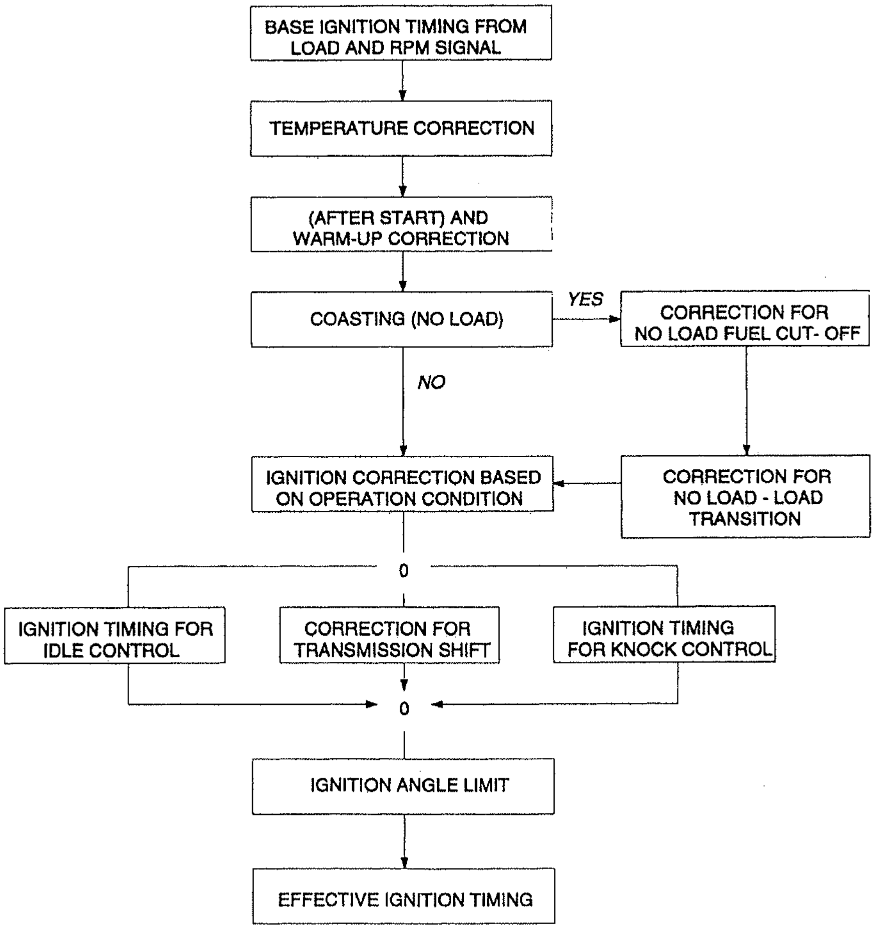

BASE IGNITION TIMING FROM LOAD AND RPM SIGNAL
TEMPERATURE CORRECTION
(AFTER START) AND WARM-UP CORRECTION
COASTING (NO LOAD)
YES
CORRECTION FOR NO LOAD FUEL CUT- OFF
NO
IGNITION CORRECTION BASED ON OPERATION CONDITION
CORRECTION FOR NO LOAD - LOAD TRANSITION
0
IGNITION TIMING FOR IDLE CONTROL
CORRECTION FOR TRANSMISSION SHIFT
IGNITION TIMING FOR KNOCK CONTROL
0
IGNITION ANGLE LIMIT
EFFECTIVE IGNITION TIMING

FIGURE 12.4 Determination of effective ignition timing.

Dwell Angle Control. The dwell angle performance map stored in the electronic control unit controls the charging time of the ignition coil, depending on RPM and battery voltage. The dwell angle is controlled so that the desired primary current is reached at the end of the charging time just prior to the ignition point. This assures the necessary primary current, even with quick transients in RPM. A limit on the charge time in the upper RPM ranges allows for the necessary spark duration.

Knock Control. The ignition timing for optimization of torque, fuel economy, and exhaust emissions is in close proximity to the ignition timing that results in engine knock. Engine knock occurs when the ignition timing is advanced too far for the engine operating conditions and causes uncontrolled combustion that can lead to engine damage, depending on the severity and frequency. If a factor of safety was used when developing the base timing map for all conditions that contribute to knock, such as fuel quality and variations in compression ratio, the ignition timing would be significantly retarded from the optimum level, resulting in a significant loss in torque and fuel economy. To avoid this, a knock sensor (one or more) is installed on the engine to detect knocking (see Chap. 8). Knock sensors are usually acceleration sensors that provide an electric signal to the electronic control unit. From this signal, the engine control unit algorithm determines which cylinder or cylinders are knocking. Ignition timing is modified (retarded) for those cylinders until the knock is no longer detected. The ignition timing is then advanced again until knock is detected. (See Fig. 12.5.) Information on the amount of spark retard required to eliminate the knock for each cylinder under each load/RPM condition is saved in the electronic control unit RAM. This allows for quick access to the appropriate "learned" ignition timing for each condition. With this control system, the base timing can be more advanced for improved fuel economy and torque.

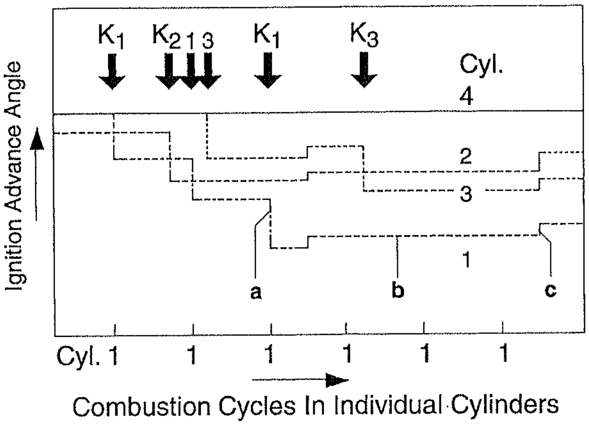


FIGURE 12.5 Knock control. Control algorithm for ignition adjustments for a four-cylinder engine. $K_{1...3}$ (knock-in cylinders 1 . . . 3), cylinder number four (no knock), (*a*) (ignition retard), (*b*) (delay before return to original point, (*c*) (spark advance).

Evaporative Emissions Control. Hydrocarbon (HC) emissions in the form of fuel vapors escaping from the vehicle, primarily from the fuel tank, are closely regulated by federal statutes. There are two principal causes of fuel vapor in the fuel tank: increasing ambient temperature and return of unused hot fuel from the engine. In order to control the release of these emissions to the atmosphere, the evaporative emissions control system was developed.

Evaporative Emissions Control System. A vapor ventilation line exits the fuel tank and enters the fuel vapor canister. The canister consists of an active charcoal element which absorbs the vapor and allows only air to escape to the atmosphere. Only a certain volume of fuel vapor can be contained by the canister. The vapors in the canister must therefore be purged from and burned by the engine so that the canister can continue to store vapors as they are generated. To accomplish this, another line leads from the charcoal canister to the intake manifold. Included in this line is the canister purge solenoid valve. Figure 12.6 shows a layout of a typical evaporative emissions control system.

During engine operation, vacuum in the intake manifold causes flow through the charcoal canister because the canister vent opening, at the charcoal filter end, is at atmospheric pres-

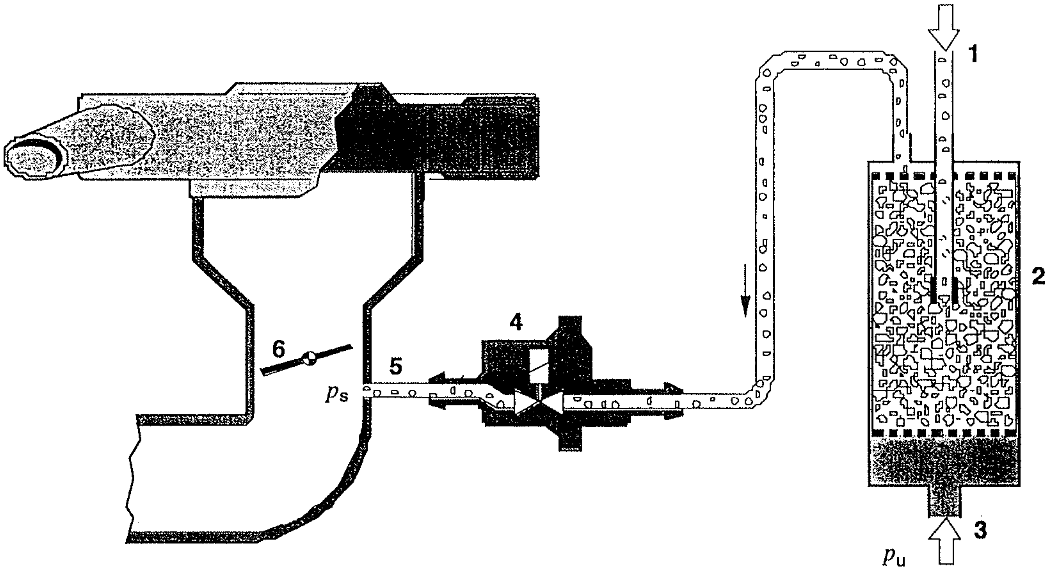


FIGURE 12.6 Evaporative emission control system: fuel vapor from fuel tank (1), charcoal canister (2), ambient air (3), canister purge control valve (4), purge line to intake manifold (5), throttle valve (6), p_s is intake manifold vacuum, and p_u is atmospheric pressure.

sure. The canister purge valve meters the amount of flow from the canister. The amount of fuel vapor in the canister and, therefore, contained in the flow stream, is not known. Therefore, it is critical that the lambda control system is operating and adjusting the fuel requirement as the vapors are being purged. Purge vapors could otherwise result in up to a 30 percent increase in air/fuel mixture richness in the engine.

Purge Valve Control. Control of the purge valve must allow for two criteria:

- There must be enough vapor flow so that the charcoal canister does not become saturated and leak fuel vapors to the atmosphere
- Purge flow must generally occur under lambda closed-loop control so that the effect of the purge vapors on A/F ratio can be detected and the fuel metering corrected

When the electronic control unit commands the purge valve to meter vapor from the canister, it requests a duty cycle (ratio of ON time to total ON and OFF time). This allows the amount of vapor flow to be regulated depending on the engine operating conditions. When lambda control is not operating, only low duty cycles and, therefore, small amounts of purge vapors, are allowed. Under deceleration fuel cutoff, the purge valve is closed entirely to minimize the possibility of unburned HCs in the exhaust.

Turbocharger Boost Pressure Control. The exhaust turbocharger consists of a compressor and an exhaust turbine arranged on a common shaft. Energy from the exhaust gas is converted to rotational energy by the exhaust turbine, which then drives the compressor. The compressed air leaves the compressor and passes through the air cooler (optional), throttle valve, intake manifold, and into the cylinders. In order to achieve near-constant air charge pressure over a wide RPM range, the turbocharger uses a circuit that allows for the bypass of the exhaust gases away from the exhaust turbine. The valve that regulates the bypass opens at a specified air charge pressure and is known as the wastegate.

Engines that have turbochargers benefit significantly from electronic boost pressure control. If only a pneumatic-mechanical wastegate is used, only one boost pressure point for the entire operating range is used to divert the exhaust gas. This creates a compromise for part-load conditions, which results in increased exhaust backpressure, more turbocharger work, more residual exhaust gas in the cylinders, and higher-charge air temperatures.

By controlling the wastegate with a pulse-width modulated solenoid valve, different wastegate opening pressures can be specified, depending on the engine operating conditions (Fig. 12.7). Therefore, only the level of air charge pressure required is developed. The electronic control unit uses information on engine load from either manifold pressure or the air meter and RPM and throttle position. From this information, a data table is referenced and the proper boost pressure (actually a duty cycle of the control valve) is determined. On systems using manifold pressure sensors, a closed-loop control system can be developed to compare the specified value with the measured value.

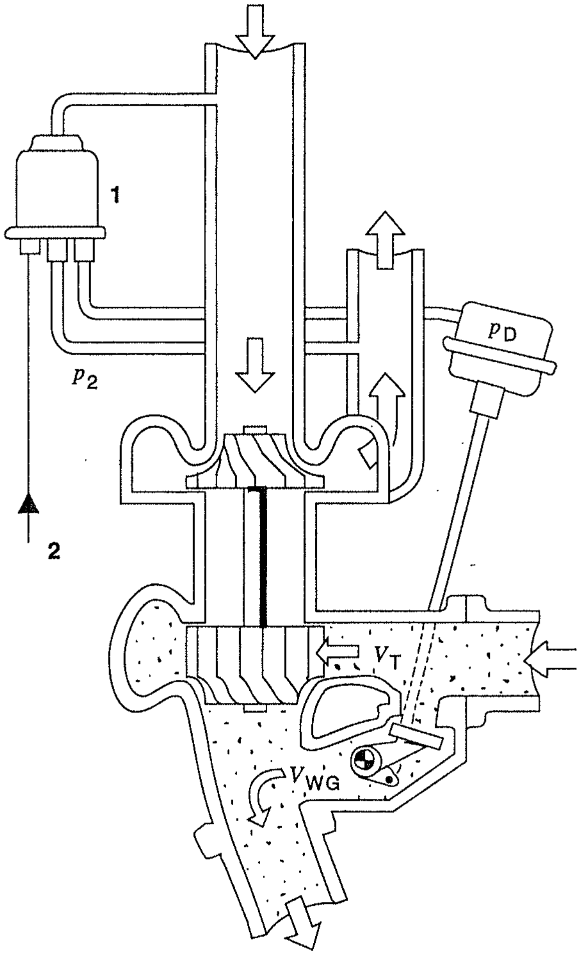


FIGURE 12.7 Electronic turbocharger boost control: solenoid valve (1), control signal from ECU (2), boost pressure (p_D), volume flow through turbine (V_T), volume flow through wastegate (V_{WG}).

The boost pressure control system is usually used in combination with knock control for turbocharged engines. When the ignition timing is retarded due to knock, an increase in already high exhaust temperatures for turbocharged engines occurs. To counteract the temperature increase, the boost pressure is reduced when the ignition timing is retarded past a predetermined threshold.

Engine/Vehicle Speed Control. Using the inputs of engine RPM and vehicle speed to the electronic control unit, thresholds can be established for limiting these variables with fuel cutoff. When the maximum speed is achieved, the fuel injectors are shut off. When the speed decreases below the threshold, fuel injection resumes.

EGR Control. By mixing a portion of the exhaust gas with the fresh intake air/fuel mixture, oxides of nitrogen (NO_x) can be reduced by lowering the peak combustion temperatures. However, the addition of exhaust gas can degrade driveability by causing combustion instability, especially at idle and low speeds and with a cold engine. The ECU references an engine RPM/load table of optimal EGR valve openings. The data table is developed on the engine dynamometer by analyzing the exhaust emissions. With increasing EGR, a point is reached where hydrocarbon (HC) emissions begin to increase. The optimal percent of EGR is just prior to that point.

The electronic control unit regulates a pneumatic- or solenoid-type valve to meter a certain quantity of exhaust gas back to the intake manifold. Typically, an engine coolant temperature threshold is also required before EGR is activated to avoid poor driveability. Under acceleration and at idle, EGR is deactivated.

Camshaft Control. There are two types of camshaft controls: phasing (i.e., overlap or intake/exhaust valve opening point) and valve lift and opening duration.

Camshaft Phasing Control. Valve overlap is a function of the rotation of the intake camshaft with respect to the exhaust camshaft. Overlap can be controlled by an electrohydraulic actuator. At idle and at high RPM, it is desirable to have the intake valves open and close later, which reduces the overlap. For idle, this reduces the residual exhaust gases that return with the fresh charge air and improves idle stability. At high RPM, late closing of the intake valve provides the best condition for maximum cylinder filling and, therefore, maximum output. For partial loads, a large valve overlap, where the intake opens early, is desirable. This allows for an increase in residual exhaust gas for improved exhaust emissions (Fig. 12.8).

Valve Lift and Opening Duration Control. Control of the valve lift and opening duration is accomplished by switching between two camshaft profiles. An initial cam specifies the optimal lift and duration for the low to middle RPM range. A second cam profile controls a higher valve lift and duration for high-RPM operation. By monitoring engine load and RPM, the ECU actuates the electrohydraulic device that switches from one cam profile to the other (Fig. 12.9).

Variable Intake Manifold Control. The goal of the engine design is to achieve the highest possible torque at low engine RPM as well as high output at high engine RPM. The torque curve of an engine is proportional to the air charge at any given engine speed. Therefore, a primary influence on the torque is the intake manifold geometric design. The simplest type of air charging uses the dynamics of the drawn-in air. The standard intake manifolds for multipoint engines consist of several intake runners and collectors converging at the throttle valve.

In general, short intake runners result in a high output at high RPM with a simultaneous loss of torque at low RPM. Long intake runners have the opposite effect. Due to intake valve and piston dynamics, pressure waves occur that oscillate within the intake manifold. Proper selection of runner lengths and collector sizes can result in the pressure waves arriving at the intake valves just before they are closing. This has a supercharging effect. The limitation of this method is that, for a given intake manifold configuration, the tuning peak can only occur at one operating point.

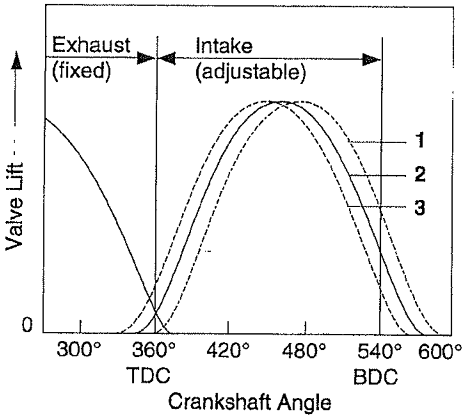


FIGURE 12.8 Adjustment angle for intake camshaft: retard (1), standard (2), advance (3).

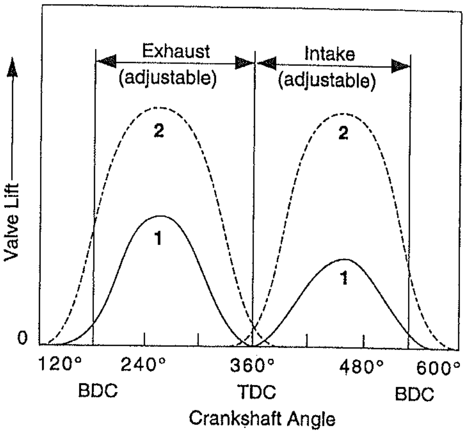


FIGURE 12.9 Selective camshaft lobe actuation: base cam lobe (1), auxiliary cam lobe (2).

Variable Intake Systems. To optimize the benefits of intake manifold charging, several systems have been developed that allow for changes in runner length and collector volume, depending on engine operating conditions. This allows for tuning peaks at more than one operating point. One method developed uses electronically controlled valves to close off areas of the intake manifold (Fig. 12.10). Inputs of engine load, RPM, and throttle angle determine the position of the valves.

12.2.2 Engine Control Modes

Engine Crank and Start. During engine cranking, the goal is to get the engine started with the minimal amount of delay. To accomplish this, fuel must be delivered that meets the requirements for starting for any combination of engine coolant and ambient temperatures. For a cold engine, an increase in the commanded air/fuel ratio is required due to poor fuel vaporization and "wall wetting," which decreases the amount of usable fuel. Wall wetting is the condensation of some of the vaporized fuel on the cold metal surfaces in the intake port and combustion chamber. It is critical that the fuel does not wet the spark plugs, which can reduce the effectiveness of the spark plug and prevent the plug from firing. Should plug wetting occur, it may be impossible to start the engine.

Fuel Requirement. Within the ECU ROM there are specific data tables to establish cold-start fuel based on engine coolant temperature. For two reasons, the lambda sensor output cannot be used during crank: the lambda sensor is below its minimum operating temperature and the air/fuel ratio required is outside the lambda sensor control window.

Many starting sequences use a front-loading strategy for fueling whereby the quantity of fuel is reduced after a speed threshold (RPM) is achieved, after a certain number of revolutions or at a defined time after the initial crank. Some systems also switch over from simultaneous injection to sequential injection after a speed threshold is achieved. For cold temperature starting, the fuel mixture may remain richer than $\lambda = 1$ after starting, due the continuing poor mixture formation in the cold induction system.

Ignition Timing Requirement. Ignition timing is controlled by the ECU during crank and is determined by engine coolant temperature and cranking speed. For a cold engine with low cranking speeds, ideal timing is near TDC. For higher cranking speeds, a slightly more advanced timing is optimal. Timing advance must be limited during cranking to avoid igniting

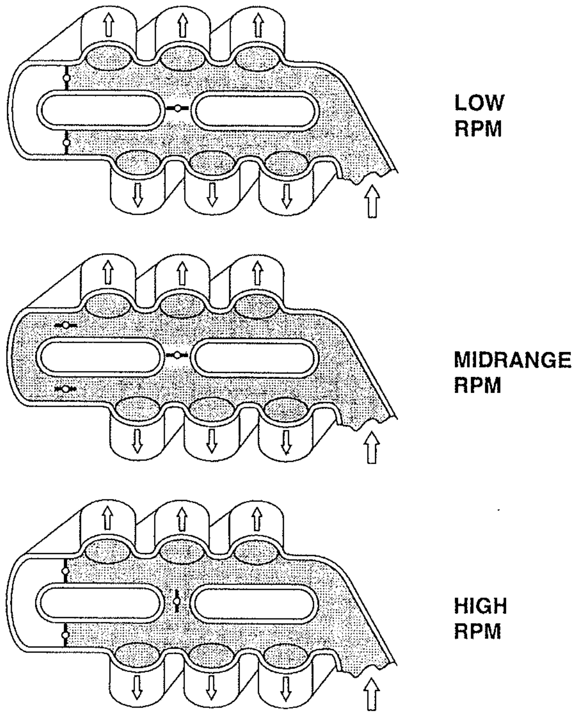


FIGURE 12.10 Variable configuration intake manifold.

the air/fuel mixture before the crankshaft reaches top dead center (TDC). A damaging torque reversal could occur that would damage the starter. After the engine starts, ignition timing is advanced to improve cold engine running as well as to reduce the need for fuel enrichment.

Engine Warm-Up. During the warm-up phase, there are three conflicting objectives: keep the engine operating smoothly (i.e., no stalls or driveability problems), increase exhaust temperature to quickly achieve operational temperature for catalyst (light-off) and lambda sensor so that closed-loop fuel control can begin operating, and keep exhaust emissions and fuel consumption to a minimum. The best method for achieving these objectives is very dependent on the specific engine application.

If the engine is still cold, fuel enrichment will be required to keep the engine running smoothly due, again, to poor fuel vaporization and wall wetting effects. The amount of enrichment is dependent on engine temperature and is a correction factor to the injector pulse width. This enrichment, combined with secondary air injection, also helps achieve the desired increase in catalyst temperature. To provide secondary air injection, an external air pump delivers fresh air downstream of the exhaust valves for a short time after start. The excess air causes oxidation (burning) of the excess HC and CO from the rich mixture in the exhaust manifold, which rapidly increases the temperature of the catalytic converter. The oxidation also removes harmful pollutants from the exhaust stream.

It is possible to increase the exhaust temperature by increasing the idle speed during warm-up. The increased idle speed may also be combined with a slightly retarded ignition timing, which increases temperatures in the exhaust, thereby promoting rapid warm-up of the catalyst.

Transient Compensation. During transitions such as acceleration or deceleration, the objective of the engine control system is to provide a smooth transition from one engine operating condition to another (i.e., no hesitations, stalls, bumps, or other objectionable driveability concerns), and keep exhaust emissions and fuel consumption to a minimum.

Acceleration Enrichment. When an increase in engine load and throttle angle occurs, a corresponding increase in fuel mixture richness is required to compensate for the increased wall wetting. The sudden increase in air results in a lean mixture that must be corrected swiftly to obtain good transitional response. The rate of change of engine load and throttle angle are used to determine the quantity of fuel during acceleration enrichment. The amount of fuel must be enough to provide the desired performance, but not so much as to degrade exhaust emissions and fuel economy.

During acceleration enrichment, the ignition timing is set for maximum torque without knocking. Additionally, when a large change in engine load occurs, some systems delay the ignition timing advance briefly to prevent engine knock, which may arise from a momentary lean mixture or from transient ignition timing errors.

Deceleration Enleanment. During deceleration modes, such as coasting or braking, there is no torque requirement. Therefore, the fuel may be shut off until either an increase in throttle angle is detected or the engine speed falls to a speed slightly above the idle RPM. Fuel shutoff or cutoff can decrease exhaust emissions by eliminating unburned HC and CO and may also improve fuel consumption. Fuel cutoff is also used to protect the catalytic converter from extreme high temperatures during extended overrun conditions. During transition to fuel cutoff, the ignition timing is retarded from its current setting to reduce engine torque and to assist in engine braking. The fuel is then shut off. During the transition, the throttle bypass valve or the main throttle valve may remain open for a short period to allow fresh air to oxidize the remaining unburned HC and CO to further reduce exhaust emissions. During development of the fuel cutoff strategy, the advantage of reduced emission effects and catalyst temperature control must be balanced against driveability requirements. The use of fuel cutoff may change the perceived amount of engine braking felt by the driver. In addition, care must be taken to avoid a "bump" feel when entering the fuel cutoff mode, due to the change in torque.

Full Load. Under steady state full-load conditions, such as for climbing a grade, it is desirable to control the air/fuel mixture and ignition timing to obtain maximum power and to also limit engine and exhaust temperatures. The best engine torque is typically delivered at about $\lambda = 0.9$ to 0.95. When the ECU determines the engine is operating at full load via the throttle valve sensor (at WOT), the commanded air/fuel mixture, if required, can be enriched. The lambda sensor signal cannot be used to provide correction to the air/fuel mixture because the rich operating point lies outside the lambda control window.

The ignition timing at full load is set to achieve the maximum torque without knocking. This initial value is determined through engine dynamometer testing. With a knock control system (see Sec. 12.2.1), the ignition timing is modified (retarded) when engine knock occurs. The modification required to eliminate the knock may be saved in the ECU so that the next time that engine RPM/load point occurs, less knocking will occur and less correction will be required.

Idle Speed Control. The objectives of the engine control system during idle are:

- Provide a balance between the engine torque produced and the changing engine loads, thus achieving a consistent idle speed even with various load changes due to accessories (i.e., air conditioning, power steering, and electrical loads) being turned on and off and during engagement of the automatic transmission. In addition, the idle speed control must be able to compensate for long-term changes in engine load, such as the reduction in engine friction that occurs with engine break-in.
- Provide the lowest idle speed that allows smooth running to achieve the lowest exhaust emissions and fuel consumption (up to 30 percent of a vehicle's fuel consumption in city driving occurs during idling).

To control the idle speed, the ECU uses inputs from the throttle position sensor, air conditioning, automatic transmission, power steering, charging system, engine RPM, and vehicle speed. There are currently two strategies used to control idle speed: air control and ignition control.

Air Control. The amount of air entering the intake manifold is controlled either by a bypass valve or by an actuator acting directly on the throttle valve. The bypass valve uses, for example, an electronically controlled motor controlled by the ECU that opens or closes a fixed amount. For large throttle valves, it may be desirable to use a bypass valve because a small change in throttle angle may result in a large change in air flow and, therefore, idle speed may be difficult to control. Using engine RPM feedback input, the ECU adjusts the air flow to increase or decrease the idle speed. A disadvantage to air control is that the response to load changes is relatively slow. To overcome this, air control is often combined with ignition timing control to provide acceptable idle speed control. The fuel quantity required at idle is determined by engine load and RPM. During closed-loop operation, this value is optimized by the lambda sensor closed-loop control.

Ignition Timing Control. Engine torque may be increased or decreased by advancing or retarding the ignition timing within an established window. This principle can be employed to help control idle speed. Ignition timing control is particularly desirable for responding to idle load changes because engine torque output changes more rapidly in response to a change in ignition timing than to a change in air valve position. Using the same inputs as for air control, the ECU adjusts the spark advance to either raise or lower the idle speed.

Anticipating Accessory Loads. Specific electric inputs to the ECU, such as a pressure switch located in the power steering system, are used to anticipate accessory loads so that the idle control system can compensate more quickly. This "feed forward" strategy allows better idle control than a strictly feedback system which does not respond until the idle speed begins to fall. When an accessory can be controlled by the ECU, further improvement in idle speed control is obtained. By delaying the load briefly after it is requested, the compensation sequence can begin before the load is actually applied. Such a load delay strategy is effective for controlling air conditioning compressor loads, for example. In this case, when the air conditioner is requested, the ECU begins to increase the idle speed first and then activates the A/C compressor.

12.2.3 Engine Control Diagnostics

The purpose of system diagnostics is to provide a warning to the driver when the control system determines that a malfunction of a component or system has occurred and to assist the service technician in identifying and correcting the failure (see Chap. 22). In many cases, to the driver, the engine may appear to be operating correctly, but excessive amounts of pollutants may be emitted. The ECU determines that a malfunction has occurred when a sensor signal received during normal engine operation or during a system test indicates there is a problem. For critical operations such as fuel metering and ignition control, if a required sensor input is faulty, a substitute value may be used by the ECU so that the engine will continue to operate, but likely not at optimal performance. It is also possible to apply an emergency measure if the failure of a component may result in engine or emission system damage. For example, if repeated misfires are detected in one cylinder, perhaps due to an ignition failure, the fuel injector feeding that cylinder can be shut off to avoid damage to the catalytic converter. When a failure occurs, the malfunction indicator light (MIL), visible to the driver, is illuminated. Information on the failure is stored in the ECU. A service technician can retrieve the information on the failure from the ECU and correct the problem.

Air Mass Sensor. For air mass measurement systems, the pulse width of the fuel injectors is calculated in the ECU from the air mass sensor input. As a comparison, the pulse width is also calculated from the throttle valve sensor and the engine RPM. If the pulse width values deviate by a predetermined amount, the discrepancy is stored in the ECU. Then, while the vehicle is being driven, plausibility tests determine which input is incorrect. When this has been determined, the appropriate failure code is saved in the ECU.

Misfire Detection. Misfiring is the lack of combustion in the cylinder. Misfiring can be caused by several factors including fouled or worn spark plugs, poor fuel metering, or faulty electrical connections. Even a small number of misfires may result in excessive exhaust emissions due to the unburned mixture. Increased misfire rates can damage the catalytic converter.

To determine if the engine is experiencing a misfire, the crankshaft speed fluctuation is monitored. If a misfire occurs, no torque is created during the power stroke of the cylinder(s) that is misfiring. A small decrease in the rotational speed of the crankshaft occurs. Because the change in speed is very small, highly accurate sensing of the crankshaft speed is required. In addition, a fairly complicated calculation process is required in order to distinguish misfiring from other influences on crankshaft speed. As was mentioned previously, if a cylinder repeatedly misfires, it is possible to shut off the fuel to that cylinder to prevent damage to the catalytic converter.

Catalytic Converter Monitoring. During the useful life of a catalytic converter, its efficiency decreases. If subjected to engine misfire, the decrease in efficiency occurs more rapidly. A loss in efficiency results in an increase in exhaust pollutants. For this reason, the catalytic converter is monitored. A properly operating catalytic converter transforms O_2, HC, CO, and NO_x into H_2O, CO_2, and N_2. The incoming air/fuel ratio oscillates from rich to lean due to the lambda closed-loop control strategy discussed in Sec. 12.2.1. Only a properly functioning catalytic converter is able to dampen these oscillations by storing and converting the incoming components. As the catalyst ages, this storage effect is diminished. To monitor the catalytic converter, an additional lambda sensor is installed downstream of the catalyst. The ECU compares the signal of the lambda sensor upstream with the lambda sensor downstream and determines if the catalytic converter is operating properly. If not, the ECU illuminates the malfunction indicator light (MIL) and stores a failure code.

Lambda Sensor Monitoring. To minimize exhaust emissions, the engine must operate within the catalytic converter window for air/fuel ratio (see Sec. 12.1.1 for a detailed description of the catalytic converter window). Output from the lambda sensor serves as feedback to the ECU to control the fuel within that window. When a lambda sensor is exposed to high heat for a long period of time, it may respond more slowly to changes in the air/fuel mixture. This can cause a deviation in the air/fuel mixture from the window, which would affect the exhaust emissions.

If the upstream lambda sensor operation is determined to be too slow, which can be detected by the system operation frequency, the ECU illuminates the malfunction indicator light (MIL) and a failure code is stored. Additionally, the ECU compares the output signal of the additional lambda sensor downstream of the catalytic converter with the lambda sensor signal upstream. With this, the ECU is able to detect deviations of the average value in air/fuel ratio.

For heated lambda sensors, the electric current and voltage of the heater circuit is monitored. To accomplish this, the heater is directly controlled by the ECU, not through a relay.

Fuel System Monitoring. To provide the correct air/fuel ratio, the ECU uses a preset data map with the optimal fuel required for each load and RPM point. The lambda closed-loop control system (see Sec. 12.2.1) provides feedback to the ECU on the necessary correction to the preset data points. The corrected information is stored in the ECU's RAM so that the next time that operating point is reached, less correction of the air/fuel ratio will be required. If the ECU correction passes a predetermined threshold, it is an indication that some component in the fuel supply system is outside of its operating range. Some examples are defective pressure regulator, defective manifold pressure sensor, intake system leakage, or exhaust system leakage. When the ECU determines a problem exists, the MIL is illuminated and a code is stored in the ECU.

Exhaust Gas Recirculation (EGR) Monitoring. There are currently two methods used to monitor EGR operation. One method confirms that hot exhaust gases are returning to the intake manifold during EGR operation by use of a temperature sensor in the intake manifold. The second method requires the EGR valve to be fully opened during coast operation, where high intake manifold vacuum occurs. The exhaust gas flowing into the manifold causes a measurable increase in pressure. Thus, if a measured increase in pressure does not occur, the EGR system is not operating.

Evaporative Emissions Control System (EECS) Monitoring. In general, a valve will be installed at the atmospheric side of the purge canister. During idle, this valve would close and the purge valve would open. Intake manifold vacuum would occur in the entire EECS. A pressure sensor in the fuel tank would provide a pressure profile during this test to the ECU, which would then determine if a leak existed in the system.

12.2.4 Fuel Delivery Systems

Overview. Fuel management in the spark ignition engine consists of metering the fuel, formation of the air/fuel mixture, transportation of the air/fuel mixture, and distribution of the air/fuel mixture. The driver operates the throttle valve, which determines the quantity of air inducted by the engine. The fuel delivery system must provide the proper quantity of fuel to create a combustible mixture in the engine cylinder. In general, two fuel delivery system configurations exist: *single-point* and *multipoint* (Fig. 12.11).

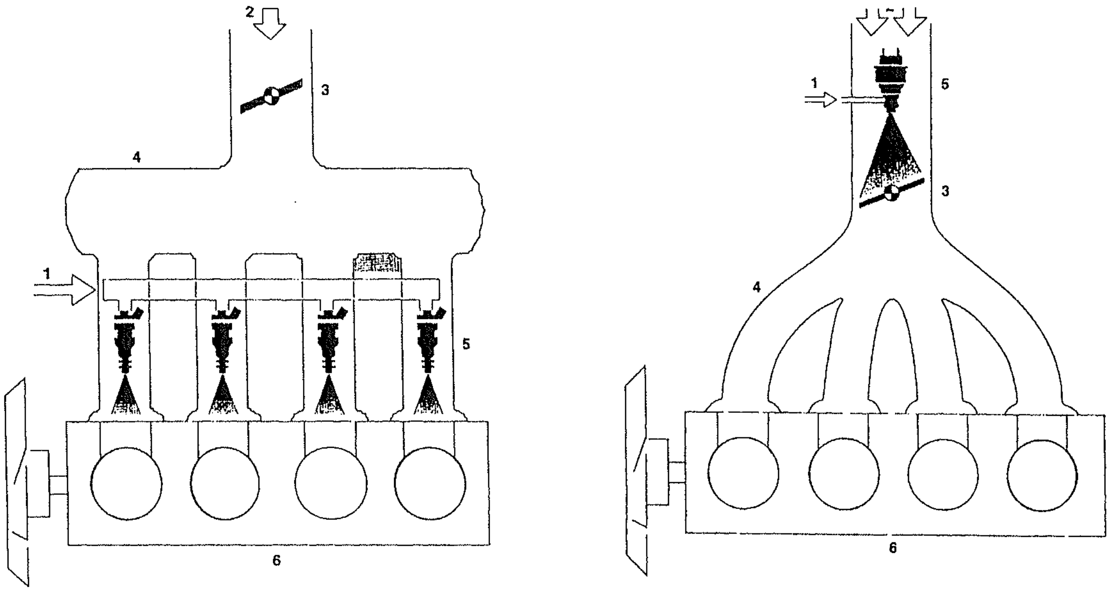


FIGURE 12.11 Air-fuel mixture preparation: right, single-point fuel injection; left, multipoint fuel injection with fuel (1), air (2), throttle valve (3), intake manifold (4), injector(s) (5), and engine (6).

For single-point systems such as carburetors or single-point fuel injection, the fuel is metered in the vicinity of the throttle valve. Mixture formation occurs in the intake manifold. Some of the fuel droplets evaporate to form fuel vapor (desirable) while others condense to form a film on the intake manifold walls (undesirable). Mixture transport and distribution is a function of intake manifold design. Uniform distribution under all operating conditions is difficult to achieve in a single-point system.

For multipoint systems, the fuel is injected near the intake valve. Mixture formation is supplemented by the evaporation of the fuel on the back of the hot intake valve. Mixture transport and distribution occurs only in the vicinity of the intake valve. The influence of the intake manifold design on uniform mixture distribution is minimized. Since mixture transport and distribution is not an issue, the intake manifold design can be optimized for air flow.

Single-Point Injection Systems A single-point injection system uses one or, in some cases, two electronic fuel injectors to inject fuel into the intake air stream. The main component is the fuel injection unit which is located upstream of the intake manifold.

Component Description. An electric fuel pump provides fuel at a medium pressure (typically 0.7 to 1.0 bar) to the electronic fuel injection unit (Fig. 12.12). The fuel injection unit houses the solenoid-operated fuel injector, which is located in the intake air flow above the throttle valve. This allows for homogeneous mixture formation and distribution. The injector spray pattern is designed to allow fuel to pass between the throttle valve and the throttle bore. To prevent vapor lock of the injector, fuel flows through the injector at all times. Fuel not used by the engine is returned to the fuel tank. The injector is activated in relation to the speed of the engine, typically once per ignition event. The length of the pulse width determines the quantity of fuel provided.

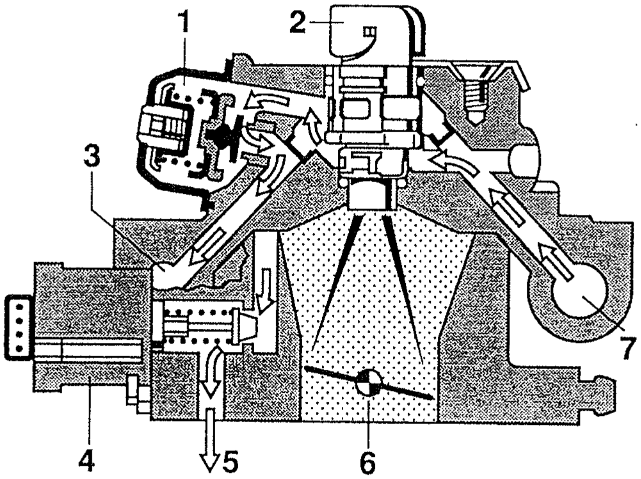


FIGURE 12.12 Single-point injection unit: pressure regulator (1), injector (2), fuel return (3), stepper motor for idle speed control (4), to intake manifold (5), throttle valve (6), and fuel inlet (7).

The electronic injection unit also houses the throttle position sensor and, in some cases, an inlet air temperature sensor which provides operating condition information to the ECU. The throttle valve actuator and fuel pressure regulator are also mounted on the injection unit. In addition, some units contain an air bypass valve for idle speed control. Engine temperature, battery voltage, and engine speed via the ignition system are all inputs to the ECU. The single-point injection system also uses lambda closed-loop fuel control to optimize fuel metering within the lambda control window (see Sec. 12.2.1).

Adaptation to Operating Conditions. For cold-start and engine warm-up, the ECU uses engine temperature information to determine the correct amount of fuel and commands the fuel injector via a pulse width. Due to wall wetting and poor fuel vaporization when the engine is cold, an increase in mixture richness is required. As the engine warms up to operating temperature, the commanded pulse width is reduced.

During an acceleration transition, the ECU adds a correction factor (an increase) to the commanded injector pulse width. The sudden increase in air results in a lean mixture which must be corrected swiftly to obtain good transitional response. During a deceleration transition, the fuel can be shut off by simply not providing a pulse width signal to the injector to minimize exhaust emissions and fuel consumption.

During full-load operation, the air/fuel mixture can be enriched ($\lambda < 1$) to deliver maximum torque. The ECU determines full-load operation by the throttle position sensor (at or near wide-open throttle) and adds a correction to the injector pulse width to achieve the desired air/fuel mixture richness.

The single-point system can control the idle speed by ECU control of either a throttle valve actuator or a bypass valve. Idle speed is a function of engine operating temperature, whether the transmission is in drive, and what accessories are in use. Fuel metering at idle is determined by engine RPM and load as well as lambda closed-loop control.

Multipoint Fuel Injection Systems. A multipoint fuel injection system supplies fuel to each cylinder individually via a mechanical or solenoid-operated fuel injector located just upstream of the intake valve. Advantages of this system type compared to SPI systems are numerous:

- *Increased fuel economy.* On an SPI engine, due to the intake manifold configuration, mixture formation will differ at each cylinder. To provide adequate fuel for the leanest cylinder, too much fuel must be metered overall. In addition, during engine load changes, a film of fuel is deposited on the intake manifold walls. This leads to further variations in mixture from cylinder to cylinder. Multipoint injection provides the same quantity of fuel to each cylinder.
- *Higher power output.* With the fuel being injected near the intake valve, the rest of the intake manifold can be optimized for maximum air flow. The result is increased torque.
- *Improved throttle response.* Because the fuel is injected onto the intake valves, responses to increases in throttle position are swift. With an SPI system, the increased fuel required must travel the length of the intake manifold before entering the cylinder.
- *Lower exhaust emissions.* As was discussed for fuel economy, mixture variation in an SPI system creates increased exhaust emissions. Metering of the fuel at the intake valve decreases this variation. In addition, the system transport time is reduced, increasing the frequency at which the lambda closed-loop control system can switch air/fuel ratio. Catalytic converter efficiency is increased.

Although there are numerous advantages of the MPI systems over the SPI systems, there is still one important advantage the SPI systems have over the MPI systems. In general, SPI systems have better fuel preparation, similar to a carburetor.

Mechanically Controlled Continuous Injection System. This type of system meters the fuel as a function of the intake air quantity and injects it continuously onto the intake valves. This is accomplished by measuring the air flow as it passes through the air flow meter by means of deflection of a meter plate. The fuel is supplied through a fuel accumulator to the fuel distributor by an electric fuel pump. A primary-pressure regulator in the fuel distributor maintains constant fuel pressure. The fuel distributor, through its interface with the air flow meter and warm-up regulator, meters fuel to the continuously flowing fuel injectors.

Component Description

Mixture control unit. The mixture control unit houses the air flow meter and the fuel distributor. In the air flow meter, the measurement of the intake air serves as the basis for determining the amount of fuel to be metered to the injectors. The air flow meter is located upstream of the throttle valve so that it measures all the air entering the engine. It consists of an air funnel, in which a sensor plate is free to pivot. Intake air flowing through the funnel causes a deflection of the sensor plate. The sensor plate is mechanically linked to a control plunger and movement of the plate results in movement of the control plunger. The control plunger movement determines the amount of fuel to be injected.

In the fuel distributor, the control plunger moves up and down in a cylindrically shaped device (barrel) with rectangular openings (metering slit), one for each engine cylinder. Increased air flow causes the control plunger to move upward, uncovering a larger area of the metering slit and increasing the fuel metered. Downstream of each metering slit is a differential-pressure valve that maintains a constant pressure drop across the metering slits at different flow rates. Due to the constant pressure, the fuel flow through the slits is directly proportional to the position of the control plunger (Fig. 12.13).

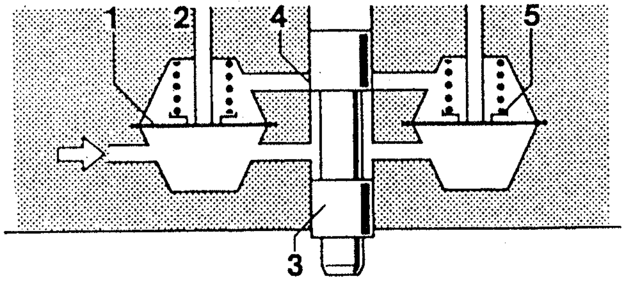


FIGURE 12.13 Fuel distributor for mixture control unit: diaphragm (1), to injector (2), control plunger (3), metering slot (4), differential pressure regulator (5).

Warm-up regulator. The warm-up regulator is used to richen the fuel mixture under cold engine conditions. It consists of a diaphragm valve and an electrically heated bimetallic spring. Under cold conditions, the warm-up regulator lowers the control pressure on the control plunger. The control pressure acts on the opposite end of the plunger from the air flow meter plate. A lower control pressure results in a lower force required to move the meter plate. Therefore, the same air flow causes the meter plate and control plunger to move a greater distance and additional fuel is metered to the injectors.

Fuel injectors. The injectors open at a pressure of approximately 3.6 bar. Atomization of the fuel occurs through oscillation (audible chatter) of the valve needle caused by the fuel flowing through it. The injectors remain open as long as fuel is provided above the opening pressure. Fuel is injected continuously into the intake port. When the intake valve opens, the mixture is drawn into the cylinder.

Auxiliary air valve. The auxiliary air valve provides additional air to the engine by bypassing the throttle valve during cold engine operation. This creates an increase in the idle speed needed during cold operation.

Thermo-time switch. The thermo-time switch controls the cold start valve as a function of time and engine temperature. Fuel enters the intake manifold from the cold start valve and further enriches the mixture to improve cold-starting at low ambient temperatures. When the engine is warm, the contacts in the thermo-time switch open and the cold-start valve is not used in starting the engine.

Lambda sensor. With the addition of a lambda sensor in the exhaust stream, a frequency valve, a modified fuel distributor, and an electronic control unit, the mechanically controlled fuel system can operate under lambda closed-loop control. The lambda sensor signal is read by the ECU. The ECU outputs electric pulses to an electromagnetic (frequency) valve. The frequency valve modulates the pressure to the lower chambers of the differential-pressure valves in the fuel distributor. This results in a modification of the pressure drop across the metering slits, effectively increasing or decreasing the amount of fuel injected. Figure 12.14 is a schematic of a typical mechanically controlled continuous injection system.

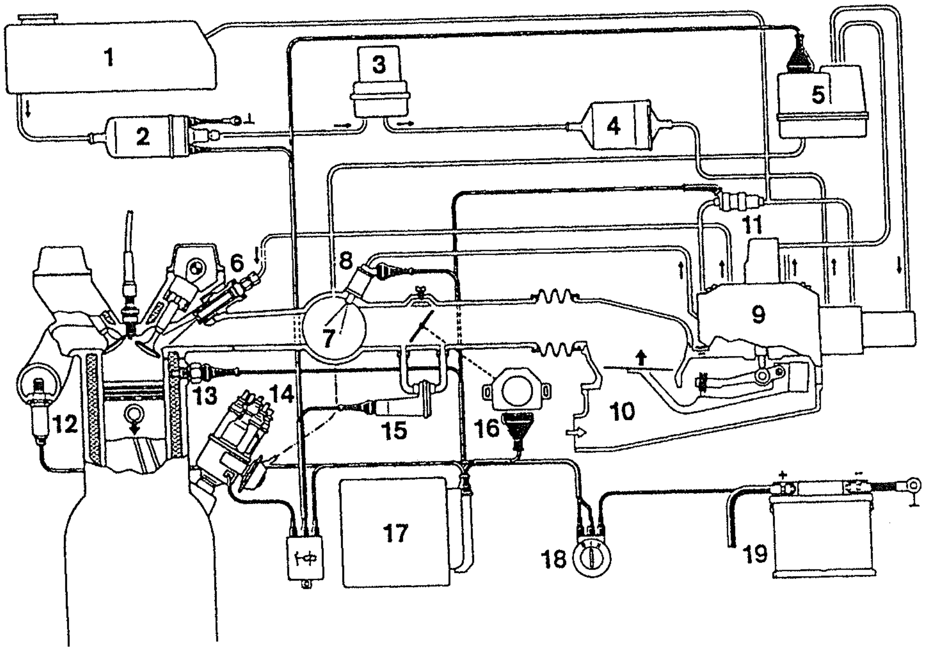


FIGURE 12.14 Schematic of mechanically controlled continuous injection system: fuel tank (1), electric fuel pump (2), fuel accumulator (3), fuel filter (4), warm-up regulator (5), injector (6), intake plenum (7), cold-start valve (8), fuel distributor (9), air flow sensor (10), electrohydraulic pressure actuator (11), lambda sensor (12), thermo-time switch (13), ignition distributor (14), auxiliary air valve (15), throttle switch (16), ECU (17), ignition switch (18), and battery (19).

Depending on the engine temperature, the cold-start valve injects extra fuel into the intake manifold for a limited period during cold start. The injection period is determined by a combination of time and temperature and is controlled by the thermo-time switch. As the engine temperature increases, this additional enrichment is no longer required and the thermo-time switch turns off the cold-start valve. For repeated start attempts or long cranking, the thermo-time switch turns off the cold-start injector after a given time. This minimizes engine flooding when engine start has not occurred.

As the engine continues to warm up, wall wetting and poor fuel vaporization still occur and mixture enrichment is still required until the engine reaches operating temperature. This enrichment is controlled as a function of temperature by the warm-up regulator. As the temperature increases, the warm-up regulator commands less and less additional fuel by increasing the control pressure.

For acceleration response, the air flow sensor "overswings" during quick throttle increases. This causes an additional quantity of fuel to be injected for acceleration enrichment. For full-load enrichment to achieve maximum power, a special warm-up regulator that uses intake manifold pressure is required. At increased manifold pressures, i.e., during wide-open throttle, the warm-up regulator lowers the control pressure, which results in an increase in fuel delivery. Deceleration fuel shutoff is accomplished by diverting all intake air through an air bypass around the air flow sensor plate. With no air flow past the air flow sensor plate, the fuel pressure to the injectors is decreased below the opening pressure.

Idle speed for the cold-running engine is increased by the auxiliary air valve. The amount of additional air varies with engine temperature until the auxiliary air valve is closed and the idle speed is then controlled only be the air passing the throttle valve.

Electronically Controlled Continuous Injection. The basis of the electronically controlled continuous injection is still the mechanical hydraulic injection system discussed previously. This is supplemented by an electronic control unit (ECU) that allows for an increase in flexibility and the use of additional functions. This system incorporates additional sensors for detecting the engine temperature, the throttle valve position (load signal), and the air flow sensor plate deflection. This information is processed by the ECU, which then commands an electrohydraulic pressure actuator to adapt the injected fuel quantity for the present operating conditions.

In contrast to the mechanical system mentioned previously, the control pressure or counterpressure on the control plunger is not varied by a warm-up regulator. The control pressure remains constant and is the same as the primary pressure. The function of the warm-up regulator is now handled by the ECU and the electrohydraulic pressure actuator. Figure 12.15 is a schematic of a typical electronically controlled continuous injection system.

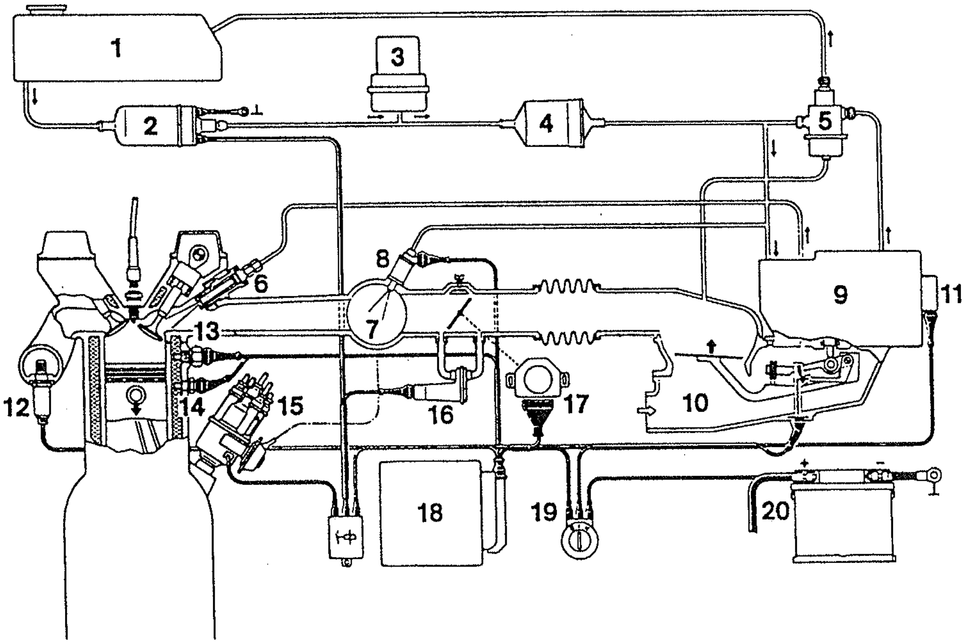


FIGURE 12.15 Schematic of an electronically controlled continuous fuel injection system: fuel tank (1), electric fuel pump (2), fuel accumulator (3), fuel filter (4), fuel pressure regulator (5), injector (6), intake plenum (7), cold-start valve (8), fuel distributor (9), air flow sensor (10), electrohydraulic pressure actuator (11), lambda sensor (12), thermo-time switch (13), coolant temperature sensor (14), ignition distributor (15), auxiliary air valve (16), throttle valve switch (17), ECU (18), ignition switch (19), and battery (20).

Component description—electrohydraulic pressure actuator. The main difference in the componentry between the purely mechanical system and the electronically controlled system is the addition of the electrohydraulic actuator and the elimination of the warm-up regulator. With the addition of the ECU control of fuel metering, the purely mechanical warm-up regulator is no longer required. Depending on the signal received from the ECU, the electrohydraulic pressure actuator varies the pressure in the lower chambers of the differential pressure valves. This changes the amount of fuel delivered to the injectors.

Lambda closed-loop control. As with the mechanical system, the lambda sensor signal is processed by the ECU to determine mixture composition. The difference is that the ECU now commands the electrohydraulic actuator to modify the fuel metered, as opposed to the separate frequency valve, which is no longer necessary.

Adaptation to operating conditions. Depending on the engine temperature, the cold-start valve injects extra fuel into the intake manifold for a limited period during cold-start. The quantity to be injected is controlled by the ECU and is a function of engine temperature (from the engine temperature sensor). The thermo-time switch controls how long the cold-start valve remains active, depending on engine temperature and time.

Acceleration enrichment is controlled by the ECU. Input from the air flow sensor plate position sensor provides the ECU with information on how quickly the engine load has increased. The ECU commands additional enrichment via the electrohydraulic pressure actuator. For full-load enrichment for maximum power, the ECU receives input from the throttle position sensor that the throttle is wide open. The ECU then commands additional enrichment via the electrohydraulic pressure actuator. Deceleration fuel shutoff is also controlled by the ECU when the throttle valve switch indicates the throttle is closed and the engine speed is above idle RPM. The ECU signals the electrohydraulic pressure actuator to interrupt fuel delivery to the injectors.

Idle speed control can be a closed-loop function with the addition of the idle actuator valve. This valve is ECU-controlled and the RPM signal from the ignition, combined with the engine temperature signal, is used to determine its position for the correct idle speed.

Pulsed Fuel Injection Systems. Pulsed fuel injection systems are a further enhancement of the continuous injection systems. Today, most continuous injection systems have been replaced with pulsed fuel injection systems. Instead of injecting fuel continuously and controlling the quantity of fuel by modifying the delivery volume flow rate, the fuel quantity is controlled by the open time of the solenoid-operated injectors. The injectors are controlled directly by the ECU. For most systems, the fuel pressure drop across the injector, from the fuel rail to the intake manifold, is kept constant by using intake manifold air pressure to compensate the fuel pressure regulator. This type of system allows for still greater precision of fuel control and is usually coupled with an equally precise ignition timing control system.

Component description. Several multipoint pulsed injection systems exist in various configurations. The components discussed here serve as a general outline of this system type. Figure 12.16 is a schematic of a typical pulsed fuel injection system.

- *Inlet air sensing.* The inlet air charge can be measured directly using either an air flow meter or a mass air flow meter. The air flow meter is a vane-type meter, which uses the force of the incoming air to move a flap through a defined angle. The angular movement is converted by a potentiometer to a voltage ratio. The air flow meter requires an air inlet temperature sensor to correct for air density changes. The air mass flow meter measures the air mass directly by hot-wire or hot-film element. As the inlet air flow passes the heated element, a bridge circuit keeps the element at a constant temperature above the inlet temperature. The heating current required by the bridge circuit to maintain the element at a constant temperature is measured and converted to an air density value.

  The air charge can also be measured indirectly by measuring the inlet air temperature, intake manifold pressure, and engine RPM and then calculating the air charge (see Sec. 12.2.1 for further discussion on the calculation method which is called speed density air measurement).

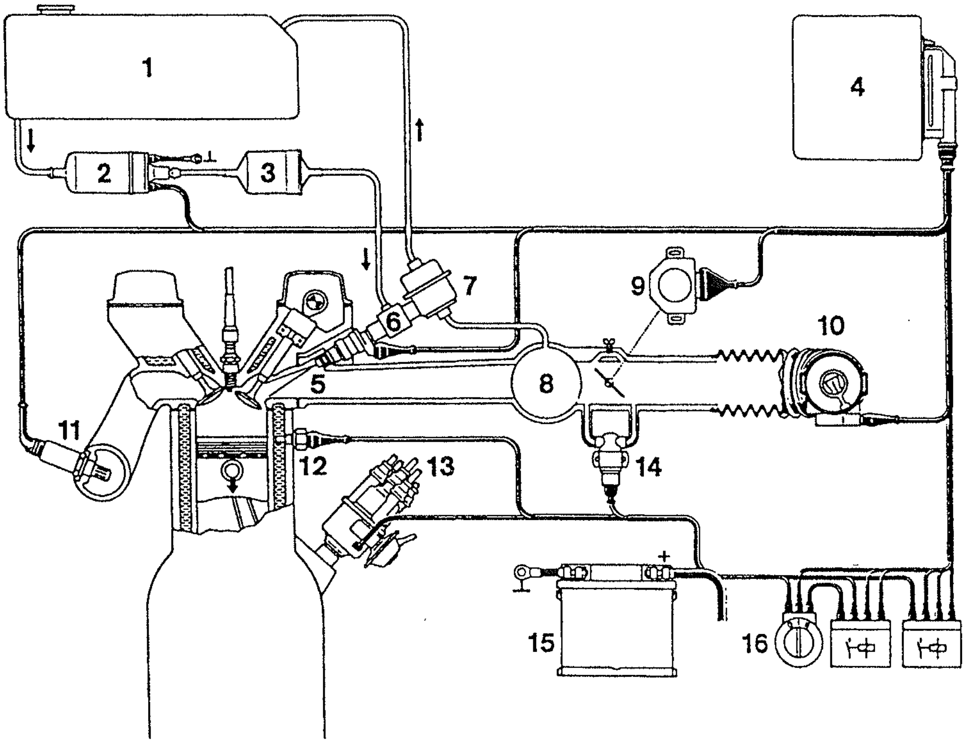


FIGURE 12.16 Schematic of a pulsed fuel injection system: fuel tank (1), electric fuel pump (2), fuel filter (3), ECU (4), injector (5), fuel distributor (6), fuel pressure regulator (7), intake plenum (8), throttle valve switch (9), hot-wire mass air flow sensor (10), lambda sensor (11), coolant temperature sensor (12), ignition distributor (13), idle speed actuator (14), battery (15), and ignition switch (16).

- *Fuel metering.* The fuel supply system includes an electric fuel pump, fuel filter, fuel rail, pressure regulator, and solenoid-operated injectors. The fuel pump provides more fuel than the maximum required by the engine. Fuel not used by the engine is returned to the fuel tank. The fuel rail supplies all injectors with an equal quantity of fuel and ensures the same fuel pressure at all injectors.

  The pressure regulator keeps the pressure differential across the injectors constant. It contains a diaphragm that has intake manifold pressure on one side and fuel rail pressure on the other. Normally, it is mounted at the outlet end of the fuel rail. The diaphragm operates a valve which opens at a differential pressure between 2.0 and 3.5 bar and allows excess fuel to return to the fuel tank.

  The fuel injectors are solenoid-operated valves that are opened and closed by means of electric pulses from the ECU. The injectors are mounted in the intake manifold and spray onto the back of the intake valves. In general, one injector is used for each cylinder.

  In addition, some systems also use a separate cold-start injector mounted in the intake manifold just downstream of the throttle valve. This injector ensures good fuel vaporization during cold-start and supplies the additional enrichment needed to start the cold engine. Control of the cold-start valve is either by the ECU directly or in conjunction with a thermo-time switch.

- *Lambda closed-loop control.* The lambda sensor signal is processed by the ECU. The ECU determines the required injector pulse width to maintain the air/fuel ratio within the lambda control window (see Sec. 12.2.1 for further discussion on lambda closed-loop control).

Adaptation to operating conditions. For cranking, the fuel required is determined by a data table in the ECU with reference to engine temperature. The ECU then commands a pulse width for the fuel injectors. The air/fuel mixture is greatly enriched due to poor fuel vaporization and wall wetting, which reduces the amount of usable fuel. After start, the fuel mixture remains rich due to continuing poor air/fuel mixture formation. The amount of enrichment should be minimized to obtain good emission results. The target is to stay close to lambda (λ) = 1.

For acceleration enrichment, the throttle valve position sensor indicates that the throttle has moved rapidly. The ECU adds a correction factor to increase the pulse width so that a smooth transition occurs. For deceleration, the ECU uses input from the throttle position sensor and engine RPM to indicate that the throttle has closed and the engine speed is above the idle speed. Since no torque is required under this condition, the ECU provides no pulse width to the injectors and they are therefore closed. For full-load enrichment, when necessary, the ECU can provide an injector pulse width that would result in the engine achieving its maximum torque (roughly $\lambda = 0.9$). Fuel metering during idle is primarily controlled by lambda closed-loop control when the engine has reached operating temperature.

12.2.5 Ignition Systems

Overview. The purpose of the ignition system in the spark ignition engine is to initiate combustion of the air/fuel mixture by delivering a spark at precisely the right moment. The spark consists of an electrical arc generated across the electrodes of the spark plug. Two important factors for proper ignition are the energy of the spark and the point in the four-stroke cycle when the spark occurs (ignition timing).

Electrical Energy. The energy required for a spark to ignite an air/fuel mixture at stoichiometry depends on specific engine conditions. If there is insufficient energy available to ignite the air/fuel mixture, misfiring will occur. Misfiring will result in poor engine operation, high exhaust emissions, and possible catalytic converter damage. Therefore, the amount of ignition energy available must always exceed the amount necessary to ensure ignition even under adverse conditions.

Some of the conditions that affect ignitability of the air/fuel mixture are fuel atomization, access of the mixture to the spark, spark duration, and spark physical length. Fuel atomization is controlled by the fuel system and the engine design. Access to the spark depends on combustion chamber and spark plug design. Spark duration is a function of the ignition system. Spark physical length is determined by the spark plug dimensions (gap).

Ignition Timing. The ignition timing must be selected to meet the following objectives: maximize engine performance, limit fuel consumption, minimize engine knock, and minimize exhaust emissions. Unfortunately, all of these objectives cannot be achieved simultaneously under all operating conditions and compromises must be made.

It is desirable in the SI engine to have ignition of the combustible mixture occur prior to the piston reaching TDC on the compression stroke to achieve the best engine performance. The ignition spark must occur early enough to ensure that the peak cylinder combustion pressure occurs at the correct point after top dead center (ADC) under all operating conditions. Figure 12.17 is a graph of ignition angle vs. combustion pressure. The length of the combustion process from initial ignition to final combustion is approximately 2 ms. This combustion time remains relatively constant with respect to engine speed. Therefore, as the engine speed increases, the ignition spark must occur earlier in terms of crankshaft angle to ensure complete combustion.

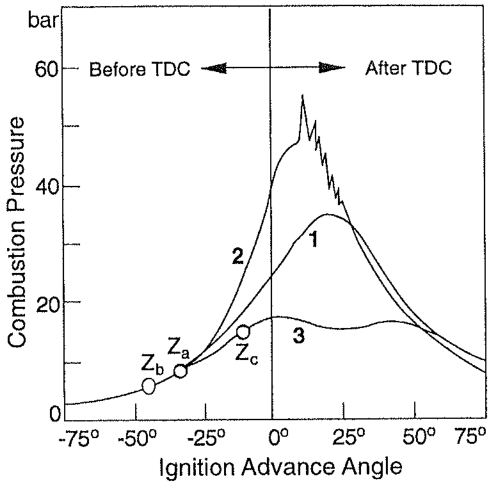


FIGURE 12.17 Combustion pressure curve for various ignition timing points: correct ignition advance Z_a (1), excessive ignition advance Z_b (2), and excessive ignition retard Z_c (3).

At low engine loads, the lower air charge and the residual gas content, due to valve overlap, serve to lengthen the time required for complete combustion. To compensate for this effect, the ignition timing is advanced at low loads to ensure that complete combustion occurs.

Ignition timing influences exhaust emissions and fuel consumption. With more advanced timing, the emission of unburned hydrocarbons (HC) and of oxides of nitrogen (NO_x) increases. Carbon monoxide (CO) emissions are not influenced greatly by ignition timing. To achieve improvements in fuel consumption, the air/fuel mixture must be lean. To ensure complete combustion for a lean mixture, the ignition timing must be advanced. However, as previously stated, advanced timing increases emissions of HC and NO_x.

Spark Ignition Systems. The general configuration of an ignition system consists of the following components: energy storage device, ignition timing mechanism, ignition triggering mechanism, spark distribution system, and spark plugs and high tension wires.

Inductive ignition systems use an ignition coil as the energy storage device. The coil also functions as a transformer, boosting the secondary ignition voltage. A typical turns-ratio of the primary to secondary winding is 1:100. Electrical energy is supplied to the coil's primary winding from the vehicle electrical system. Before the ignition point, the coil is charged during the dwell period to its interruption current. Open- or closed-loop dwell angle control ensures a sufficient interruption current even at high speeds. Sufficient ignition energy at the interruption current is ensured by an adequate coil design. At the ignition point, the primary current will be interrupted. The rapid change of the magnetic field induces the secondary voltage in the secondary winding. A distribution system assigns the high voltage to the corresponding spark plug. After exceeding the arcing over voltage at the spark plug, the coil will be discharged during the spark duration.

The ignition timing mechanism, ignition triggering mechanism, and the spark distribution system differ between ignition systems. Further discussion of these will occur within the discussion of each ignition system type.

The spark plugs provide the ignition energy via the high-tension wires to the air/fuel mixture in the cylinder to initiate combustion. The voltage required at the spark plug can be more than 30 kV. Because the spark plug extends into the combustion chamber, it is exposed to extreme temperature and pressure conditions. Spark plug design and materials are chosen to ensure long-term operation under tough operating conditions.

A typical spark plug consists of a pair of electrodes, called a center and ground electrode, separated by a gap. The size of the gap is important and is specified for each plug type and engine. The center electrode is electrically connected to the top terminal of the plug which is attached to the high-tension wire. The electrical energy travels through the high-tension wire to the top terminal and down to the center electrode. The ground electrode is part of the threaded portion of the spark plug that is installed in the cylinder head. The ground electrode is at electrical ground potential because the negative terminal of the battery is also connected to the engine. The spark is produced when the high-voltage pulse travels to the center electrode and jumps the gap to the ground electrode.

Ignition System Types. Table 12.1 summarizes the various ignition systems used on SI engines.

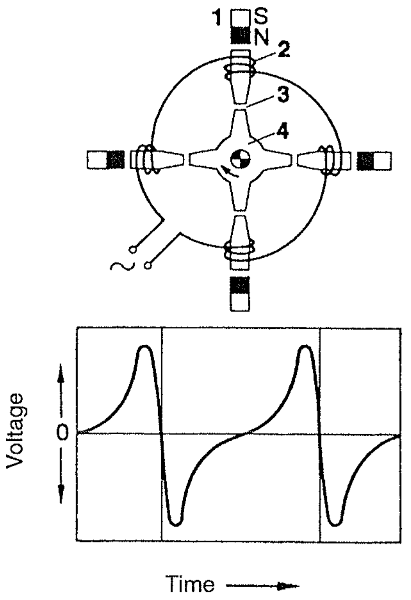


FIGURE 12.18 Induction-type pulse generator: permanent magnet (1), induction winding with core (2), variable air gap (3), trigger wheel (4).

Coil ignition. Breaker-triggered coil ignition systems have been replaced by breakerless transistorized ignition systems and are no longer installed as original equipment.

On breakerless transistorized ignition systems, the contact breaker's function is replaced by a magnetic pulse generator. The pulse generator is installed in the distributor and turns with the distributor shaft. There are commonly two types of pulse generators: induction-type and Hall-type. Induction-type pulse generators consist of a stator and a trigger wheel (Fig. 12.18). The stator consists of a permanent magnet, inductive winding, and core, and remains fixed. The trigger wheel teeth correspond to the number of cylinders, and the trigger wheel turns with the distributor shaft. The operating principle is that as the air gap changes between the stator and the rotor, the magnetic flux changes. The change in magnetic flux induces an ac voltage in the inductive winding. The frequency and magnitude of the alternating current increases with increasing engine speed. The electronic control unit or trigger box uses this information to trigger the ignition timing.

TABLE 12.1 Overview of Various Ignition Systems

	Ignition designation				
Ignition function	Coil system	Transistorized coil system	Capacitor discharge system	Electronic system with distributor	Electronic distributorless system
Ignition triggering	Mechanical	Electronic	Electronic	Electronic	Electronic
Ignition timing	Mechanical	Mechanical	Electronic	Electronic	Electronic
High-voltage generation	Inductive	Inductive	Capacitive	Inductive	Inductive
Spark distribution to appropriate cylinder	Mechanical	Mechanical	Mechanical	Mechanical	Electronic

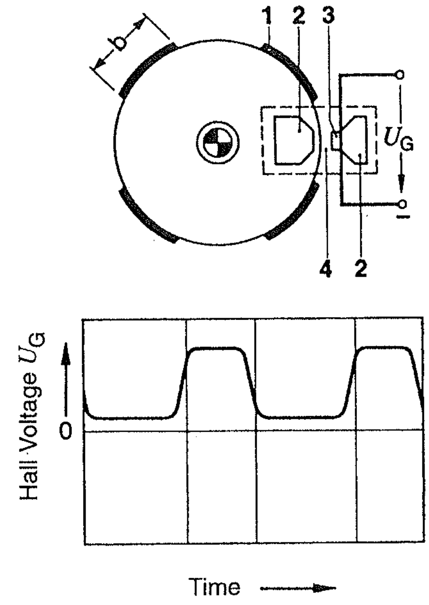


FIGURE 12.19 Hall-type pulse generator: vane (1), soft magnetic conductive elements (2), Hall IC (3), and air gap, UG-Hall sensor voltage (4).

Hall-type pulse generators utilize the Hall effect (Fig. 12.19). As the distributor shaft turns, the vanes of the rotor move through the air gap of the magnetic barrier. When the vane is not in front of the Hall IC, the sensor is subjected to a magnetic field. The magnetic flux density is high and thus the voltage U_G is at a maximum. As soon as the rotor vane enters the air gap, the magnetic flux runs through the vane area and is largely prevented from reaching the Hall layer. The voltage U_G is at a minimum. The resulting pulses switch the primary current off and on.

The distributor disburses the ignition pulses to the spark plugs via the high-tension wires in a specific sequence. It also adjusts the ignition timing by means of spark advance mechanisms. The distributor rotor is turned by the engine at one-half the crankshaft speed. The electrical energy is fed to the center of the rotor. While the rotor turns, the rotor electrode aligns with the outer electrodes that are connected to the high-tension wires. One outer electrode and high-tension wire connection exists for each cylinder. When alignment occurs between the center and outer electrode, the spark is distributed to that particular cylinder.

The spark advance mechanisms advance the ignition timing by rotating the distributor plate relative to the distributor shaft. The centrifugal advance increases the spark advance with increasing engine speed. The vacuum advance, using intake manifold vacuum, increases the spark advance at low engine speeds.

Capacitor discharge ignition system. The capacitive discharge system differs from the coil-type ignition systems previously discussed. Ignition energy is stored in the electrical field of a capacitor. Capacitance and charge voltage of the capacitor determine the amount of energy that is stored. The ignition transformer converts the primary voltage discharged from the capacitor to the required high voltage.

Electronic ignition—with distributor. Electronic ignition calculates the ignition timing electronically (Fig. 12.20). This replaces the function of the centrifugal advance and vacuum advance in the distributor discussed on the previous coil ignition systems. Because the ignition timing is not limited by mechanical devices, the optimal timing can be chosen for each engine operating point. Figure 12.21 is a comparison of an ignition map from a mechanical advance system and a map of an electronically optimized system. Also, additional influences such as engine knock detection can be used to modify the ignition timing. The engine speed input and crankshaft position input can be obtained from a sensor mounted near the crankshaft. Precision is improved over using the distributor-mounted trigger. This input is provided to the electronic control unit (ECU) along with the engine temperature and engine load. The ECU references data tables to determine the optimal spark advance for each engine operating condition. Additional corrections to the spark timing, such as for EGR usage or knock sensor detection, are made in the ECU.

Electronic ignition—distributorless. On distributorless ignition systems, the high voltage distribution is accomplished by using either single or double spark ignition coils. Ignition timing is determined by the ECU, as discussed for electronic ignition with distributor. For the double spark ignition coils, one coil exists for two corresponding cylinders. Two high-tension

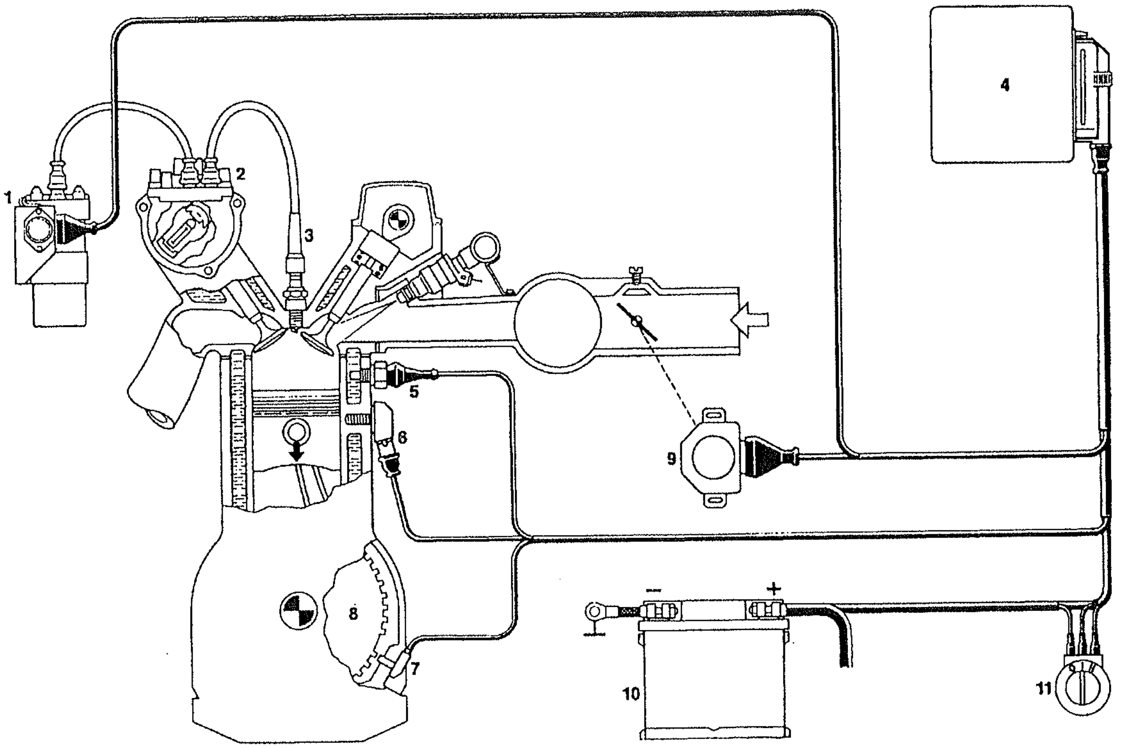


FIGURE 12.20 Schematic of an electronic ignition system with distributor: ignition coil (1), high-voltage distributor (2), spark plug (3), ECU (4), coolant temperature sensor (5), knock sensor (6), engine speed and crankshaft reference sensor (7), sensor wheel (8), throttle valve (9), battery (10), and ignition switch (11).

wires are routed from each coil to two cylinders, which are 360° out of phase. When the coil output stage is triggered via the ECU, a spark is delivered to both cylinders. One cylinder will be on the compression stroke, the other on the exhaust stroke. Because both cylinders are fired together, for a given crankshaft rotation, one will always be on the compression stroke and the other on the exhaust stroke. Therefore, there is no need to know which cylinder is compressing the ignitable mixture.

On single spark ignition coils, one coil exists for each cylinder. Each coil triggers only once during the four-stroke cycle. Because of this, it must be known which cylinder is on the compression stroke. Synchronization with the camshaft must occur. The information needed on camshaft position is supplied by a phase sensor mounted on the camshaft.

12.3 COMPRESSION IGNITION ENGINES

12.3.1 Engine Control Functions

Electronic engine controls are now being used on compression ignition (diesel) engines. These controls offer greater precision and control of fuel injection quantity and timing, engine speed, EGR, turbocharger boost pressure, and auxiliary starting devices. The following inputs

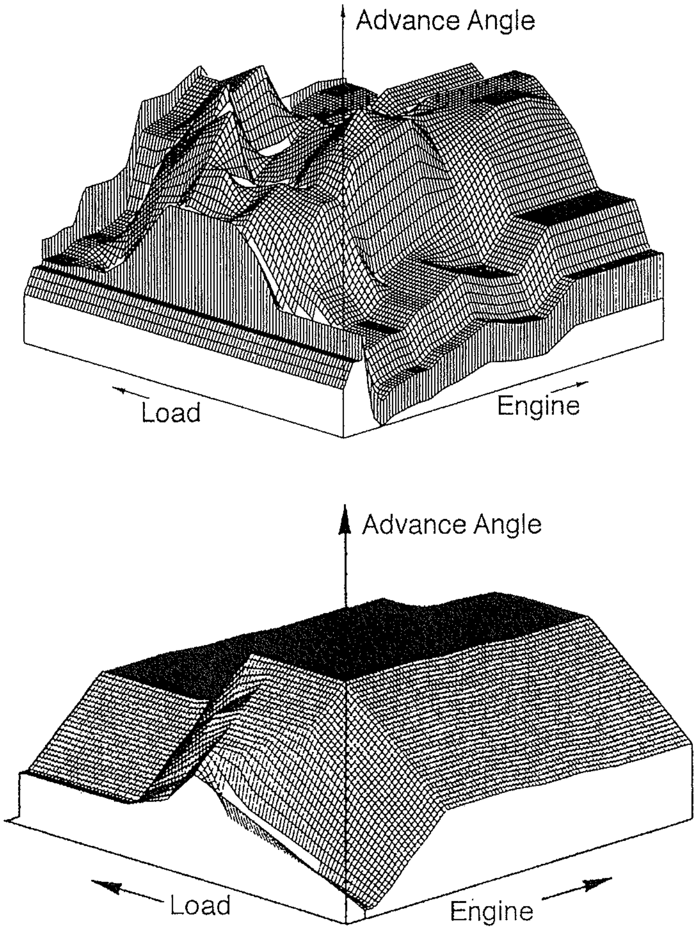


FIGURE 12.21 Ignition timing maps: electronically optimized (*above*) and mechanical advance system (*below*).

are used to provide the ECU with information on current engine operating conditions: engine speed; accelerator position; engine coolant, fuel, and inlet air temperatures; turbocharger boost pressure, vehicle speed, control rack, or control collar position (for control of fuel quantity); and amospheric pressure. Figure 12.22 is a schematic of an electronic engine control system on an in-line diesel fuel injection pump application.

Fuel Quantity and Timing. The fuel quantity alone controls a compression ignition engine's speed and load. The intake air is not throttled as in a spark ignition engine. The quantity of fuel to be delivered is changed by increasing or decreasing the length of fuel delivery time per injection. On the injection pump, the delivery time is controlled by the position of the control rack on in-line pumps and the position of the control collar on distributor-type pumps. An ECU-controlled actuator is used to move the control rack or the collar to increase or decrease the fuel delivery time. The ECU determines the correct length of delivery time (expressed as a function of control rack or collar position) using performance maps based on engine speed and calculated fuel quantity. Corrections and/or limitations as functions of

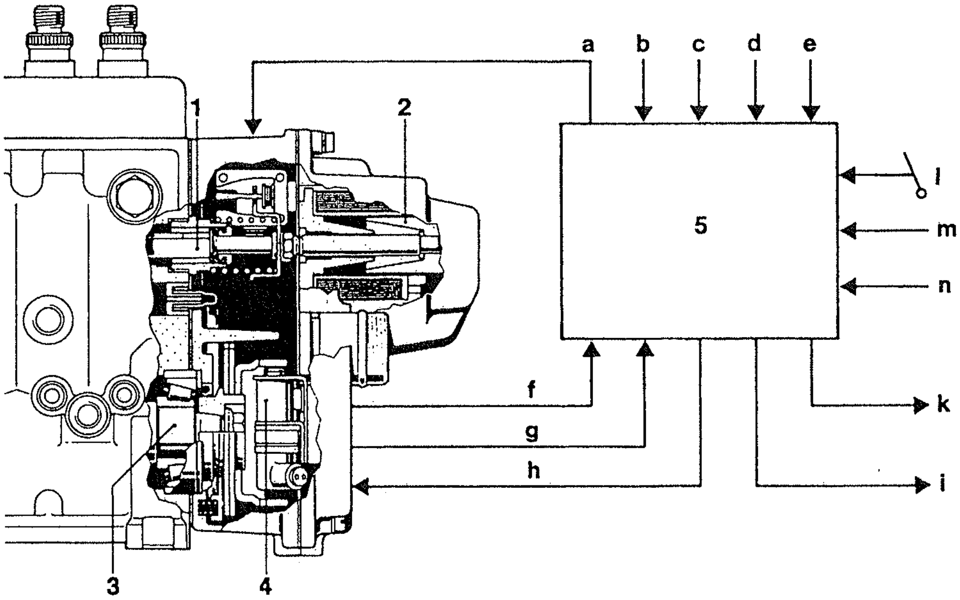


FIGURE 12.22 Electronic engine control system for an in-line injection pump: control rack (1), actuator (2), camshaft (3), engine speed sensor (4), ECU (5). Input/output: redundant fuel shutoff (a), boost pressure (b), vehicle speed (c), temperature—water, air, fuel (d), intervention in injection fuel quantity (e), speed (f), control rack position (g), solenoid position (h), fuel consumption and engine speed display (i), system diagnosis information (k), accelerator position (l), preset speed (m), and clutch, brakes, engine brake (n).

engine speed, temperature, and turbocharger boost pressure are used to modify the delivery time. In addition, the control rack or collar actuator contains a position sensor that provides feedback to the ECU on controller position. If the requested position differs from the commanded position, the ECU continues to move the controller via the actuator until the commanded and actual position are the same.

The start of injection time of the fuel at the cylinder is a function of the wave propagation speed (i.e., the speed of sound) of the fuel from the fuel injection pump to the injector. Because this time remains a constant, at increasing engine speed the delivery of fuel at the cylinder would be delayed with reference to crankshaft angle. Therefore, the timing at the injection pump must be advanced with increasing engine speed so that the start of injection occurs at the same crankshaft angle at higher engine speeds. Selection of injection timing has a large impact on exhaust emissions and engine noise. Delaying the start of injection reduces NO_x emissions, but excessive delay increases HCs in the exhaust. A 1° deviation in injection timing can increase NO_x emissions by 5 percent and HC emissions by as much as 15 percent. Therefore, precise control of the start of injection is essential.

Although many systems use mechanical devices to control injection timing, electronic control of injection timing is being used on some pump types. The advantage of electronic control is that a sophisticated timing data map can be used that provides the best injection timing for exhaust emissions under various operating conditions. On electronic control systems, the start of injection is monitored at the injector nozzle by a needle-motion sensor. The ECU uses this information to determine and control the injection timing. The timing is then modified by control of a pulse-width modulated solenoid valve. The valve varies the pressure exerted on the spring-loaded timing device plunger. The plunger rotates the pump's collar ring (for distributor type pumps) in the opposite direction of the pump's rotation which advances the timing.

Speed Control. As was mentioned previously, for a CI engine, fuel quantity alone controls the engine's speed and load. Therefore, presuming adequate injected fuel quantity, an unloaded CI engine can speed up out of control and destroy itself. Because of this, a governor is required to limit the engine's maximum speed. In addition, governors are also used for low idle and cruise control to maintain a constant engine or vehicle speed and meter the correct fuel for cold-starting. Fuel is also controlled as a function of speed and boost pressure to limit smoke levels, engine torque, and exhaust gas temperatures. On an electronically controlled CI engine, the governor's functions are controlled by the fuel delivery system described previously. Engine speed is provided by an RPM sensor that monitors the periods of angular segments between the reference marks on the engine's flywheel or in the in-line injection pump.

EGR Control. Rerouting of exhaust gases into the intake air stream is known as exhaust gas recirculation (EGR). EGR reduces the amount of oxygen in the fresh intake charge while increasing its specific heat. This lowers combustion temperatures and results in lower NO_x emissions. However, excessive amounts of EGR result in higher emissions of soot (particulates), CO, and HCs all due to insufficient air. Also, the introduction of EGR can have an adverse affect on driveability during cold-engine operation, full-load operation, and at idle. It is best, therefore, to control the EGR valve with the ECU. Both pneumatically controlled and solenoid-controlled EGR valves are in use. The ECU determines when and how much EGR will occur based on engine temperature and accelerator position.

Turbocharger Boost Pressure Control. Engines that have turbochargers benefit significantly from electronic boost pressure control. If only a pneumatic-mechanical wastegate is used, only one boost pressure point for the entire operating range is used to divert the exhaust gas away from the turbine side of the turbocharger. This creates a compromise for part-load conditions because all the exhaust gases must pass the turbine. The result is increased exhaust backpressure, more turbocharger work, more residual exhaust gas in the cylinders, and higher charge air temperatures.

By controlling the wastegate with a pulse-width-modulated solenoid valve, the wastegate can be opened at different pressures depending on the engine operating conditions. Therefore, only the level of air charge pressure required is developed. The electronic control unit uses information on engine speed and accelerator position to reference a data table and the proper boost pressure (actually, duty cycle of the control valve) is determined. On systems using intake manifold pressure sensors, a closed-loop control system can be developed to compare the specified value with the measured value.

Glow Plug Control. Electronic control of the glow plug duration can be handled by the ECU or a separate control unit. Input for determining glow time is from an engine coolant temperature sensor. At the end of the specified glow period, the controller turns out the start indicator light to signal the driver that the engine can be started. The glow plugs remain energized while the starter is engaged. An engine load monitor is used to switch off the glow process after start. To limit the loads on the battery and the glow plugs, a safety override is also used.

12.3.2 Fuel Delivery Systems

The diesel fuel delivery system comprises a low- and high-pressure side. On the low-pressure side is the fuel tank, fuel filter, fuel supply pump, overflow valve, and fuel supply lines. The high-pressure side is initiated in the plunger and barrel assembly and continues through the delivery valve, high-pressure injection lines, and injection nozzle.

The fuel injection pump must deliver fuel at a pressure between 350 and 1200 bar, depending on the engine's combustion configuration. The quantity and timing of injection must be precisely controlled to achieve good mixture quality and to minimize exhaust emissions.

Fuel Injection Process. An engine-driven camshaft (in-line pump) or cam plate (distributor pump) drives the injection pump's plunger in the supply direction, creating pressure in the high-pressure gallery. The delivery valve responds to the increase in pressure by opening. This sends a pressure wave to the injection nozzle at the speed of sound. The needle valve in the nozzle overcomes the spring force of the injection nozzle spring and lifts from its seat when the opening pressure is reached. Fuel is then injected from the spray orifices into the engine's combustion chamber. The injection process ends with the opening of the spill port in the plunger and barrel assembly. This causes the pressure in the pump chamber to collapse, which then causes the delivery valve to close. Due to the action of the delivery valve relief collar, the pressure in the injection line is reduced to the "stand-by pressure." The stand-by pressure is determined to ensure that the injector nozzle closes quickly to eliminate fuel dribble, and the residual pressure waves in the lines prevent the nozzles from reopening.

ABOUT THE AUTHORS

GARY C. HIRSCHLIEB is chief engineer, engine management systems, for the Robert Bosch Corp. He previously held various engineering and sales responsibilities with Bosch. In his earlier career, he worked as a senior engineer in powertrain development for Ford tractor operations, and as a sales engineer with GTE, and as an engineer in plant engineering for GM Truck and Coach.

GOTTFRIED SCHILLER is engineering manager, engine management systems, for Robert Bosch Corp. His previous positions with Bosch included applications engineering, engine management systems; application engineer, diesel systems; and development engineer, diesel products.

SHARI STOTTLER is now a self-employed technical writer, but, until 1993, she was a senior application engineer with Robert Bosch Corp. Prior to that she had been an engineering project coordinator for Honda of North America, Manufacturing, and a product engineer with General Motors Corp.

CHAPTER 13
TRANSMISSION CONTROL

Kurt Neuffer, Wolfgang Bullmer, and Werner Brehm
Robert Bosch GmbH

13.1 INTRODUCTION

In North America and Japan, 80 to 90 percent of all passenger cars sold have automatic transmissions (ATs), but in Europe only 10 to 15 percent of passenger cars sold have ATs. There are two main reasons for the difference. In Europe, drivers tend to view ATs, compared to manual transmissions, as detrimental to driveability and responsible for a somewhat higher fuel consumption. But implementation of electronic control concepts has invalidated both of those arguments.

Since the introduction of electronic transmission controls units (TCUs) in the early 1980s by Renault and BMW (together with a four-speed transmission from Zahnradfabrik Friedrichshafen, or ZF), the acceptance of the AT rose steeply, even in Europe. For this reason, all new ATs are designed with electronic control. The market for ATs is divided into stepped and continuously variable transmissions (CVTs). For both types the driver gets many advantages. In stepped transmissions, the smooth shifts can be optimized by the reduction of engine torque during gear shift, combined with the correctly matched oil pressure for the friction elements (clutches, brake bands). The reduction of shift shocks to a very low or even to an unnoticeable level has allowed the design of five-speed ATs where a slightly higher number of gear shifts occur. In today's standard systems, the driver can choose between sport and economic drive programs by operating a selector switch. In highly sophisticated newer systems, the selection can be replaced by the self-adaptation of shift strategies. This leads not only to better driveability but also to a significant reduction in fuel consumption. Additionally, a well-matched electronic control of the torque converter lockup helps to improve the yield of the overall system. Both automotive and transmission manufacturers benefit from the reduced expense resulting from the application of different car/engine combinations. Different shift characteristics are easy to implement in software, and much adaptation can be achieved by data change, leaving the transmission hardware and TCU unchanged. The reduction of power losses in friction elements increases the life expectancy and enables the optimization of transmission hardware design.

With the CVT, one of the biggest obstacles to the potential reduction in fuel consumption by operating the engine at its optimal working point is the power loss from the transmission's oil pump. Only with electronic control is it possible to achieve the required yield by matching the oil mass-stream and oil pressure for the pulleys to the actual working conditions.

To guarantee the overall economic solution for an electronically controlled transmission, either stepped or CVT, the availability of precision electrohydraulic actuators is imperative.

13.2 SYSTEM COMPONENTS

The components of an electronic transmission control system are a transmission which is adapted to the electronic control requirements and an electronic control unit with corresponding inputs and outputs and attached sensor elements.

13.2.1 Transmission

The greatest share of electronically controlled transmissions currently on the market consists of four- or five-speed units with a torque converter lockup clutch, commanded by the control unit. Market share for five-speed transmissions is continuously increasing. With electronically controlled transmissions there are numerous possibilities to substitute mechanical and hydraulic components with electromechanical or electrohydraulic components. One basic method is to substitute only the shift point control. In a conventional pure hydraulic AT, the gear shifts are carried out by mechanical and hydraulic components. These are controlled by a centrifugal governor that detects the vehicle speed, and a wire cable connected to the throttle plate lever. With an electronic shift point control, on the other hand, an electronic control unit detects and controls the relevant components. In the transmission's hydraulic control unit, mechanical and hydraulic components are replaced by electrohydraulic controlling elements, usually in the form of electrohydraulic on/off solenoids. This way the number of solenoids, as well as the control logic, can be varied over a wide range. For example, for each gear, one specific solenoid can operate the relevant clutch for this gear shift. Alternatively, there can be one solenoid for each gear change, which is switched corresponding to the shift command. In this way, only three solenoids are required in a four-speed transmission. In some current designs, the gears are controlled by a logical combination of solenoid states. This design needs only two gear-controlling solenoids for a four-speed transmission. For five-speed applications, accordingly, three solenoids are required (Table 13.1)

TABLE 13.1 Example of a Gear-Solenoid Combination for a Five-Speed Transmission Application

	Solenoid 1	Solenoid 2	Solenoid 3
1st gear	on	on	on
2nd gear	on	on	off
3rd gear	on	off	off
4th gear	off	off	off
5th gear	off	on	off

The hydraulic pressure is controlled in this basic application by a hydraulic proportional valve which is, in turn, controlled by a wire cable connected to the throttle plate lever. With this design, the shift points can be determined by the electronic TCU, resulting in a wide range of freely selectable driving behaviors regarding the shift points. It is also possible to use different shift maps according to switch or sensor signals. The influence on driving comfort during gear shifting in this electronic transmission control application has important restrictions. The only possible way to control shift smoothness is with an interface to the electronic engine management. This way, the engine output torque is influenced during gear shifting. A systematic wide-range control of the hydraulic pressure during and after the gear shift necessitates the replacement of the hydraulic pressure governor with an electronically controlled hydraulic solenoid. This design allows the use of either a pulse-width-modulated (PWM) solenoid or a pressure regulator. The choice of which type of pressure control solenoid to use results from the requirements concerning shift comfort under all driving conditions. For present-day designs with high requirements for shift comfort during the entire life of the transmission, at all temperatures, and with varying oil quality, the analog pressure control solenoid is superior to the usual PWM solenoid, providing there is no pressure sensor in operation as a guideline for pressure regulation. This application usually uses one central controlling element in the transmission for the pressure regulation to control the shift quality.

In other transmission developments, the shift quality is further increased using electronically controllable brake elements (brake bands) for some specific gear changes. In this case, the flywheel effect of the revolving elements is limited by an electronic control of a brake band according to an algorithm or special timing conditions.

The most sophisticated transmission application to date is so designed that overrunning clutches are eliminated and gear changes are exclusively controlled by the electronic control unit with pressure regulator solenoids.[1] This application is characterized by extremely high demands on the electronic TCU concerning real-time behavior and data handling. The relationship between weight, transmission outline, and transferrable torque has reached a high level. Compared to transmissions with overrun clutches, the necessary fitting dimensions are reduced.

Present electronically controlled ATs usually have an electronically commanded torque converter clutch, which can lock up the torque converter between the engine output and the transmission input. The torque converter clutch is activated under certain driving conditions by a solenoid controlled by the electronic TCU. The solenoid design, depending on the requirements of TCC functions and shift comfort, can either be an on/off solenoid, a PWM solenoid, or a pressure regulator. Locking up the torque converter eliminates the slip of the converter, and the efficiency of the transmission system is increased. This results in an even lower fuel consumption for cars equipped with AT.

13.2.2 Electronic Control Unit

Another important component in electronic transmission control is the electronic control unit, which is designed according to the requirements of the transmission and the car environments. The electronic control unit can be divided into two main parts: the hardware and the corresponding software.

Hardware. The hardware of the electronic control unit consists of the housing, the plug, the carrier for the electronic devices, and the devices themselves. The housing, according to the requirements, is available as an unsealed design for applications inside the passenger compartment or within the luggage compartment. It is also possible to have sealed variants for mounting conditions inside the engine compartment or at the bulkhead. The materials for the housing can be either various plastics or metals. There are many different nonstandardized housings on the market. The various outlines and plug configurations differ, depending upon the manufacturer of the electronic unit. The plug configuration, i.e., the number of pins and the shape, depends on the functions and the requirements of the automotive manufacturer. The number of pins is usually less than 100. Some control unit manufacturers try to standardize their plugs and housings throughout all their electronic control units, such as engine management, ABS, traction control, and others. This is important to simplify and to standardize the unit production and the tests during manufacturing.

The carrier for the electronic devices is usually a conventional printed circuit board (PCB). The number of layers on the PCB depends on the application. For units with a complex device structure and high demands for electromagnetic compatibility, multilayer applications are in use. In special cases, it is possible to use ceramics as a carrier. There are usually some parts of the electronic circuit, resistors for example, designed as a thick-film circuit on the hybrid. In this case the electronic unit is manufactured as a solder hybrid or as a bond hybrid with direct-bonded integrated circuit devices. Some single applications exist with a flex-foil as a carrier for the electronic devices. These applications are limited to very special requirements.

The transmission control area requires some specially designed electronic devices, in particular, the output stages for the actuators of pressure regulation and torque converter clutch control. These actuators for pressure control have extremely high demands regarding accuracy of the actuator current over the whole temperature range and under all conditions independent of battery voltage and over the entire lifetime. There are some known applications of customer-specific integrated circuits or devices. Here, special attention paid to quality and reliability over the entire lifetime is necessary to meet the continuously increasing quality requirements of the automotive market. Currently, there is an increasing spread of surface-mounted devices in transmission control applications. This is why the unit size is continuously decreasing despite an increasing number of functions.

On the functional side, the hardware configuration can be divided into power supply, input signal transfer circuits, output stages, and microcontroller, including peripheral components and monitoring and safety circuits (Fig. 13.1). The power supply converts the vehicle battery voltage into a constant voltage required by the electronic devices inside the control unit. Accordingly, special attention must be paid to the protection of the internal devices against destruction by transients from the vehicle electrical system such as load dump, reverse battery polarity, and voltage peaks. Particular attention is also necessary in the design of the elec-

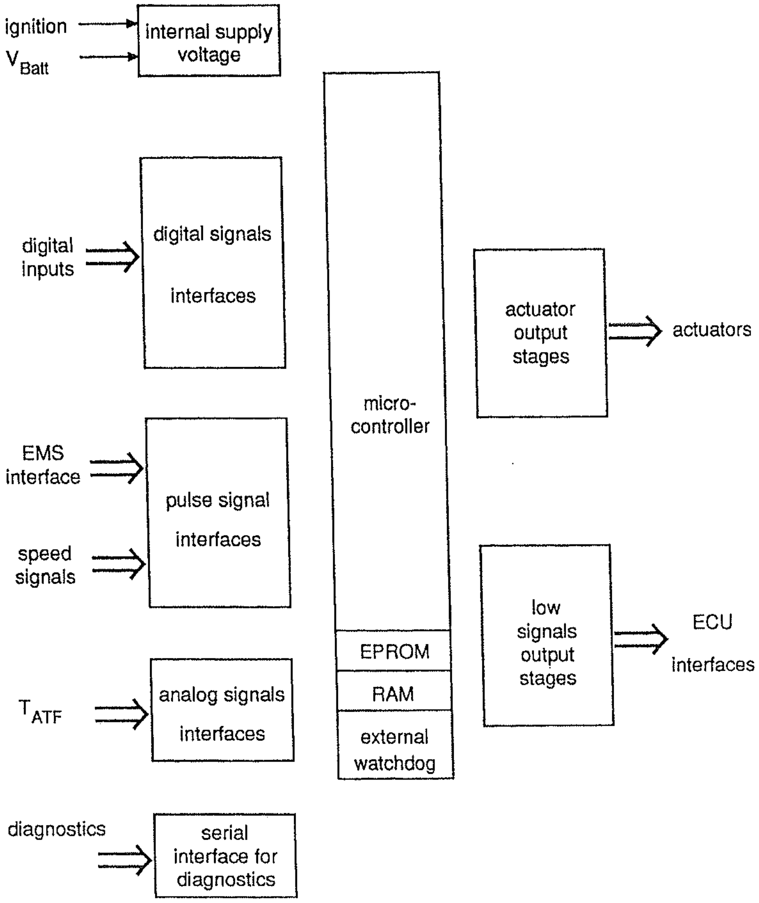


FIGURE 13.1 Overview of hardware parts.

tronic ground concept for the control unit, especially where the electromagnetic compatibility and RF interference is concerned. This is very important to prevent undesired gear shifting that may be troublesome for drivers. One part of the input circuit is the preparation of the digital signals, such as position switch, program selector, and kickdown switch. A second part is the transfer of the analog signals like ATF temperature and voltages according to potentiometer states. The third part is the interface to other electronic control units, especially to the engine management system. Here the single signal lines between the control units will be increasingly substituted by bus systems like CAN. The fourth part is the preparation of the transmission-specific signals from the speed sensors inside the transmission.

The calculators inside the control units are usually microcontrollers. The real-time requirements and the directly addressable program storage size of the selected microcontrollers are determined by the functions of the transmission control and the car environment. In present applications, either 8-bit or 16-bit microcontrollers are in use. There are systems with 32-bit microcontrollers in development for new, highly sophisticated control systems with increasing functional and extreme real-time requirements originating from the transmission concept. The memory devices for program and data are usually EPROMS. Their storage capacity is, in present applications, up to 64 Kbytes. Future applications will necessitate storage sizes up to 128 Kbytes. The failure storages for diagnostics and the storage for adaptive data are in conventional applications, battery voltage-supplied RAMs. These are increasingly being replaced by EEPROMs.

There are usually watchdog circuits in various configurations in use regarding safety and monitoring. These can be either a second, low-performance microcontroller, a customer-specific circuit, or a circuit with common available devices. The output stages can be divided into high-power stages for the transmission actuator control and low-power stages like lamp drivers or interfaces to other electronic control units. The low-power output stages are mostly conventional output drivers either in single or in multiple applications, which are mainly protected against short circuits and voltage overloads.

For the transmission solenoid control, special output stages are necessary, and they are specialized for operation with inductive actuators. The pressure regulation during shifting in some applications requires high accuracy and current-regulated output stages are needed. These are mainly designed as customer-specific devices. The type and number of solenoid output stages depend on the control philosophy of the transmission: they are generally of a special design for specific transmission applications. During the preparation of the speed sensor signals, attention must be paid to the electromagnetic compatibility and radio frequency interference conditions.

Software. The software within the electronic transmission control system is gaining increasing importance due to the increasing number of functions which, in turn, requires increasing software volume. The software for the control unit can be divided into two parts: the program and the data. The program structure is defined by the functions. The data are specific for the relevant program parts and have to be fixed during the calibration stage. The most difficult software requirements result from the real-time conditions coming from the transmission design. This is also the main criterion for the selection of the microcontroller (Fig. 13.2).

The program is generally made up in several parts:

- Software according to the special microcontroller hardware; e.g., I/O preparation and filter, driver functions, initialization of the microcontroller and the control unit, internal services for the controller peripheral devices, and internal software services like operating systems.
- Software coming from the defined functions, originating from specific transmission and car functions.
- Parts concerning safety functions like output switch-off, substitute values for the input signals, and safety states of the microcontroller environment in case of failures. Depending on the requirements, there can be a software watchdog or a hardware-configured watchdog circuit in use. The watchdog instruction is also part of the security software.

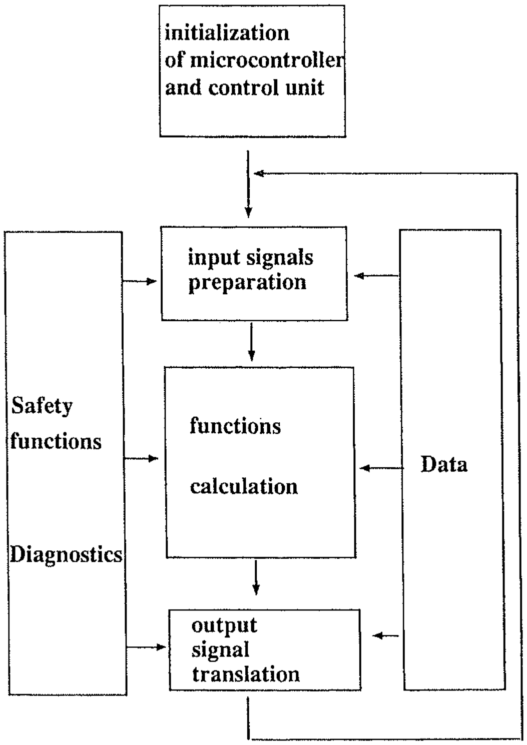


FIGURE 13.2 Software structure overview.

- Diagnostic and communication software for the self-test of the control unit and also the test of the control unit environment.

These functions are related to the defined functions of the electronic control system. Parts of the software component are usually the output stages monitoring, the input monitoring, and the diagnosis of the microcontroller environment. Failure handling and storage is gaining importance as system complexity increases. These diagnostic functions are also very useful for the service station to determine the reason for eventual problems. Part of these functions is reserved for the communication software needed for the test equipment to read the failures stored during car service. Current protocols are standardized communication protocols like ISO 9141. There is an increasing share of bus systems for communication with other electronic control units, using standardized protocols like CAN, VAN, or J1850. These bus systems allow an increasing unit function by changing software when other control units are added to the bus.

Most software models are written directly in an assembler to meet the real-time requirements and because there is a limited memory size in common mass production units. The number of powerful, cost-effective microcontrollers is continuously increasing. The availability of memory components with larger storage sizes suitable for automotive use is also rising, making it possible to use a higher programming language in future developments. This allows an ingenious structure of software models and an application of operating systems. This can be followed by an effective distribution of functions during and outside gear shifting with related time requirements and event management. This type of program structure improves the function of the electronic TCU because of the accelerated handling of time-critical functions during gear shifting.

The second software part, data, can be divided into fixed data, which is related to fixed attributes of the system; e.g., the number of actuators, and calibration data for system tuning. The calibration data can be adapted to changing parameters of the system such as the engine, vehicle, and transmission characteristics. The fixing of calibration data takes place during the tuning stage of the vehicle and has to be redetermined for each type of vehicle and transmission. With some applications, the calibration data are added to a basic program during the vehicle production according to different types of cars by the so-called end-of-line programming. This means that the units can be programmed with the calibration data with closed housings by a special interface. The share of software development in relation to the total development time is increasing continuously. The requirements for real-time behavior and memory size are rising in accordance with the considerably increasing demands for shift comfort and self-learning functions. This requires an ingenious structure of the software and an event-related distribution of software models, especially during gear shifting. The rising software complexity with simultaneously increasing quality requirements causes higher demands for software quality control.

13.2.3 Actuators

Electrohydraulic actuators are important components of the electronic transmission control systems.[2] Continuously operating actuators are used to modulate pressure, while switching actuators function as supply and discharge valves for shift-point control. Figure 13.3 provides a basic overview of these types of solenoids.

Important qualities for the use of actuators in ATs are low hydraulic resistance to achieve high flow rates, operation temperature range from –40 to +150 °C, small power loss, minimized heat dissipation in the ECU's output stages, small size and low weight, highest reliability in heavily contaminated oils, maximum accuracy and repeatability over lifetime, short reaction times, pressure range up to 2000 kPa, maximum vibration acceleration of 300 m/s^2, and high number of switch operations.

A very important aspect is that the hardware and software of the ECU be developed, taking into account the electrical specifications of the solenoid to obtain an optimized complete system concerning performance and cost.[6,7] For further details in design and application, refer to Sec. 10.3.5.

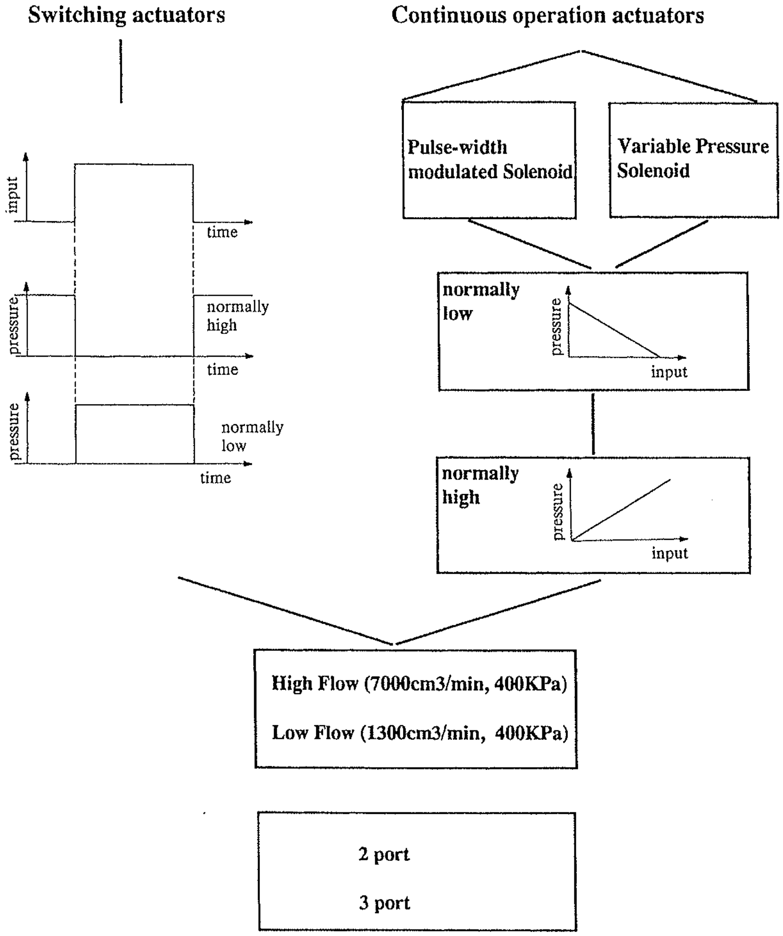
It should be noted that these characteristics can be varied over a wide range and that many other types of solenoids exist or are in development for the special requirements of new applications.

13.3 SYSTEM FUNCTIONS

Functions can be designated at systems functions if the individual components of the total electronic transmission control system cooperate efficiently to provide a desired behavior of the transmission and the vehicle. There are different stages of functionality which have different effects on driving behavior and shift characteristics (Fig. 13.4). In general, there is an increasing complexity of the system relating to all components to improve the translation of driver behavior into transmission action. That means that the expense of actuators, sensors, and links to other control units is increasing, as is the expense of the TCU software and hardware in the case of high-level requirements regarding driveability and shift comfort. Figure 13.4 shows three main areas. These will be discussed in detail in the following material.

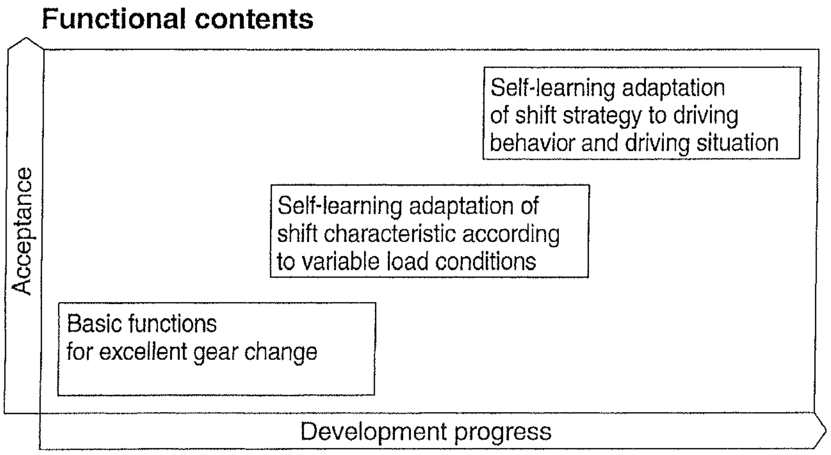
13.3.1 Basic Functions

The basic functions of the transmission control are the shift point control, the lockup control, engine torque control during shifting, related safety functions, and diagnostic functions for



Switching actuators
Continuous operation actuators
Pulse-width modulated Solenoid
Variable Pressure Solenoid
input
time
pressure
normally high
time
pressure
normally low
time
normally low
pressure
input
normally high
pressure
input
High Flow (7000cm3/min, 400KPa)
Low Flow (1300cm3/min, 400KPa)
2 port
3 port

FIGURE 13.3 Electrohydraulic actuators for automatic transmissions.



Functional contents
Self-learning adaptation of shift strategy to driving behavior and driving situation
Self-learning adaptation of shift characteristic according to variable load conditions
Basic functions for excellent gear change
Acceptance
Development progress

FIGURE 13.4 Relationship between driving characteristic and function complexity.

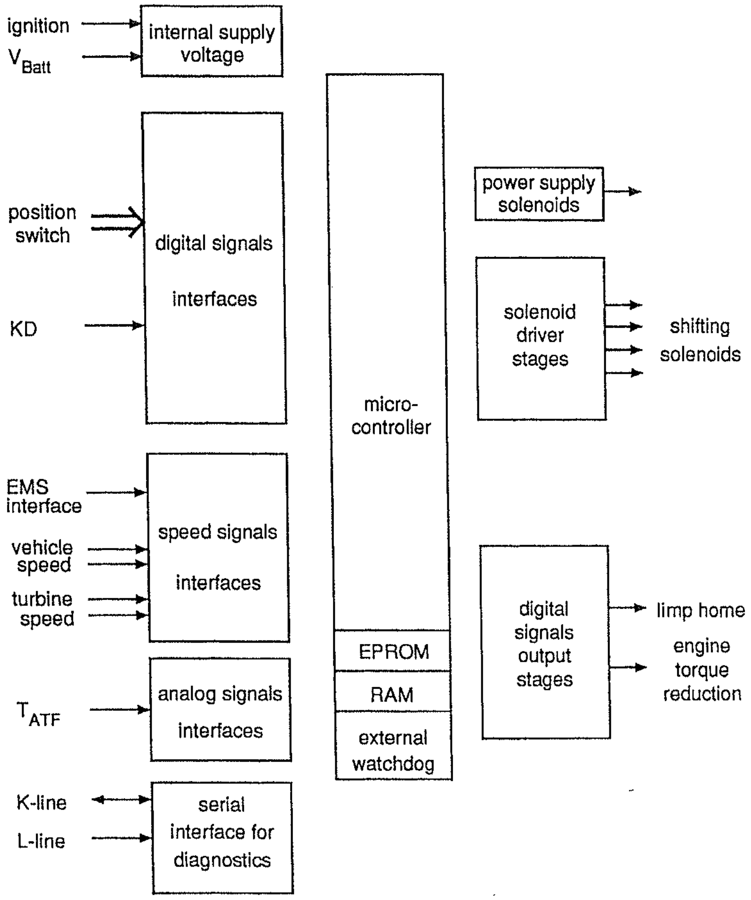


FIGURE 13.5 Structure of a basic transmission electronic control unit.

vehicle service. The pressure control in transmission systems with electrical operating possibilities for the pressure during and outside shifting can also be considered as a basic function. Figure 13.5 shows the necessary inputs and outputs as well as the block diagram of an electronic TCU suitable for the basic functions.

Shift Point Control. The basic shift point control uses shift maps, which are defined in data in the unit memory. These shift maps are selectable over a wide range. The shift point limitations are made, on the one hand, by the highest admissible engine speed for each application and, on the other hand, by the lowest engine speed that is practical for driving comfort and noise emission. The inputs of the shift point determination are the throttle position, the accelerator pedal position, and the vehicle speed (determined by the transmission output speed). Figure 13.6 shows a typical shift map application of a four-speed transmission.

To prevent overly frequent shifting between two gears, a hysteresis between the upshift and the downshift characteristic is incorporated. The hysteresis is determined by the desired shifting habit of the transmission and, alternatively, the car behavior. In the event that the particular shift characteristic is crossed by one of either of the two input valves, the electronic ECU releases the shift by activating the related actuators. This can be a direct shift into the

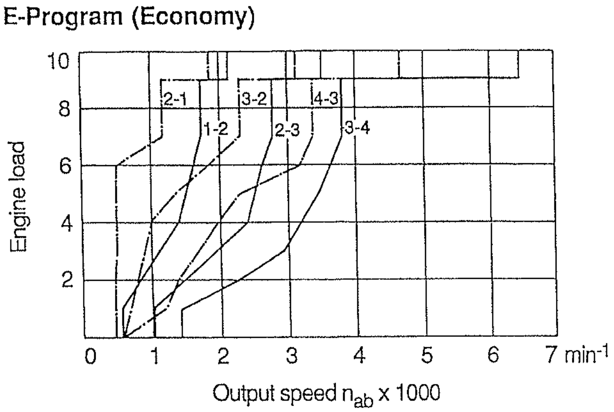


FIGURE 13.6 Shift characteristics of a four-speed application.

target gear or by a serial activation of specific actuators in a fixed sequence to the target gear, depending on the transmission hardware design.

Lockup Control/Torque Converter Clutch.[8] The torque converter clutch connects both functional components of the hydraulic converter, the pump and the turbine. The lockup of the clutch reduces the power losses coming from the torque converter slip. This is a permanent slip because it is necessary in principle to have a slip between the pumpwheel and the turbine to translate torque from the engine output to the transmission input. To increase the efficiency of the lockup, it is necessary to close the clutch as often as possible. On the other hand, the torque converter is an important component to prevent vibrations of the powertrain. The activation of the lockup is, therefore, a compromise between low fuel consumption and high driving comfort. The shift points of the lockup are determined in the same way as the determination of the shift point in the gear shift point control. Usually there is one separate characteristic curve for the lockup for each gear. To prevent powertrain vibrations, it is advisable to open the lockup during coasting to use the damping effect of the torque converter. In the case of a high positive gradient of the accelerator pedal with low engine speed, the converter clutch has to open to use the torque gain of the converter for better acceleration of the car. In some applications, the lockup is opened during shifting for improved shift comfort. After shifting, the lockup can be closed again. When driving in first gear, the lockup is usually open, because the time spent in first gear is usually very low and, therefore, the frequency of lockup shifting versus gear shifting becomes very high. This may result in decreased driving comfort. A second reason is the improved acceleration of the car in first gear when using the converter gain for wheel torque.

Engine Torque Control During Shifting.[8] The engine torque control requires an interface to an electronic engine management system. The target of the engine torque control, torque reduction during shifting, is to support the synchronization of the transmission and to prevent shift shocks.

In conventional applications, the engine torque reduction originates from an ignition angle control. The timing and absolute value of the ignition control depends on the operating conditions concerning actual engine torque and shifting type.

Upshift. The upshift occurs without an interruption of the tractive power. The engine torque reduction may be activated if the clutch of the target gear stays with the translation of torque. The beginning of the engine torque reduction is determined by the course of engine or transmission input speed. There it is important to detect a decreasing speed. The start of the torque reduction is characterized by a specific speed difference. The end of the torque reduction is activated at an applicable speed lead before reaching the synchronous speed of the new gear.

The power losses, which have to be picked up by the clutches, are dependent on the engine torque and the slipping time

$$Q = f \times (M_{\text{eng}} \times t_s + Q_{\text{kin}}) \tag{13.1}$$

where Q = power losses
M_{eng} = engine torque
t_s = slipping time
Q_{kin} = kinetic energy of revolving elements

It is possible to reduce the temperature stress to the clutches by reducing the engine torque and, consequently, by increasing the slipping time at a fixed possible maximum power loss Q [Eq. (13.1)]. Figure 13.7 shows a typical upshift characteristic.

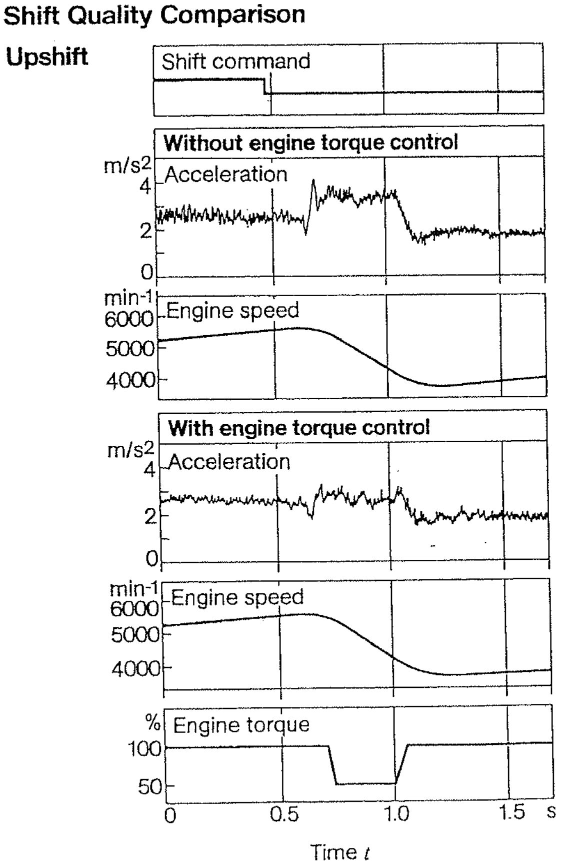


FIGURE 13.7 Timing of engine torque reduction during upshift.

Downshift. Downshift under driving conditions results in a short interruption of the tractive power. At the synchronous point, the tractive power is in operation. The higher revolving energy, on the other hand, results in undesired vibrations of the powertrain. To prevent such vibrations, it is necessary to reduce the engine output torque before reaching the synchronous point of the new gear. When the transmission input speed reaches the synchronous speed of the new gear, the engine torque has to increase to the nominal value. The increase is usually applied as a torque ramp. Figure 13.8 shows a typical characteristic of a downshift. The values and timing of the engine torque reduction are generally part of the special calibration data for each combination of vehicle, engine, and transmission.

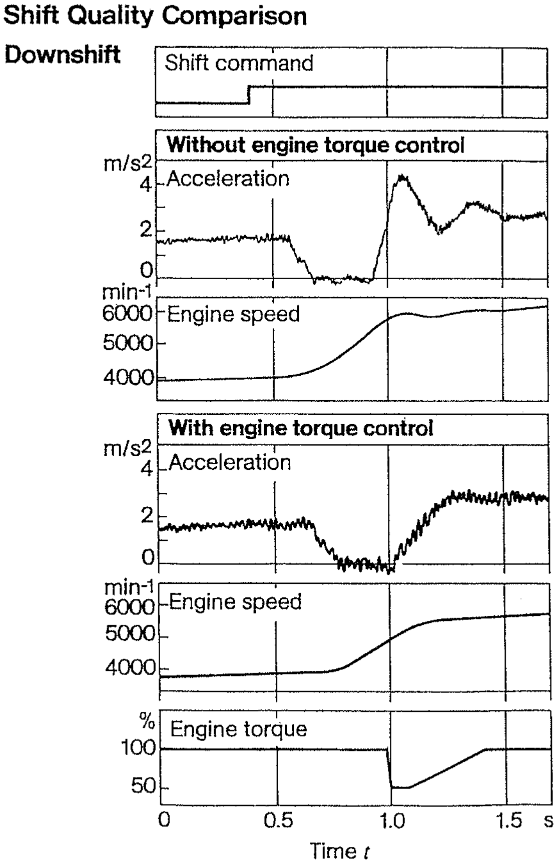


FIGURE 13.8 Timing of engine torque reduction during downshift.

Pressure Control.[8] The timing and absolute values of the pressure, which is responsible for the torque translation of the friction elements, is, aside from the engine torque reduction, the most important influence to shift comfort. The electronic TCU offers additional possibilities for better function than a conventional hydraulic system.

The pressure values during and outside shifting can be calculated by different algorithms or can be determined by characteristic maps. The inputs for a pressure calculation are engine torque, transmission input speed, turbine torque, throttle position, and so on. The inputs depend on the special signal availability in different systems as well as the requirement concerning shift comfort. The variable pressure components are usually added to a constant pressure value according to the different transmission designs. Equation (13.2) gives a typical algorithm for a pressure calculation.

$$P_{\text{mod}} = P_{\text{const}} + k_n \times P_n + k_{\text{tor}} \times P_{\text{tor}} + k_s \times P_s \tag{13.2}$$

where P_{mod} = pressure
P_{const} = constant pressure value
k_n = adaptation factor for input speed
P_n = pressure component dependent on the revolution signal
k_{tor} = adaptation factor for engine torque
P_{tor} = pressure component dependent on torque
k_s = adaptation factor for vehicle speed
P_s = pressure component dependent on vehicle speed

During applications, the factors must be defined in the calibration phase. In general, the determination of these factors requires many vehicle tests, because the dynamic characteristic of the total system has an important influence on shift comfort. Another possibility for the pressure determination is to use characteristic maps which have to be defined during the calibration phase. This kind of pressure determination allows an improved selection of the optimum pressure at various extreme points independent of an algorithm.

Safety and Diagnostic Functions. The functions, which are usually known as diagnostic functions of the electronic TCU, can be divided into real safety functions to prevent critical driving conditions and diagnostic functions which affect an increasing availability of the car and a better failure detection for servicing. The boundary between safety and diagnostic functions depends on the philosophy of the automotive manufacturer. In the category of real safety functions belong all security functions that prevent uncontrollable shifting, especially unintended downshifting. One section is the monitoring of the microcontroller and its related peripheral devices. The monitoring of the transmission, like gear ratio detection, is also a part of this functional block, as are the actuator and speed sensor monitoring. The microcontroller monitor is usually a watchdog circuit. One possibility is to use the controller internal watchdog. In common applications, it is necessary to use an external watchdog circuit for safety reasons. This can be done with a second, low-performance microcontroller or by a separate hardware watchdog designed as an ASIC or as a conventional circuit device. Usually there is a safety logic circuit connected to the watchdog, which, in the case of a microcontroller breakdown, activates the failure signal and switches the outputs for the transmission actuators to a safety condition.

For the detection of the watchdog, it is necessary to test the watchdog function after each power-on during the electronic initialization. The monitoring of the controller peripheral components, in general EPROM, RAM, and chip-select circuits, works continuously with specific algorithms; e.g., by writing fixed data values to the storage cells and following comparison with the read value or by checksum comparison with fixed sum values. The actuator monitoring includes detection of short circuit to supply voltage and ground, as well as open-load conditions.

In case of actuator malfunction, the limp home mode is selected. This means that the transmission runs in a fixed, safe gear, depending on the driving conditions. The safe state of the actuators is the noncurrent condition, which is secured by the electronic control unit. The control unit can put the output stages into the noncurrent stage separate for each output or by a common supply switch, usually a relay or a transistor. There are some applications that use a combination of both the watchdog and safety circuits.

The monitoring of the transmission-specific sensors, such as input speed, output speed, and oil temperature, works as a plausability check. For example the transmission input speed can be calculated as a combination of the transmission output speed and the gear ratio. In case of a detected speed sensor malfunction, the limp home mode is generally required. With a temperature sensor failure, the TCU usually works with a substitute value.

The diagnostic functions, which facilitate the finding of failures in the service station, contain the failure storage and the communication to the service tester, which allow the stored failures of the electronic TCU to be read. The communication between the control unit and the service tester is mainly car manufacturer-specific and must be defined by the car manufacturer before going into series production. The communication runs on a bidirectional, separate communication link.

The failure storage takes place in a nonvolatile memory device; e.g., in a permanent supplied RAM or in an EEPROM. It is also possible to store sporadic failures to detect such problems during the next service. The failure codes, number of stored failures, the handling of the failure storage, as well as the reaction of the TCU in case of a particular failure, is manufacturer-specific and is part of the unit specification. The real safety functions are part of the basic functions of an electronic TCU. The diagnostic functions concerning service tester and communication protocols are, over a wide range, manufacturer-specific. These range from a simple blink code up to a real self-test of the electronic unit, including all peripheral components.

13.3.2 Improvement of Shift Control

In a second development stage, the basic functions can be revised by a modification of the software functions and by adding new parts to the basic functions. This action results in a significant enhancement of the driving and shifting comfort. By a revision of the basic safety and diagnostic functions with so-called substitute functions, it is possible to increase the availability of the vehicle with AT as well as the driveability in case of a malfunction.

Shift Point Control. The basic function can be improved significantly by adding a software function, the so-called adaptive shift point control.[8] This function requires only signals which are available in an electronic TCU with basic functions. The adaptive shift point control is able to prevent an often-criticized attribute, the tendency for shift hunting especially when hill climbing and under heavy load conditions.

The adaptive function calculates the vehicle acceleration from the transmission output speed over time. The value of the actual acceleration in relation to a set value of the acceleration is the input for the shift point correction. The set value is given by the traction resistance characteristic. For a certain difference between set and actual value, the adaptation of the shift point occurs. The dimension of the shift point correction can be determined by calibration data and depends in general also on the actual vehicle speed and the engine load.

The shift point correction leads to higher hysteresis between upshift and downshift characteristics. With a high difference between set and actual values, it is also possible to forbid certain gears. The return to the basic shift point control is organized by software and can be fixed by calibration data. Usually, in the case of power-on, the adaptive shift point control is reset (Figs. 13.9 and 13.10).

In addition to these functions, different shift maps can be implemented into the data field of the TCU. For example, it is possible to have one shift map for low fuel consumption, which has shift points in the range of the best efficiency of the engine, and additionally to have another map for power operation, where the shift points are placed at points of highest engine output power. The character and number of different shift maps can be selected over a wide range. The choice of the different shift maps can be done by a selector push button or switch commanded by the driver. In further applications, the changing of the different shift programs is possible by self-learning strategies. It is also possible to implement a manual program in which fixed gears are specific to predetermined positions of the selector lever.

Lockup Control. There are some additional functions which can improve considerably the shift comfort of the lockup. In a first step, it is possible to replace the on/off control of the lockup actuator by a pulse control during opening and closing. This can be achieved using conventional hardware only by a software modification. In a further step, the on/off solenoid is replaced by a pressure regulator or a PWM solenoid.

By coordinating intelligent control strategies and corresponding output stages within the electronic TCU, a considerable improvement of the shift behavior of the lockup results. Here

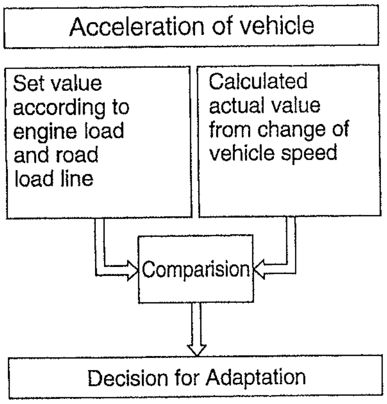

Acceleration of vehicle
Set value according to engine load and road load line
Calculated actual value from change of vehicle speed
Comparision
Decision for Adaptation

FIGURE 13.9 Basic principle of adaptive shift point control.

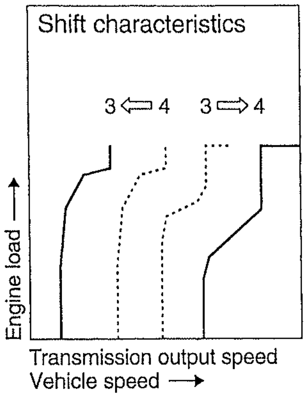


FIGURE 13.10 Shift characteristics before (- - -) and after (—) adaptation.

it is possible to close the lockup at low engine speed and low engine load with good shift comfort, resulting in decreased fuel consumption.

Engine Torque Reduction During Gear Shifting. By an improved interface design to the engine management system, it is possible to extend the engine torque reduction function. It is necessary to use a PWM signal with related fixed values or a bus interface. The engine torque reduction is controlled directly by the TCU. The advantage of such an interface is an independent calibration of the TCU data over a wide range without changing the engine management data. A further advantage is the improved possibility for the coordination of the engine torque reduction and the pressure control within the TCU. The improvement of this interface can be extended up to a real torque interface, especially when using a bus communication link.

Pressure Control.[8] The pressure control can be improved in a similar way as the shift point control with an adaptive software strategy. The required inputs for the adaptive pressure control are calculated from available signals in the transmission control. The main reasons for the implementation of the adaptive pressure control are the variations of the attributes of the transmission components like clutch surfaces and oil quality as well as the changing engine output torque over the lifetime of the car.

The principle of adaptive pressure control is a comparison of a set value for the shift time with an actual value, measured by the transmission input speed course. At a specific difference of the set value to the actual value, the pressure value is corrected by a certain increment in the positive or negative direction. The original adaptation time and the pressure value increment were fixed during the calibration phase. For safety reasons, the total deviation of the pressure value from a given value is limited, depending on the particular application. Usually the correction values are stored in the nonvolatile memory to have the correct values available after power-on of the electronic TCU.

Safety and Diagnostic Functions. The safety functions extend over better monitoring of the selector lever and functions concerning misuse by the driver. With a corresponding transmission hardware design, the implementation of a reverse gear inhibit function is possible; i.e, above a certain vehicle speed, the position R is blocked hydraulically by a single solenoid or by a particular solenoid combination commanded by the electronic TCU. This function prevents the destruction of the transmission in the event of an unintentional shift to the reverse gear. Downshift prevention is part of the safety function, especially during manual shifting by the driver. Here the synchronous speed of the new gear is calculated and compared with the admissable maximum engine speed. In the case of a calculated synchronous speed above the maximum engine speed, the downshift is prohibited by the TCU. This function can be supported by an overrun safeguard which releases the limp home mode in case of exceeding the admissable maximum engine speed.

All of those functions can be extended and have to be defined during the development stage by the automobile transmission and electronic TCU manufacturers. To increase availability of the AT system, even with the failure of certain signals, it is possible to provide a substitute operation with better drivability than in the limp home mode. This can be done by substitute functions. The electronic TCU falls back on substitute values or signals in the case of a breakdown of certain interfaces. There is, for example, the possibility to run with a programmable fixed throttle value with a breakdown of the throttle position signal. This results in a reduction of the shift characteristics to shift points. Shifting into all gears is possible, however, with reduced shift comfort. A further method is to use secondary signals in case the original signals break down. For example, the calculation of vehicle speed can be from wheel speed during breakdown of the transmission output speed signal. This technique usually requires a connection between ABS and transmission control. The third variant is the canceling of certain functions if the necessary input signals are missed. For example, in the case of a kickdown switch failure, the kickdown function is canceled. This results in no downshift after operation of the kickdown. Downshifts are nevertheless still possible via the full-throttle opening point according to the full-load shift characteristic.

The availability and driveability of automobiles equipped with electronic TCU in case of system failures can be improved significantly with the implementation of substitute functions. This results in a considerable increase in acceptance by the drivers of automobiles with electronic transmission control.

13.3.3 Adaptation to Driver's Behavior and Traffic Situations

In certain driving conditions, some disadvantages of the conventional AT can be prevented by using self-learning strategies.[9] This is especially valid when improving the compromise in the shift characteristics regarding gear selection under particular driving conditions and under difficult environmental conditions. The intention of the self-learning functions is to provide the appropriate shift characteristic suitable to the driver under all driving conditions. Additionally, the behavior of the car under special conditions can be improved by suitable functions. Available input signals of the car, provided by the related electronic TCUs from interfaces and communication links, are processed by the TCU with specific algorithms. The self-learning functions can be divided into a long respectively medium term adaptation for driver's style detection and into a short-term adaptation which reacts to the present driving situation, such as hills or curves.

The core of the adaptive strategies is the driver's style detection. The driver's style can be detected by monitoring of the accelerator pedal movements. The inputs are operation speed, operation frequency, and the rating position of the accelerator pedal. These inputs are processed depending on priorities with special algorithms related to the desired driving behavior of the car. The calculated driver style is related to certain shift maps. There is a large choice of shift maps available. With the currently known applications, there are mostly four different shift maps ranging from fuel economic to extremely sporty vehicle behavior. The calculated driver's style can also depend on the actual vehicle speed and the share of constant driving conditions during a certain driving cycle. These self-learning functions can be calibrated by the car manufacturer, depending on his philosophy and target market. In this way, the number of shift maps and the speed of the adaptation have the main influence. A further possibility to match the driver's style is by rating the accelerator pedal operation during vehicle start, for example, after a red light stop. In this way the operation speed and frequency of the accelerator pedal below a certain vehicle speed can be interpreted and calculated as part of the driver's style rating. In the event of kickdown, the shift maps of the driver's style rating are shut down by a priority command. The driver has the usual behavior of the car during kickdown, generally a downshift, providing no other safety function is in operation.

To prevent shift hunting, the self-learning functions are carried out over a long respectively medium term adaptation with the adaptation timer ranging from several seconds up to one minute. The second part of the self-learning functions is the driving condition detection. There is a correlation between the input signals of the transmission control and the driving condition.

One of the main disadvantages of a conventional electronic transmission control is the upshifting at constant vehicle speed by crossing the upshift characteristic with a reduction of the accelerator pedal angle. This results in an unintended gear shift, especially when cornering and when approaching a crossing or an obstacle. To prevent these gear shifts it is possible to use so-called upshift prevention. Cornering can be detected by the acceleration of the car along the driving direction related to the vehicle speed. The vehicle speed is calculated from the transmission output speed. The acceleration can be detected by an acceleration sensor or by the difference between the nondriven wheel speeds. In this way it is possible to prevent the upshift when cornering, resulting in a considerable improvement in vehicle stability.

The detection of a crossing or obstacle approach is possible by the detection of a fast off condition of the accelerator pedal. At a certain gradient of the pedal position, the upshift is prevented. This is a considerable advantage especially when overtaking low-speed vehicles. With this strategy the correct gear is available without a shift delay.

Another part of the driving situation detection is the recognition of uphill driving and full-load conditions. This is possible by adding special functions to the adaptive gear shift control. When driving downhill, it makes sense to support the engine braking effect for a better deceleration of the car. Downhill driving can be detected by a comparison of throttle position and vehicle speed gradient. An upshift is prevented and, in some special cases, a downshift is activated by the electronic TCU.

A further section of the self-learning functions is the environmental monitoring with related shift strategies. A special application can be a self-learning winter program. The wheel slip of the driven wheels is compared with a set value of a combination of given wheel torque and vehicle speed. When exceeding a set limit of wheel slip, a special shift strategy is chosen. For example, the vehicle starts off in second gear or an upshift takes place at lower engine speeds.

The development of adaptive shift strategies started a few years ago and is currently one of the main areas in electronic transmission development. The efficiency of the self-learning functions has led to a wide acceptance of AT-equipped vehicles. The future development concerning new adaptive functions and an improvement of the already known functions is an important area in control development. This can be supported by an increasing share of electronic units and interfaces for the communication between units. With multiple use of sensors providing the necessary input signals, the total system gains increased functionality, especially with bus systems.

At present, an increasing share of manual programs with an AT can be registrated. The driver instructs the AT to shift via a switch or a push button. In this manner, the driver can operate the AT like a manual gearbox independently of other shift maps, with only the safety functions in operation. This has led to a broad acceptance, especially in the sports car market. These functions can all be calibrated and applied by the car manufacturer with data relating to his philosophy and to the target market. The result is the prevention of the known disadvantages of the conventional AT control without canceling the advantages in driving comfort and safety.

13.4 COMMUNICATIONS WITH OTHER ELECTRONIC CONTROL UNITS

With the existence of electronic control units for various applications in vehicles, many opportunities exist to link these ECUs and to establish communications between them. The main partner of the TCU is the engine management system. Due to the coupling of engine and transmission within the vehicle powertrain, it is necessary to have an interface between these ECUs for a functional coupling and an interchange of signals. It is essential for the pressure control inside the transmission control to sensor the engine load, the engine speed, and the throttle position. The engine torque reduction during shifting is also important to establish a good shift comfort and a satisfactory lifetime for the clutches. By handing over certain signals like position lever state, lockup condition, or shift commands to the engine management, the driving comfort of the vehicle can be improved significantly. An interface to ABS and traction control is useful for some self-learning functions in the transmission control when using the wheel speeds.

It is possible to implement certain shift strategies in the transmission control as an active support for ABS and traction control. A link to the electronic throttle control or cruise control makes it possible to optimize certain functions for the total vehicle. By interfaces between the ECUs, a reduction of the sensor expense results by a multiple use via communications. Suitable links include, especially, PWM or bus configurations for trouble-free communication. Bus systems in particular have the advantage of the link-up of additional ECUs without changing their existing hardware. Additional coupling requires only a software change. The interchange of required supplementary signals for new functions is possible without any problems. An example of powertrain management by coupling the powertrain ECUs to achieve lower fuel consumption, simultaneously improving the driveability, is described as follows.

13.5 OPTIMIZATION OF THE DRIVETRAIN

The newest generation of transmission controllers has overcome the former disadvantage regarding fuel efficiency. Adaptive functions in cooperation with carefully designed torque converter clutch control,[8] which allows the clutch to be closed even at low gears, have improved fuel consumption significantly. Based on the driver's behavior, together with an adaptive shift strategy as previously described, part of the TCU's adaptive program software may select an economy or even super-economy shift strategy whenever possible. There is, however, still more potential for fuel economy by optimization of the drivetrain.

The concept called Mastershift[10] is shown in Fig. 13.11. The basic idea is to interpret the accelerator pedal position as an acceleration request. That acceleration request, or a request for wheel torque, has to be converted by operating the engine at high torque, i.e., open throttle and low rpm values. In order to realize this, it is necessary to use an electronic throttle control system. The communication between the electronic throttle, the engine, and transmission is shown in Fig. 13.12.

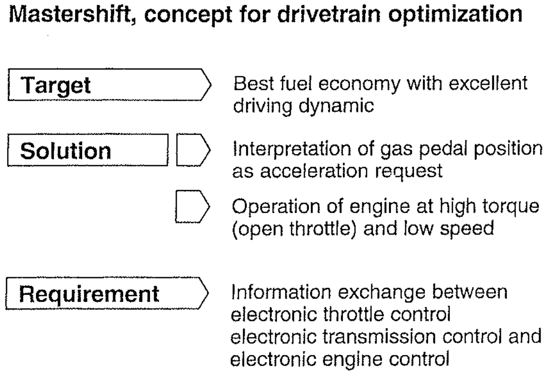

Mastershift, concept for drivetrain optimization
Target
Best fuel economy with excellent driving dynamic
Solution
Interpretation of gas pedal position as acceleration request
Operation of engine at high torque (open throttle) and low speed
Requirement
Information exchange between electronic throttle control electronic transmission control and electronic engine control

FIGURE 13.11 Drivetrain operation

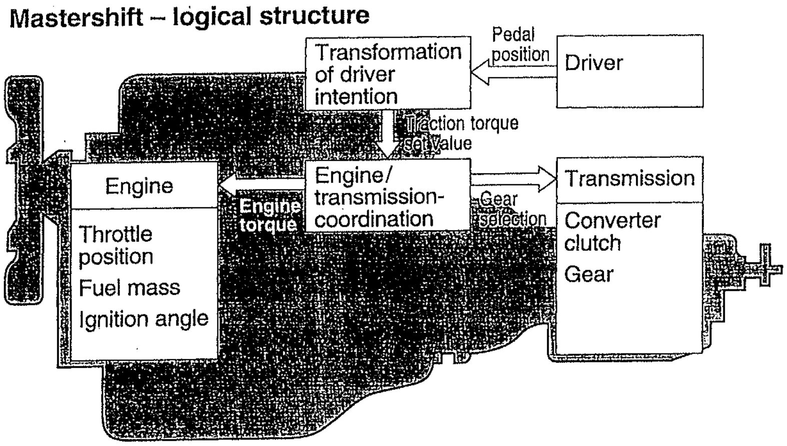


FIGURE 13.12 Mastershift: logical structure and communication between different control systems.

In such a system, a well-defined coordination between the engine torque, mainly given by throttle position (air mass), fuel mass, and ignition angle on one side and selection of the appropriate gear including torque converter clutch on the other side, is imperative. Depending on the type of engine, fuel consumption can be reduced further by 5 to 10 percent with this optimized Mastershift concept. Because the average engine operation is at higher torque levels compared to standard systems, a greater number of gear shifts may occur. This is important to guarantee optimal shift comfort. Figure 13.13 shows how that can be accomplished by using the additional degree of freedom given by the electronic throttle control. It is possible to operate the throttle angle during the gear shift in such a way as to achieve constant wheel torque before and after downshifts.

13.6 FUTURE DEVELOPMENTS

In future years, development work will be concentrated on redesign of hardware components for cost reduction, improvement of yield to reduce fuel consumption, and improvement of driveability. A good approach to meet cost targets on the electronic hardware side would be to integrate two or more individual control modules into a common housing. Regarding the electronic components, one could continue using two separate microcontrollers. This would have the advantage that the software development and application could be done individually for two different systems, for example engine and transmission controllers. Another approach could be to mount the TCU on the transmission housing itself. This could lead to a significant reduction in the expense for the wiring harness. Here, however, the problem of hostile ambient temperatures on electronic components has to be solved. Today's stand-alone actuators could be integrated into a common housing similar to the solution shown by Chrysler Corp. in its A 604 transmission.

The improvement of the yield is a main topic for designers of ATs. Oil pumps and torque converters are a major source of energy losses. A significant improvement of yield will be possible as soon as torque converter clutches are available with the capability for continuous slip operation. The torque converter clutch can then be operated in low gears and at low engine speeds without facing problems from drivetrain oscillations and/or noise emission.

The driveability is the most important feature for the drivers' acceptance of ATs. In addition to the self-adaptive functions described, the implementation of shift strategies benefiting from control algorithms using fuzzy theory may further improve driveability.

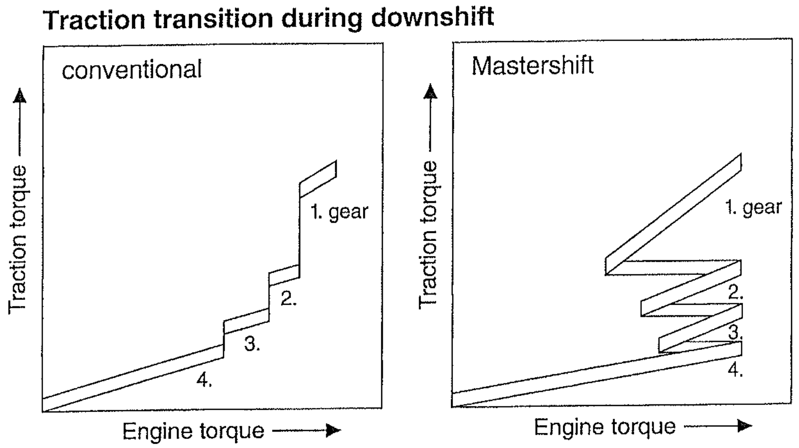


FIGURE 13.13 Constant traction torque by operation of throttle opening during gear shift.

GLOSSARY

ASIC Application-specific integrated circuit.

AT Automatic transmission.

ATF Automatic transmission fluid.

CAN Controller area network.

CVT Continuously variable transmission.

EEPROM Electrically erasable and programmable read-only memory.

EMC Electromagnetic compatibility.

EPROM Erasable programmable read-only memory.

PWM Pulse-width modulation.

RAM Random access memory.

RFI Radio frequency interference.

TCC Torque converter clutch.

TCU Transmission control unit.

REFERENCES

1. F. Kucukay and Lorenz, K., "Das neue Fünfgang-Automatikgetriebe für V8-Motoren in der 7er Baureihe von BMW," *ATZ Automobiltechnische Zeitschrift 94,* Heft 7/8, 1992.
2. K. Neuffer, "Recent development of AT-control: adaptive functions and actuators," Symposium No. 9313, *Advanced Technologies in Automotive Propulsion Systems,* Society of Automotive Engineers of Japan Inc., 1993, pp. 42–49.

3. J. G. Eleftherakis and Khalil, A., "Development of a laboratory test contaminant for transmissions," SAE Paper 90 0561, Society of Automotive Engineers, Warrendale, Pa.

4. B. Aldefeld, "Numerical calculation of electromagnetic actuators," *Archiv für Elektrotechnik,* Bd. 61, 1979, pp. 347–352

5. K. Hasuuaka, Takagi, K., and Sinji, W., "A study on electro-hydraulic control for automatic transmissions," SAE Paper 89 2000, Society of Automotive Engineers, Warrendale, Pa.

6. P. C. Sen, "Principles of electric machines and power electronics," J. Wiley, New York, 1989.

7. "Method and Apparatus to Convert an Electrical Value into a Mechanical Position by Using an Electromagnetic Element Subject to Hysteresis," U. S. Patent 4,577,143 March 18, 1986.

8. K. Neuffer, "Electronische Getriebesteuerung von Bosch," *ATZ Automobiltechnische Zeitschrift 94,* Heft 9, 1992, pp. 442–449.

9. A. Welter, et al., "Die Adaptive Getriegesteuerung für Automatikgetriebe der BMW-Fahrzeuge mit Zwolfzylindermotor," *ATZ Automobiltechnische Zeitschrift 94,* 1992, pp. 428–436.

10. H. M. Streib and R. Leonhard, "Hierarchical control strategy for powertrain function," *XXIV Fisita Congress,* London, 1992.

ABOUT THE AUTHORS

KURT NEUFFER is responsible at Robert Bosch GmbH for the development of electronic control units for automatic transmissions and also for the development of actuators. He was educated in electronics engineering at the University of Stuttgart and holds a Dr. Ing. in the field of basic semiconductor research. He has been in the field of automotive component development for 10 years.

WOLFGANG BULLMER is responsible at Bosch for systems and software development of electronic control units for automatic transmissions. He was educated in electronics engineering at the University of Stuttgart. He has been working in the area of transmission control unit development for eight years.

WERNER BREHM is a Bosch section manager for the design of electrohydraulic actuators used in electronically controlled automatic transmissions. He was educated in mechanical engineering at the University of Stuttgart and has worked on components engineering for antilock braking systems in passenger cars.

CHAPTER 14
CRUISE CONTROL

Richard Valentine
Motorola Inc.

14.1 CRUISE CONTROL SYSTEM

A vehicle speed control system can range from a simple throttle latching device to a sophisticated digital controller that constantly maintains a set speed under varying driving conditions. The next generation of electronic speed control systems will probably still use a separate module (black box), the same as present-day systems, but will share data from the engine, ABS, and transmission control systems. Futuristic cruise control systems that include radar sensors to measure the rate of closure to other vehicles and adjust the speed to maintain a constant distance are possible but need significant cost reductions for widespread private vehicle usage.

The objective of an automatic vehicle cruise control is to sustain a steady speed under varying road conditions, thus allowing the vehicle operator to relax from constant foot throttle manipulation. In some cases, the cruise control system may actually improve the vehicle's fuel efficiency value by limiting throttle excursions to small steps. By using the power and speed of a microcontroller device and fuzzy logic software design, an excellent cruise control system can be designed.

14.1.1 Functional Elements

The cruise control system is a closed-loop speed control as shown in Fig. 14.1. The key input signals are the driver's speed setpoint and the vehicle's actual speed. Other important inputs are the faster-accel/slower-coast driver adjustments, resume, on/off, brake switch, and engine control messages. The key output signals are the throttle control servo actuator values. Additional output signals include cruise ON and service indicators, plus messages to the engine and/or transmission control system and possibly data for diagnostics.

14.1.2 Performance Expectations

The ideal cruise system features would include the following specifications:

- *Speed performance:* ±0.5 m/h control at less than 5 percent grade, and ±1 m/h control or vehicle limit over 5 percent grade.
- *Reliability:* Circuit designed to withstand overvoltage transients, reverse voltages, and power dissipation of components kept to minimum.

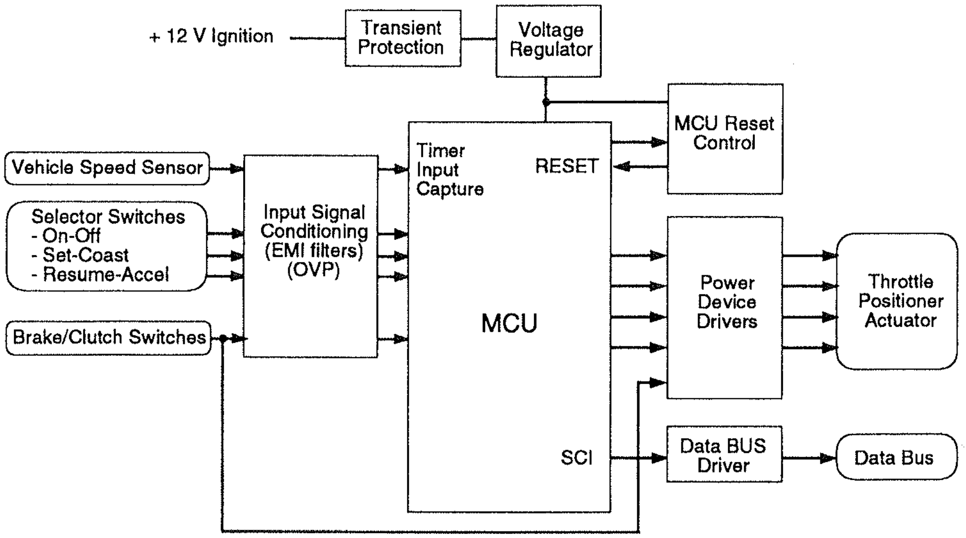


FIGURE 14.1 Cruise control system.

- *Application options:* By changing EEPROM via a simple serial data interface or over the MUX network, the cruise software can be upgraded and optimized for specific vehicle types. These provisions allow for various sensors, servos, and speed ranges.
- *Driver adaptability:* The response time of the cruise control can be adjusted to match the driver's preferences within the constraints of the vehicle's performance.
- *Favorable price-to-performance ratio:* The use of integrated actuator drivers and a high-functionality MCU reduce component counts, increase reliability, and decrease the cruise control module's footprint.

14.1.3 Safety Considerations (Failsafe)

Several safety factors need to be considered for a vehicle speed control design. The most basic is a method designed into the throttle control circuit to insure a failsafe mode of operation in the event that the microcontroller or actuator drivers should fail. This electronic failsafe circuit shuts off the control servos so that the throttle linkage will be released when the brake switch or cruise off switch is activated, no matter the condition of the MCU or servo actuator control transistors. (This assumes the actuators are mechanically in good shape and will release.)

Other safety-related items include program code to detect abnormal operating conditions and preserving into memory the data points associated with the abnormal condition for later diagnostics. Abnormal conditions, for example, could be an intermittent vehicle speed sensor, or erratic driver switch signals. A test could also be made during the initial ignition "key on time" plus any time the cruise is activated to verify the integrity of the cruise system, with any faults resulting in a warning indicator to the driver. Obviously, the most serious fault to avoid is runaway acceleration. Continuous monitoring of the MCU and key control elements will help minimize the potential for this type of fault.

14.2 MICROCONTROLLER REQUIREMENTS FOR CRUISE CONTROL

The MCU for cruise control applications requires high functionality. The MCU would include the following:

- a precise internal timebase for the speed measurement calculations
- A/D inputs
- PWM outputs
- timer input capture
- timer output compares
- serial data port (MUX port)
- internal watchdog
- EEPROM
- low-power CMOS technology

14.2.1 Input Signals

The speed sensor is one of the most critical parts in the system, because the microcontroller calculates the vehicle speed from the speed sensor's signal to within ⅟₃₂ m/h. Any speedometer cable whip or oscillation can cause errors to be introduced into the speed calculation. An averaging routine in the speed calculations can minimize this effect. The speedometer sensor drives the microcontroller's timer input capture line or the external interrupt line. The MCU then calculates the vehicle's speed from the frequency of the sensor signals and the MCU internal timebase. The vehicle's speed value is continually updated and stored into RAM for use by the basic speed control program. Speed sensors traditionally have been a simple ac generator located in the transmission or speedometer cable. The ac generator produces an ac voltage waveform with its frequency proportional to the sensor's rpm and vehicle speed. Optical sensors in the speedometer head can also be incorporated. Usually the speed sensor produces a number of pulses or cycles per km or mile. With the increasing ABS system usage, a backup speed sensor value could be obtained from the ABS wheel speed sensors. The ABS speed data could be obtained by way of a MUX network.

The user command switch signals could either be single MCU input lines to each switch contact or a more complex analog resistor divider type to an A/D input line. Other input signals of interest to the cruise system program would be throttle position, transmission or clutch status, A/C status, actuator diagnostics, engine status, etc., which could be obtained over the MUX data network.

14.2.2 Program Flow

The microcontroller is programmed to measure the rate of vehicle speed and note how much, and in which direction, the vehicle speed is drifting. The standard PI (proportional-integral) method produces one output signal *p* that is proportional to the difference between the set-speed and actual vehicle speed (the error value) by a proportional gain block *Kp*. Another signal *i* is generated that ramps up or down at a rate set by the error signal magnitude. The gains of both *Ki* and *Kp* are chosen to provide a quick response, but with little instability. In effect, the PI system adds up the error rate over time, and, therefore, if an underspeed condition occurs as in a long uphill grade, the error signal will begin to greatly increase to try to compensate. Under level driving conditions, the integral control block *Ki* will tend toward zero

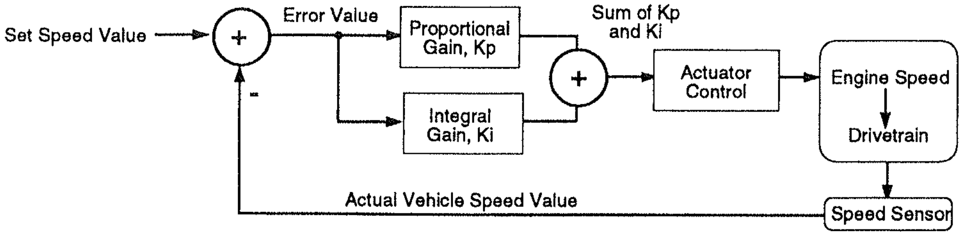


FIGURE 14.2 PI speed error control.

because there is less error over time. The vehicle's weight, engine performance, and rolling resistance all factor in to determine the PI gain constants. In summary, the PI method allows fast response to abrupt grades or mountains and stable operation under light grades or hills. Figure 14.2 shows the traditional PI cruise control diagram.

14.2.3 Output Controls

When the error signal has been computed, an output signal to the servo actuators is generated to increase, hold, or decrease the throttle position. The servo is updated at a rate that is within the servo's mechanical operating specifications, which could be several milliseconds. The error signal can be computed at a much faster rate and, therefore, gives extra time for some averaging of the vehicle speed sensor signal.

Throttle positioning is traditionally either a vacuum type servo or motor. The vacuum supply to the vacuum servo/actuator is discharged as a failsafe measure whenever the brake system is engaged in addition to the normal turn-off of the actuator driver coils. Electric servo type motors require more complex drive electronics and some type of mechanical failsafe linked backed to the brake system.

14.3 CRUISE CONTROL SOFTWARE

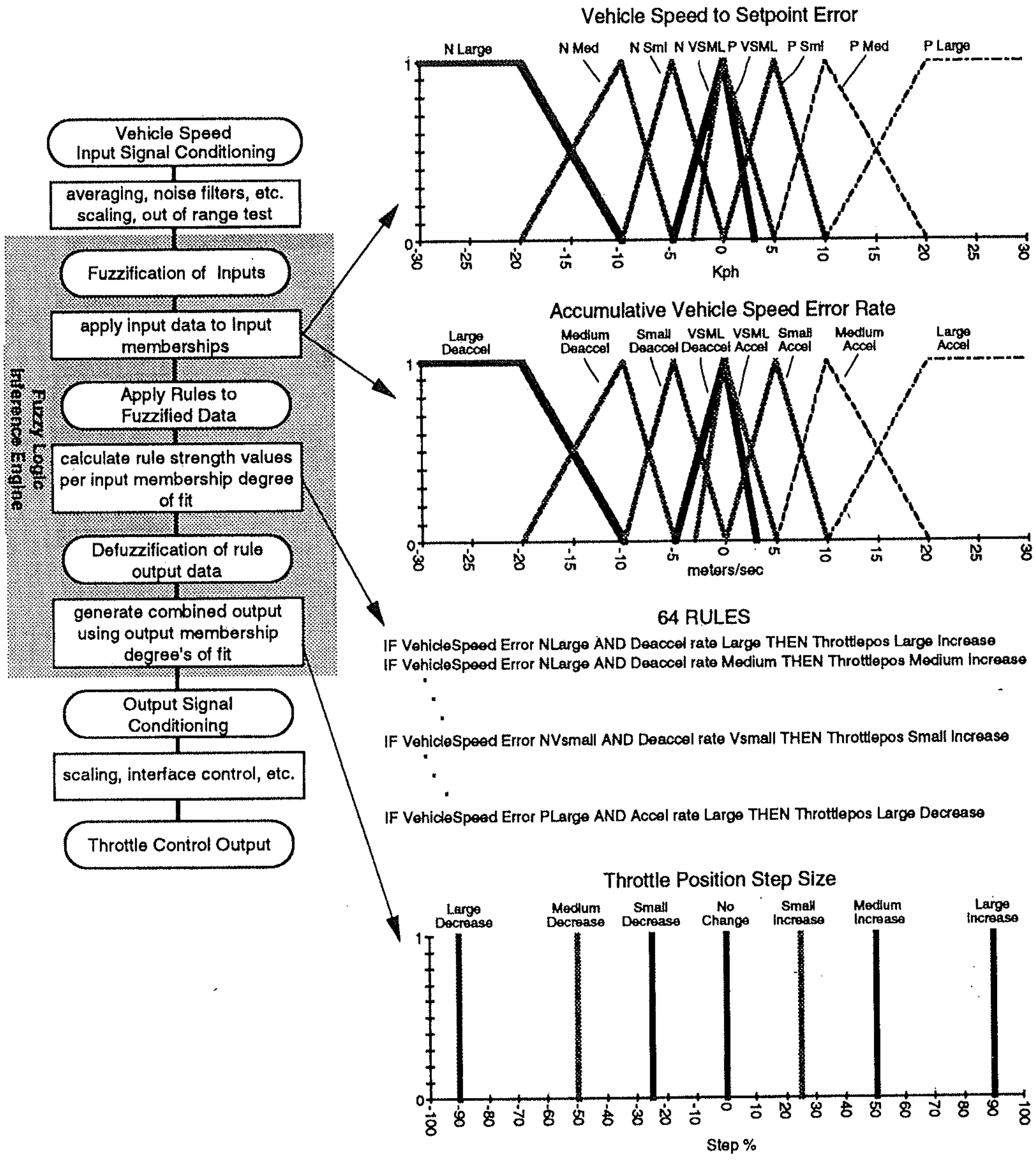
The cruise error calculation algorithm can be designed around traditional math models such as PI or fuzzy logic.

14.3.1 Fuzzy Logic Examples

Fuzzy logic allows somewhat easier implementation of the speed error calculation because its design syntax uses simple linguistics. For example: IF speed difference negative and small, THEN increase throttle slightly.

The output is then adjusted to slightly increase the throttle. The throttle position update rate is determined by another fuzzy program which looks for the driver's cruise performance request (slow, medium, or fast reaction), the application type (small, medium, or large engine size), and other cruise system factory preset parameters. Figure 14.3 shows one part of a fuzzy logic design for computing normal throttle position. Other parts would compute the effects of other inputs, such as resume, driver habits, engine type, and the like.

Other program design requirements include verification that the input signals fall within expected boundaries. For example, a broken or intermittent speed sensor could be detected.


Vehicle Speed Input Signal Conditioning
averaging, noise filters, etc. scaling, out of range test
Fuzzy Logic Inference Engine
Fuzzification of Inputs
apply input data to input memberships
Apply Rules to Fuzzified Data
calculate rule strength values per input membership degree of fit
Defuzzification of rule output data
generate combined output using output membership degree's of fit
Output Signal Conditioning
scaling, interface control, etc.
Throttle Control Output
Vehicle Speed to Setpoint Error
N Large
N Med
N Sml
N VSML
P VSML
P Sml
P Med
P Large
1
0
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
Kph
Accumulative Vehicle Speed Error Rate
Large Deaccel
Medium Deaccel
Small Deaccel
VSML Deaccel
VSML Accel
Small Accel
Medium Accel
Large Accel
meters/sec
64 RULES
Throttle Position Step Size
Large Decrease
Medium Decrease
Small Decrease
No Change
Small Increase
Medium Increase
Large Increase
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
0
10
20
30
40
50
60
70
80
90
100
Step %

IF VehicleSpeed Error NLarge AND Deaccel rate Large THEN Throttlepos Large Increase
IF VehicleSpeed Error NLarge AND Deaccel rate Medium THEN Throttlepos Medium Increase

IF VehicleSpeed Error NVsmall AND Deaccel rate Vsmall THEN Throttlepos Small Increase

IF VehicleSpeed Error PLarge AND Accel rate Large THEN Throttlepos Large Decrease

FIGURE 14.3 Fuzzy speed error program flow.

A heavily loaded vehicle with a small engine may not be able to maintain a high setpoint speed up a steep grade, and the cruise control needs to be disengaged to protect the engine from sustained full-throttle operation under a heavy load. This could be preset to occur 20 percent below the setpoint speed. Another program can test the vehicle speed to resume setpoint speed and prevent unsafe acceleration under certain conditions. For example, if a high-performance vehicle (>200-kW or 268-hp engine) has a setpoint speed of 125 km/h (78 mi/h), and drives from the freeway into heavy city traffic doing 48 km/h (30 mi/h) and the vehicle's driver fortuitously hits the cruise resume switch at this low speed, the cruise control invokes a near full-throttle action, and an accident is likely. A fuzzy design can limit the acceleration upon resume using simple rules such as IF resume and big speed error, THEN increase throttle slightly.

14.3.2 Adaptive Programming

The response time and gain of the cruise system can be adjusted to match individual drivers. For example, some drivers may prefer to allow the vehicle to slow down somewhat when climbing a grade and then respond quickly to maintain a setspeed; other drivers may prefer a constant speed at all times, while still other drivers may prefer a very slow responding cruise system to maximize fuel efficiency. The cruise system can be adapted either by a user selection switch (slow, medium, fast) or by analyzing the driver's acceleration/deacceleration habits during noncruise operation. Once these habits are analyzed, they can be grouped into the three previously mentioned categories. One drawback of a totally automatic adaptive cruise system is when various drivers with vastly different driving preferences operate the vehicle on the same trip. The cruise system would have to be "retrained" for each driver.

14.4 CRUISE CONTROL DESIGN

Many of the required elements of a cruise control can be integrated into one single-chip MCU device. For example, the actuator drivers can be designed in the MCU if their power requirements are on the low side.

14.4.1 Automatic Cruise System

Figure 14.4 shows an experimental system design for a cruise control based upon a semicustom 8 or 16-bit single-chip MCU that incorporates special high-power output driver elements and a built-in voltage regulator.

14.4.2 Safety Backup Examples

The design of a cruise control system should include many safeguards:

- A test to determine vehicle speed conditions or command inputs that do not fall within the normal conditions for operation of the cruise control function.
- A test to determine if the vehicle speed has decreased below what the cruise routine can compensate for.
- Speed setpoint minimums and maximums (30 km/h min to 125 km/h max, for example) are checked and, if exceeded, will cause the cruise function to turn off.
- Speedometer cable failure is detected by checking for speed sensor electrical output pulses over a 100-ms time period and, if these pulses are absent, the system is disengaged.
- Software program traps should also be scattered throughout the program and, if memory permits, at the end of each program loop. These will catch an out-of-control program and initiate a vector restart.

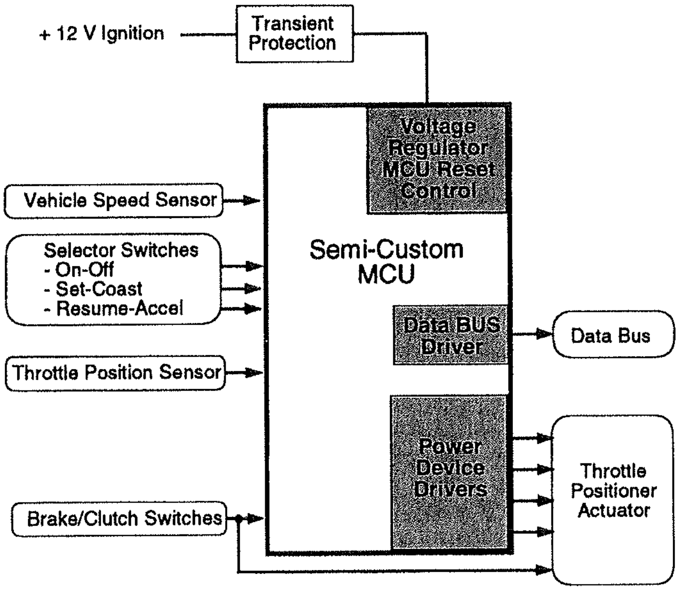


FIGURE 14.4 Automatic cruise control.

14.4.3 EMI and RFI Noise Problems

As with any electronic design, consideration must be given to suppressing RFI (radio frequency interference) from the circuit, besides minimizing effects of external EMI (electromagnetic interference) and RFI to the circuit's normal operation. It is not uncommon that the circuit must operate in RF fields up to 200 V/m intensity. This requires careful layout of the module's PCB (printed circuit board) and RF filters on all lines going in or out of the module. The module case may even have to contain some type of RF shielding. Minimizing generated RFI from the cruise circuit can be accomplished by operating the MCU's crystal oscillator at a minimal power level (this is controlled mostly by the MCU internal design), careful PCB trace layout of the MCU oscillator area, metal shielding over the MCU, ground planes on the PCB under the MCU, and setting the actuator switching edge transition times to over 10 ms. (See Chaps. 27 and 28.)

14.5 *FUTURE CRUISE CONCEPTS*

Several research projects are underway to develop a crash avoidance system that could be interconnected with a cruise system. The development of a low-cost distance sensor that can measure up to a few hundred meters away with a tight focal point in all weather conditions is proving to be a challenge. When a practical vehicular distance sensor is available, the cruise control can be programmed to maintain either constant speed or constant distance to another vehicle. Other methods of cruise control could include receiving a roadside signal that gives an optimum speed value for the vehicle when travelling within certain traffic control areas.

14.5.1 Road Conditions Integration with IVHS

The IVHS (Intelligent Vehicle-Highway System) network may be a more practical approach to setting optimum cruise speed values for groups of vehicles. The IVHS can monitor road conditions, local weather, etc., and broadcast optimal speed data values for vehicles in its zone. (See Chap. 29.)

GLOSSARY

Analog input Sensors usually generate electrical signals that are directly proportional to the mechanism being sensed. The signal is, therefore, analog or can vary from a minimum limit to a maximum limit. Normally, an 8-bit MCU A/D input using a 5-V reference, the analog input resolution is 1 bit, which is 1/256 of 5 V or 0.0193 V.

Defuzzification The process of translating output grades to analog output values.

Fuzzification The process of translating analog input values to input memberships or labels.

Fuzzy logic Software design based upon a reasoning model rather than fixed mathematical algorithms. A fuzzy logic design allows the system engineer to participate in the software design because the fuzzy language is linguistic and built upon easy-to-comprehend fundamentals.

Inference engine The internal software program that produces output values through fuzzy rules for given input values. The inference process involves three steps: fuzzification, rule evaluation, and defuzzification.

Input memberships The input signal or sensor range is divided into degrees of membership, i.e., low, medium, high or cold, cool, comfortable, warm, hot. Each of these membership labels is assigned numerical values or grades.

Output memberships The output signal is divided into grades such as off, slow, medium, fast, and full-on. Numerical values are assigned to each grade. Grades can be either singleton (one value) or Mandani (a range of values per grade).

Rule evaluation Output values are computed per the input memberships and their relationship to the output memberships. The number of rules is usually set by the total number of input memberships and the total number of output memberships. The rules consist of IF inputvarA is *x*, AND inputvarB is *y*, THEN outvar is *z*.

Semicustom MCU An MCU (microcontroller unit) that incorporates normal MCU elements plus user-specified peripheral devices such as higher-power port outputs, special timer units, etc. Mixed semiconductor technologies, such as high-density CMOS (HCMOS) and bipolar analog, are available in a semicustom MCU. Generally, HCMOS is limited to 10 V, whereas bipolar-analog is usable to 60 V.

BIBLIOGRAPHY

Bannatyne, R., "Fuzzy logic—A new approach to embedded control solutions," Motorola Semiconductor Design Concept, DC410, 1992.

Catherwood, M., "Designing for electromagnetic compatibility (EMC) with HCMOS microcontrollers," Motorola Semiconductor Application Note, AN1050, 1989.

Chaudhuri, et al., "Speed control integrated into the powertrain computer," *New Trends in Electronic Management and Driveline Controls,* SAE SP-653, 1986, pp. 65–72.

Hosaka, T., et al., "Vehicle control system and method therefore," U.S. Patent 4809175, May 29, 1990.

Hosaka, T., et al., "Vehicle control system," U.S. Patent 4930084, Feb. 28, 1989.

Mamdani, E. H., "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Transactions on Computers,* C-26-12, 1977, pp. 1182–1191.

Ribbens, W., "Vehicle Motion Control," *Understanding Automotive Electronics,* 4th ed., 1992, pp. 247–257.

Takahashi, Hioshi, "Automatic speed control device using self-tuning fuzzy logic," *IEEE Workshop on Automotive Applications of Electronics,* 88THO321, 1988, pp. 65–71.

Self, Kevin, "Designing with fuzzy logic," *IEEE Spectrum,* Nov. 1990, pp. 42–44, 105.

Sibigtroth, J., "Implementing fuzzy expert rules in hardware," *AI Expert,* April 1992.

Stefanides, E. J., "Cruise control components packaged as one unit," *Design News,* Oct. 1, 1990, pp. 162–163.

Zadeh, L. A., "Fuzzy sets, information and control," vol. 8, 1965, pp. 338–353.

ABOUT THE AUTHOR

Richard J. Valentine is a principal staff engineer at Motorola SPS in Phoenix, Ariz. His present assignments include engineering evaluation of advanced semiconductor products for emerging automotive systems. He holds two patents and has published 29 technical articles during his 24 years at Motorola.

CHAPTER 22
ON- AND OFF-BOARD DIAGNOSTICS

Wolfgang Bremer, Frieder Heintz, and Robert Hugel
Robert Bosch GmbH

22.1 WHY DIAGNOSTICS?

The desire for greater safety, driving comfort, and environmental compatibility is leading to a rapid increase in electronic control units and sensors in upper class, medium-sized, and compact vehicles. Additional functions and their corresponding equipment in today's cars create a bewildering tangle of cables and confusing functional connections. As a result, it has become more and more difficult to diagnose faults in such systems and to resolve them within a reasonable period.

22.1.1 Diagnostics in the Past and Today

On-board diagnosis has been limited thus far to a few error displays and fault storage achieved by relatively simple means. It has been left more or less to each manufacturer to decide to what extent diagnosis would be carried out. Diagnosis always means the working together of man and machine and consists essentially of three major components: registration of the actual condition, knowledge of the vehicle and its nominal condition, and strategy—how to find the smallest exchangeable deficient component by means of combining and comparing both the nominal and actual conditions.

All three points are inseparably connected. Only the means to the end have changed over time. The oldest and simplest method of diagnosis is that done with the help of our sense organs, but the limits of this kind of diagnosis are obvious. In fact, the objective in the development of diagnostic techniques is the extension of human abilities with the aid of diagnostic tools in order to be able to measure more precisely and more directly, to compare more objectively, and to draw definite conclusions.

The development of control techniques was essentially determined by the following items: the development of automotive engineering; the structure of workshops—that is, essentially the relation between the costs of labor and materials; and the development of electronics and data processing.

For a long time, motor diagnosis was limited to ignition control and timing. In the 1960s, new exhaust-gas measuring instruments for fuel injection adjustment were developed, but the mechanic still had to make the diagnosis. In the 1980s, the introduction of electronics in the vehicle was followed by a new generation of measuring instruments in the workshops. Not

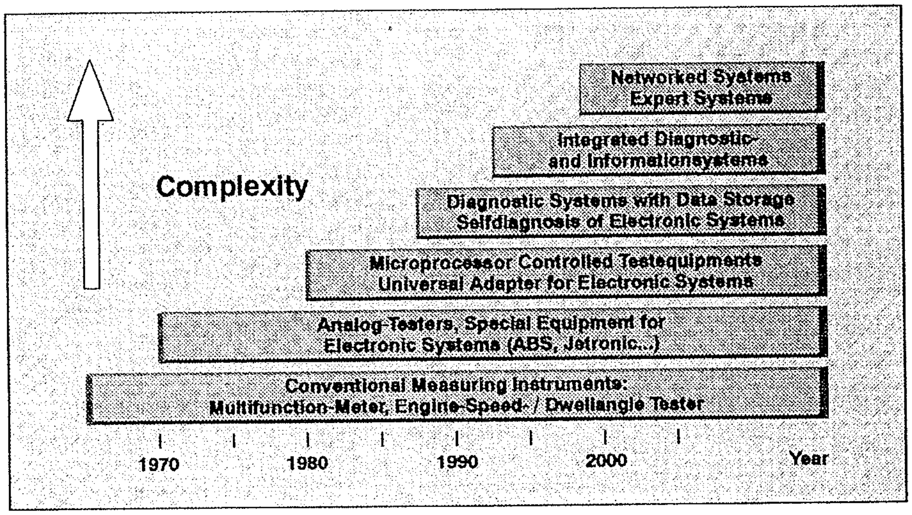


FIGURE 22.1 Evolution of diagnostic test equipment.

only were separate measurements combined with comprehensive test procedures, but also the information about the nominal condition of the vehicle was stored in a data memory.[1] A view of the development is shown in Fig. 22.1.

As more and more electronic systems were added to cars, the more difficult it became to determine the actual condition in case of a defect. Soon a multitude of connecting cables and adapters were required to reach the necessary measuring points. Moreover there was an increasing amount of information needed to make an effective diagnosis. In the majority of workshops, diagnosis is carried out as shown in Fig. 22.2. The most important test points of

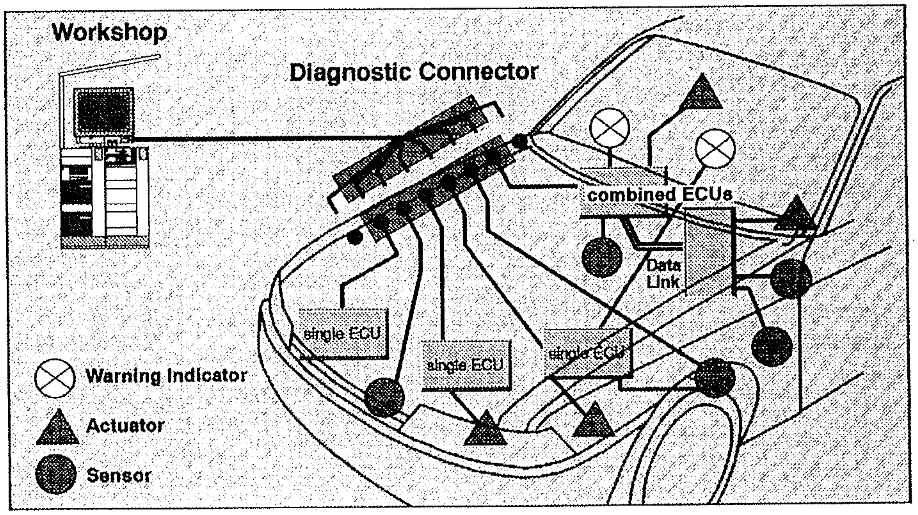


FIGURE 22.2 Present-day diagnostic connector installation in a vehicle.

control units and sensors are tied to a diagnostic connector which is plugged into the measuring instrument with a corresponding adapter for the respective vehicle. Because of the permanently increasing amount of electronic functions, it is necessary to develop connectors with more and more contacts. It is evident that this method soon will become too unwieldy.

Modern electronics in vehicles support diagnosis by comparing the registered actual values with the internally stored nominal values with the help of control units and their self-diagnosis, thus detecting faults. By interconnecting the measuring instruments, a detailed survey of the entire condition of the vehicle is available and an intelligent on-board diagnostic system is able to carry out a more precise and more definite localization of the defect.[2] With the help of an interconnection and standardization of the interface leading to the external tester, the many different complex and expensive adapters have become superfluous. Modern diagnosis will look like what is shown in Fig. 22.3.

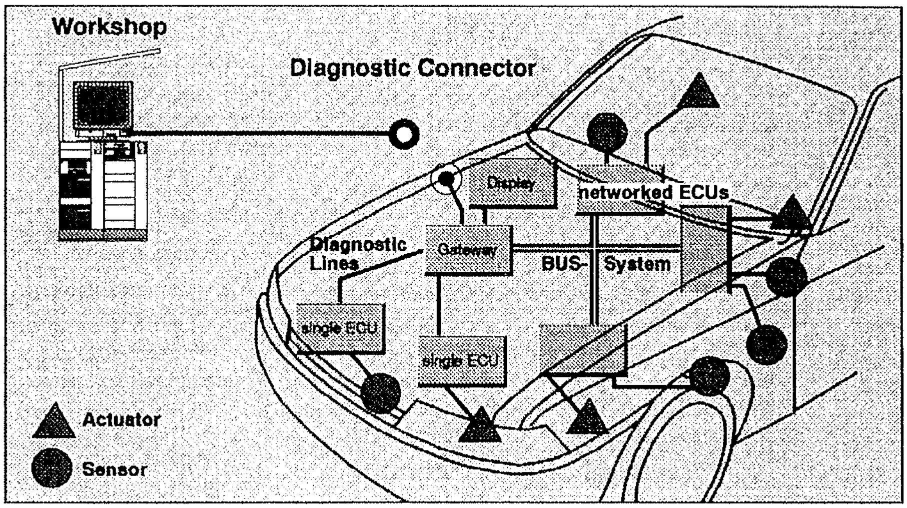


FIGURE 22.3 Future diagnostic connector installation in a vehicle.

Instead of a multiplicity of adapters there is only a single standardized interface, provided by the diagnostic processor. By means of interconnection, the diagnostic processor is provided with all available data and the condition of the vehicle is known. With the help of the diagnostic processor, the external measuring instrument has access to the measuring and diagnostic values of the sensors and is able to directly reach the actuator for measuring purposes.[3]

Such a diagnosis also demands a certain change in the functional structure of a vehicle. Corresponding hierarchical models have already been presented.[4]

22.1.2 Reasons for Diagnostics in Vehicles

Which are the most important reasons for diagnostics as demanded and desired in today's vehicles?

Existing Diagnostic Problems. A number of diagnostic problems must be resolved:

- Early diagnostic information was related only to single components and control units. In case of a defective comprehensive system, every unit, component, sensor, and connecting cable of the system had to be tested and controlled. This was a very time consuming and expensive process.

- Because of the single component and control unit checks, it was impossible to analyze all the additional data correlated with a particular defect.
- In the case of a defect in single sensors or units, the car was often inoperable. Taking into consideration all available information about the vehicle, it is possible to use alternative parameters and procedures in order to achieve at least a so-called limp-home function and sometimes continue the use of the vehicle under only slightly limited operating conditions.
- Usually there was only a global error display with an often ambiguous warning light available for the driver. Drivers desire more detailed information and especially guidelines for what procedures should be followed.
- The multitude of adapter cables, plugs, diagnostic equipment, and communication interfaces in a workshop has become so complex that the effectiveness decreased dramatically, with the repair costs increasing disproportionally.

New Legal Proposals. Worldwide new legal proposals and governmental regulations [e.g., California Air Resources Board (CARB), On Board Diagnostics II (OBDII), Environmental Protection Agency (EPA)] are forcing manufacturers and subcontractors to seek more profitable, effective, and convincing diagnosis of vehicles.

Serial Data Networks. New serial data networks for the connection of control units and vehicle body components, installed in the vehicle, offer the possibility of absolutely new optimum approaches and even anticipate maintenance and diagnosis up to the introduction of autodidactic data processing systems and external data bases.[5,6,7,8]

International Initiatives for Standardization. Initiated by legislative and governmental demands for better diagnostics in the area of emission control, initiatives for standardization in the entire diagnostic field in vehicles were launched during recent years to achieve worldwide standardization of tools, interfaces, connectors, and protocols.

22.1.3 Diagnostic Tasks in Vehicles

In order to minimize the number of defects or even to completely avoid them, a vehicle requires regular checks. In case of an inevitable defect, a clear and directed diagnosis is required and has to be followed by a prompt, reliable, and inexpensive repair. Therefore appropriate diagnostic systems are being developed considering the following targets: simplification of maintenance, fault indication in time, guidelines for the driver in case of a defect, and safer and faster repairs with the help of a specific fault indication.

In addition to technical considerations, environmental aspects are now being taken into consideration as reflected in the diagnostic concepts. In the future, only perfect systems will be accepted, in order to keep environmental pollution to a minimum. It is understandable, therefore, that legislators insist on increased monitoring standards, particularly for exhaust-related components.

As an example of the new monitoring standards, consider the requirements of CARB and EPA in the United States and the resulting consequences for diagnosis. At the moment, the extent of such a detailed monitoring has to be a compromise between the different requirements and the possible technical and economical solutions, but the environmental aspects will gain more and more importance. The increased amount of available data will certainly permit a considerably higher rate of in-depth fault localization and will also allow clear fault identification without interactive outside intervention. Having knowledge of the functional interrelationships and access to all essential data, a picture of the defect can be created with the help of individual pieces of information. The driver and the workshop can then be provided with appropriate instructions. In this context, on-board expert systems are being considered.

For an effective and successful diagnosis today and in the future the following tasks and targets can be defined.

Fault Storage with Boundary Conditions. A very important aspect of modern diagnosis is the clear and reliable analysis of the respective fault. During the self-diagnosis, it is absolutely necessary to store not only the respective fault information but also all relevant marginal parameters in the control unit, e.g., ambient temperature, velocity, engine speed, engine knock, and so on. The additional data can be stored when a defect occurs as well as during specified intervals around the moment of a defect. Such additional data is called "freeze frame" data.[9]

Fault Localization. Mechanics must be able to locate a defective control unit quickly and then determine which component of that control unit is at fault so that it can be replaced.

Data Correlation, Recognition of Imminent Faults. A large amount of data useful for the analysis of a vehicle is now available and even more will be available in the future. These data will have to be evaluated and compared with the help of modern data processing techniques, including fuzzy logic, neural networks, autodidactic systems, and expert systems. These techniques will not only enable the diagnosis of the actual condition of the vehicle but will also determine future maintenance needs. As a result, the reliability and availability of a vehicle will be increased and the possible consequences of a defect kept to a minimum. The driver can also be forewarned about imminent problems and can then take appropriate steps before starting on a trip.

Parameter Substitution. The breakdown of a sensor in modern diagnostic procedures is not necessarily followed by a lack of the respective information. After having diagnosed a fault, the diagnostic computer—with the aid of the available information—is often able to compute an auxiliary parameter to replace the original one. As a result, either a limp-home condition is possible or else the nominal function can be assured but under slightly limited conditions. Simple examples for such a calculated parameter are vehicle speed (considering the gear and the synchronous speed, or the antilock braking information, or the data of the navigation system), motor temperature (considering the outside temperature and the operating time), and the amount of remaining fuel (considering the last actual fuel content and the calculated consumption).

Providing Guidelines. As mentioned earlier, a diagnostic system has to provide clear information to the driver in case of a defect. A global warning indication is not sufficient. The driver needs to learn the extent of the defect and its consequences by appropriate text, graphics, or synthetic voice. In addition, the driver needs to be told the steps that have to be taken (e.g., "refill cooling water," "minimum speed to the next service station, risk of engine breakdown," "stop, brake system out of order").[10]

The diagnostic monitoring system can also be used, if there is no service station nearby, as a substitutional off-board system. The defect is then localized by an interactive working together of the indicating system and an appropriate input medium.

External Diagnostic Access. For off-board diagnosis, the diagnostic system of the vehicle has to provide a standardized access to all relevant components, control units, and stored information. This standardized access might also be used by the vehicle manufacturer, legislator, application engineer, and the end-of-the-line programmer. The access itself has to be controlled with the help of an appropriate mechanism to prevent possible abuse.[11]

Logbook Function. The control unit or the diagnostic computer of the vehicle is supposed to store every repair that has been carried out in the format of a logbook. It should contain the time and name of the workshop, every exchanged and newly installed element, every inspection carried out, and so forth.

22.2 ON-BOARD DIAGNOSTICS

The more complex automobiles became, the greater the number of electronic systems and the more difficult became the registration of the actual condition in case of a defect. To reach the necessary measuring points, many connecting cables and adapters were required. In addition, much data about the different systems and their working together was needed to allow a system-specific diagnosis. Modern electronics with self-diagnosis supports the service mechanic by registrating the actual values, comparing them with the nominal values, and diagnosing faults that are stored for repair purposes. Actually, the internal functions are checked whenever an ECU is turned on.

First, the checksum of the program memory is checked together with its function and the correct version. Then a read and write test of the RAM cells is performed. Special peripheral elements (e.g., AD converters) are also checked within this test cycle. During the entire operating time of the vehicle, the ECUs are constantly supervising the sensors they are connected to. With the help of an adequate interpretation of the hardware, controllers are able to determine whether a sensor has a short circuit to ground or battery voltage, or if a cable to the sensor is interrupted. By comparing the measured values and the stored technical data, a controller is able to determine whether the measured values exceed the limits, drift away, or are still within the tolerable limits. The combination of information provided by other sensors allows the monitoring for plausibleness of the measured values.

Sensors are tested similarly to the way actuators are monitored for short circuits or interruptions of cables. The check is carried out by measuring the electric current or reading the diagnostic output of intelligent driver circuits. The function of an actuator under certain conditions can be tested by powering the actuator and observing the corresponding reaction of the system. If discrepancies to the nominal values are diagnosed, the information is stored in an internal fault memory together with relevant outside parameters, e.g., the motor temperature or the engine speed. Thus, defects that appear once or under certain conditions can be diagnosed. If a fault occurs only once during several journeys, it is deleted. The fault memory can be read later in the workshop and provides valuable information for the mechanic.

In case of a detected defective sensor, the measured values are replaced by nominal values or an alternative value is formed using the information of other sensors to provide at least a limp-home function.

With the help of an appropriate interface, a tester can communicate with the ECUs, read the fault memory and the measured values, and send signals to the actuators. In order to be able to use self-diagnosis as universally as possible, manufacturers aim at the standardization of the interface and the determination of appropriate protocols for data exchange.

Another task of self-diagnosis is the indication of a defect to the driver. Faults are mostly indicated by one or more warning lights on the dashboard. Modern developments aim at more comprehensive information using displays for text and graphics, which provide priority-controlled information for the driver. Legal regulations concerning exhaust-gas gave rise to an essential extension of self diagnosis. The control units have to be able to control all exhaust-relevant functions and components and to clearly indicate a defective function or the exceeding of the permissible exhaust limits. Some of the demanded functions require an enormous amount of additional instructions; therefore, the extent of self-diagnosis already reaches up to 40 percent of the entire software of the control unit.

22.3 OFF-BOARD DIAGNOSTICS

The continual increase in the use of electronics within the broad range of different vehicles represents one of the major challenges for customer service and workshop operations. Modern diagnosis and information systems must cope with this challenge and manufacturers of test equipments must provide instruments that are flexible and easy to handle. Quick and reliable fault diagnosis in modern vehicles requires extensive technical knowledge, detailed vehicle information, and up-to-date testing systems.

Due to the different demands of the service providers, there are many different test equipments on the market. They can be subdivided into two main categories: handheld or portable instruments and stationary equipments. Handheld instruments are commonly used for the control of engine functions like ignition or fuel injection and the request of error codes of the electronic control units (ECUs). Stationary test equipment, on the other hand, covers the whole range of function and performance checks of the engine, gear, brakes, chassis, and exhaust monitoring.

Most of the common testers are used for the diagnosis of the engine. The Bosch MOT 250, for example, offers the following functions:

- Engine speed by means of the top dead center (TDC) transmitter, cylinder 1 or terminal 1 signal
- Ignition timing with TDC sensor or stroboscope
- Dwell angle in percent, degrees, or dwell time
- On/off-ratio in percent
- Injection timing or other times measured at the valve or other suitable measuring points
- Electric cylinder balance in absolute or relative terms
- Voltage to ground or floating potential including lambda-sensor voltages or dynamic voltage at terminal 1
- Current with two test adapters for maximum 20 A and 600 A
- Resistances from milliohms to megohms
- Temperature with oil-temperature sensor

For most variables, a maximum of four blocks of measured variables can be stored and recalled one after the other. Twelve blocks can be stored for the cylinder balance function. A digital storage oscilloscope records and stores up to 32 oscillograms of ignition voltages, alternator ripple, and current or voltage transients in the electric or electronic systems. Two RS232 interfaces are provided for documentation purposes and data exchange.

For repair, service, and maintenance, many different manuals and microfiches are stored in the workshops. It is a time-consuming task to collect all the necessary information, especially when vehicles of different makes have to be repaired. To avoid unnecessary paper, information and communication systems among workshop, dealer, and manufacturer are built up. The corresponding manuals have to be standardized and distributed on electronic data processing media, preferably on CD-ROMs.

Every garage or workshop, equipped with the appropriate data system (basically a tester connected to a PC), will receive servicing aids and updates via telephone line or by periodic receipt of updated CDs. A committee of the SAE is preparing rules for the standardization of manuals. There are already published draft international standards (DIS) for terms and definitions (J1930) used in the manuals, for diagnostic codes/messages (J2012), or electronic access/service information (J2008) (see the following). Most of the available test equipment is capable of storing operator manuals within its memory and offers menu-guided assistance to the service personnel. Automatic vehicle and component identification by the tester and the availability of corresponding data at the workbench eases troubleshooting and repairs.

22.4 LEGISLATION AND STANDARDIZATION

22.4.1 CARB, EPA, OBD II

The following is an abstract of the California Air Resource Board (CARB) Regulations for On-Board-Diagnosis two(OBDII):

> All 1994 and subsequent model-year passenger cars, light-duty trucks, and medium-duty vehicles shall be equipped with a malfunction indicator light (MIL) located on the instrument panel that will automatically inform the vehicle operator in the event of a malfunction of any power train component which can affect emission and which provide input to, or receive output from, the on-board computer(s) or of the malfunction of the on-board computer(s) itself. The MIL shall not be used for any other purpose.
>
> . . . .
>
> All 1994 and subsequent model-year passenger cars, light-duty trucks, and medium-duty vehicles required to have MIL pursuant to paragraph above shall also be equipped with an on-board diagnostic system capable of identifying the likely area of the malfunction by means of fault codes stored in the computer memory. These vehicles shall be equipped with a standardized electrical connector to provide access to the stored fault codes . . . Starting with model-year 1995, manufacturers of non-complying systems shall be subject to fines pursuant to section 43016 of the California Health and Safety Code for each deficiency identified, after the second, in a vehicle model. For the third deficiency and every deficiency thereafter identified in a vehicle model, the fines shall be in the amount of $50 per deficiency per vehicle for non-compliance with any of the monitoring requirements . . .

Systems to Be Monitored

OBD II Functions. These include catalyst monitoring, misfire monitoring, evaporative system monitoring, secondary air system monitoring, fuel systems monitoring, oxygen sensor monitoring, exhaust-gas-recirculation (EGR) system monitoring, and comprehensive component monitoring.

Catalyst. Legal requirements (CARB excerpt): "The diagnostic system shall individually monitor the front catalyst or catalysts which receive untreated engine out exhaust-gas for malfunction. A catalyst is regarded as malfunctioning when the average hydrocarbon conversion efficiency falls between 50 and 60 percent."

Technical solution: In addition to the oxygen sensor upstream the catalyst, another sensor is mounted downstream.

A properly working catalyst shows a storage effect so that the oscillation of the lambda-controller appears damped at the downstream lambda probe. A worn-out catalyst has a reduced damping effect and the signals of up- and downstream sensors are equivalent.

The ratio of the signal amplitudes is a measure of the conversion efficiency. The electronic system that controls the fuel injection monitors these signals together with other relevant engine conditions to derive the catalyst efficiency.

Misfire Detection. Legal requirements (CARB excerpt): "To avoid catalyst damage, the diagnostic system shall monitor engine misfire and identify the specific cylinder experiencing misfire."

Technical solution: Misfire can be caused by worn-out spark plugs or defective electrical wiring. Unburned fuel reaches the catalyst and may destroy it by overheating. Even the least amount of misfire rates influences the emission and therefore single misfire events must be detected.

The speed of the engine is measured very precisely. In case of misfire, the momentum, which is normally produced by the combustion, is lacking. Thus abnormal variations of speed-changes at steady state conditions may be considered as misfire. To distinguish clearly between misfire and other malfunctions, complicated calculations have to be carried out.

If a certain percentage of misfires within 200 or 1000 revolutions is detected, a fault code is stored in the control unit and the fault is indicated to the driver.

Oxygen Sensor. Legal requirements (CARB excerpt): "The diagnostic system shall monitor the output voltage, the response rate, and any other parameter which can affect emission and all fuel control oxygen sensors for malfunction."

Technical solution: The control unit has a special input circuit for detecting shorts or breaks and monitors the switching frequency of the control loop.

By means of a second lambda probe behind the catalyst, it is possible to monitor the lambda probe in front of the catalyst for its correct position. A lambda probe which is subject to an increased temperature for extensive periods may react slower on variations of the air/fuel mixture, thus increasing the period of the lambda-probe regulation. The diagnostic system of the control unit controls the regular frequency and indicates slow sensors to the driver by means of a warning light.

Heated sensors are monitored for correct heater current and voltage by hardware means within the control unit.

Evaporative System. Legal requirements (CARB excerpt): "The diagnostic system shall control the air flow of the complete evaporative system. In addition, the diagnostic system shall also monitor the complete evaporative system for the emission of HC vapor into the atmosphere by performing a pressure or vacuum check of the complete evaporative system. From time to time, manufacturers may occasionally turn off the evaporative purge system in order to carry out a check."

Technical solution: At idle position, the canister purge valve is activated and the lambda controller is monitored for its reaction. For leak detection of the evaporative system, the output to the active carbon filter is shut off and the canister pressure is decreased to about −1.5 kPa. Then the complete system is turned off and the pressure within the canister is monitored for variation with time. The pressure gradient, together with other parameters like the amount of fuel, may indicate possible leaks.

Secondary Air System. Legal requirements: "Any vehicle equipped with any form of a secondary air delivery system shall have the diagnostic system monitor the proper functioning of the secondary air delivery system and any air switching valve."

Technical solution: The lambda controller is monitored for correlated deviations when the secondary air flow is changed.

Fuel System. Legal requirements: "The diagnostic system shall monitor the fuel delivery system for its ability to provide compliance with emission standards."

Deviations of the stochiometric ratio which last for a longer time are stored within the adaptive mixture controller. If these values exceed defined limits, components of the fuel system obviously do not correspond to the specification.

Exhaust-Gas Recirculation (EGR) System. Legal requirement: "The diagnostic system shall monitor the EGR system on vehicles for low and high flow rate malfunctions."

Technical solution: (1) At overrun, the fuel is cut off and the EGR valve is completely opened. The flow of exhaust gas to the manifold raises the manifold pressure, which is recorded and allows statements about the function of the EGR valve. (2) Another possibility is to control the increase of the manifold intake temperature when the EGR valve is opened.

In a conclusion to the previously described OBD II requirements and technical solutions, we can define the following four quality demands for electronic control units:

- Guarantee for exhaust-gas-relevant components with repair costs >$300 for seven years or 70,000 miles for all 1990 and subsequent model-year vehicles (CARB).
- Guarantee for exhaust-gas-relevant components with repair costs >$200 for eight years or 80,000 miles for all 1994 and subsequent model-year vehicles (EPA/Clean Air Act).
- Guarantee protocols in case of a reclamation rate of exhaust-gas-relevant components higher than 1 percent (CARB).
- Recall of vehicles in case of a calculated reclamation rate of more than 20,000 ppm within a period of five years/50,000 miles (CARB).

22.4.2 International Standardizations

Because of the manifold requirements on modern diagnostics, the national and international standardization committees soon came to the conclusion that with the help of appropriate and, if possible, international agreements about protocols, connectors, tools and auxiliaries, the process of diagnosis can be standardized, thus reducing time and costs.

Figure 22.4 shows how, in a standardized graphic, control units and diagnostic tools are connected and diagnostic data exchanged.

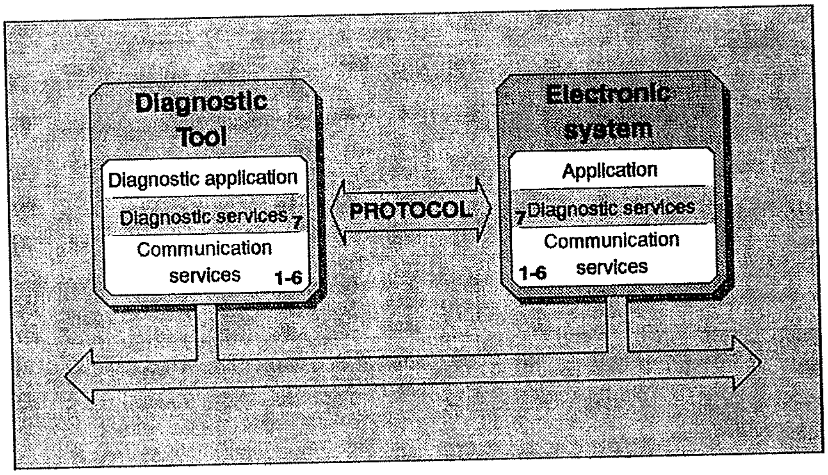


FIGURE 22.4 Standardized testing link according to the OSI model.

For data exchange, electronic systems are structured and described according to a seven-layer model (OSI model, open system interconnection) developed by the ISO (International Standardization Organization). Every unit connected to a data network can be structured with the help of this model—control units as well as diagnostic tools.

The diagnostic services that the controller may use during the diagnostic process are regulated in the seventh layer. Diagnostic service means definite instructions, which actuate determined and standardized diagnostic procedures, e.g. "start diagnostic session," "read diagnostic trouble codes," "read freeze frame data," and so on. There are different sequences of bits and bytes code for such instructions. On the hardware level (plugs, cables, potentials), the sequences are finally transmitted from unit to unit. The ISO and the SAE (Society of Automotive Engineers) developed corresponding standards in the area of service definition

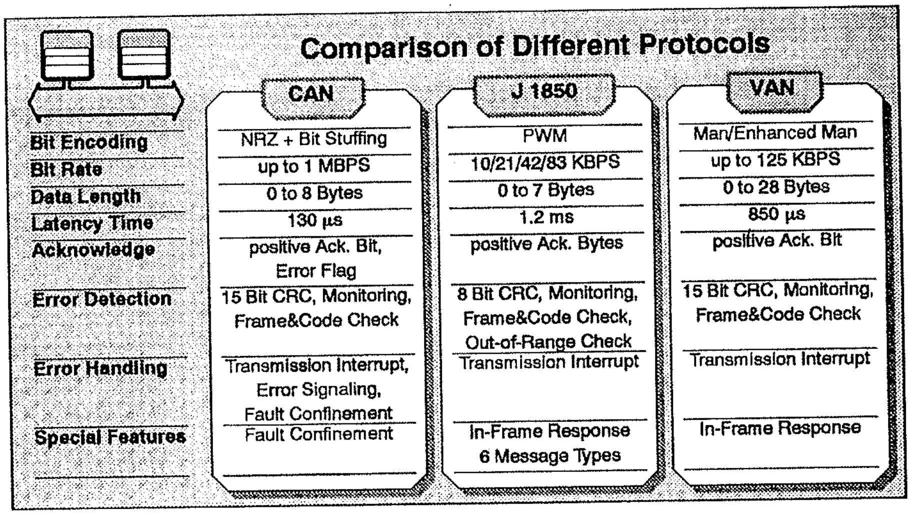
TABLE 22.1 ISO Diagnostic Services

Diagnostic management
StartDiagnosticSession
StopDiagnosticSession
SecurityAccess
TesterPresent
EcuReset
ReadEcuIdentification
DisableNormalMessageTransmission
EnableNormalMessageTransmission
Data transmission
ReadDataByLocalIdentifier
ReadDataByGlobalIdentifier
ReadMemoryByAddress
WriteDataByLocalIdentifier
WriteDataByGlobalIdentifier
WriteMemoryByAddress
SetDataRates
StopRepeatedDataTransmission
Input/output control
InputOutputControlByGlobalIdentifier
InputOutputControlByLocalIdentifier
Stored data transmission
ReadNumberOfDiagnosticTroubleCodes
ReadDiagnosticTroubleCode
ReadDiagnosticTroubleCodesByStatus
ReadStatusOfDiagnosticTroubleCodes
ReadFreezeFrameData
ClearDiagnosticInformation
Remote activation of routine
StartRoutineByLocalIdentifier
StartRoutineByAddress
StopRoutineByLocalIdentifier
StopRoutineByAddress
RequestRoutineResultsByLocalIdentifier
RequestRoutineResultsByAddress
Upload download
RequestDownload
RequestUpload
TransferData
RequestTransferExit

as well as in the area of communication. Table 22.1 shows the diagnostic services as proposed by the ISO.

Figure 22.5 presents the determined standards with some essential technical details as developed for the field of communication.

Unfortunately the whole spectrum of available standards has become very complex and difficult to use. The following explanations try to provide a unified system for the existing standards in the area of diagnosis.



Comparison of Different Protocols

	CAN	J 1850	VAN
Bit Encoding	NRZ + Bit Stuffing	PWM	Man/Enhanced Man
Bit Rate	up to 1 MBPS	10/21/42/83 KBPS	up to 125 KBPS
Data Length	0 to 8 Bytes	0 to 7 Bytes	0 to 28 Bytes
Latency Time	130 µs	1.2 ms	850 µs
Acknowledge	positive Ack. Bit, Error Flag	positive Ack. Bytes	positive Ack. Bit
Error Detection	15 Bit CRC, Monitoring, Frame&Code Check	8 Bit CRC, Monitoring, Frame&Code Check, Out-of-Range Check	15 Bit CRC, Monitoring, Frame&Code Check
Error Handling	Transmission Interrupt, Error Signaling, Fault Confinement	Transmission Interrupt	Transmission Interrupt
Special Features	Fault Confinement	In-Frame Response 6 Message Types	In-Frame Response

FIGURE 22.5 In-vehicle networks.

Figure 22.6 shows a general model for diagnostic concepts. The three main levels comprehensively describe the whole area of diagnostics. The three levels are hierarchically structured, closely linked together with flowing transition from one level to the other. Although there are certain similarities between this model and the seven-layer model of the OSI, both models do not correlate.

The upper level comprises the elements, which are essential for the user or generator of diagnostic applications. The term "user" includes the driver, the legislator, the mechanic, and the manufacturer. This upper level can be subdivided into three main fields of activities: user

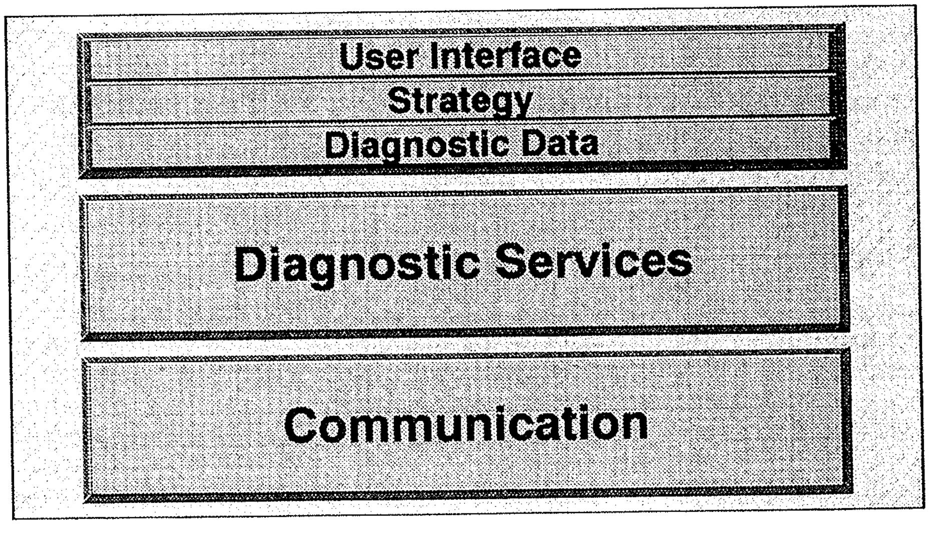


FIGURE 22.6 Model for diagnostic concept.

interface, strategy, and diagnostic data. Although presented as layers, these activities do not correlate hierarchically, but each is associated with a service or group of services.

The "user interface" describes how information flows between the user and the diagnostic service. This includes a functional description of scan tools, handheld testers, monitoring systems, and so on.

The term "strategies" stands for strategic details, which are essential for the diagnosis or repair of a vehicle, including communication access, diagnostic data and information.

The term "diagnostic data" includes the data that are necessary for the diagnosis itself. The details concerning parameters, trouble codes, and so on are described here.

The intermediate level describes the diagnostic services, defining a set of services and a set of commands for general purpose, which allow the diagnosis of a vehicle. The set of commands is supposed to cover the needs of users concerning repair and maintenance as described by the strategies and diagnostic data.

The lowest level deals with the communication area. It describes every technical detail that is necessary for communication and provides the information about how to start communication (initialization). It also specifies the appropriate Baud rate, the suitable protocol, and the necessary hardware (connector, cable, and so on).

This model offers a general description of the essential fields of diagnostic interest and allows the categorization of all ISO and SAE standardization activities in the three main levels of the diagnostic concept model.

Figures 22.7 and 22.8 are presented in the same graphic form (three-level structure). They provide a summary of the concrete standardization activities of the SAE and ISO. Figure 22.7 shows the existing standards or drafts of automotive diagnosis for general purpose.

The user interface for general purposes is undefined. The SAE J2186 (Data Link Security) and the SAE J2008 (Electronic Access/Service Information) are strategic documents, though most strategies are not standardized and diagnostic data is described in documents SAE J2012 (Diagnostic Codes and Messages) and SAE J2190-2 (Parameters—in preparation).

On the level of diagnostic services, the standardization activities can be divided in two fields called service definition and service implementation. The term "service definition" describes a set of useful diagnostic services, which enable the user to run a diagnostic session

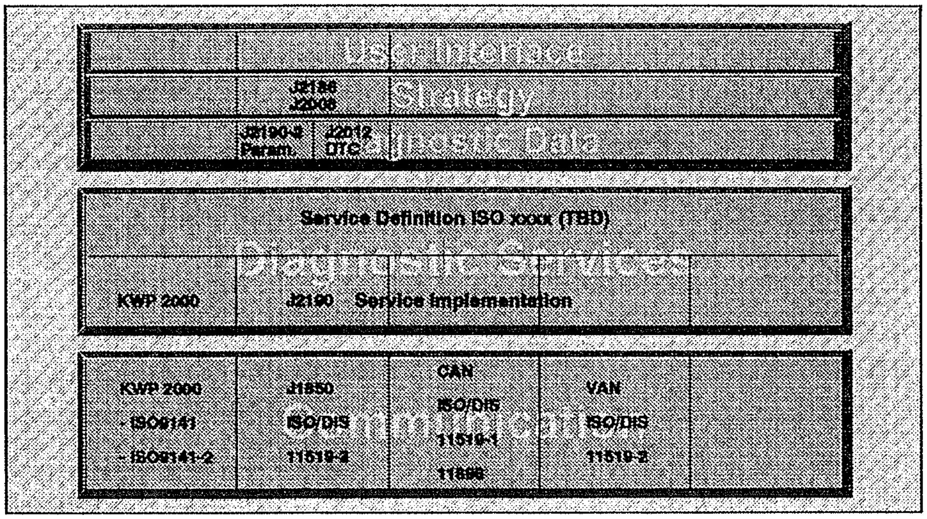


FIGURE 22.7 Realization for general automotive diagnosis.

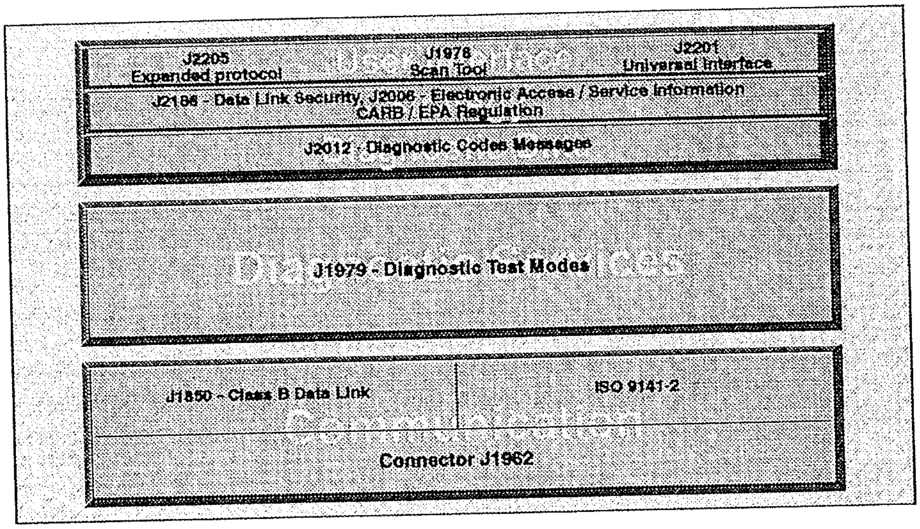

User Interface
J2205
Expanded protocol
J1978
Scan Tool
J2201
Universal Interface
J2186 - Data Link Security, J2008 - Electronic Access / Service Information
CARB / EPA Regulation
J2012 - Diagnostic Codes Messages
Diagnostic Services
J1979 - Diagnostic Test Modes
Communication
J1850 - Class B Data Link
ISO 9141-2
Connector J1962

FIGURE 22.8 Realization for CARB and EPA requirements.

independently of the knowledge of any technical detail in the communication area as described in the level below.

This set of diagnostic services for general purpose can now be mapped on different protocols. Any bit representation of the different services can be built up. This is called service implementation. At the moment, there are two implementations available, the SAE J2190 (Diagnostic Test Modes) and the KWP 2000 (ISO Draft: Keyword Protocol 2000). The lowest level (the Communication level) shows the standardized details of communication such as the data formats and the physical layers; e.g., the KWP 2000 uses the physical layer of ISO 9141 or ISO 9141-2, the SAE J2190 uses the SAE J1850 Class B network (ISO/DIS 11519-3). It is shown that communication can also be built up with a CAN or a VAN network.

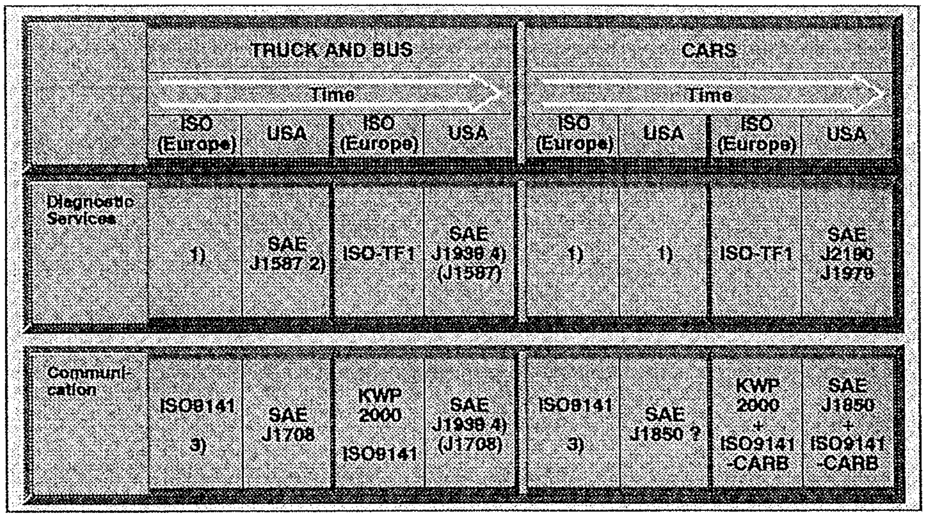
Figure 22.8 shows the standardization activities for the special requirements of the CARB and the EPA using the same three-level-concept.

The user interface, a generic scan tool, is standardized within the SAE J1978, including the SAE J2205 (Expanded Diagnostic Protocol) and the SAE J2201 (Universal Interface). Some aspects of the diagnostic strategy are described in the SAE J2186 (Data Link Security), the SAE J2008 (Electronic Access/Service Information), and some in the regulations. The diagnostic data is described in the document SAE J2012—Diagnostic Codes and Messages.

The level of diagnostic services defines one SAE J1979 standard—Diagnostic Test Modes. This standard is a closely linked combination of a service definition and a service implementation (referring to the SAE as "modes").

In the field of communication, the possible networks are described in the SAE J1850 (Class B Data Network) and the ISO 9141-2 (CARB Requirements for Interchange of Digital Information).

A standard for the physical connector (SAE J1962) has also been developed. Figure 22.9 shows the status of diagnostic standards for trucks and buses and for passenger cars in Europe and in the United States. It shows also a time schedule for the development of standards. A comparison of the communication and diagnostic services levels has already been realized. The titles of the different SAE and ISO numbers are shown in Tables 22.2 and 22.3, where all ISO and SAE papers, relevant for diagnostics, are listed. Table 22.3 offers a detailed list of trucks and bus activities (J1939).



	TRUCK AND BUS				CARS			
	Time				Time			
	ISO (Europe)	USA	ISO (Europe)	USA	ISO (Europe)	USA	ISO (Europe)	USA
Diagnostic Services	1)	SAE J1587 2)	ISO-TF1	SAE J1939 4) (J1587)	1)	1)	ISO-TF1	SAE J2190 J1979
Communi-cation	ISO9141 3)	SAE J1708	KWP 2000 ISO9141	SAE J1939 4) (J1708)	ISO9141 3)	SAE J1850 ?	KWP 2000 + ISO9141 -CARB	SAE J1850 + ISO9141 -CARB

FIGURE 22.9 Status of diagnostic standards.

22.5 FUTURE DIAGNOSTIC CONCEPTS

As yet, most vehicle manufacturers have installed a diagnostic connector in the engine compartment in order to offer essential electric signals for diagnostic purposes. Due to the multitude of different equipments and philosophies of car makers, the connectors have different shapes and contact arrangements. Therefore, a workshop has to keep a lot of different expensive cables and adaptors in store.

For future diagnostic systems, the connection between control unit and vehicle is supposed to be realized with the help of a standardized connector. A connector for the legally demanded exhaust-gas diagnosis was defined by an SAE draft (J1962), concerning form, contact arrangement, and installation position. (Fig. 22.10)

With this connector and a so-called generic scan tool, anyone is able to read the fault-memory in regard to exhaust-gas-relevant defects. The interconnection of the control units allows the access to the entire electronics of the vehicle.

The necessary protocols are partly defined and developed further in standardization committees of the ISO. At the moment, there are two actual standards available:

1. *ISO 9141-2:* Determination of the requirements on hardware and communication protocols. The requirements on hardware are essentially determined by the maximum Baud rate of data transfer and the maximum number of control units simultaneously connected with the diagnostic cable.

   Communication is started by means of a trigger address, and is followed by a synchronization byte of the control unit(s), which is necessary for the automatic setting of the Baud rate. The trigger address calls either a particular control unit or a function, that may also address several control units.

   After transmission of the synchronization byte, the control unit waits for the tester to set the Baud rate, then sends two key-bytes that inform the tester about the suitable data transfer protocol. The tester responds with the last inverted key-byte, in order to confirm the correct receipt. The connection between tester and control unit is now established.

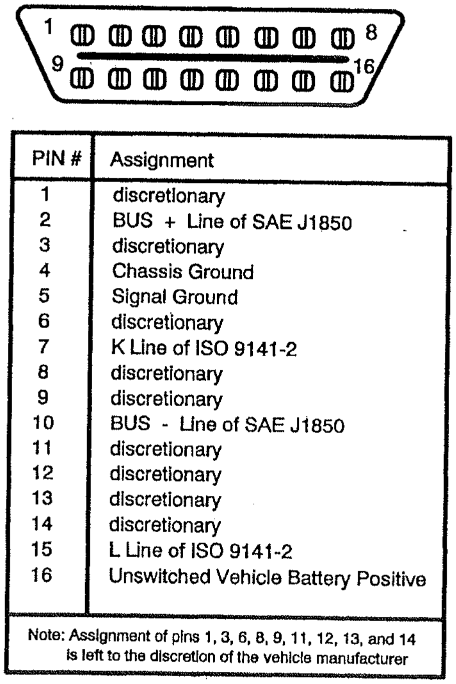
TABLE 22.2 ISO and SAE Documents

ISO 9141		Road Vehicles—Diagnostic System—Requirements for Interchange of Digital Information
ISO/DIS 9141-2		Road Vehicles—Diagnostic System—Part 2: CARB Requirements for Interchange of Digital Information
ISO/DIS 11519-1		Road Vehicles—Low-Speed Serial Data Communication—Part 1: General Definitions
ISO/DIS 11519-2		Road Vehicles—Low-Speed Serial Data Communication—Part 2: Low Speed Controller Area Network (CAN)
ISO/DIS 11519-3		Road Vehicles—Low-Speed Serial Data Communication—Part 3: Vehicle Area Network (VAN)
ISO/DIS 11519-4		Road Vehicles—Low-Speed Serial Data Communication—Part 4: Class B Data Communication Network Interface (J1850)
ISO/DIS 11898		Road Vehicles—Interchange of Digital Information—Controller Area Network (CAN) for High-Speed Communication
ISO/WD 14229		Diagnostic Systems—Diagnostic Services Specification
ISO/WD 14230		Diagnostic Systems—Keyword Protocol 2000 (3 parts: 1: Physical Layer, 2: Data Link Layer, 3: Implementation)
SAE J 1213/1	IR	Glossary of Vehicle Networks for Multiplexing and Data Communications
SAE J 1583	IR	Controller Area Network (CAN), An In-Vehicle Serial Communication Protocol
SAE J 1587	RP	Joint SAE/TMC Electronic Data Interchange Between Microcomputer Systems in Heavy-Duty Vehicle Applications
SAE J 1699	RP	J 1850 Verification Test Procedures
SAE J 1708	RP	Serial Data Communications Between Microcomputer Systems in Heavy-Duty Vehicle Application
SAE J 1724		Vehicle Electronic Identification (New Task Force)
SAE J 1850	RP	Class B Data Communication Network Interface
SAE J 1930	RP	Electrical/Electronic Systems Diagnostic Terms, Definitions, Abreviations and Acronyms
SAE J1939/xx		Truck + Bus, Details next page
SAE J 1962	RP	Diagnostic Connector
SAE J 1978	RP	OBD II Scan Tool
SAE J 1979	RP	E/E Diagnostic Test Modes
SAE J 2008	RP	Electronic Access/Service Information
SAE J 2012	RP	Diagnostic Trouble Code Definitions
SAE J 2037	IR	Off-Board Diagnostic Message Formats
SAE J 2054	IR	E/E Diagnostic Data Communications
SAE J 2056/1	RP	Class C Application Requirement Considerations (Part 2: IR: Survey of Known Protocols, Part 3: IR: Selection of Transmission Media)
SAE J 2057/1	IR	Class A Application/Definition (Part 3: IR: Class A Multiplexing Sensors, Part 4: IR: Class A Multiplexing Architecture Strategies)
SAE J 2106	IR	Token Slot Network for Automotive Control
SAE J 2112	IR	Diagnostic Technician Questionnaire Summary
SAE J 2178	RP	Class B Data Communication Network Messages (Part 1: Detailed Header Formats and Physical Address Assignments, Part 2: Data Parameter Definitions, Part 3: Frame Ids for Single Byte Forms of Headers, Part 4: Message Definition for Three Byte Headers)
SAE J 2186	RP	E/E Data Link Security
SAE J 2190	RP	Enhanced E/E Diagnostic Test Modes
SAE J 2201	RP	Universal Interface for OBD II Scan Tool
SAE J 2205	RP	Diagnostic Specific Functionality Protocol
SAE J 2216	RP	Application of the Clean Air Act Amendment of 1990 (Section 207, Paragraph M5)

RP = Recommended Practice, IR = Information Report

TABLE 22.3 SAE Truck and Bus Documents

SAE J 1939	RP	Serial Control and Communication Vehicle Network (Class C)
SAE J 1939/01		Truck and Bus Control and Communication Vehicle Network (Class C)
SAE J 1939/02		Agricultural Equipment Control and Communication Network
SAE J 1939/1x		Physical Layer, x refers to a specific version
SAE J 1939/11		Physical Layer, 250 kBaud, Twisted Shielded Pair
SAE J 1939/12		Physical Layer, 125 kBaud, Twisted Pair
SAE J 1939/13		Physical Layer, 250 kBaud, Twisted Pair with Ground
SAE J 1939/14		Physical Layer, 1 MBaud, Fiber Optic
SAE J 1939/15		Physical Layer, 50 kBaud, German Agricultural
SAE J 1939/21		CAN 29 Bit Identifier Data Link Layer
SAE J 1939/3x		Network Layer, x refers to a specific version
SAE J 1939/31		Truck + Bus Network Layer
SAE J 1939/4x		Transport Layer, x refers to a specific version
SAE J 1939/5x		Session Layer, x refers to a specific version
SAE J 1939/6x		Presentation Layer, x refers to a specific version
SAE J 1939/7x		Application Layer, x refers to a specific version
SAE J 1939/71		Truck, Bus, Agricultural and Construction Equipment Application Layer
SAE J 1939/72		Virtual Terminal
SAE J 1939/73		Application Layer—Diagnostics
SAE J 1939/81		Network Management
SAE J 1939/??		Tractor-Trailer-Interface



PIN #	Assignment
1	discretionary
2	BUS + Line of SAE J1850
3	discretionary
4	Chassis Ground
5	Signal Ground
6	discretionary
7	K Line of ISO 9141-2
8	discretionary
9	discretionary
10	BUS - Line of SAE J1850
11	discretionary
12	discretionary
13	discretionary
14	discretionary
15	L Line of ISO 9141-2
16	Unswitched Vehicle Battery Positive

Note: Assignment of pins 1, 3, 6, 8, 9, 11, 12, 13, and 14 is left to the discretion of the vehicle manufacturer

FIGURE 22.10 SAE J1962 diagnostic connector.

2. *Interface according to the SAE J1850 (Class B Data Communication Network Interface):* The SAE J1850 defines means and methods for serial data exchange for automotive application at the physical and data link layer of the OSI model. It is used for networked systems and for diagnostic purposes.

   Two implementations are characterized: pulse-width modulation (PWM) at 41.6 kbps transmitted on twisted pair wires, and variable pulse-width modulation (VPM) at 10.4 kbps, transmitted on a single wire.[12]

A generic scan tool, as mentioned, therefore, has to handle the three different interfaces.

A new protocol, *Keyword 2000,* is prepared by the ISO committees. It is supposed to combine the protocols that have been used up to now.

With the introduction of more and more diagnostic functions and networked systems in the vehicle, the functional structure will be modified (Fig. 22.11).

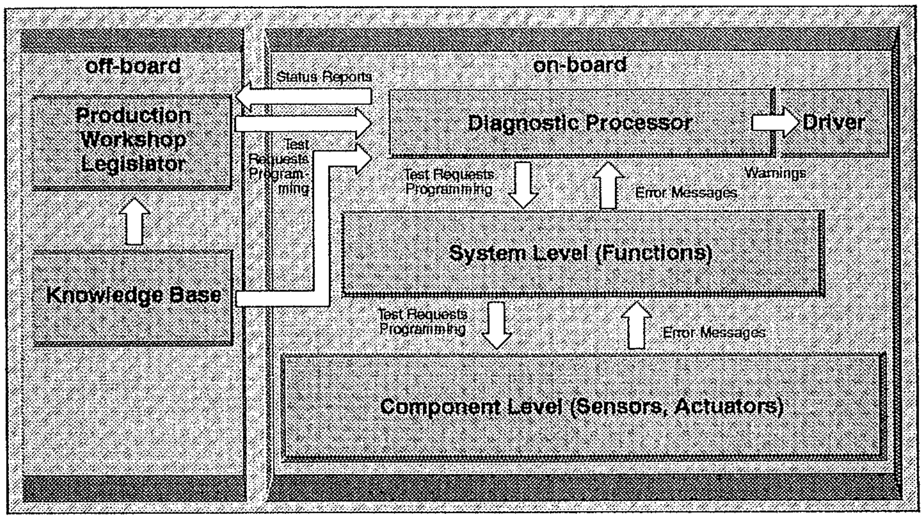


FIGURE 22.11 Logical structure for future diagnosis.

A diagnostic processor on top of a hierarchical structure of functions has access to every system via the network. It can request status information of the functions of the levels below, or of the sensors and actuators, and receives warning messages if problems are detected by the self-diagnosis of the different subsystems. The diagnostic processor serves as a man-machine interface to the driver and as a gate to the outside. It is the only secure access to the entire system of the vehicle.

GLOSSARY

CAN Controller Area Network (standardized protocol developed by Bosch for networked systems).

CARB California Air Resources Board.

CD-ROM Compact disk read only memory, a data storage medium.

DIS Draft International Standard.

ECU Electronic control unit.

EGR Exhaust-gas recirculation.

EPA Environmental Protection Agency.

Freeze frame Faults stored together with various related parameters.

HC Hydrocarbon.

ISO International Standardization Organization.

ISO 9141-2 Standardized protocol for data exchange between ECUs and testers.

Lambda controller Electronic system for controlling the air/fuel ratio.

Lambda sensor A sensor for air/fuel ratio (oxygen sensor).

MIL Malfunction indicator lamp (indicates emission-related faults to the driver).

OBDII On Board Diagnostics II.

Off-board diagnosis Diagnosis performed by means outside a vehicle.

On-board diagnosis Diagnosis performed by means within a vehicle.

OSI Open System Interconnection.

PC Personal computer.

PWM Pulse-width modulation.

RS 232 Standardized data link (hardware).

Scan tool Small tester that can be connected to the diagnostic connector to interrogate emission-related fault codes.

SAE Society of Automotive Engineers.

TDC Top dead center.

Terminal 1 Connection to a signal related to ignition timing.

VAN Vehicle Area Network (French proposal for network protocol).

VPM Variable pulse-width modulation.

REFERENCES

1. W. Bremer and Heintz, F., "Was bieten die Karosserie-Zentral-Elektronik mit Kabelbaum-Multiplex und die On-Board-Diagnose?," (Advantages of a central body-electronic with multiplex and on-board diagnosis), Bosch Customer Information, 1977.
2. W. Bremer, Heintz, F., and Hugel, R., "Diagnosis in Networked Systems," *International Conference Automotive Diagnostics,* London, 1990.
3. K. Dieterich and Unruh, J., "CAN—A Bus System for Serial Data Transfer," *23. FISITA Congress "The promise of new technology in the automotive industry,"* 1990.
4. T. Goelzer and Leonhard, R., "A new Architecture for Car Electronics", *International Symposium Vehicle Electronics Integration ATA-EL 91,* 1991.

5. W. Kremer and Kaminski, D., "Integrated Vehicle Electronics with CAN," *23. FISITA Congress "The promise of new technology in the automotive industry,"* 1990.
6. T. Kühner and Kaminski, D., "Structuring of vehicle electronics in cars for an effective on-board-diagnosis system," *VDI Reports "Electronic in Vehicles,"* Nr. 819, 1990.
7. E. Hipp, Jung, C., and Morizur, P., "On Board Diagnosis as the Central Interface for Modern Vehicle Electronics," *International Symposium Vehicle Electronics Integration ATA-EL 91,* 1991.
8. W. Botzenhardt, Litschel, M., and Unruh, J., "Bussysteme für Kfz-Sterugeräte" (Bus systems for vehicle electronic control units), *VDI Reports "Electronic in Vehicles,"* Nr. 612, 1986.
9. B. Przbylla, "Eigendiagnose von elektronischen Steuergeräten im Kraftfahrzeug" (Self-diagnosis of electronic control units in the vehicle), *VDI Reports "Electronic in Vehicles,"* Nr. 612, 1986.
10. W. Bremer, "Möglichkeiten der fahrzeugfesten Überwachung mit Kabelbaum-Multiplex," (Possibilities of On-Board Diagnostics with Multiplex), *Status-seminar of the German Ministry of Research and Technology,* 1978.
11. D. Nemec, "Möglichkeiten komfortabler Testgeräte zur Auswertung der Eigendiagnose von Steuergeräten" (Possibilities of comfortable test equipment for the evaluation of selfdiagnostic data of control systems), *VDI Reports "Electronic in Vehicles,"* Nr. 687, 1988.
12. A. W. Millsap, Lowden, M. T., Folkerts, M. A., Unruh, J., and Dais, S., "Mapping J1850 Messages into CAN Version 2.0," *SAE International Congress,* Paper 930437, 1993.

ABOUT THE AUTHORS

Wolfgang Bremer studied electrical communication techniques from 1962 to 1968 at the University of Karlsruhe (Germany). Afterwards he was engaged with Siemens AG in Karlsruhe in the development of high-precision electronic balances. Since 1970, he has been working in the advanced engineering department of measurement and information techniques of Robert Bosch GmbH and is responsible for the development of serial communication and diagnostic systems in the chassis area of vehicles. He is working in several ISO and SAE committees and working groups in the field of diagnosis.

Frieder Heintz studied electrical communication techniques from 1954 to 1959 at the University of Karlsruhe. Afterwards he was engaged in research and development of process-control computers with Siemens AG in Karlsruhe, Munich, and New York. Since 1969, he has been the head of the advanced engineering department for measurement and information techniques at Robert Bosch GmbH. For 20 years, he has managed national and international working groups for diagnosis and serial data transfer in vehicles within the ISO (International Organization for Standardization).

Robert Hugel studied physics at the University of Karlsruhe from 1967 to 1972. Then he was engaged for six years in the development, production, and sales of infrared spectrometers, two years in the development of laser-based measuring systems, and the last 13 years in developing various automotive electronics at Robert Bosch in the advanced engineering department for measurement and information techniques.