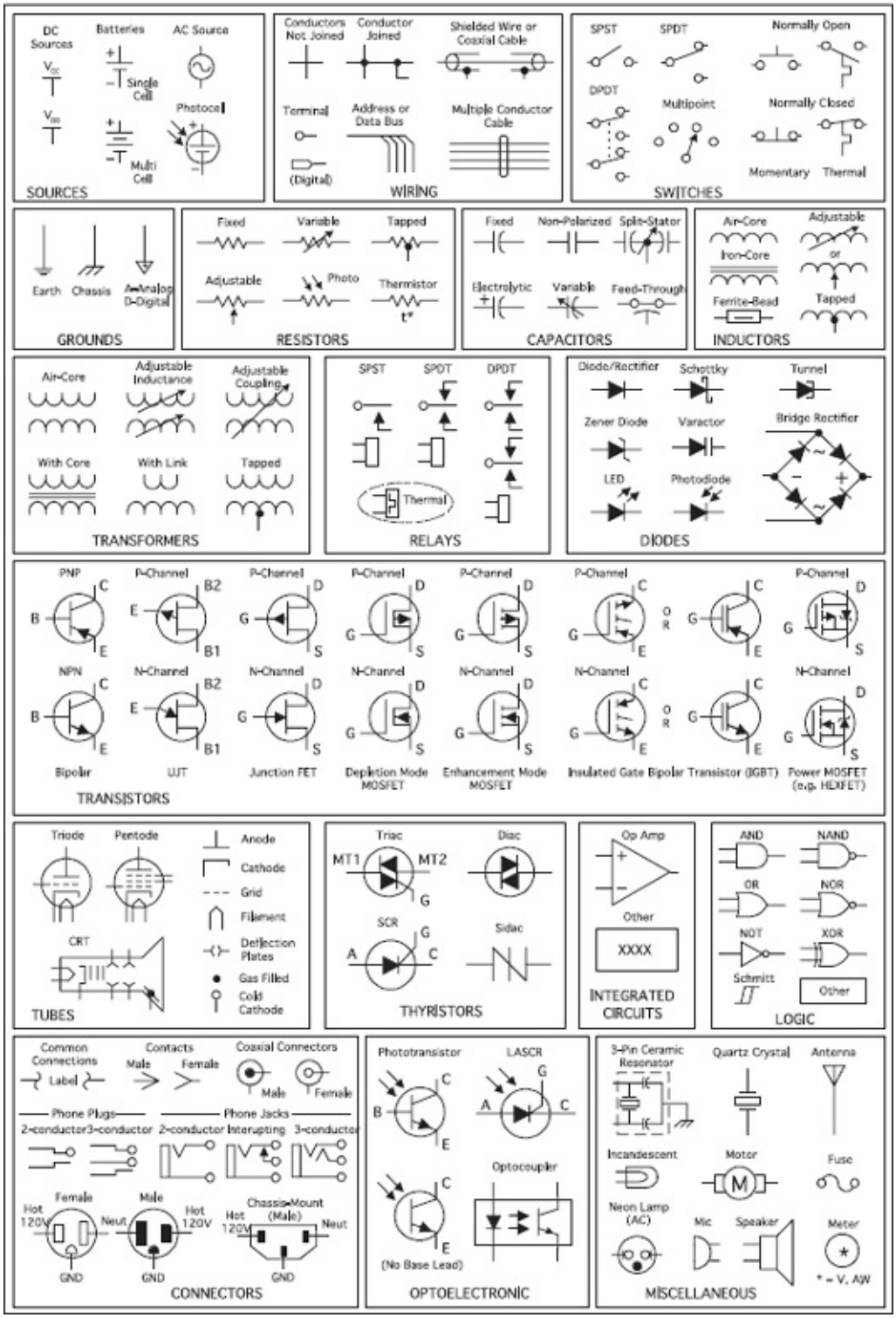
THIRD EDITION

PRACTICAL ELECTRONICS FOR INVENTORS



Paul Scherz and Simon Monk

TAB
Make Great Stuff!


DC Sources
Batteries
AC Source
Single Cell
Photocell
Multi Cell
SOURCES
Conductors Not Joined
Conductor Joined
Shielded Wire or Coaxial Cable
Terminal
(Digital)
Address or Data Bus
Multiple Conductor Cable
WIRING
SPST
SPDT
Normally Open
DPDT
Multipoint
Normally Closed
Momentary
Thermal
SWITCHES
Earth
Chassis
Analog
Digital
GROUNDS
Fixed
Variable
Tapped
Adjustable
Photo
Thermistor
t°
RESISTORS
Fixed
Non-Polarized
Split-Stator
Electrolytic
Variable
Feed-Through
CAPACITORS
Air-Core
Adjustable
or
Iron-Core
Tapped
Ferrite-Bead
INDUCTORS
Air-Core
Adjustable Inductance
Adjustable Coupling
With Core
With Link
Tapped
TRANSFORMERS
SPST
SPDT
DPDT
Thermal
RELAYS
Diode/Rectifier
Schottky
Tunnel
Zener Diode
Varactor
Bridge Rectifier
LED
Photodiode
DIODES
PNP
P-Channel
NPN
N-Channel
Bipolar
UJT
Junction FET
Depletion Mode MOSFET
Enhancement Mode MOSFET
Insulated Gate Bipolar Transistor (IGBT)
Power MOSFET (e.g. HEXFET)
TRANSISTORS
Triode
Pentode
CRT
Anode
Cathode
Grid
Filament
Deflection Plates
Gas Filled
Cold Cathode
TUBES
Triac
MT1
MT2
G
Diac
SCR
A
C
Sidac
THYRISTORS
Op Amp
Other
XXXX
INTEGRATED CIRCUITS
AND
NAND
OR
NOR
NOT
XOR
Schmitt
Other
LOGIC
Common Connections
Label
Contacts
Male
Female
Coaxial Connectors
Phone Plugs
2-conductor
3-conductor
Phone Jacks
2-conductor
Interupting
3-conductor
Female
Male
Chassis-Mount (Male)
Hot 120V
Neut
GND
CONNECTORS
Phototransistor
LASCR
Optocoupler
(No Base Lead)
OPTOELECTRONIC
3-Pin Ceramic Resonator
Quartz Crystal
Antenna
Incandescent
Motor
Fuse
Neon Lamp (AC)
Mic
Speaker
Meter
* = V, A, W
MISCELLANEOUS

Resistor Labels

Conversion Calculator

k = 1,000 ; M = 1,000,000

1MΩ = 1,000,000 Ω = 1 x 10^6Ω

1kΩ = 1,000 Ω = 1 x 10^3Ω

Examples:

3.3 kΩ = 3,300 Ω = 3.3 x 10^3Ω

22 kΩ = 22,000 Ω = 22 x 10^3Ω

2 MΩ = 2,000,000 Ω = 2 x 10^6Ω

1.68 MΩ = 1,680,000 Ω = 1.68 x 10^6Ω

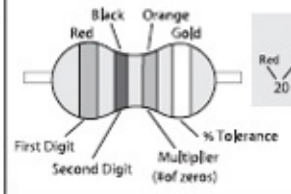
Resistor Color Code

Color	Sig. Fig.	Decimal Multiplier	Tolerance (%)
Black	0	1	-
Brown	1	10	1
Red	2	100	2
Orange	3	1,000	*
Yellow	4	10,000	-
Green	5	100,000	0.5
Blue	6	1,000,000	0.25
Purple	7	10,000,000	0.1
Grey	8	100,000,000	-
White	9	1,000,000,000	*
Gold	-	0.1	5
Silver	*	0.01	10
No Color	-	-	20

Body Color

The body color of a resistor typically doesn't carry meaning, except in some instances where it may specify temperature coefficient. However, if you find resistors within a circuit that are white/gray or blue in color, they may be non-flammable or fusible resistors. Care must be taken when replacing such resistors—don't substitute ordinary resistors in their place.

4-Band Resistor Code (Most Common)

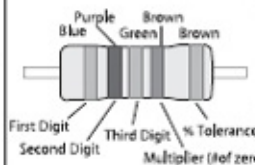


Label Meaning

Red Black Orange Gold

20 x 1,000 = 20k Ω ± 5%

5-Band Resistor Code (3-digit)



Label Meaning

Purple Brown

Blue Green Brown

675 x 10 = 6750 Ω ±1%

5-Band Resistor Code (Reliability)



Label Meaning

Yellow Green Silver

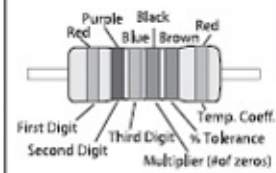
Purple

47 x 100,000 = 4.7 MΩ ± 10%

1% Reliability/1000 Hr — Brown

Color	Reliability (%/1000 Hr)
Brown	1
Red	0.1
Orange	0.01
Yellow	0.001

6-Band Resistor Code



Label Meaning

Purple Black

Red Blue Brown

276 x 1 = 276Ω ± 1%

TC of 50 ppm — Red

Color	Temp. Coeff
Brown	100 ppm
Red	50 ppm
Orange	15 ppm
Yellow	25 ppm

Surface Mount Resistor Code

3-digit Label

Label Meaning

101	10 and 1 zero = 100 Ω
105	10 and 5 zero = 1,000,000 Ω
224	22 and 4 zeros = 220,000 Ω
1R0	1.0 and no zeros = 1 Ω
22R	22.0 and no zeros = 22 Ω
R10	0.1 and no zeros = 0.1 Ω

The first two digits represent significant figures; the last digit specifies the multiplier. For values under 100 Ω, the letter R is substituted for one of the significant digits and represents a decimal point.

4-digit Label

Label Meaning

1001	100 and 1 zero =1000 Ω
22R0	22.0 and no zeros = 22 Ω

The first three digits represent significant figures; the last digit specifies the multiplier. R represents a decimal point

Tolerance Label

Label Meaning

101F	100 Ω ± 1%
1R0D	1.0 Ω ± 0.5%

Letter	Tolerance
D	±0.5 %
F	±1.0 %
G	±2.0 %
J	±5.0 %

Capacitor Markings

Capacitance Conversion Calculator

$1\ F = 1 \times 10^{6}\ \mu F = 1 \times 10^{9}\ nF = 1 \times 10^{12}\ pF$
$1\ \mu F = 1 \times 10^{-6}\ F = 1 \times 10^{3}\ nF = 1 \times 10^{6}\ pF$
$1\ nF = 1 \times 10^{-9}\ F = 1 \times 10^{-6}\ \mu F = 1 \times 10^{3}\ pF$
$1\ pF = 1 \times 10^{-12}\ F = 1 \times 10^{-6}\ \mu F = 1 \times 10^{-3}\ nF$

F = Farad, μ = micro, n = nano, p = pico

$1000\ \mu F = 1{,}000{,}000\ nF = 10 \times 10^{8}\ pF$
$100\ \mu F = 100{,}000\ nF = 10 \times 10^{7}\ pF$
$10\ \mu F = 10{,}000\ nF = 10 \times 10^{6}\ pF$
$1\ \mu F = 1{,}000\ nF = 10 \times 10^{5}\ pF$
$0.1\ \mu F = 100\ nF = 10 \times 10^{4}\ pF$
$0.01\ \mu F = 10\ nF = 10 \times 10^{3}\ pF$
$0.001\ \mu F = 1\ nF = 10 \times 10^{2}\ pF$

Tantalum



Label meaning 1
- 1st significant figure in μF
- 2nd significant figure in μF
- Multiplier
- Voltage (See table)

Color	S.F.	Multiple	Voltage
Black	0	1	10V
Brown	1	10	
Red	2	100	
Orange	3	1000	
Yellow	4		6.3V
Green	5		16V
Blue	6		20V
Violet	7		
Grey	8	0.01	25V
White	9	0.1	3V
Pink			35V



Label meaning 2

Marking	Actual
22	22μF, 16 V

Mylar (Polyester Film)
Polypropylene
Dipped Mica



Label meaning

Marking	Actual
.001K*	0.001 μF, ± 10%
104K	0.1 μF, ± 10%
.22J*	0.22 μF, ± 5%
472K	0.0047 μF, ± 10%
221J	220 pF, ± 5%
470J	47pf, ± 5%
102J	1000pF, ± 5%
103F	0.01 μF, ± 1%
223F	0.022 μF, ± 1%
104F	0.1 μF, ± 1%

Labels:
1st digit, 2nd digit, multiplier in pF (or μF if decimal before digits), and tolerance.

Metallized Polyester Film



Label meaning

Marking	Actual
2μ2	2.2 μF
μ22	0.22 μF
68n	68 nF
4n7	4.7 nF

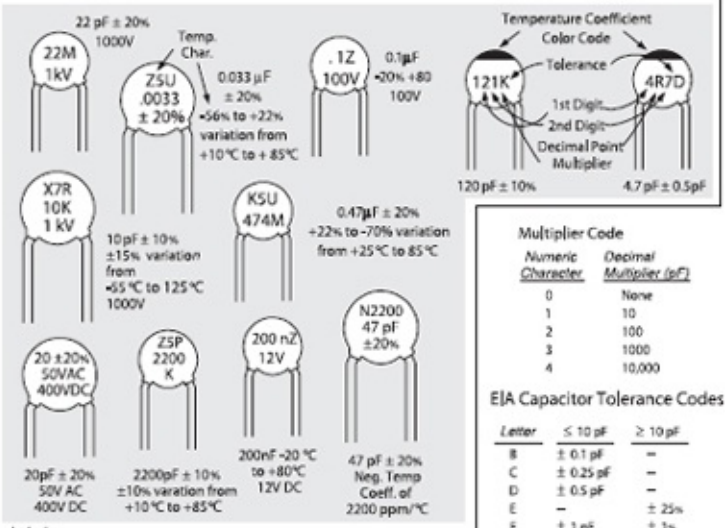
Label:
"μ" place of decimal in microfarads
"n" place of decimal in nanofarads

Polyester Color Coded



1st digit (pF)
2nd digit (pF)
Multiplier — Standard color code
Tolerance — Black ± 20%, White ± 10%, Green ± 5%
Voltage — Brown 100, Red 250, Yellow 400

Ceramic Disc Capacitors



Label:
Varies widely according to manufacturer. Usually given in pF (see multiplier code table) but may be given in μF when there is a decimal before digits. See other tables for temperature and tolerance markings.

Multiplier Code

Numeric Character	Decimal Multiplier (pF)
0	None
1	10
2	100
3	1000
4	10,000

EIA Capacitor Tolerance Codes

Letter	≤ 10 pF	≥ 10 pF
B	± 0.1 pF	–
C	± 0.25 pF	–
D	± 0.5 pF	–
E	–	± 25%
F	± 1 pF	± 1%
G	–	± 2%
H	–	± 2.5%
J	–	± 5%
K	–	± 10%
M	–	± 20%
P	–	+0 + 100%
S	–	-20 + 50%
W	–	+0 + 200%
X	–	-20 + 40%
Z	–	+20 + 80%

Ceramic Disc (European Markings)



Label Meaning

Marking	Actual	Marking	Actual
p68	0.68 pF	22p	22 pF
1p0	1.0 pF	n10	0.1 nF
4p7	4.7 pF	n27	0.27 nF

Label: p = picofarads, n = nanofarads; location of p or n signifies decimal point.

EIA Temperature Characteristic Codes

Minimum temperature		Maximum temperature		Max cap. change over temp. range	
X	−55°C	2	+45°C	A	±1.0%
Y	−35°C	4	+65°C	B	±1.5%
Z	+10°C	5	+85°C	C	±2.2%
		6	+105°C	D	±3.3%
		7	+125°C	E	±4.7%
				F	±7.5%
				P	±10%
				R	±15%
				S	±22%
				T	-33%, +22%
				U	-56%, +22%
				V	-82%, +22%

Fixed Ceramic Color Code



Color	S.F.	Multiple	Tolerance >10pF	Tolerance <10pF	Temp. Coeff. ppm/°C
Black	0	1	± 20%	2.0 pF	-0
Brown	1	10	± 1%		-30
Red	2	100	± 2%		-60
Orange	3	1000			-150
Yellow	4				-220
Green	5		± 5%	0.5pF	-330
Blue	6				-470
Violet	7				-750
Grey	8	0.01		0.25pF	30
White	9	0.1	± 10%	1.0pF	500

Surface Mount Capacitors

SMD Ceramic



Label meaning

Marking	Actual
N1	33 pF
A4	0.01 μF
S6	4.7 μF

SMD Electrolytic



Label meaning 1

Marking	Actual
10 6V	10 μF, 6V

Label meaning 2

Marking	Actual
A47S	4.7 μF, 10V

Significant Figure Code

Char.	S. F.	Char.	S. F.
A	1.0	T	5.1
B	1.1	U	5.6
C	1.2	V	6.2
D	1.3	W	6.8
E	1.5	X	7.5
F	1.6	Y	8.2
G	1.8	Z	9.1
H	2.0	a	2.5
J	2.2	b	3.5
K	2.4	d	4.0
L	2.7	e	4.5
M	3.0	f	5.0
N	3.3	m	6.0
P	3.6	n	7.0
Q	3.9	t	8.0
R	4.3	y	9.0
S	4.7		

Multiplier Code

Numeric Character	Decimal Multiplier (pF)
0	1
1	10
2	100
3	1,000
4	10,000
5	100,000
6	1,000,000
7	10,000,000
8	100,000,000
9	0.1

Label 2:
Voltage (see table below), 1st digit, 2nd digit, multiplier (pF).

Char.	Voltage
e	2.5
G	4
J	6.3
A	10
C	16
D	20
E	25
V	35
H	50

EIA Temperature Coefficient Color Codes

Color	Temp. Coeff. Industry	EIA
Black	N90	C0G
Brown	N030/N033	S1G
Red	N075/N080	U1G
Orange	N 150	P2G
Yellow	N 220	R2G
Green	N 330	S2H
Blue	N 470	U2J
Violet	N 750	
Gray		
White	P 100	
Red/Violet	P 100	

Electrolytic Capacitors



1 μF, 50V

Label: Usually self-explanatory

Practical Electronics for Inventors

Third Edition

ABOUT THE AUTHORS

Paul Scherz is a physicist/mechanical engineer who received his BS in Physics from the University of Wisconsin. His area of interest in physics focuses on elementary particle interactions. Paul is an inventor/hobbyist in electronics, an area he grew to appreciate through his experience at the University's Department of Nuclear Engineering and Engineering Physics and the Department of Plasma Physics.

Dr. Simon Monk has a bachelor's degree in Cybernetics and Computer Science and a PhD in Software Engineering. Simon spent several years as an academic before he returned to industry, cofounding the mobile software company Momote Ltd. Simon is now a full-time author and has published a number of books in the McGraw-Hill *Evil Genius* series, as well as books on programming the Arduino and Raspberry Pi. He has also published books on IOIO and .NET Gadgeteer.

ABOUT THE TECHNICAL EDITORS

Michael Margolis has more than 40 years of experience developing and delivering hardware and software solutions. He has worked at senior levels with Sony, Lucent/Bell Labs, and a number of start-up companies. Michael is the author of two books, *Arduino Cookbook* and *Make an Arduino-Controlled Robot: Autonomous and Remote-Controlled Bots on Wheels*.

Chris Fitzer is a solutions architect and technical manager, who received his PhD in Electrical and Electronic Engineering from the University of Manchester Institute of Science and Technology (UMIST) in 2003 and a first class honors degree (BSc) in 1999. Chris currently leads a global team, developing and deploying Smart Grid technologies around the world. Previous positions have seen him drive the European interests of the ZigBee Smart Energy (ZSE) profile and lead the development of the world's first certified Smart Energy In Premise Display (IPD) and prototype smart meter. He has also authored or co-authored numerous technical journal papers within the field of Smarter Grids.

Practical Electronics for Inventors

Third Edition

Paul Scherz

Simon Monk



New York Chicago San Francisco Lisbon
London Madrid Mexico City Milan New Delhi
San Juan Seoul Singapore Sydney Toronto

Copyright © 2013 by The McGraw-Hill Companies. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

ISBN: 978-0-07-177134-4

MHID: 0-07-177134-4

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-177133-7,

MHID: 0-07-177133-6.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. To contact a representative please e-mail us at bulksales@mcgraw-hill.com.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

Information has been obtained by McGraw-Hill from sources believed to be reliable. However, because of the possibility of human or mechanical error by our sources, McGraw-Hill, or others, McGraw-Hill does not guarantee the accuracy, adequacy, or completeness of any information and is not responsible for any errors or omissions or the results obtained from the use of such information.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. (“McGrawHill”) and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill’s prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED “AS IS.” McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

CONTENTS

Preface

Acknowledgments

CHAPTER 1 Introduction to Electronics

CHAPTER 2 Theory

2.1 Theory of Electronics
2.2 Electric Current
2.2.1 Currents in Perspective
2.3 Voltage
2.3.1 The Mechanisms of Voltage
2.3.2 Definition of Volt and Generalized Power Law
2.3.3 Combining Batteries
2.3.4 Other Voltage Sources
2.3.5 Water Analogies
2.4 A Microscopic View of Conduction (For Those Who Are Interested)
2.4.1 Applying a Voltage
2.5 Resistance, Resistivity, Conductivity
2.5.1 How the Shape of a Conductor Affects Resistance
2.5.2 Resistivity and Conductivity
2.6 Insulators, Conductors, and Semiconductors
2.7 Heat and Power
2.8 Thermal Heat Conduction and Thermal Resistance
2.8.1 Importance of Heat Production
2.9 Wire Gauges
2.10 Grounds
2.10.1 Earth Ground
2.10.2 Different Types of Ground Symbols
2.10.3 Loose Ends on Grounding
2.11 Electric Circuits
2.12 Ohm's Law and Resistors
2.12.1 Resistor Power Ratings
2.12.2 Resistors in Parallel
2.12.3 Resistors in Series
2.12.4 Reducing a Complex Resistor Network
2.12.5 Multiple Voltage Dividers
2.13 Voltage and Current Sources
2.14 Measuring Voltage, Current, and Resistance
2.15 Combining Batteries
2.16 Open and Short Circuits
2.17 Kirchhoff's Laws
2.18 Superposition Theorem
2.19 Thevenin's and Norton's Theorems
2.19.1 Thevenin's Theorem
2.19.2 Norton's Theorem
2.20 AC Circuits

2.20.1 Generating AC
2.20.2 Water Analogy of AC
2.20.3 Pulsating DC
2.20.4 Combining Sinusoidal Sources
2.20.5 AC Waveforms
2.20.6 Describing an AC Waveform
2.20.7 Frequency and Period
2.20.8 Phase
2.21 AC and Resistors, RMS Voltage, and Current
2.22 Mains Power
2.23 Capacitors
2.23.1 Determining Capacitance
2.23.2 Commercial Capacitors
2.23.3 Voltage Rating and Dielectric Breakdown
2.23.4 Maxwell's Displacement Current
2.23.5 Charge-Based Model of Current Through a Capacitor
2.23.6 Capacitor Water Analogy
2.23.7 Energy in a Capacitor
2.23.8 RC Time Constant
2.23.9 Stray Capacitance
2.23.10 Capacitors in Parallel
2.23.11 Capacitors in Series
2.23.12 Alternating Current in a Capacitor
2.23.13 Capacitive Reactance
2.23.14 Capacitive Divider
2.23.15 Quality Factor
2.24 Inductors
2.24.1 Electromagnetism
2.24.2 Magnetic Fields and Their Influence
2.24.3 Self-Inductance
2.24.4 Inductors
2.24.5 Inductor Water Analogy
2.24.6 Inductor Equations
2.24.7 Energy Within an Inductor
2.24.8 Inductor Cores
2.24.9 Understanding the Inductor Equations
2.24.10 Energizing LR Circuit
2.24.11 Deenergizing LR Circuit
2.24.12 Voltage Spikes Due to Switching
2.24.13 Straight-Wire Inductance
2.24.14 Mutual Inductance and Magnetic Coupling
2.24.15 Unwanted Coupling: Spikes, Lightning, and Other Pulses
2.24.16 Inductors in Series and Parallel
2.24.17 Alternating Current and Inductors
2.24.18 Inductive Reactance
2.24.19 Nonideal Inductor Model
2.24.20 Quality Factor
2.24.21 Inductor Applications
2.25 Modeling Complex Circuits
2.26 Complex Numbers
2.27 Circuit with Sinusoidal Sources
2.27.1 Analyzing Sinusoidal Circuits with Complex Impedances
2.27.2 Sinusoidal Voltage Source in Complex Notation
2.27.3 Odd Phenomena in Reactive Circuits

2.28 Power in AC Circuits (Apparent Power, Real Power, Reactive Power)
2.28.1 Power Factor
2.29 Thevenin's Theorem in AC Form
2.30 Resonant Circuits
2.30.1 Resonance in RLC Circuits
2.30.2 Q (Quality Factor) and Bandwidth
2.30.3 Bandwidth
2.30.4 Voltage Drop Across Components in RLC Resonant Circuit
2.30.5 Capacitor Losses
2.30.6 Parallel-Resonant Circuits
2.30.7 The Q of Loaded Circuits
2.31 Lecture on Decibels
2.31.1 Alternative Decibel Representations
2.32 Input and Output Impedance
2.32.1 Input Impedance
2.32.2 Output Impedance
2.33 Two-Port Networks and Filters
2.33.1 Filters
2.33.2 Attenuators
2.34 Transient Circuits
2.34.1 Series RLC Circuit
2.35 Circuits with Periodic Nonsinusoidal Sources
2.35.1 Fourier Series
2.36 Nonperiodic Sources
2.37 SPICE
2.37.1 How SPICE Works
2.37.2 Limitations of SPICE and Other Simulators
2.37.3 A Simple Simulation Example

CHAPTER 3 Basic Electronic Circuit Components

3.1 Wires, Cables, and Connectors
3.1.1 Wires
3.1.2 Cables
3.1.3 Connectors
3.1.4 Wiring and Connector Symbols
3.1.5 High-Frequency Effects Within Wires and Cables
3.2 Batteries
3.2.1 How a Cell Works
3.2.2 Primary Batteries
3.2.3 Comparing Primary Batteries
3.2.4 Secondary Batteries
3.2.5 Battery Capacity
3.2.6 Note on Internal Voltage Drop of a Battery
3.3 Switches
3.3.1 How a Switch Works
3.3.2 Describing a Switch
3.3.3 Kinds of Switches
3.3.4 Simple Switch Applications
3.4 Relays
3.4.1 Specific Kinds of Relays
3.4.2 A Few Notes about Relays
3.4.3 Some Simple Relay Circuits
3.5 Resistors

3.5.1 Resistance and Ohm's Law
3.5.2 Resistors in Series and Parallel
3.5.3 Reading Resistor Labels
3.5.4 Real Resistor Characteristics
3.5.5 Types of Resistors
3.5.6 Variable Resistors (Rheostats, Potentiometers, Trimmers)
3.5.7 Potentiometer Characteristics
3.6 Capacitors
3.6.1 Capacitance
3.6.2 Capacitors in Parallel
3.6.3 Capacitors in Series
3.6.4 RC Time Constant
3.6.5 Capacitive Reactance
3.6.6 Real Capacitors
3.6.7 Capacitor Specifications
3.6.8 Types of Capacitors
3.6.9 Capacitor Applications
3.6.10 Timing and Sample and Hold
3.6.11 RC Ripple Filter
3.6.12 Arc Suppression
3.6.13 Supercapacitor Applications
3.6.14 Problems
3.7 Inductors
3.7.1 Inductance
3.7.2 Constructing Inductors
3.7.3 Inductors in Series and Parallel
3.7.4 RL Time Constant
3.7.5 Inductive Reactance
3.7.6 Real Inductors
3.7.7 Inductor Specifications
3.7.8 Types of Inductors
3.7.9 Reading Inductor Labels
3.7.10 Inductor Applications
3.7.11 EMI/EMC Design Tips
3.8 Transformers
3.8.1 Basic Operations
3.8.2 Transformer Construction
3.8.3 Autotransformers and Variable Transformers
3.8.4 Circuit Isolation and the Isolation Transformer
3.8.5 Various Standard and Specialized Transformers
3.8.6 Transformer Applications
3.9 Fuses and Circuit Breakers
3.9.1 Types of Fuses and Circuit Breakers

CHAPTER 4 Semiconductors

4.1 Semiconductor Technology
4.1.1 What Is a Semiconductor?
4.1.2 Applications of Silicon
4.2 Diodes
4.2.1 How p-n Junction Diodes Work
4.2.2 Diode Water Analogy
4.2.3 Kinds of Rectifiers/Diodes
4.2.4 Practical Considerations

4.2.5 Diode/Rectifier Applications
4.2.6 Zener Diodes
4.2.7 Zener Diode Applications
4.2.8 Varactor Diodes (Variable Capacitance Diodes)
4.2.9 PIN Diodes
4.2.10 Microwave Diodes (IMPATT, Gunn, Tunnel, etc.)
4.2.11 Problems
4.3 Transistors
4.3.1 Introduction to Transistors
4.3.2 Bipolar Transistors
4.3.3 Junction Field-Effect Transistors
4.3.4 Metal Oxide Semiconductor Field-Effect Transistors
4.3.5 Unijunction Transistors
4.4 Thyristors
4.4.1 Introduction
4.4.2 Silicon-Controlled Rectifiers
4.4.3 Silicon-Controlled Switches
4.4.4 Triacs
4.4.5 Four-Layer Diodes and Diacs
4.5 Transient Voltage Suppressors
4.5.1 Lecture on Transients
4.5.2 Devices Used to Suppress Transients
4.6 Integrated Circuits
4.6.1 IC Packages

CHAPTER 5 Optoelectronics

5.1 A Little Lecture on Photons
5.2 Lamps
5.3 Light-Emitting Diodes
5.3.1 How an LED Works
5.3.2 Kinds of LEDs
5.3.3 More on LEDs
5.3.4 LED Applications
5.3.5 Laser Diodes
5.4 Photoresistors
5.4.1 How a Photoresistor Works
5.4.2 Technical Stuff
5.4.3 Applications
5.5 Photodiodes
5.5.1 How a Photodiode Works
5.5.2 Basic Operations
5.5.3 Kinds of Photodiodes
5.6 Solar Cells
5.6.1 Basic Operations
5.7 Phototransistors
5.7.1 How a Phototransistor Works
5.7.2 Basic Configurations
5.7.3 Kinds of Phototransistors
5.7.4 Technical Stuff
5.7.5 Applications
5.8 Photothyristors
5.8.1 How LASCRs Work
5.8.2 Basic Operation

5.9 Optoisolators
5.9.1 Integrated Optoisolators
5.9.2 Applications
5.10 Optical Fiber

CHAPTER 6 Sensors

6.1 General Principals
6.1.1 Precision, Accuracy, and Resolution
6.1.2 The Observer Effect
6.1.3 Calibration
6.2 Temperature
6.2.1 Thermistors
6.2.2 Thermocouples
6.2.3 Resistive Temperature Detectors
6.2.4 Analog Output Thermometer ICs
6.2.5 Digital Thermometer ICs
6.2.6 Infrared Thermometers/Pyrometers
6.2.7 Summary
6.3 Proximity and Touch
6.3.1 Touch Screens
6.3.2 Ultrasonic Distance
6.3.3 Optical Distance
6.3.4 Capacitive Sensors
6.3.5 Summary
6.4 Movement, Force, and Pressure
6.4.1 Passive Infrared
6.4.2 Acceleration
6.4.3 Rotation
6.4.4 Flow
6.4.5 Force
6.4.6 Tilt
6.4.7 Vibration and Mechanical Shock
6.4.8 Pressure
6.5 Chemical
6.5.1 Smoke
6.5.2 Gas
6.5.3 Humidity
6.6 Light, Radiation, Magnetism, and Sound
6.6.1 Light
6.6.2 Ionizing Radiation
6.6.3 Magnetic Fields
6.6.4 Sound
6.7 GPS

CHAPTER 7 Hands-on Electronics

7.1 Safety
7.1.1 Lecture on Safety
7.1.2 Damaging Components with Electrostatic Discharge
7.1.3 Component Handling Precautions
7.2 Constructing Circuits
7.2.1 Drawing a Circuit Schematic
7.2.2 A Note on Circuit Simulator Programs
7.2.3 Making a Prototype of Your Circuit

7.2.4 The Final Circuit
7.2.5 Making a PCB
7.2.6 Special Pieces of Hardware Used in Circuit Construction
7.2.7 Soldering
7.2.8 Desoldering
7.2.9 Enclosing the Circuit
7.2.10 Useful Items to Keep Handy
7.2.11 Troubleshooting the Circuits You Build
7.3 Multimeters
7.3.1 Basic Operation
7.3.2 How Analog VOMs Work
7.3.3 How Digital Multimeters Work
7.3.4 A Note on Measurement Errors
7.4 Oscilloscopes
7.4.1 How Scopes Work
7.4.2 Interior Circuitry of a Scope
7.4.3 Aiming the Beam
7.4.4 Scope Usage
7.4.5 What All the Little Knobs and Switches Do
7.4.6 Measuring Things with Scopes
7.4.7 Scope Applications
7.4.8 Measuring Impedances
7.5 The Electronics Laboratory
7.5.1 Work Area
7.5.2 Test Equipment
7.5.3 Multimeters
7.5.4 DC Power Supplies
7.5.5 Oscilloscope
7.5.6 Oscilloscope Probes
7.5.7 General-Purpose Function Generator
7.5.8 Frequency Counter
7.5.9 Computer
7.5.10 Miscellaneous Test Equipment
7.5.11 Multifunction PC Instruments
7.5.12 Isolation Transformers
7.5.13 Variable Transformers, or Variacs
7.5.14 Substitution Boxes
7.5.15 Test Cables, Connectors, and Adapters
7.5.16 Soldering Equipment
7.5.17 Prototyping Boards
7.5.18 Hand Tools
7.5.19 Wires, Cables, Hardware, and Chemicals
7.5.20 Electronics Catalogs
7.5.21 Recommended Electronics Parts
7.5.22 Electronic CAD Programs
7.5.23 Building Your Own Workbench

CHAPTER 8 Operational Amplifiers

8.1 Operational Amplifier Water Analogy
8.2 How Op Amps Work (The “Cop-Out” Explanation)
8.3 Theory
8.4 Negative Feedback
8.5 Positive Feedback

8.6 Real Kinds of Op Amps
8.7 Op Amp Specifications
8.8 Powering Op Amps
8.9 Some Practical Notes
8.10 Voltage and Current Offset Compensation
8.11 Frequency Compensation
8.12 Comparators
8.13 Comparators with Hysteresis
8.13.1 Inverting Comparator with Hysteresis
8.13.2 Noninverting Comparator with Hysteresis
8.14 Using Single-Supply Comparators
8.15 Window Comparator
8.16 Voltage-Level Indicator
8.17 Applications

CHAPTER 9 Filters

9.1 Things to Know Before You Start Designing Filters
9.2 Basic Filters
9.3 Passive Low-Pass Filter Design
9.4 A Note on Filter Types
9.5 Passive High-Pass Filter Design
9.6 Passive Bandpass Filter Design
9.7 Passive Notch Filter Design
9.8 Active Filter Design
9.8.1 Active Low-Pass Filter Example
9.8.2 Active High-Pass Filter Example
9.8.3 Active Bandpass Filters
9.8.4 Active Notch Filters
9.9 Integrated Filter Circuits

CHAPTER 10 Oscillators and Timers

10.1 *RC* Relaxation Oscillators
10.2 The 555 Timer IC
10.2.1 How a 555 Works (Astable Operation)
10.2.2 Basic Astable Operation
10.2.3 How a 555 Works (Monostable Operation)
10.2.4 Basic Monostable Operation
10.2.5 Some Important Notes About 555 Timers
10.2.6 Simple 555 Applications
10.3 Voltage-Controlled Oscillators
10.4 Wien-Bridge and Twin-T Oscillators
10.5 *LC* Oscillators (Sinusoidal Oscillators)
10.6 Crystal Oscillators
10.7 Microcontroller Oscillators

CHAPTER 11 Voltage Regulators and Power Supplies

11.1 Voltage-Regulator ICs
11.1.1 Fixed Regulator ICs
11.1.2 Adjustable Regulator ICs
11.1.3 Regulator Specifications
11.2 A Quick Look at a Few Regulator Applications
11.3 The Transformer

11.4 Rectifier Packages
11.5 A Few Simple Power Supplies
11.6 Technical Points About Ripple Reduction
11.7 Loose Ends
11.8 Switching Regulator Supplies (Switchers)
11.9 Switch-Mode Power Supplies
11.10 Kinds of Commercial Power Supply Packages
11.11 Power Supply Construction

CHAPTER 12 Digital Electronics

12.1 The Basics of Digital Electronics
12.1.1 Digital Logic States
12.1.2 Number Codes Used in Digital Electronics
12.1.3 Clock Timing and Parallel Versus Serial Transmission
12.2 Logic Gates
12.2.1 Multiple-Input Logic Gates
12.2.2 Digital Logic Gate ICs
12.2.3 Applications for a Single Logic Gate
12.2.4 Combinational Logic
12.2.5 Keeping Circuits Simple (Karnaugh Maps)
12.3 Combinational Devices
12.3.1 Multiplexers (Data Selectors) and Bilateral Switches
12.3.2 Demultiplexers (Data Distributors) and Decoders
12.3.3 Encoders and Code Converters
12.3.4 Binary Adders
12.3.5 Binary Adder/Subtractor
12.3.6 Arithmetic Logic Units
12.3.7 Comparators and Magnitude Comparator ICs
12.3.8 Parity Generator/Checker
12.3.9 A Note on Obsolescence and the Trend Toward Microcontroller Control
12.4 Logic Families
12.4.1 TTL Family of ICs
12.4.2 CMOS Family of ICs
12.4.3 I/O Voltages and Noise Margins
12.4.4 Current Ratings, Fanout, and Propagation Delays
12.4.5 A Detailed Look at the TTL and CMOS Subfamilies
12.4.6 A Look at a Few Other Logic Series
12.4.7 Logic Gates with Open-Collector Outputs
12.4.8 Schmitt-Triggered Gates
12.4.9 Interfacing Logic Families
12.5 Powering and Testing Logic ICs
12.5.1 Power Supply Decoupling
12.5.2 Unused Inputs
12.5.3 Logic Probes and Logic Pulsers
12.6 Sequential Logic
12.6.1 SR Flip-Flops
12.6.2 SR Flip-Flop ICs
12.6.3 D-Type Flip-Flops
12.6.4 Quad and Octal D Flip-Flops
12.6.5 JK Flip-Flops
12.6.6 Practical Timing Considerations with Flip-Flops
12.6.7 Digital Clock Generators and Single-Pulse Generators
12.6.8 Automatic Power-Up Clear (Reset) Circuits

12.6.9 More on Switch Debouncers
12.6.10 Pullup and Pulldown Resistors
12.7 Counter ICs
12.7.1 Asynchronous Counter (Ripple Counter) ICs
12.7.2 Synchronous Counter ICs
12.7.3 A Note on Counters with Displays
12.8 Shift Registers
12.8.1 Serial-In/Serial-Out Shift Registers
12.8.2 Serial-In/Parallel-Out Shift Registers
12.8.3 Parallel-In/Serial-Out Shift Registers
12.8.4 Ring Counter (Shift Register Sequencer)
12.8.5 Johnson Shift Counter
12.8.6 Shift Register ICs
12.8.7 Simple Shift Register Applications
12.9 Three-State Buffers, Latches, and Transceivers
12.9.1 Three-State Octal Buffers
12.9.2 Three-State Octal Latches and Flip-Flops
12.9.3 Transceivers
12.10 Analog/Digital Interfacing
12.10.1 Triggering Simple Logic Responses from Analog Signals
12.10.2 Using Logic to Drive External Loads
12.10.3 Analog Switches
12.10.4 Analog Multiplexer/Demultiplexer
12.10.5 Analog-to-Digital and Digital-to-Analog Conversion
12.10.6 Analog-to-Digital Converters
12.11 Displays
12.11.1 LED Displays
12.11.2 Liquid-Crystal Displays
12.12 Memory Devices
12.12.1 Read-Only Memory
12.12.2 Simple ROM Made Using Diodes
12.12.3 Memory Size and Organization
12.12.4 Simple Programmable ROM
12.12.5 ROM Devices
12.12.6 RAM

CHAPTER 13 Microcontrollers

13.1 Basic Structure of a Microcontroller
13.2 Example Microcontrollers
13.2.1 The ATtiny85 Microcontroller
13.2.2 The PIC16Cx Microcontrollers
13.2.3 32-Bit Microcontrollers
13.2.4 Digital Signal Processing
13.3 Evaluation/Development Boards
13.4 Arduino
13.4.1 A Tour of Arduino
13.4.2 The Arduino IDE
13.4.3 Arduino Board Models
13.4.4 Shields
13.4.5 The Arduino C Library
13.4.6 Arduino Example Project
13.4.7 Taking the Arduino Offboard
13.5 Interfacing with Microcontrollers

13.5.1 Switches
13.5.2 Analog Inputs
13.5.3 High-Power Digital Outputs
13.5.4 Sound Interfaces
13.5.5 Serial Interfaces
13.5.6 Level Conversion
13.5.7 LED Display Interfaces

CHAPTER 14 Motors

14.1 DC Continuous Motors
14.2 Speed Control of DC Motors
14.3 Directional Control of DC Motors
14.4 RC Servos
14.5 Stepper Motors
14.6 Kinds of Stepper Motors
14.7 Driving Stepper Motors
14.8 Controlling the Driver with a Translator
14.9 A Final Word on Identifying Stepper Motors

CHAPTER 15 Audio Electronics

15.1 A Little Lecture on Sound
15.2 Microphones
15.3 Microphone Specifications
15.4 Audio Amplifiers
15.4.1 Inverting Amplifier
15.4.2 Noninverting Amplifier
15.4.3 Digital Amplifiers
15.4.4 Reducing Hum in Audio Amplifiers
15.5 Preamplifiers
15.6 Mixer Circuits
15.7 A Note on Impedance Matching
15.8 Speakers
15.9 Crossover Networks
15.10 Simple ICs Used to Drive Speakers
15.11 Audible-Signal Devices
15.12 Miscellaneous Audio Circuits

CHAPTER 16 Modular Electronics

16.1 There’s an IC for It
16.2 Breakout Boards and Modules
16.2.1 Radio Frequency Modules
16.2.2 Audio Modules
16.3 Plug-and-Play Prototyping
16.4 Open Source Hardware

APPENDIX A Power Distribution and Home Wiring

A.1 Power Distribution
A.2 A Closer Look at Three-Phase Electricity
A.3 Home Wiring
A.4 Electricity in Other Countries

APPENDIX B Error Analysis

B.1 Absolute Error, Relative Error, and Percent Error

B.2 Uncertainty Estimates

APPENDIX C Useful Facts and Formulas

C.1 Greek Alphabet

C.2 Powers of 10 Unit Prefixes

C.3 Linear Functions (y = mx + b)

C.4 Quadratic Equation (y = ax2 + bx + c)

C.5 Exponents and Logarithms

C.6 Trigonometry

C.7 Complex Numbers

C.8 Differential Calculus

C.9 Integral Calculus

Index

PREFACE

Inventors in the field of electronics are individuals who possess the knowledge, intuition, creativity, and technical know-how to turn their ideas into real-life electrical gadgets. It is my hope that this book will provide you with an intuitive understanding of the theoretical and practical aspects of electronics in a way that fuels your creativity.

This book is designed to help beginning inventors invent. It assumes little to no prior knowledge of electronics. Therefore, educators, students, and aspiring hobbyists will find this book a good initial text. At the same time, technicians and more advanced hobbyists may find this book a useful resource.

Notes About the Third Edition

The third edition of *Practical Electronics for Inventors* includes the following new chapters:

- Chapter 6, "Sensors," covers a wide range of discrete and IC sensors, such as temperature sensors, accelerometers, rotary encoders, and Geiger–Müller tubes.
- Chapter 13, "Microcontrollers," includes in-depth details of Atmel and Microchip microcontrollers, as well as the popular Arduino and BASIC Stamp prototyping platforms. There is also a comprehensive section on interfacing with microcontrollers and serial communications, including I^2C, SPI, and 1-Wire.
- Chapter 16, "Modular Electronics," covers the wide range of prebuilt electronic modules as useful prototyping tools. This chapter lists some of the most useful modules available, and also provides details of useful special-purpose ICs and an introduction to the plug-together approach of .NET Gadgeteer.

Throughout the book, there have been numerous changes, bringing the material up to date and including new topics such as GPS modules, digital amplification, LED Charlieplexing, and introductions to the use of modern software tools like EAGLE PCB Design and CircuitLab's online simulation software.

ACKNOWLEDGMENTS

We would like to thank the many people who have helped in the production of this book. Special thanks are due to the technical reviewers Michael Margolis, Chris Fitzer, and David Buckley.

We have been able to greatly improve the accuracy of the book thanks to the very detailed and helpful errata from the second edition that was collated by Martin Ligare at Bucknell University. Contributors to this errata were Steve Baker (Naval Postgraduate School), George Caplan (Wellesley College), Robert Drehmel, Earl Morris, Robert Strzelczyk (Motorola), Lloyd Lowe (Boise State University), John Kelty (University of Nebraska), Perry Spring (Cascadia Community College), Michael B. Allen, Jeffrey Audia, and Ken Ballinger (EIT). Thank you all for taking the time to make this a better book.

Many thanks to Roger Stewart, Patty Mon, and everyone from McGraw-Hill, for their support and skill in converting this manuscript into a great book.

—*Paul Scherz* and *Simon Monk*

CHAPTER 1

Introduction to Electronics

Perhaps the most common predicament newcomers face when learning electronics is figuring out exactly what it is they must learn. What topics are worth covering, and in which general order should they be covered? A good starting point for answering these questions is the flowchart presented in Fig. 1.1. This chart provides an overview of the basic elements that go into designing practical electrical gadgets and represents the information you will find in this book. This chapter introduces these basic elements.

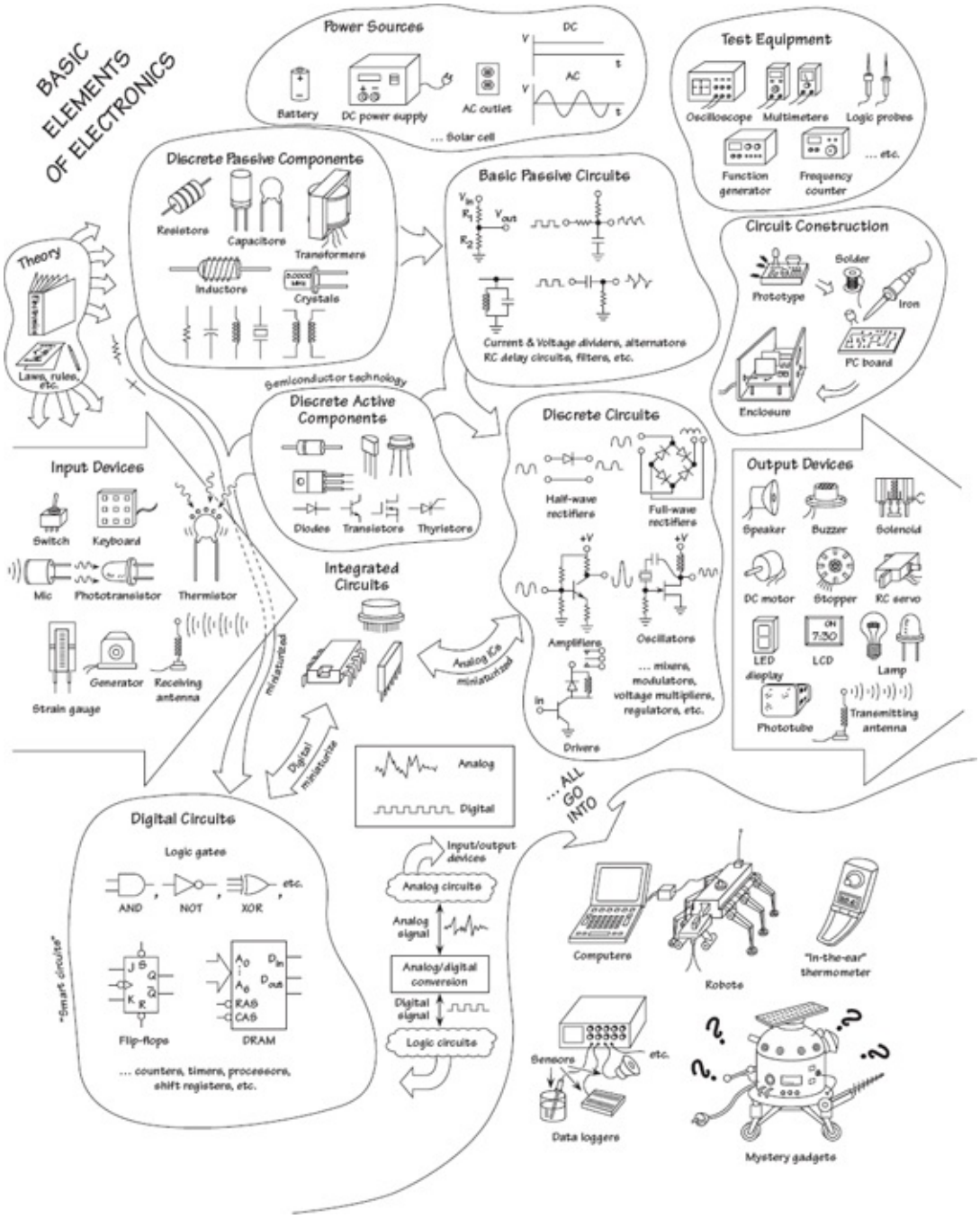


FIGURE 1.1

At the top of the chart comes the theory. This involves learning about voltage, current, resistance, capacitance, inductance, and various laws and theorems that help predict the size and direction of voltages and currents within circuits. As you learn the basic theory, you will be introduced to basic passive components such as resistors, capacitors, inductors, and transformers.

Next down the line are discrete passive circuits. Discrete passive circuits include current-limiting networks, voltage dividers, filter circuits, attenuators, and so on. These simple circuits, by themselves, are not very interesting, but they are vital ingredients in more complex circuits.

After you have learned about passive components and circuits, you move on to discrete active devices, which are built from semiconductor materials. These devices consist mainly of diodes (one-way current-flow gates) and transistors (electrically controlled switches/amplifiers).

Once you have covered the discrete active devices, you get to discrete active/passive circuits. Some of these circuits include rectifiers (ac-to-dc converters), amplifiers, oscillators, modulators, mixers, and voltage regulators. This is where things start getting interesting.

Throughout your study of electronics, you will learn about various input/output (I/O) devices (transducers). Input devices (sensors) convert physical signals, such as sound, light, and pressure, into electrical signals that circuits can use. These devices include microphones, phototransistors, switches, keyboards, thermistors, strain gauges, generators, and antennas. Output devices convert electrical signals into physical signals. Output devices include lamps, LED and LCD displays, speakers, buzzers, motors (dc, servo, and stepper), solenoids, and antennas. These I/O devices allow humans and circuits to communicate with one another.

To make things easier on the circuit designer, manufacturers have created integrated circuits (ICs), which contain discrete circuits (like the ones mentioned in the previous paragraph) that are crammed onto a tiny chip of silicon. The chip is usually housed within a plastic package, where little internal wires link the chip to external metal terminals. ICs such as amplifiers and voltage regulators are referred to as *analog devices*, which means that they respond to and produce signals of varying degrees of voltage. (This is unlike *digital* ICs, which work with only two voltage levels.) Becoming familiar with ICs is a necessity for any practical circuit designer.

Digital electronics comes next. Digital circuits work with only two voltage states: *high* (such as 5 V) or *low* (such as 0 V). The reason for having only two voltage states has to do with the ease of processing data (numbers, symbols, and control information) and storage. The process of encoding information into signals that digital circuits can use involves combining bits (1s and 0s, equivalent to *high* and *low* voltages) into discrete-meaning "words." The designer dictates what these words will mean to a specific circuit. Unlike analog electronics, digital electronics uses a whole new set of components, which at the heart are all integrated in form.

A huge number of specialized ICs are used in digital electronics. Some of these ICs are designed to perform logical operations on input information; others are designed to count; while still others are designed to store information that can be retrieved later on. Digital ICs include logic gates, flip-flops, shift registers, counters, memories, processors, and so on. Digital circuits are what give electrical gadgets "brains." In order for digital circuits to interact with analog circuits, special analog-to-digital (A/D) conversion circuits are needed to convert analog signals into strings of 1s and 0s. Likewise, digital-to-analog conversion circuits are used to convert strings of 1s and 0s into analog signals.

With an understanding of the principals behind digital electronics, we are free to explore the world of microcontrollers. These are programmable digital electronics that can read values from sensors and control output devices using the I/O pins, all on a single IC controlled by a little program.

And mixed in among all this is the practical side of electronics. This involves learning to read schematic diagrams, constructing circuit prototypes using breadboards, testing prototypes (using multimeters, oscilloscopes, and logic probes), revising prototypes (if needed), and constructing final

circuits using various tools and special circuit boards.

In the next chapter, we will start at the beginning by looking at the theory of electronics.

CHAPTER 2

Theory

2.1 Theory of Electronics

This chapter covers the basic concepts of electronics, such as current, voltage, resistance, electrical power, capacitance, and inductance. After going through these concepts, this chapter illustrates how to mathematically model currents and voltage through and across basic electrical elements such as resistors, capacitors, and inductors. By using some fundamental laws and theorems, such as Ohm's law, Kirchhoff's laws, and Thevenin's theorem, the chapter presents methods for analyzing complex networks containing resistors, capacitors, and inductors that are driven by a power source. The kinds of power sources used to drive these networks, as we will see, include direct current (dc) sources, alternating current (ac) sources (including sinusoidal and nonsinusoidal periodic sources), and nonsinusoidal nonperiodic sources. We will also discuss transient circuits, where sudden changes in state (such as flipping a switch within a circuit) are encountered. At the end of the chapter, the approach needed to analyze circuits that contain nonlinear elements (diodes, transistors, integrated circuits, etc.) is discussed.

I recommend using a circuit simulator program if you're just starting out in electronics. The web-based simulator CircuitLab (www.circuitlab.com) is extremely easy to use and has a nice graphical interface. There are also online calculators that can help you with many of the calculations in this chapter. Using a simulator program as you go through this chapter will help crystallize your knowledge, while providing an intuitive understanding of circuit behavior. Be careful—simulators can lie, or at least they can appear to lie when you don't understand all the necessary parameters the simulator needs to make a realistic simulation. It is always important to get your hands dirty—get out the breadboards, wires, resistors, power supplies, and so on, and construct. It is during this stage that you gain the greatest practical knowledge that is necessary for an inventor.

It is important to realize that components mentioned in this chapter are only "theoretically" explained. For example, in regard to capacitors, you'll learn how a capacitor works, what characteristic equations are used to describe a capacitor under certain conditions, and various other basic tricks related to predicting basic behavior. To get important practical insight into capacitors, however, such as real-life capacitor applications (filtering, snubbing, oscillator design, etc.), what type of real capacitors exist, how these real capacitors differ in terms of nonideal characteristics, which capacitors work best for a particular application, and, more important, how to read a capacitor label, requires that you jump to Chap. 3, Section 3.6, which is dedicated to these issues. This applies to other components mentioned in this theory portion of the book.

The theoretical and practical information regarding transformers and nonlinear devices, such as diodes, transistors, and analog and digital integrated circuits (ICs), is not treated within this chapter. Transformers are discussed in full in Chap. 3, Section 3.8, while the various nonlinear devices are treated separately in the remaining chapters of this book.

A word of advice: if the math in a particular section of this chapter starts looking scary, don't worry. As it turns out, most of the nasty math in this chapter is used to prove, say, a theorem or law or to give you an idea of how hard things can get if you do not use some mathematical tricks. The actual amount of math

you will need to know to design most circuits is surprisingly small; in fact, basic algebra may be all you need to know. Therefore, when the math in a particular section in this chapter starts looking ugly, skim through the section until you locate the useful, nonugly formulas, rules, and so on, that do not have weird mathematical expressions in them. You don't have to be a mathematical whiz to be able to design decent circuits.

2.2 Electric Current

Electric current is the total charge that passes through some cross-sectional area A per unit time. This cross-sectional area could represent a disk placed in a gas, plasma, or liquid, but in electronics, this cross-sectional area is most frequently a slice through a solid material, such as a conductor.

If ΔQ is the amount of charge passing through an area in a time interval Δt, then the *average current* I_{ave} is defined as:

$$I_{ave} = \frac{\Delta Q}{\Delta t}$$

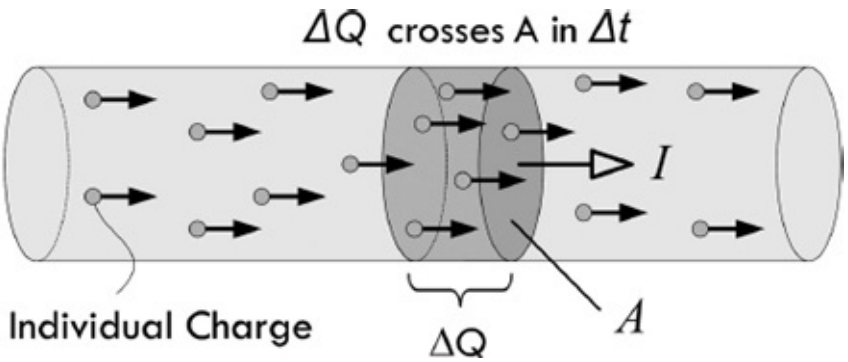


FIGURE 2.1

If the current changes with time, we define the *instantaneous current* I by taking the limit as $\Delta t \rightarrow 0$, so that the current is the instantaneous rate at which charge passes through an area:

$$I = \lim_{\Delta t \to 0} \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt} \tag{2.1}$$

The unit of current is coulombs per second, but this unit is also called the *ampere* (A), named after Andre-Marie Ampere:

$$1 \text{ A} = 1 \text{ C/s}$$

To sound less nerdy, the term *amp* can be used in place of ampere. Because the ampere is a rather large unit, current is also expressed in *milliamps* ($1 \text{ mA} = 1 \times 10^{-3}$ A), *microamps* ($1 \mu\text{A} = 1 \times 10^{-6}$ A), and *nanoamps* ($1 \text{ nA} = 1 \times 10^{-9}$ A).

Within conductors such as copper, electrical current is made up of free electrons moving through a lattice of copper ions. Copper has one free electron per copper atom. The charge on a single electron is given by:

$$Q_{electron} = (-e) = -1.602 \times 10^{-19} \text{ C} \tag{2.2.a}$$

This is equal to, but opposite in sign of, the charge of a single copper ion. (The positive charge is a result of the atom donating one electron to the "sea" of free electrons randomly moving about the lattice. The

loss of the electron means there is one more proton per atom than electrons.) The charge of a proton is:

$$Q_{\text{proton}} = (+e) = +1.602 \times 10^{-19}\ \text{C} \tag{2.2.b}$$

The conductor, as a whole, is neutral, since there are equal numbers of electrons and protons. Using Eq. 2.2, we see that if a current of 1 A flows through a copper wire, the number of electrons flowing by a cross section of the wire in 1 s is equal to:

$$1\ \text{A} = \left(\frac{1\ \text{C}}{1\ \text{s}}\right)\left(\frac{\text{electron}}{-1.602 \times 10^{-19}\ \text{C}}\right) = -6.24 \times 10^{18}\ \text{electrons/s}$$

Now, there is a problem! How do we get a negative number of electrons flowing per second, as our result indicates? The only two possibilities for this would be to say that either electrons must be flowing in the opposite direction as the defined current, or positive charges must be moving in our wire instead of electrons to account for the sign. The last choice is an incorrect one, since experimental evidence exists to prove electrons are free to move, not positive charges, which are fixed in the lattice network of the conductor. (Note, however, there are media in which positive charge flow is possible, such as positive ion flow in liquids, gases, and plasmas.) It turns out that the first choice—namely, electrons flowing in the opposite direction as the defined current flow—is the correct answer.

Long ago, when Benjamin Franklin (often considered the father of electronics) was doing his pioneering work in early electronics, he had a convention of assigning positive charge signs to the mysterious (at that time) things that were moving and doing work. Sometime later, a physicist by the name of Joseph Thomson performed an experiment that isolated the mysterious moving charges. However, to measure and record his experiments, as well as to do his calculations, Thomson had to stick with using the only laws available to him—those formulated using Franklin's positive currents. But these moving charges that Thomson found (which he called electrons) were moving in the opposite direction of the conventional current *I* used in the equations, or moving against convention. See Fig. 2.2.

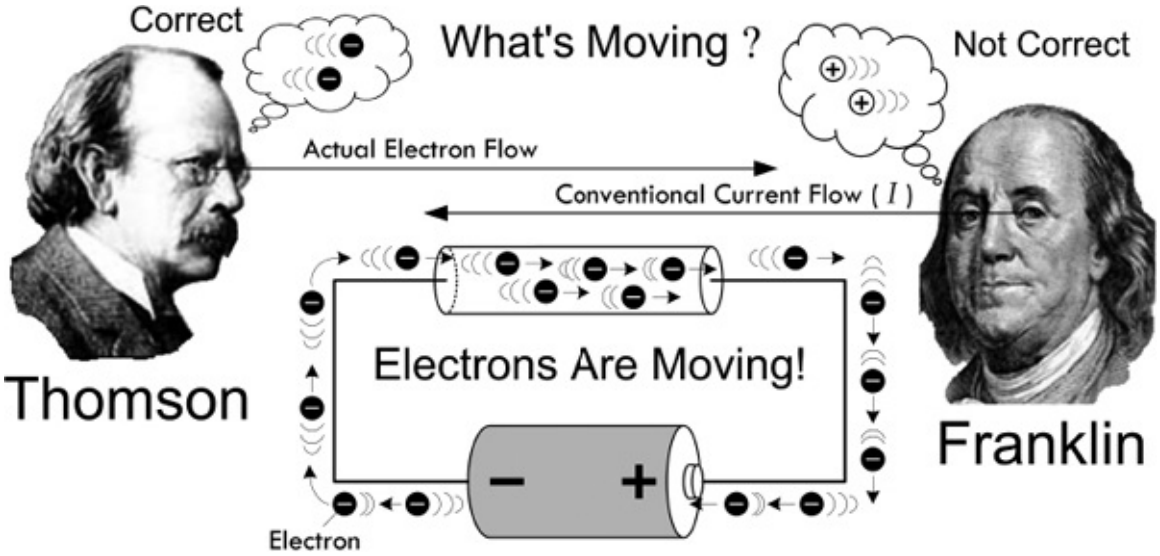
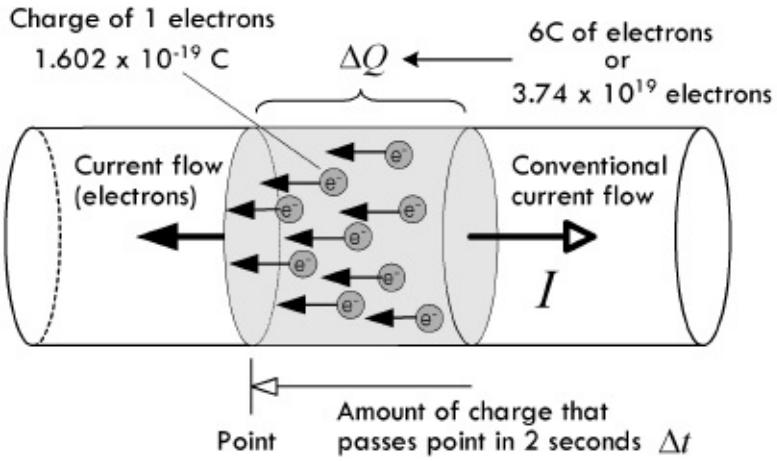


FIGURE 2.2 Thomson changed the notion that positive charges were what were moving in conductors, contrary to Franklin's notion. However, negative electrons going one way is equivalent to positive charges going the opposite direction, so the old formulas still work. Since you deal with the old formulas, it's practical to adopt Franklin's conventional current—though realize that what's actually moving in conductors is electrons.

What does this mean to us, to those of us not so interested in the detailed physics and such? Well, not too much. We could pretend that there were positive charges moving in the wires and various electrical

devices, and everything would work out fine: negative electrons going one way are equivalent to positive charges going in the opposite direction. In fact, all the formulas used in electronics, such as Ohm's law ($V = IR$), "pretend" that the current I is made up of positive charge carriers. We will always be stuck with this convention. In a nutshell, it's convenient to pretend that positive charges are moving. So when you see the term *electron flow*, make sure you realize that the conventional current flow I is moving in the opposite direction. In a minute, we'll discuss the microscopic goings-on within a conductor that will clarify things a bit better.

Example 1: How many electrons pass a given point in 3 s if a conductor is carrying a 2-A current?



$$I_{ave} = \frac{\Delta Q}{\Delta t} = \frac{6\text{C}}{2\text{s}} = 3\text{A}$$

FIGURE 2.3

Answer: The charge that passes a given point in 3 s is:

$$\Delta Q = I \times \Delta t = (2\text{ A})(3\text{ s}) = 6\text{ C}$$

One electron has a charge of 1.6×10^{-19} C, so 6 C worth of electrons is:

$$\#\text{ Electrons} = 6\text{ C}/1.602 \times 10^{-19}\text{ C} = 3.74 \times 10^{19}$$

Example 2: Charge is changing in a circuit with time according to $Q(t) = (0.001\text{ C}) \sin[(1000/\text{s})\,t]$. Calculate the instantaneous current flow.

$$I = \frac{dQ}{dt} = \frac{d}{dt}[(0.001\text{C})\sin(1000/\text{s}\cdot t)] = (0.001\text{C})(1000/\text{s})\cos(1000/\text{s}\cdot t)$$
$$= (1\text{A})\cos(1000/\text{s}\cdot t)$$

Answer: If we plug in a specific time within this equation, we get an instantaneous current for that time. For example, if $t = 1$, the current would be 0.174 A. At $t = 3$ s, the current would be − 0.5 A, the negative sign indicating that the current is in the opposite direction—a result of the sinusoidal nature.

Note: The last example involved using calculus—you can read about the basics of calculus in App. C if you're unfamiliar with it. Fortunately, as we'll see, rarely do you actually need to work in units of charge when doing electronics. Usually you worry only about current, which can be directly measured using an ammeter, or calculated by applying formulas that usually require no calculus whatsoever.

2.2.1 Currents in Perspective

What's considered a lot or a little amount of current? It's a good idea to have a gauge of comparison when you start tinkering with electronic devices. Here are some examples: a 100-W lightbulb draws about 1 A; a microwave draws 8 to 13 A; a laptop computer, 2 to 3 A; an electric fan, 1 A; a television, 1 to 3 A; a toaster, 7 to 10 A; a fluorescent light, 1 to 2 A; a radio/stereo, 1 to 4 A; a typical LED, 20 mA; a mobile (smart) phone accessing the web uses around 200 mA; an advanced low-power microchip (individual), a few μA to perhaps even several pA; an automobile starter, around 200 A; a lightning strike, around 1000 A; a sufficient amount of current to induce cardiac/respiratory arrest, around 100 mA to 1 A.

2.3 Voltage

To get electrical current to flow from one point to another, a voltage must exist between the two points. A voltage placed across a conductor gives rise to an *electromotive force* (EMF) that is responsible for giving all *free electrons* within the conductor a push.

As a technical note, before we begin, voltage is also referred to as a *potential difference* or just *potential*—they all mean the same thing. I'll avoid using these terms, however, because it is easy to confuse them with the term *potential energy*, which is not the same thing.

A simple flashlight circuit, consisting of a battery connected to a lamp, through two conductors and a switch, is shown in Fig. 2.4. When the switch is open ("off"), no current will flow. The moment the switch is closed, however, the resistance of the switch falls to almost zero, and current will flow. This voltage then drives all free electrons, everywhere within the circuit, in a direction that points from negative to positive; conventional current flow, of course, points in the opposite direction (see Benjamin Franklin).

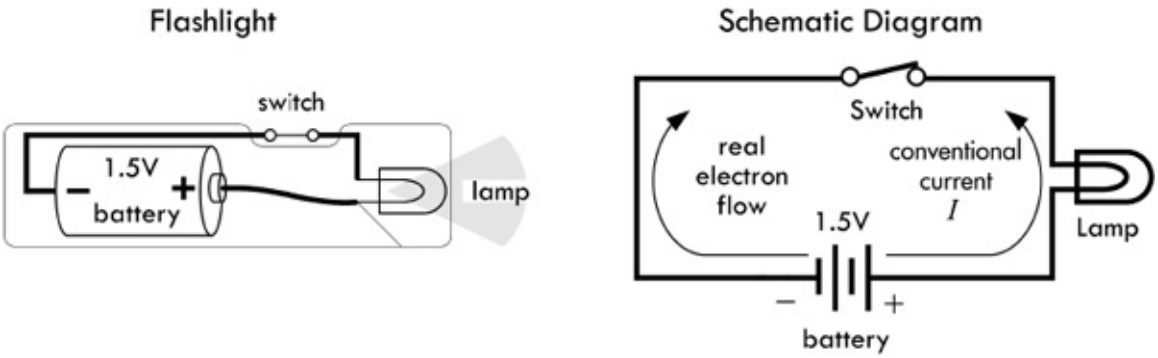


FIGURE 2.4

It is important to note that the battery needs the rest of the circuit, just as the rest of the circuit needs the battery. Without the linkage between its terminals, the chemical reactions within the battery cannot be carried out. These chemical reactions involve the transfer of electrons, which by intended design can only occur through a link between the battery's terminals (e.g., where the circuit goes). Figure 2.5 shows this process using an alkaline dry cell battery. Notice that the flow of current is conserved through the circuit, even though the nature of the current throughout the circuit varies—ionic current within sections of the battery, electron current elsewhere.

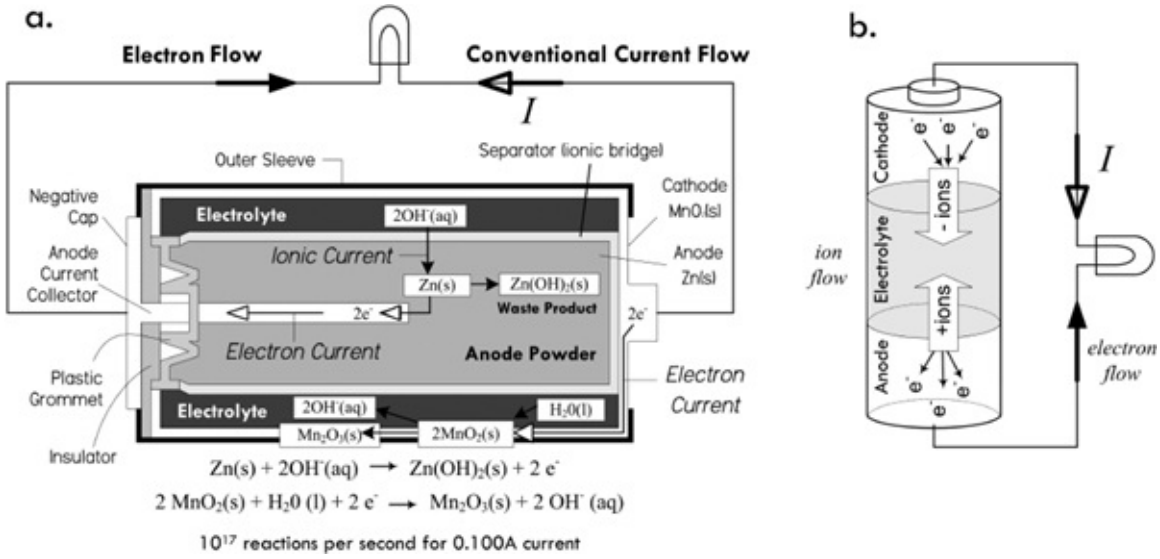


FIGURE 2.5

As free electrons within the lamp filament experience an EMF due to the applied voltage, the extra energy they gain is transferred to the filament lattice atoms, which result in heat (filament atomic vibrations) and emitted light (when a valence electron of a lattice atom is excited by a free electron and the bound electron returns to a lower energy configuration, thus releasing a photon).

A device that maintains a constant voltage across it terminals is called a *direct current voltage source* (or dc voltage source). A battery is an example of a dc voltage source. The schematic symbol for a battery is ⊣⁞⊢.

2.3.1 The Mechanisms of Voltage

To get a mental image of how a battery generates an EMF through a circuit, we envision that chemical reactions inside yield free electrons that quickly build in number within the negative terminal region (anode material), causing an electron concentration. This concentration is full of repulsive force (electrons repel) that can be viewed as a kind of “electrical pressure.” With a load (e.g., our flashlight lamp, conductors, switch) placed between the battery’s terminals, electrons from the battery’s negative terminal attempt to alleviate this pressure by dispersing into the circuit. These electrons increase the concentration of free electrons within the end of the conductor attached to the negative terminal. Even a small percentage difference in free electron concentration in one region gives rise to great repulsive forces between free electrons. The repulsive force is expressed as a seemingly instantaneous (close to the speed of light) pulse that travels throughout the circuit. Those free electrons nearest to the pumped-in electrons are quickly repulsed in the opposite direction; the next neighboring electrons get shoved, and so on down the line, causing a chain reaction, or pulse. This pulse travels down the conductor near the speed of light. See Fig. 2.6.

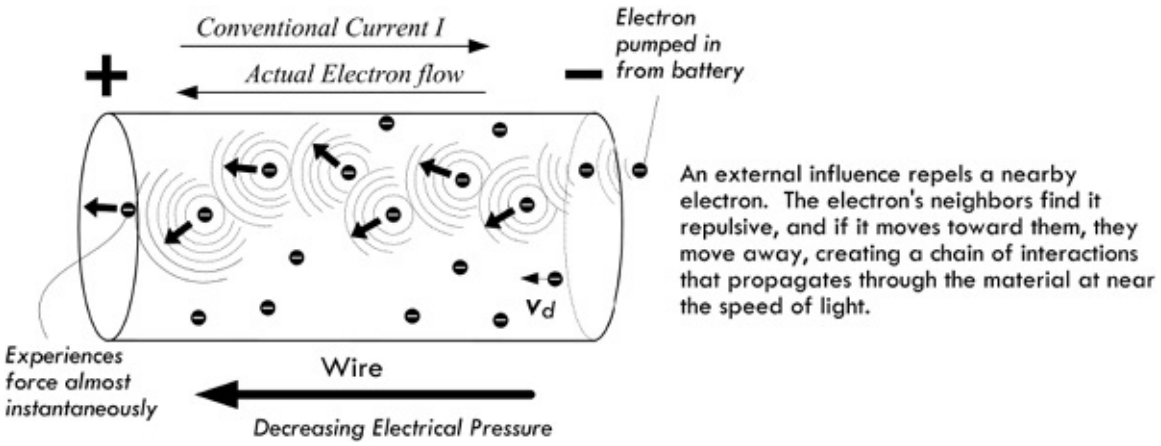


FIGURE 2.6

The actual physical movement of electrons is, on average, much slower. In fact, the drift velocity (average net velocity of electrons toward the positive terminal) is usually fractions of a millimeter per second—say, 0.002 mm/s for a 0.1-A current through a 12-gauge wire. We associate this drift movement of free electrons with current flow or, more precisely, conventional current flow *I* moving in the opposite direction. (As it turns out, the actual motion of electrons is quite complex, involving thermal effects, too—we'll go over this in the next section.)

It is likely that those electrons farther "down in" the circuit will not feel the same level of repulsive force, since there may be quite a bit of material in the way which absorbs some of the repulsive energy flow emanating from the negative terminal (absorbing via electron-electron collisions, free electron–bond electron interactions, etc.). And, as you probably know, circuits can contain large numbers of components, some of which are buried deep within a complex network of pathways. It is possible to imagine that through some of these pathways the repulsive effects are reduced to a weak nudge. We associate these regions of "weak nudge" with regions of low "electrical pressure," or voltage. Electrons in these regions have little potential to do work—they have low potential energy relative to those closer to the source of pumped-in electrons.

Voltage represents the difference in potential energy a unit charge has being at one location relative to another within a region of "electrical pressure"—the pressure attributed to new free electrons being pumped into the system. The relationship between the voltage and the difference in potential energy is expressed as:

$$V_{AB} = \frac{U_{AB}}{q} \text{ or } V_B - V_A = \frac{U_B - U_A}{q} \text{ or } \Delta V = \frac{\Delta U}{q}$$

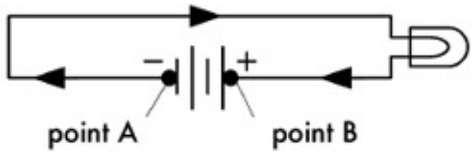


FIGURE 2.7

Implicit in the definition of *voltage* is the notion that voltage is always a measurement between two points, say point A and point B. That is the reason for the subscript "AB" in V_{AB}. The symbol ΔV means the same. Both infer that there is an absolute scale on which to measure and give individual points a specific voltage value. In electronics, we can create such a scale by picking a point, often the point where there is the lowest electrical pressure, and defining this point as the zero point, or 0-V reference. In many dc circuits, people choose the negative terminal of the battery as the 0-V reference, and let everyone know by inserting a ground symbol ⏚ (more on this later). In practice, you rarely see voltages expressed using subscripts (V_{AB}) or deltas (ΔV), but

instead you simply see the symbol V, or you may see a symbol like V_R. The "blank symbol" V, however, is always modified with a phrase stating the two points across which the voltage is present. In the second case, V_R, the subscript means that the voltage is measured across the component R—in this case, a resistor. In light of this, we can write a cleaner expression for the voltage/potential energy expression:

$$V = \frac{U}{q}$$

Just make sure you remember that the voltage and potential energy variables represent the difference in relation to two points. As we'll discover, all the big electronics laws usually assume that variables of voltage or energy are of this "clean form."

In our flashlight example, we can calculate the difference in potential energy between an electron emanating from the negative terminal of the 1.5-V battery and one entering the positive terminal.

$$\Delta U = \Delta V q = (1.5\text{ V})(1.602 \times 10^{-19}\text{ C}) = 2.4 \times 10^{-19}\text{ J}$$

Notice that this result gives us the potential energy difference between the two electrons, not the actual potential energy of either the electron emanating from the negative terminal (U_1) or the electron entering the positive terminal (U_0). However, if we make the assumption that the electron entering the positive terminal is at zero potential energy, we can figure that the electron emanating from the negative terminal has a relative potential energy of:

$$U_1 = \Delta U + U_0 = \Delta U + 0 = 2.4 \times 10^{-19}\text{ J}$$

Note: Increasing positive potential energy can be associated with similar charges getting closer together. Decreasing energy can be associated with similar charges getting farther apart. We avoided the use of a negative sign in front of the charge of the electron, because voltages are defined by a positive test charge. We are in a pickle similar to the one we saw with Benjamin Franklin's positive charges. As long as we treat the potential relative to the pumped-in electron concentration, things work out.

In a real circuit, where the number of electrons pumped out by the battery will be quite large—hundreds to thousands of trillions of electrons, depending on the resistance to electron flow—we must multiply our previous calculation by the total number of entering electrons. For example, if our flashlight draws 0.1 A, there will be 6.24×10^{17} electrons pumped into it by the battery per second, so you calculate the potential energy of all the new electrons together to be about 0.15 J/s.

What about the potential energies of free electrons at other locations throughout the circuit, such as those found in the lamp filament, those in the positive wire, those in the negative wire, and so on? We can say that somewhere in the filament of the lamp, there is an electron that has half the potential energy of a fresh pumped-in electron emanating from the negative terminal of the battery. We attribute this lower energy to the fact that other free electrons up the line have lost energy due to collision mechanisms, which in turn yields a weaker electrical repulsive pressure (shoving action) that our electron in question experiences. In fact, in our flashlight circuit, we attribute all loss in electrical pressure to be through the lamp filament as free-electron energy is converted into heat and light.

In regard to potential energies of free electrons within the conductors leading to and from the battery, we assume all electrons within the same conductor have the same potential energy. This assumes that

there is no voltage difference between points in the same conductor. For example, if you take a voltmeter and place it between any two points of a single conductor, it will measure 0 V. (See Fig. 2.8.) For practical purposes, we accept this as true. However, in reality it isn't. There is a slight voltage drop through a conductor, and if we had a voltmeter that was extremely accurate we might measure a voltage drop of 0.00001 V or so, depending on the length of the conductor, current flow, and conductor material type. This is attributed to internal resistance within conductors—a topic we'll cover in a moment.

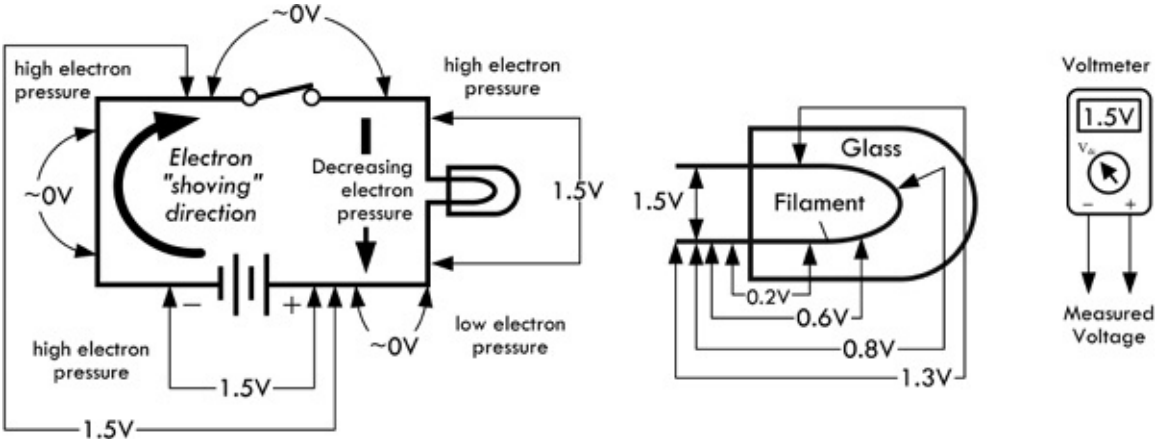


FIGURE 2.8

2.3.2 Definition of Volt and Generalized Power Law

We come now to a formal definition of the volt—the unit of measure of voltage. Using the relationship between voltage and potential energy difference $V = U/q$, we define a volt to be:

$$1 \text{ volt} = \frac{1 \text{ joule}}{1 \text{ coulomb}}, \quad 1 \text{ V} \frac{1 \text{ J}}{1 \text{ C}} = \text{J/C (Energy definition)}$$

(Be aware that the use of "V" for both an algebraic quantity and a unit of voltage is a potential source of confusion in an expression like $V = 1.5$ V. The algebraic quantity is in italic.)

Two points with a voltage of 1 V between them have enough "pressure" to perform 1 J worth of work while moving 1 C worth of charge between the points. For example, an ideal 1.5-V battery is capable of moving 1 C of charge through a circuit while performing 1.5 J worth of work.

Another way to define a volt is in terms of power, which happens to be more useful in electronics. *Power* represents how much energy per second goes into powering a circuit. According to the conservation of energy, we can say the power used to drive a circuit must equal the power used by the circuit to do useful work plus the power wasted, as in the case of heat. Assuming that a single electron loses all its potential energy from going through a circuit from negative to positive terminal, we say, for the sake of argument, that all this energy must have been converted to work—useful and wasted (heat). By definition, power is mathematically expressed as dW/dt. If we substitute the potential energy expression $U = Vq$ for W, assuming the voltage is constant (e.g., battery voltage), we get the following:

$$P = \frac{dW}{dt} = \frac{dU}{dt} = V\frac{dq}{dt}$$

Since we know that current $I = dq/dt$, we can substitute this into the preceding expression to get:

$$P = VI \tag{2.3}$$

This is referred to as the *generalized power law*. This law is incredibly powerful, and it provides a general result, one that is independent of type of material and of the nature of the charge movement. The unit of this electrical power is watts (W), with 1 W = 1 J/s, or in terms of volts and amps, 1 W = 1 VA.

In terms of power, then, the volt is defined as:

$$1\ \text{volt} = \frac{1\ \text{watt}}{1\ \text{amp}},\quad 1\ \text{V} = \frac{1\ \text{W}}{1\ \text{A}} = \text{W/A}$$

The generalized power law can be used to determine the power loss of any circuit, given only the voltage applied across it and the current drawn, both of which can easily be measured using a voltmeter and an ammeter. However, it doesn't tell you specifically how this power is used up—more on this when we get to resistance. See Fig. 2.9.

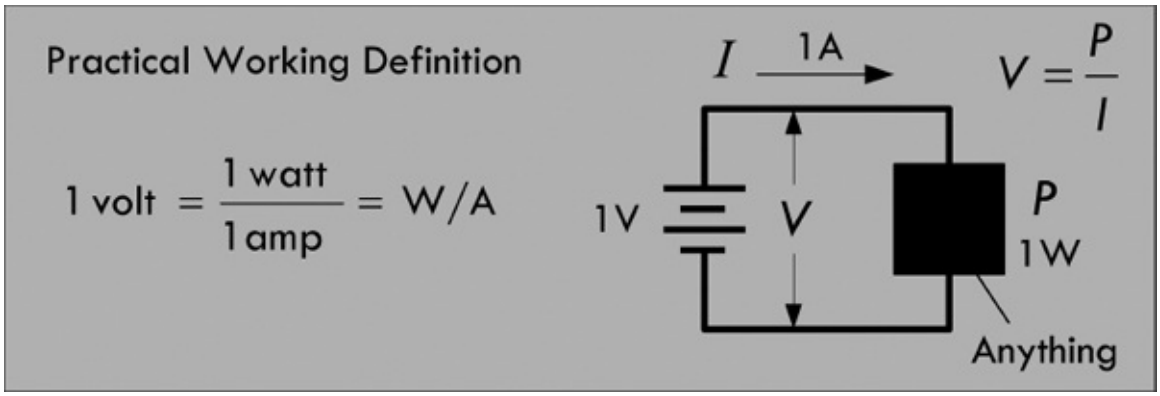


FIGURE 2.9

Example 1: Our 1.5-V flashlight circuit draws 0.1 A. How much power does the circuit consume?

Answer:

$$P = VI = (1.5\ \text{V})(0.1\ \text{A}) = 0.15\ \text{W}$$

Example 2: A 12-V electrical device is specified as consuming 100 W of power. How much current does it draw?

Answer:

$$I = \frac{P}{V} = \frac{100\ \text{W}}{12\ \text{V}} = 8.3\ \text{A}$$

2.3.3 Combining Batteries

To get a larger voltage capable of supplying more power, we can place two batteries in series (end to end), as shown in Fig. 2.10. The voltage across the combination is equal to the individual battery voltages added together. In essence, we have placed two charge pumps in series, increasing the effective electrical pressure. Chemically speaking, if the batteries are of the same voltage, we double the number of chemical reactions, doubling the number of electrons that can be pumped out into the circuit.

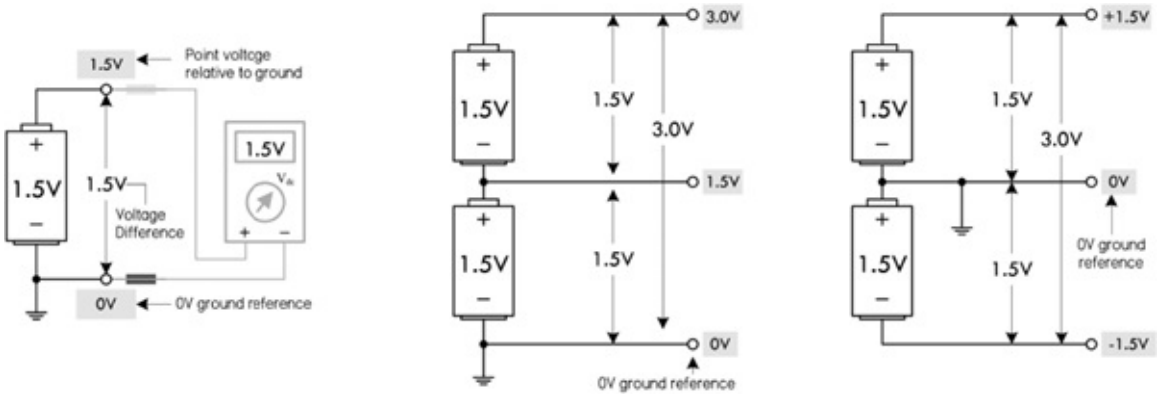


FIGURE 2.10

In Fig. 2.10, we use the notion of a *ground reference*, or 0-V reference, symbolized ⏚. Though this symbol is used to represent an earth ground (which we define a bit later), it can also be used to indicate the point where all voltage measurements are to be referenced within a circuit. Logically, whenever you create a scale of measure, you pick the lowest point in the scale to be zero—0 V here. For most dc circuits, the ground reference point is usually placed at the negative terminal of the voltage source. With the notion of ground reference point, we also get the notion of a *point voltage*, which is the voltage measured between the ground reference and a specific point of interest within the circuit. For example, the single battery shown in Fig. 2.10 has a voltage of 1.5 V. We place a ground reference at the negative terminal and give this a 0-V point voltage, and place a 1.5-V point voltage marker at the positive terminal.

In the center of Fig. 2.10, we have two 1.5-V batteries in series, giving a combined voltage of 3.0 V. A ground placed at the negative terminal of the lower battery gives us point voltages of 1.5 V between the batteries, and 3.0 V at the positive terminal of the top battery. A load placed between ground and 3.0 V will result in a load current that returns to the lower battery's negative terminal.

Finally, it is possible to create a split supply by simply repositioning the 0-V ground reference, placing it between the batteries. This creates +1.5 V and −1.5 V leads relative to the 0-V reference. Many circuits require both positive and negative voltage relative to a 0-V ground reference. In this case, the 0-V ground reference acts as a *common return*. This is often necessary, say, in an audio circuit, where signals are sinusoidal and alternate between positive and negative voltage relative to a 0-V reference.

2.3.4 Other Voltage Sources

There are other mechanisms besides the chemical reactions within batteries that give rise to an electromotive force that pushes electrons through circuits. Some examples include magnetic induction, photovoltaic action, thermoelectric effect, piezoelectric effect, and static electric effect. Magnetic induction (used in electrical generators) and photovoltaic action (used in photocells), along with chemical reactions, are, however, the only mechanisms of those listed that provide enough power to drive most circuits. The thermoelectric and piezoelectric effects are usually so small (mV range, typically) that they are limited to sensor-type applications. Static electric effect is based on giving objects, such as conductors and insulators, a surplus of charge. Though voltages can become very high between charged objects, if a circuit were connected between the objects, a dangerous discharge of current could flow, possibly damaging sensitive circuits. Also, once the discharge is complete—a matter of milliseconds—there is no more current to power the circuit. Static electricity is considered a nuisance in electrons, not a source of useful power. We'll discuss all these different mechanisms in more detail throughout the book.

2.3.5 Water Analogies

It is often helpful to use a water analogy to explain voltage. In Fig. 2.11, we treat a dc voltage source as a water pump, wires like pipes, Benjamin Franklin's positive charges as water, and conventional current flow like water flow. A load (resistor) is treated as a network of stationary force-absorbing particles that limit water flow. I'll leave it to you to compare the similarities and differences.

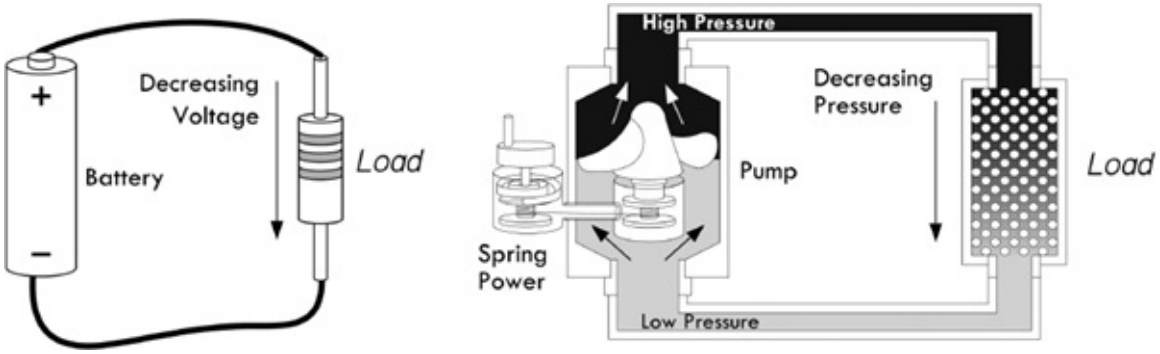


FIGURE 2.11

Here's another water analogy that relies on gravity to provide the pressure. Though this analogy falls short of being accurate in many regards, it at least demonstrates how a larger voltage (greater water pressure) can result in greater current flow.

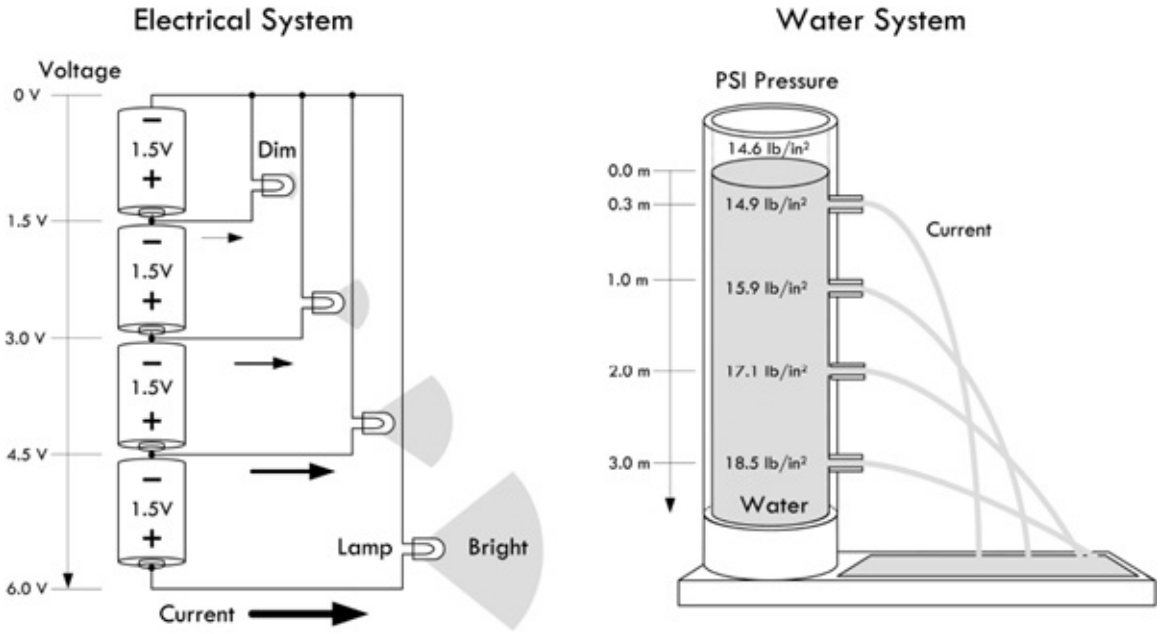


FIGURE 2.12

It's not wise to focus too much attention on these water analogies—they fall short of being truly analogous to electric circuits. Take them with a grain of salt. The next section will prove how true this is.

Example 1: Find the voltage between the various points indicated in the following figures. For example, the voltage between points A and B in Fig. 2.13a is 12 V.

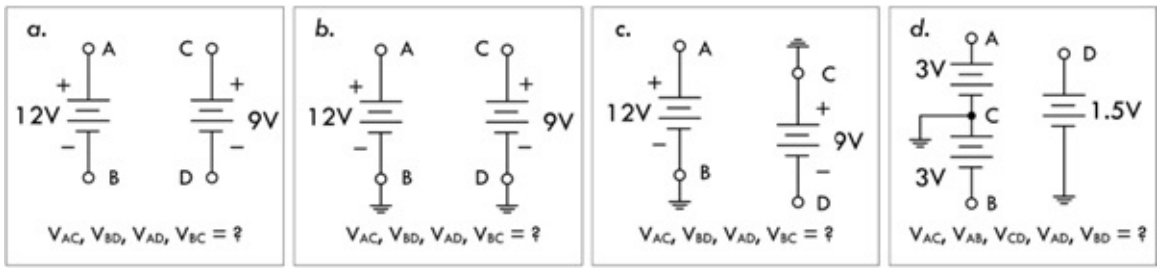


FIGURE 2.13

Answer: a. $V_{AC} = 0$, $V_{BD} = 0$, $V_{AD} = 0$, $V_{BC} = 0$. b. $V_{AC} = 3$ V, $V_{BD} = 0$ V, $V_{AD} = 12$ V, $V_{BC} = 9$ V. c. $V_{AC} = 12$ V, $V_{BD} = -9$V. $V_{AD} = -21$V, $V_{BC} = 0$ V. d. $V_{AC} = 3$ V, $V_{AB} = 6$ V, $V_{CD} = 1.5$ V, $V_{AD} = 1.5$ V, $V_{BD} = 4.5$ V.

Example 2: Find the point voltages (referenced to ground) at the various locations indicated in the following figures.

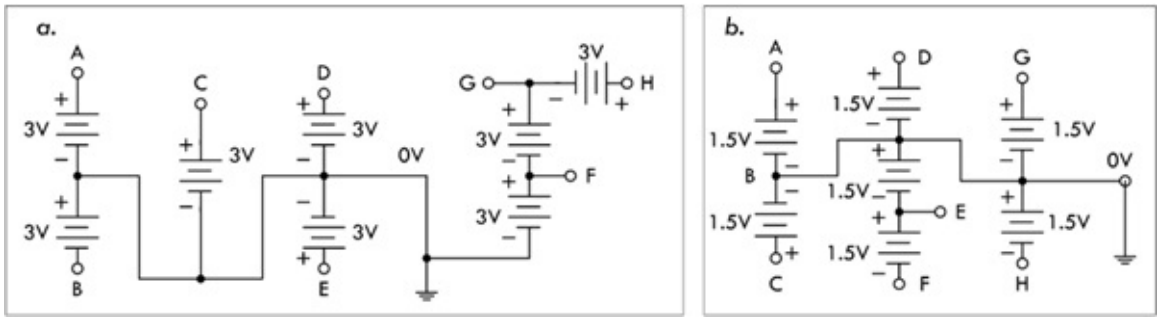


FIGURE 2.14

Answer: a. A = 3 V, B = −3 V, C = 3 V, D = 3 V, E = 3 V, F = 3 V, G = 6 V, H = 9 V. b. A = 1.5 V, B = 0 V, C = 1.5 V, D = 1.5 V, E = −1.5 V, F = −3.0 V, G = 1.5 V, H = −1.5 V.

2.4 A Microscopic View of Conduction (For Those Who Are Interested)

At a microscopic level, a copper conductor resembles a lattice of copper balls packed together in what's called a face-centered-cubic lattice structure, as shown in Fig. 2.15. For copper, as well as other metals, the bonding mechanism that holds everything together is referred to as *metallic bonding*, where outermost valence electrons from the metal atoms form a "cloud of free electrons" which fill the space between the metal ions (positively charged atoms missing an electron that became "free"—see the planetary model in Fig. 2.15b). This cloud of free electrons acts as a glue, holding the lattice metal ions together.

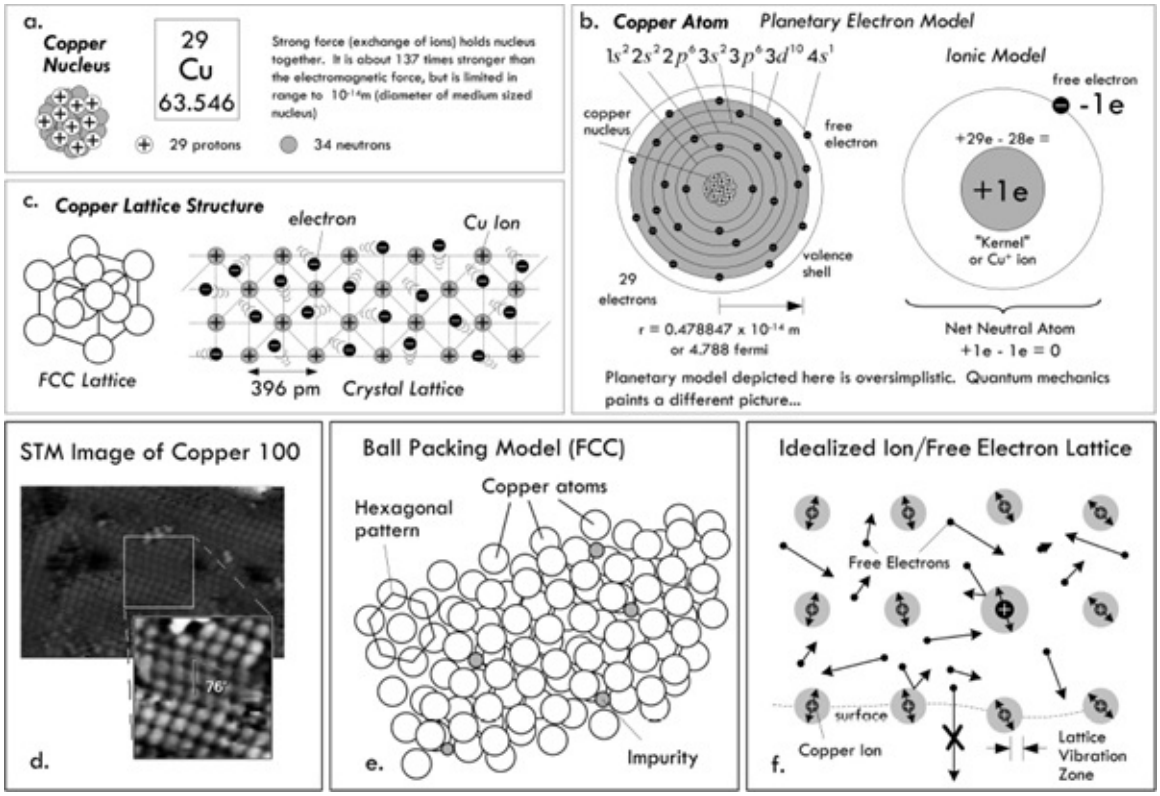


FIGURE 2.15 (a) Copper nucleus composed of protons and neutrons held together by nuclear forces that is roughly 137 times stronger than the electromagnetic force. (b) Copper atoms, as viewed by the classic planetary model, consisting of valance electrons held in orbit by electric forces. Quantum mechanics is required to explain why electrons exist in discrete energy levels, and why they don't fall into the nucleus or radiate electromagnetic energy as they orbit. (c) Copper lattice has a face-centered cubic packing arrangement. (d) Scanning tunneling electron microscope (STM) image of copper 100, courtesy of Institut für Allgemeine Physik, TU Wien. (e) Ball packing model of lattice, showing irregularities in lattice geometry, partly caused by impurities (other kinds of atoms). (f) Lattice view showing that lattice atoms vibrate due to external thermal interactions as well as interactions with free electrons. Free electrons move about randomly, at varying speeds and directions, colliding with other electrons and lattice ions. Under normal conditions, they do not leave the surface of the metal.

Each free electron within the cloud of free electrons moves about in random directions and speeds, colliding and rebounding "off" metal ions and other imperfections (impurities in lattice and grain boundary transitions, etc.). It is important to realize that this is occurring in a chunk of copper, at room temperature, without any applied voltage.

At room temperature, no free electrons ever leave the surface of the metal. A free electron cannot escape the coulomb (electric) attractive forces presented by the positive metal ions in the lattice. (We'll see later that under special conditions, using unique mechanisms, it is possible for electrons to escape.)

According to what's called the *free-electron model*—a classical model that treats free electrons as a gas of noninteracting charges—there is approximately one free electron per copper atom, giving a copper conductor a free electron concentration of $\rho_n = 8.5 \times 10^{28}$ electrons/m^3. This model predicts that, under normal conditions (a piece of copper just sitting there at room temperature), the *thermal velocity* v of electrons (or root-mean-square speed) within copper is about 120 km/s (1.2×10^5 m/s), but depends on temperature. The average distance an electron travels before it collides with something, called the *mean free path* λ, is about 0.000003 mm (2.9×10^{-9} m), with the average time between collisions τ of roughly 0.000000000000024 s (2.4×10^{-14} s). The free-electron model is qualitatively correct in many respects, but isn't as accurate as models based on quantum mechanics. (The speed, path, and time are related by $v = \lambda/\tau$.)

In quantum mechanics, electrons obey velocity-distribution laws based on quantum physics, and the movement of electrons depends on these quantum ideas. It requires that we treat electrons as though they were waves scattering from the lattice structure of the copper. The quantum view shows the thermal speed (now called Fermi velocity v_F) of a free electron to be faster than that predicted by the free-electron

model, now around 1.57×10^6 m/s, and contrarily, it is essentially independent of temperature. In addition, the quantum model predicts a larger mean free path, now around 3.9×10^{-8} m, which is independent of temperature. The quantum view happens to be the accepted view, since it gives answers that match more precisely with experimental data. Table 2.1 shows the Fermi velocities of electrons for various metals.

TABLE 2.1 Condensed Matter Properties of Various Metals

MATERIAL	FERMI ENERGY E_f (eV)	FERMI TEMPERATURE ($\times 10^4$ K)	FERMI VELOCITY (M/S) $V_f = C\sqrt{\frac{2E_f}{m_e C^2}}$	FREE-ELECTRON DENSITY ρ, (ELECTRONS/m^3)	WORK FUNCTION W (eV)
Copper (Cu)	7.00	8.16	1.57×10^6	8.47×10^{28}	4.7
Silver (Ag)	5.49	6.38	1.39×10^6	5.86×10^{28}	4.73
Gold (Au)	5.53	6.42	1.40×10^6	5.90×10^{28}	5.1
Iron (Fe)	11.1	13.0	1.98×10^6	17.0×10^{28}	4.5
Tin (Sn)	8.15	9.46	1.69×10^6	14.8×10^{28}	4.42
Lead (Pb)	9.47	11.0	1.83×10^6	13.2×10^{28}	4.14
Aluminum (Al)	11.7	13.6	2.03×10^6	18.1×10^{28}	4.08

Note: **1 eV = 1.6022×10^{-19} J, $m_e = 9.11 \times 10^{-31}$ kg, $c = 3.0 \times 10^8$ m/s**

Fermi energy and free electron density data from N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Saunders, 1976; work function data from Paul A. Tipler and Ralph A. Llewellyn, *Modern Physics*, 3rd ed., W.H. Freeman, 1999.

Also, the surface binding energy (caused by electrostatic attraction) that prevents electrons from exiting the surface of the metal, referred to as the *work function*, is about 4.7 eV for copper (1 eV = 1.6022×10^{-19} J). The only way to eject electrons is through special processes, such as thermionic emission, field emission, secondary emission, and photoelectric emission.

(*Thermionic emission:* increase in temperature provides free electrons enough energy to overcome work function of the material. The emitted electron is referred to as a thermoelectron. *Field emission:* additional energy from an electric field generated by a high-voltage conductor provides an attractive enough positive field to free electrons from the surface. This requires huge voltages [MV per cm between emitting surface and positive conductor]. *Secondary emission:* electrons are emitted from a metallic surface by the bombardment of high-speed electrons or other particles. *Photoelectric emission:* electron in material absorbs energy from incoming photon of particular frequency, giving it enough energy to overcome work function. A photon must be of the correct frequency, governed by $W = hf_0$, for this to occur [Planck's constant $h = 6.63 \times 10^{-34}$ J-s or 4.14×10^{-14} eV; f_0 is in hertz]).

2.4.1 Applying a Voltage

Next, we wish to see what happens when we apply a voltage across the conductor—say, by attaching a thick copper wire across a battery. When we do this, our randomly moving free electrons all experience a force pointing toward the positive end of the wire due to the electric field set up within the wire. (The field is due to the negative concentration of pumped-in electrons at one end relative to the neutral [positive relative to negative] concentration at the other end.) The actual influence this force has on the motion of the random free electrons is small—the thermal velocity is so large that it is difficult to change the momentum of the electrons. What you get is a slightly parabolic deviation in path, as shown in Fig. 2.16.

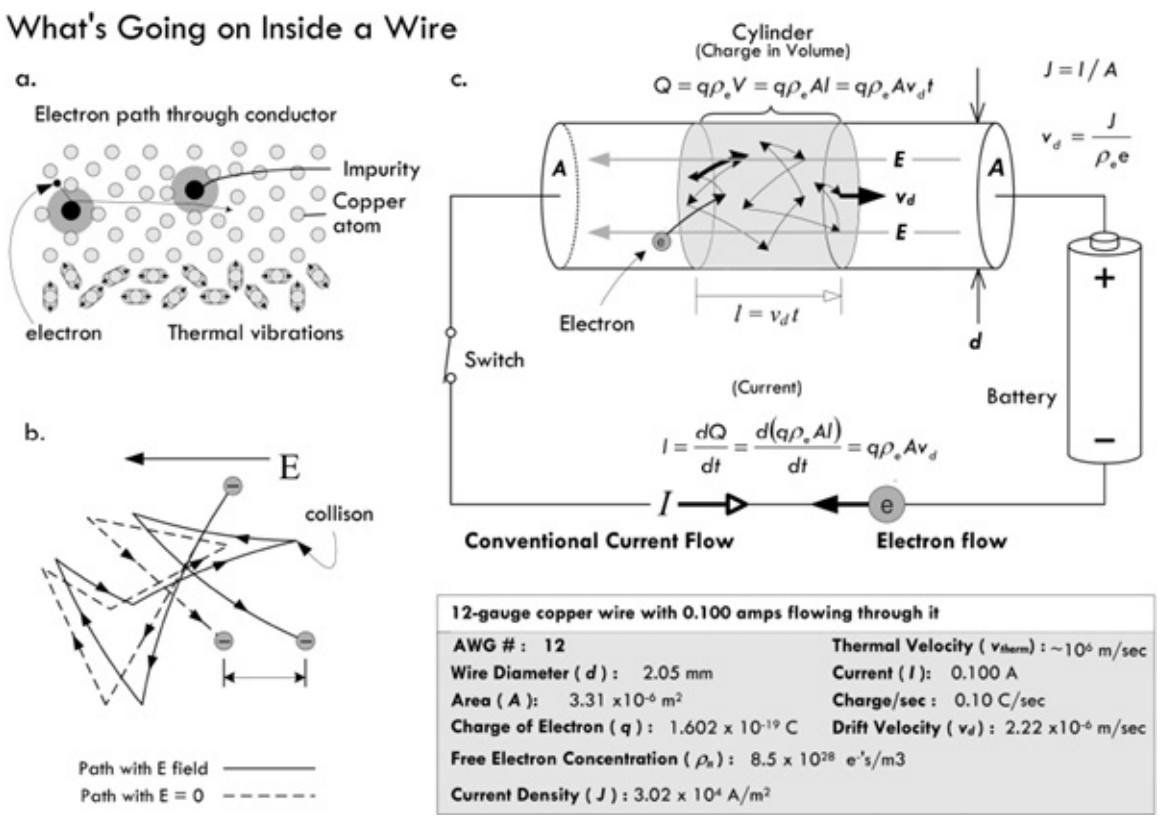


FIGURE 2.16 (a) Simplistic view of an electron randomly moving through a copper lattice, rebounding off lattice atoms and impurities. (b) An electron collides frequently with the ions and impurities in a metal and scatters randomly. In an electric field, the electron picks up a small component of velocity opposite the field. The differences in the paths are exaggerated. The electron's path in an electric field is slightly parabolic. (c) Model illustrating current density, drift velocity, charge density, thermal velocity, and current.

Normally, the field present in the wire would create a net acceleration component in the direction of the force; however, the constant collisions electrons experience create a drag force, similar to the drag experienced by a parachute. The net effect is an average group velocity referred to as the *drift velocity* v_d. Remarkably, this velocity is surprisingly small. For example, the voltage applied to a 12-gauge copper wire to yield a 0.100-A current will result in a drift velocity of about 0.002 mm per second! The drift velocity is related, determined by

$$v_d = J/(\rho_e e)$$

where J is the current density—the current flowing through an area ($J = I/A$), ρ_e is the free-electron density in the material, and e is the charge of an electron. Table 2.1 shows free-electron densities for various materials. As you can see, the drift velocity varies with current and diameter of the conductor.

The drift velocity is so slow, only fractions of a millimeter per second, that it is worth pondering how a measurable current can even flow. For example, what happens when you flip the switch on a flashlight? Of course, we don't have to wait hours for electrons to move down the conductors from the battery. When we throw the switch, the electric field of the incoming electron has a repulsive effect on its neighbor within the wire. This neighbor then moves away toward another neighbor, creating a chain of interactions that propagates through the material at near the speed of light. (See Fig. 2.17.) This reaction, however, really isn't the speed of light, but a fraction less, depending on the medium. The free electrons spreading throughout the conductor all start moving at once in response—those nearest the switch, as well as those nearest the light filament or LED. A similar effect occurs in fluid flow, as when you turn on a garden hose. Because if the hose is already full of water, the outflow starts immediately. The force of the water at the faucet end is quickly transmitted all along the hose, and the water at the open end of the hose flows almost at the moment the faucet is opened.

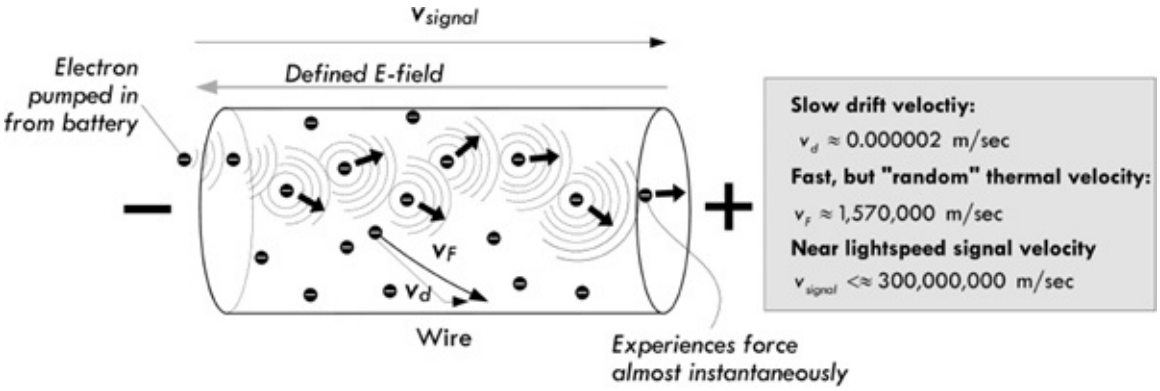


FIGURE 2.17 Illustration of how the electric field propagates down a wire as electrons are pumped into one end.

In the case of alternating current, the field reverses directions in a sinusoidal fashion, causing the drift velocity component of electrons to swish back and forth. If the alternating current has a frequency of 60 Hz, the velocity component would be vibrating back and forth 60 times a second. If our maximum drift velocity during an ac cycle is 0.002 mm/s, we could roughly determine that the distance between maximum swings in the drift distance would be about 0.00045 mm. Of course, this doesn't mean that electrons are fixed in an oscillatory position. It means only that the drift displacement component of electrons is—if there is such a notion. Recall that an electron's overall motion is quite random and its actual displacement quite large, due to the thermal effects.

2.5 Resistance, Resistivity, Conductivity

As was explained in the last section, free electrons in a copper wire at room temperature undergo frequent collisions with other electrons, lattice ions, and impurities within the lattice that limit their forward motion. We associate these microscopic mechanisms that impede electron flow with electrical resistance. In 1826, Georg Simon Ohm published experimental results regarding the resistance of different materials, using a qualitative approach that wasn't concerned with the hidden mechanisms, but rather considered only the net observable effects. He found a linear relationship between how much current flowed through a material when a given voltage was applied across it. He defined the *resistance* to be the ratio of the applied voltage divided by the resultant current flow, given by:

$$R \equiv \frac{V}{I} \tag{2.4}$$

This statement is called Ohm's law, where R is the resistance, given in volts per ampere or ohms (abbreviated with the Greek symbol omega, Ω). One ohm is the resistance through which a current of 1 A flows when a voltage of 1 V is applied:

$$1\ \Omega = 1\ \text{V}/1\ \text{A}$$

The symbol —∿∿∿— is used to designate a resistor.

Now, Ohm's law isn't really a law, but rather an empirical statement about the behavior of materials. In fact, there are some materials for which Ohm's law actually doesn't work.

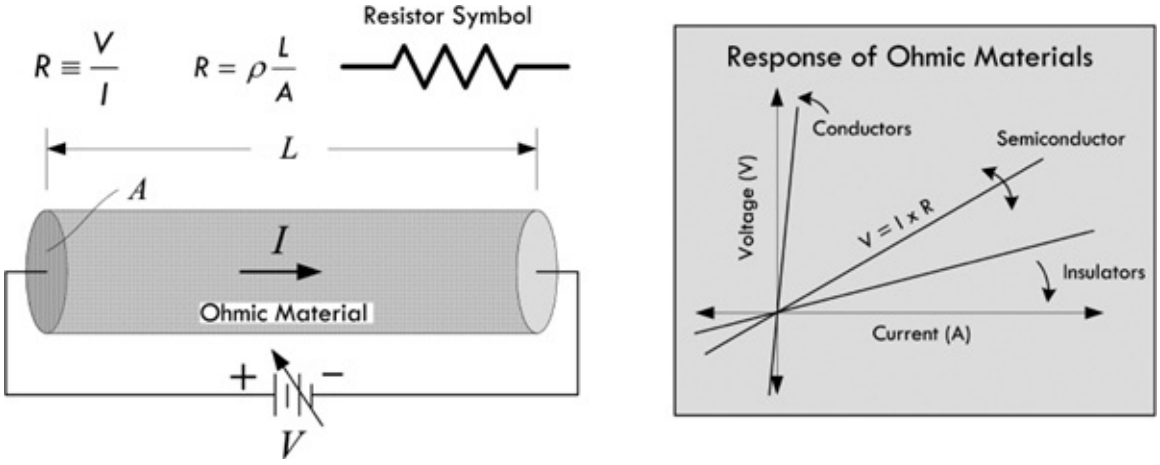


FIGURE 2.18

Ohm's law can be applied only to *ohmic materials*—materials whose resistance remains constant over a range of voltages. *Nonohmic materials*, on the other hand, do not follow this pattern; they do not obey Ohm's law. For example, a diode is a device that allows current to pass easily when the voltage is positive, but prevents current flow (creates a high resistance) when the voltage is negative.

CURIOUS NOTE ABOUT OHM'S LAW

Usually you see Ohm's law written in the following form:

$$V = I \times R$$

However, in this form it is tempting to define voltage in terms of resistance and current. It is important to realize that R is the resistance of an ohmic material and is independent of V in Ohm's law. In fact, Ohm's law does not say anything about voltage; rather, it defines resistance in terms of it and cannot be applied to other areas of physics such as static electricity, because there is no current flow. In other words, you don't define voltage in terms of current and resistance; you define resistance in terms of voltage and current. That's not to say that you can't apply Ohm's law to, say, predict what voltage must exist across a known resistance, given a measured current. In fact, this is done all the time in circuit analysis.

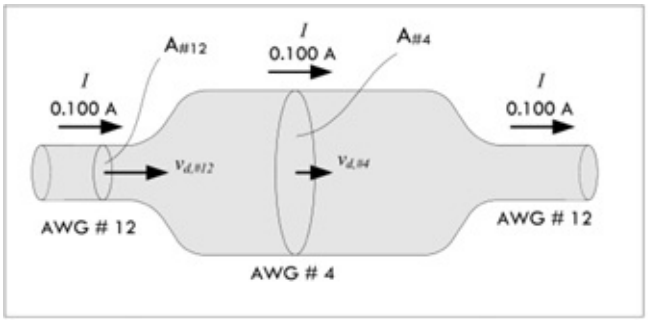
2.5.1 How the Shape of a Conductor Affects Resistance

The resistance of a conducting wire of a given material varies with its shape. Doubling the length of a wire doubles the resistance, allowing half the current to flow, assuming similar applied voltages. Conversely, doubling the cross-sectional area A has the opposite effect—the resistance is cut in half, and twice as much current will flow, again assuming similar applied voltages.

Increasing resistance with length can be explained by the fact that down the wire, there are more lattice ions and imperfections present for which an applied field (electric field instigated by added electrons pumped in by the source) must shove against. This field is less effective at moving electrons because as you go down the line, there are more electrons pushing back—there are more collisions occurring on average.

Decreasing resistance with cross-sectional area can be explained by the fact that a larger-volume conductor (greater cross-sectional area) can support a larger current flow. If you have a thin wire passing 0.100 A and a thick wire passing 0.100 A, the thinner wire must concentrate the 0.100 A through a small

volume, while the thick wire can distribute this current over a greater volume. Electrons confined to a smaller volume tend to undergo a greater number of collisions with other electrons, lattice ions, and imperfections than a wire with a larger volume. The concentration of free-electron flow, according to Benjamin Franklin's convention, represents a concentration of conventional current flow in the opposite direction. This concentration of current flow is called the *current density J*—the rate of current flows per unit area. For a wire: $J = I/A$. Figure 2.19 demonstrates how the current density is greater in a thin 12-gauge wire than in a thicker 4-gauge wire. It also shows that the drift velocity in the thick wire is less than the drift velocity in the thin wire—a result of a "decrease in electron field pressure" lowering the average "push" in the current flow direction.



0.100A Current Flowing		
AWG #	12	4
Wire Diameter (d)	2.05 mm	5.19 mm
Area (A)	3.31 x10^{-6} m^2	2.11 x10^{-5} m^2
Drift Velocity (v_d)	2.22 x10^{-6} m/sec	3.48 x10^{-7} m/sec
Current Density (J)	30,211 A/m^3	4,739 A/m^2

Conductivity (σ) : $1.7 \times 10^{-8}\ \Omega \cdot m$
Resistivity (ρ) : $6.0 \times 10^{7}\ (\Omega \cdot m)^{-1}$
Properties of copper independent of diameter:

FIGURE 2.19 Effects of wire diameter on resistance. A thinner wire has more resistance per unit length than a thicker wire.

2.5.2 Resistivity and Conductivity

We have left out the most important aspect of resistance that has nothing to do with the physical length or diameter of the material. How does the "chemistry" of the material affect the resistance? For example, if you have a copper wire with the same dimensions as a brass wire, which metal has the greater overall resistance? To answer this question, as well as provide a way to categorize materials, we adopt the concept of resistivity. Unlike resistance, resistivity is entirely independent of the dimensions of the material. Resistivity is a property unique to the material. The *resistivity* ρ is defined as follows:

$$\rho \equiv R\frac{A}{L} \tag{2.5}$$

where A is the cross-sectional area, L is the length, and R is the overall resistance of the material, as measured across its length. The units of resistivity are ohm-meters (Ωm).

For some, resistivity is too negative a concept—it tells you how "bad" something is at passing current. Optimists prefer the concept of *conductivity*—how "good" something is at passing current. Conductivity, symbolized σ, is simply the mathematical inverse of resistivity:

$$\sigma \equiv \frac{1}{\rho} \tag{2.6}$$

The units of conductivity are siemens, S = $(\Omega m)^{-1}$. (*Mathematical note:* $[\Omega m]^{-1} = 1/[\Omega m]$). Both conductivity and resistivity contain the same important underlining information. Some prefer to play with equations that use the optimistic notion of conductivity ("glass half full"); others prefer the pessimistic notion of resistivity ("glass half empty").

In terms of resistivity and conductivity, we can rewrite Ohm's law as follows:

$$V = IR = \rho \frac{L}{A} I = \frac{I \times L}{\sigma \times A} \tag{2.7}$$

Table 2.2 shows the conductivities (resistivities, for the pessimists) of various materials. (Consult a technical handbook, such as *The Handbook of Chemistry and Physics*, for a more complete list.) The conductivity of metals, such as copper and silver, is a factor of 10^{21} higher than that of a good insulator, such as Teflon. Though both copper and silver are great conductors, silver is simply too expensive for practical use. Though aluminum is a fairly good conductor and was actually used at one time in home circuitry, it quickly became apparent that it oxidized badly, inhibiting good electrical contacts and limiting current flow to channels of limited size. The result produced fire hazards.

TABLE 2.2 Conductivity of Various Materials

MATERIAL	RESISTIVITY ρ (ΩM)	CONDUCTIVITY σ $(\Omega M)^{-1}$	TEMPERATURE COEFFICIENT α $(°C^{-1})$	THERMAL RESISTIVITY $(W/cm\ °C)^{-1}$	THERMAL CONDUCTIVITY K (W/cm °C)
Conductors					
Aluminum	2.82×10^{-8}	3.55×10^{7}	0.0039	0.462	2.165
Gold	2.44×10^{-8}	4.10×10^{7}		0.343	2.913
Silver	1.59×10^{-8}	6.29×10^{7}	0.0038	0.240	4.173
Copper	1.72×10^{-8}	5.81×10^{7}	0.0039	0.254	3.937
Iron	10.0×10^{-8}	1.0×10^{7}	0.0050	1.495	0.669
Tungsten	5.6×10^{-8}	1.8×10^{7}	0.0045	0.508	1.969
Platinum	10.6×10^{-8}	1.0×10^{7}	0.003927		
Lead	0.22×10^{-6}	4.54×10^{6}		2.915	0.343
Steel (stainless)	0.72×10^{-6}	1.39×10^{6}		6.757	0.148 (312)
Nichrome	100×10^{-8}	0.1×10^{7}	0.0004		
Manganin	44×10^{-8}	0.23×10^{7}	0.00001		
Brass	7×10^{-8}	1.4×10^{7}	0.002	0.820	1.22
Semiconductors					
Carbon (Graphite)	3.5×10^{-5}	2.9×10^{4}	-0.0005		
Germanium	0.46	2.2	-0.048		
Silicon	640	3.5×10^{-3}	-0.075	0.686	1.457 (pure)
Gallium arsenide				1.692	0.591
Insulators					
Glass	10^{10}–10^{14}	10^{-14}–10^{-10}			
Neoprene rubber	10^{9}	10^{-9}			
Quartz (fused)	75×10^{16}	10^{-16}			
Sulfur	10^{15}	10^{-15}			
Teflon	10^{14}	10^{-14}			

An important feature of resistivity (or conductivity) is its temperature dependency. Generally, within a certain temperature range, the resistivity for a large number of metals obeys the following equation:

$$\rho = \rho_0[1 + \alpha(T - T_0)] \tag{2.8}$$

where ρ is the calculated resistivity based on a set reference resistivity ρ_0 and temperature T_0. Alpha α is called the *temperature coefficient of resistivity*, given in units of 1/°C or $(°C)^{-1}$. The resistivity for most metals increases with temperature because lattice atom vibrations caused by thermal energy (increased temperature) impede the drift velocity of conducting electrons.

Are Water, Air, and Vacuums Considered Insulators or Conductors?

These media require special mention. See comments presented in Table 2.3.

TABLE 2.3 Resistivity of Special Materials

MATERIAL	RESISTIVITY, ρ (ΩM)	COMMENTS
Pure water	2.5×10^5	Distilled water is a very good insulator—it has a high resistivity. The rather weak mechanism of conduction is one based on ion flow, not electron flow, as we saw in metals. Water normally has an autoionization at room temperature that gives rise to an amount of H_3O^+ and OH^- ions, which number small in comparison to H_2O molecules ($1{:}10^{-7}$). If a battery of known voltage is connected to two copper electrodes placed apart within a bucket of distilled water, and a current is measured in the hardwire section of the circuit, using Ohm's law, we find that the water presents a resistance of around 20×10^6 Ω/cm. Between the electrodes, H_3O^+ flows toward the negative electrode, and OH^- flows toward the positive electrode. When each ion makes contact with the electrodes, it deposits an electron or accepts an electron, whereby the ion recombines in solution.
Saltwater	~0.2	Adding an ionic compound in the form of common salt (NaCl) to water increases the ion concentration within solution—NaCl ionizes into Na^+ and Cl^-. A gram of salt added introduces around 2×10^{22} ions. These ions act as charge carriers, which in turn, effectively lowers the solution's resistance to below an ohm per meter. If we use the solution as a conductor between a battery and a lamp—via electrodes placed in solution—there is ample current to light the lamp.
Human skin	$\sim 5.0 \times 10^5$	Varies with moisture and salt content.
Air		Considered an insulator since it is fairly void of free electrons. However, like liquids, air often contains concentrations of positive and negative airborne ions. The formation of an air ion starts out with an electron being knocked off a neutral air molecule, such as oxygen (O_2) or nitrogen (N_2), via naturally occurring bombardment with x-rays, gamma rays, or alpha particles emitted by decaying radon atoms. (There are about 5 to 10 ion pairs produced per second per centimeter at sea level; more around radon areas.) The positive molecule of oxygen or nitrogen rapidly attracts polar molecules (from 10 to 15 in number), mostly water. The resulting cluster is called a positive air ion. The liberated electron, on the other hand, will most likely attach itself to an oxygen molecule (nitrogen has no affinity for electrons). *Note:* air is considered neutral—ions are always formed in pairs, equal parts positive and negative. For current to flow through air, you must apply an electric field across the air—say, by using two parallel plates placed apart and set to different potentials (a voltage existing between them). At low voltage, the field strength is small and you may get movement of air ions between the plates, but since the concentration of free-floating ions is so small, there is practically no current. However, if you increase the field strength, it is possible to accelerate those free electrons (those liberated by natural processes) within the air to such high velocities (toward the positive plate) that these electrons themselves can have enough energy to collide with air molecules and generate more ion pairs (positive and negative). The breakdown field strength at which to instigate these ionizations is somewhere around 3 megavolts per meter between plane electrodes in air. Two metal plates 1 cm apart need about 30,000 V for ionization to take place between the plates. However, you can increase the strength of an electric field by replacing one of the plates with a sharp metal point or thin wire. The required voltage drops to a few kilovolts. What occurs can be classified into the following:

Air	*Corona discharge:* Ionization is limited to a small region around the electrode, where the breakdown field strength is exceeded. In the rest of the field, there is just a current of slow-moving ions and even slower-moving charged particles finding their way to the counter electrode (which may be an oppositely charged plate, or simply the wall of the room or the ground). Such a discharge may be maintained as long as the breakdown field strength is exceeded in some region. *Spark discharge:* May take place between two well-rounded conductors at different potentials, one of them often grounded. Like corona discharge, the discharge starts at a point where the breakdown field strength is exceeded. But in contrast to corona discharge, in a spark the ionization takes place all the way between the electrodes. The discharge is very fast, and the energy of the discharge is confined to a narrow volume. The voltage necessary to cause breakdown is called the breakdown voltage. For example, when you become charged by walking across an insulated floor (e.g., carpet) with insulated shoes and acquire a charge due to contact and friction, then touch a grounded object, charge will flow between you, even before you make physical contact—the spark. Few people notice discharges at voltages lower than about 1000 V. Most people start to feel an unpleasant effect around 2000 V. Almost everyone will complain when exposed to discharges at voltages above 3000 V. *Brush discharge:* In between the corona discharge and the spark, which may take place, for example, between a charged material and a normally grounded electrode with a radius of curvature of some millimeters. If a brush discharge is maintained over longer periods, it may appear as irregular luminescent paths. Almost all discharges from insulators are brush discharges, like the crackle heard as you pick up a charged photocopy or that you feel when you pull a sweater over your head.
Vacuum	Considered a perfect insulator, since, by definition, it contains no free charges at all. However, this doesn't mean it isn't possible for charges to pass through it. Various mechanisms can eject electrons from a material into a so-called vacuum. (This applies for air, too.) Examples of such mechanisms include *thermionic emission,* where an increase in temperature provides free electrons enough energy to overcome surface potential barriers (the work function) of the material; *field emission,* whereby additional energy from an electric field generated by a high-voltage conductor provides an attractive enough positive field to free electrons from the surface, which requires huge voltages (MV per cm between emitting surface and positive conductor); *secondary emission,* where electrons are emitted from a metallic surface by the bombardment of high-speed electrons or other particles; *photoelectric emission,* where electrons are ejected from the surface of a material when a photon of a particular frequency strikes it (absorbed electron). Many of these mechanisms are employed in vacuum tube technology. However, without such mechanisms to source electrons, there will be no source of charge within the vacuum to support a current.

2.6 Insulators, Conductors, and Semiconductors

As we have seen, the electrical resistivities of materials vary greatly between conductors and insulators. A decent conductor has around 10^{-8} Ωm; a good insulator has around 10^{14} Ωm; a typical semiconductor from 10^{-5} to 10^{3} Ωm—depending on temperature. What is the microscopic explanation for these differences?

The answer to this question rests upon the quantum nature of electrons. In classical physics the energy of an electron in a metal can take on any value—it is said the energy values form a continuum. (Here, electron energy is considered zero at infinite distance from the nucleus and becomes more negative in energy closer to the nucleus, relative to the zero reference state. Negative energy infers that there is electric attraction between the positive nucleus and the electron—it is electric potential energy.) However, a quantum description of electrons in metals shows that the energy values of electrons are quantized, taking on discrete values. This comes from the wavelike nature of electrons—analogous to standing waves on strings existing only at discrete frequencies. Figure 2.20 shows an energy diagram illustrating the possible energy levels of an electron (ignoring lattice influences). The diagram illustrates only the possible energy levels—electrons are not necessarily in each level.

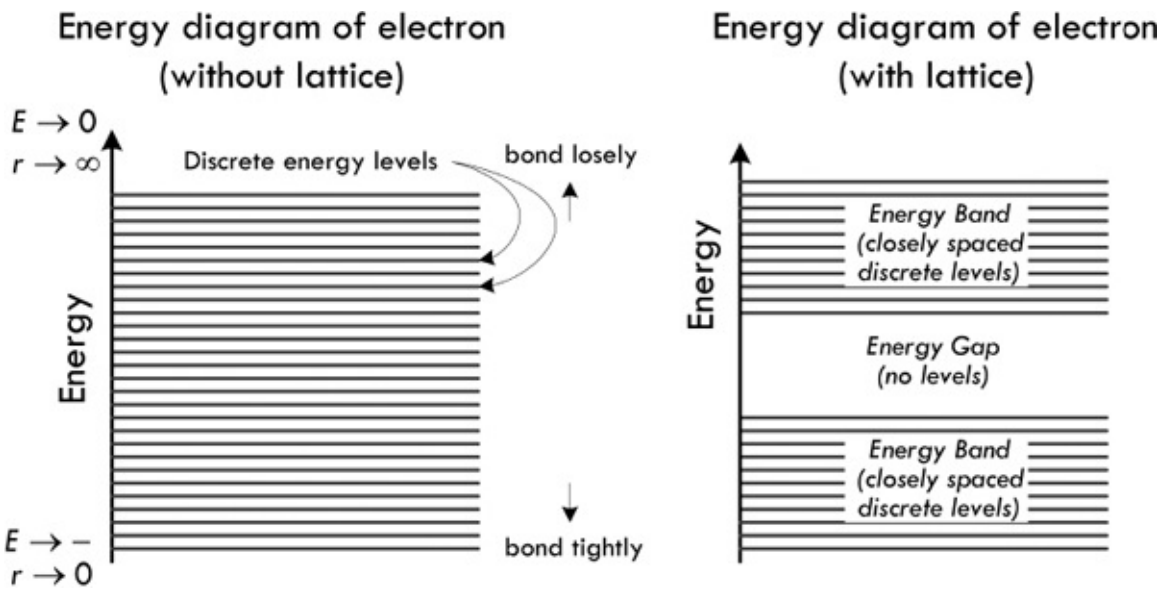


FIGURE 2.20 (Left) Energy diagram showing all possible energy levels of an electron in a solid; but it takes no account of the influence of an atomic crystalline lattice structure. (Right) Energy diagram that shows possible energy levels of an electron within a material made of a regular lattice of atoms. Electron energies are restricted to lie within allowed bands, and there is a large energy gap where no electrons are allowed. Even within the allowed bands, the possible electron energies are closely spaced discrete levels.

When a set of atoms forms a regular background lattice, the possible energy values of the electrons are altered even further. We still have discrete energy regions, called allowed bands, but we now get what are called energy gaps. Energy gaps are forbidden regions to the electron, and represent regions where no traveling wave (electron) can exist when placed in the periodic electric potential of the metal's positive lattice ions. These gaps are quite large in the scale of atomic physics—within electron-volts range. Again, energy levels presented in the band diagram specify only possible values of electron energies—they may or may not be occupied.

Now, quantum physics has an interesting property, called the *Pauli exclusion principle*, which has a critical role in determining the properties of materials. The Pauli exclusion principle says that no two electrons in an atom can be in the same quantum state. The lowest common divisor of quantum states is the spin quantum number m_s, which states that no more than two electrons with opposing spin (up or down) can be located in the same energy level. Now, if we consider a solid that has many free electrons, which is in an equilibrium state, electrons fill the lowest energy levels available in the allowed band, up to two in each level. Those electrons further down in energy are more tightly bound and are called innermost electrons. When all the electrons are placed in the lowest energy state, we are left with two possible outcomes. In the first case, the highest level to be filled is somewhere in the middle of a band. In the second case, the electrons just fill one or more bands completely. We assume that the material is at low enough temperature to prevent electrons from jumping to higher energy levels due to thermal effects.

Now if we add some energy to the free electrons by applying an electric field (attach a voltage source), for example, the electrons in the lower energy levels cannot accept that energy, because they cannot move into an already filled higher energy level. The only electrons that can accept energy are those that lie in the top levels, and then only if there are nearby empty levels into which they can move. Materials with electrons in only partly filled bands are *conductors*. When the top layer of their electrons moves freely into the empty energy levels immediately above, there is a current. The electrons that jump from a lower level to a higher level are said to be excited. The valance band is an allowed occupied band. The conduction band is an allowed empty band. The energy-band structure for conductors is shown in Fig. 2.21a and c.

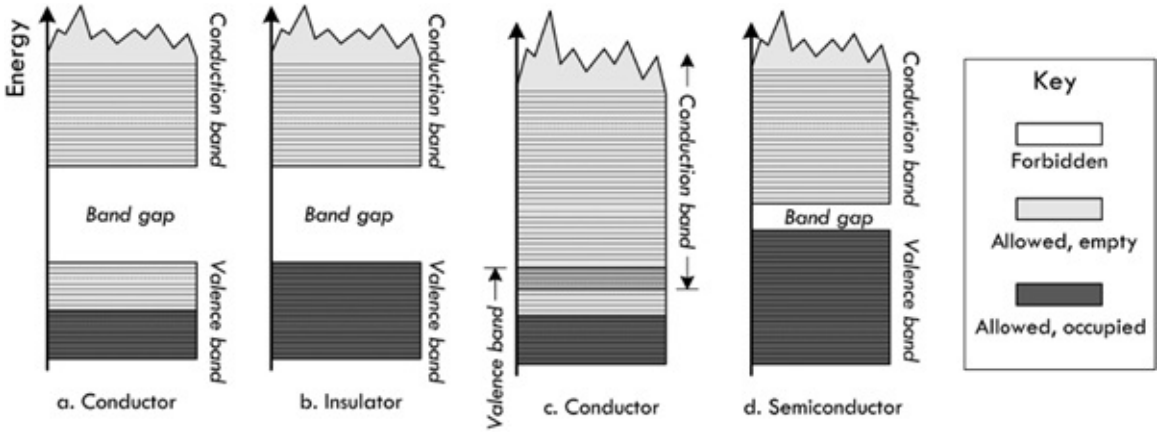


FIGURE 2.21 Four possible band structures for a solid: (a) Conductor—allowed band is only partially full, so electrons can be excited to nearby states. (b) Insulator—forbidden band with large energy gap between the filled band and the next allowed band. (c) Conductor—allowed bands overlap. (d) Semiconductor—energy gap between the filled band and the next allowed band is very small, so some electrons are excited to the conduction band at normal temperatures, leaving holes in the valence band.

If the highest-energy electrons of a material fill a band completely, then a small electric field will not give these electrons enough energy to jump the large energy gap to the bottom of the next (empty) band. We then have an *insulator* (see Fig. 2.21b). An example of a good insulator is diamond, whose energy gap is 6 eV.

In *semiconductors*, the highest-energy electrons fill a band (the valance band) at T = 0, as in insulators. However, unlike insulators, semiconductors have a small energy gap between that band and the next, the conduction band. Because the energy gap is so small, a modest electric field (or finite temperature) will allow electrons to jump the gap and thereby conduct electricity. Thus, there is a minimum electric field under the influence of which a material changes from insulator to conductor. Silicon and germanium have energy gaps of 1.1 eV and 0.7 eV, respectively, and are semiconductors. For semiconductors, an increase in temperature will give a fraction of the electrons enough thermal energy to jump the gap. For an ordinary conductor, a rise in temperature increases the resistivity, because the atoms, which are obstacles to electron flow, vibrate more vigorously. A temperature increase in a semiconductor allows more electrons into the empty band and thus lowers the resistivity.

When an electron in the valance band of a semiconductor crosses the energy gap and conducts electricity, it leaves behind what is known as a hole. Other electrons in the valence band near the top of the stack of energy levels can move into this hole, leaving behind their own holes, in which still other electrons can move, and so forth. The hole behaves like a positive charge that conducts electricity on its own as a positive charge carrier. An electron excited from the valence band to the conduction band is thus doubly effective at conducting electricity in semiconductors.

Besides the intrinsic elemental semiconductors, such as silicon and germanium, there are hybrid compounds—compounds such as gallium arsenide. Other semiconductors are made by introducing impurities into a silicon lattice. For example, an atom in the chemical group of phosphorous, arsenic, and antimony can replace one of the silicon atoms in a lattice without affecting the lattice itself too much. However, each of these impurities has one more electron in its valence level than the silicon atom has; this extra electron, for which there is no room in the valence band, takes a place in the conduction band and can conduct electricity. A semiconductor with impurities of this sort is called an *n-type semiconductor*, and the extra electrons are called donor electrons.

Atoms of elements in the same chemical group as boron, aluminum, and gallium have one less valence electron than silicon has. If an atom is added to a lattice of silicon as an impurity, there is one less electron than is needed to form a bond that holds the lattice together. This electron must be provided by the electrons of the valence band of the lattice material, and holes are created in this band. These holes act as positive charge carriers. The impurity atoms are called acceptors. A semiconductor with such

impurities is called a *p-type semiconductor*.

We will see later on how n-type and p-type semiconductors are used for making one-way gates for current flow (diodes) and voltage-controlled current switches (transistors).

2.7 Heat and Power

In Sec. 2.3, we discovered the generalized power law, which tells us that if we can measure the current entering a device, as well as the voltage across the device, we know the power that is used by the device:

$$P = VI \tag{2.9}$$

The generalized power law tells us how much power is pumped into a circuit but doesn't say anything about how this power is used up. Let's consider a two-lead black box—an unknown circuit that may contain all sorts of devices, such as resistors, lamps, motors, or transistors. If all we can do is measure the current entering the black box and the voltage across it—say, using an ammeter and voltmeter (or singularly, a wattmeter)—we could apply the generalized power law, multiply the measured current and voltage readings together, and find the power pumped into the black box. For example, in Fig. 2.22, we measure 0.1 A entering when a voltage of 10 V is applied, giving us a total consumed power of 1 W.

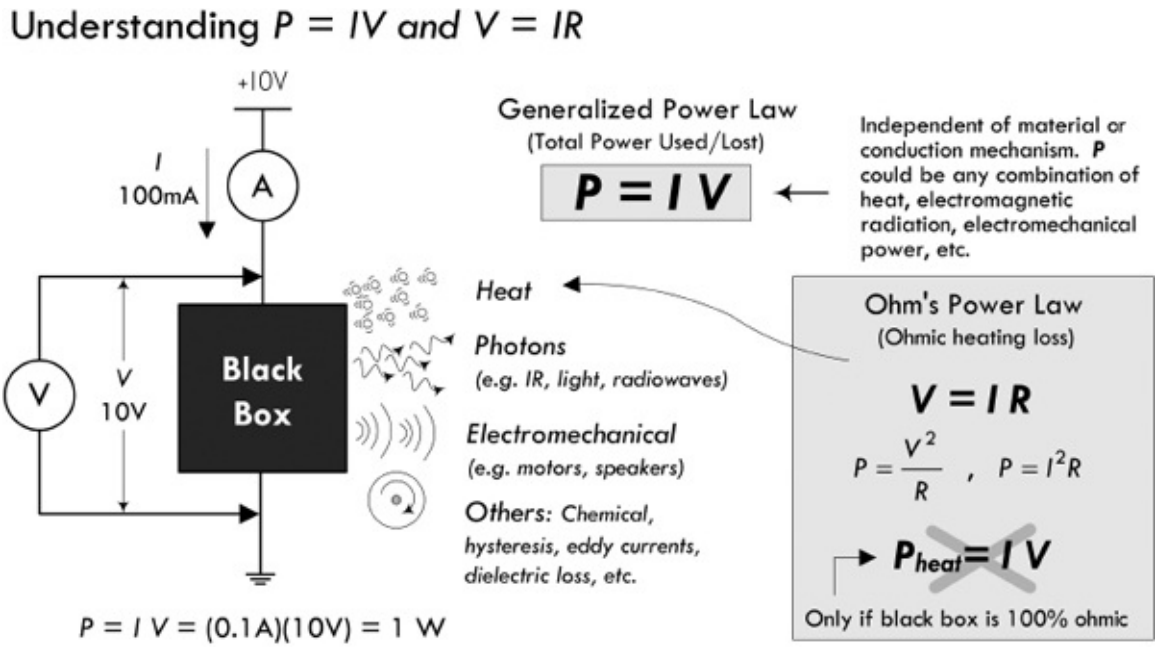


FIGURE 2.22

Knowing how much power is pumped into the black box is incredibly useful—it allows for quick power consumption measurements and often simplifies circuit analysis—as we'll see later. But let's say that we are interested in figuring out how much power is lost to heating (energy that goes into lattice vibrations, emissive radiation, etc.). We really can't say, assuming we aren't allowed to look inside the black box. There could be devices inside that take some of the initial energy and use it to do useful work, such as generating magnetic fields in the armature (rotor) and stator sections of a motor, causing the stator to rotate; or generating a magnetic field in a voice coil attached to a paper speaker cone that compresses air; or generating light energy, radio waves, and so on. There may be power converted into other weird forms not really coined as heat, such as driving chemical reactions, generating hysteresis effects, or eddy current in transformers.

The only time we can say for certain that power is totally converted into heat energy is if we assume

our black box is a perfect resistor (100 percent ohmic in nature). Only then can we substitute Ohm's law into the generalized power equation:

$$P = VI = V(V/R) = V^2/R \tag{2.10}$$

or

$$P = VI = (IR)I = I^2R$$

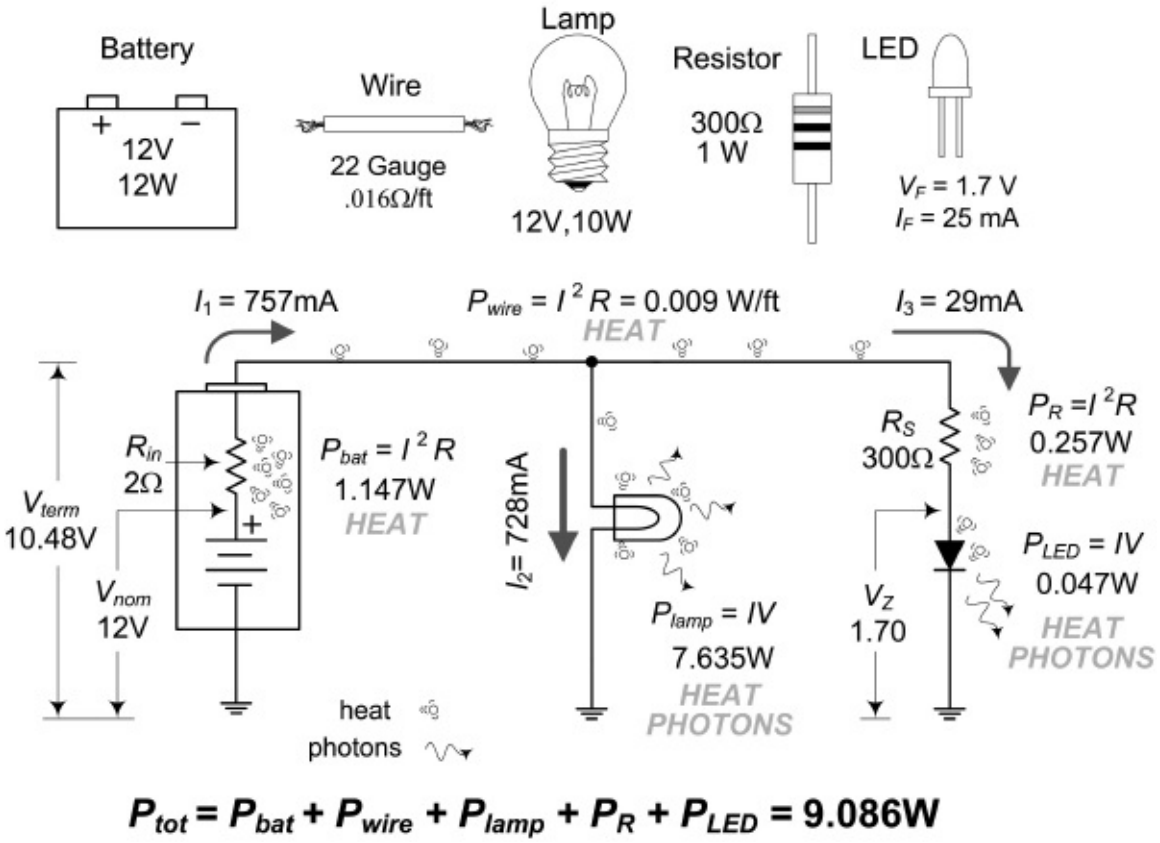
In this form, the power lost due to heating is often called *Ohmic heating*, *Joule heating*, or I^2R *loss*. Be careful how you interpret this law. For example, let's consider our black box that drew 1 W of power. Given the power and the current—which we measure—it would be easy to assume that the resistance of the black box is:

$$R = \frac{P}{I^2} = \frac{1\text{ W}}{(0.01\text{ A})} = 100\ \Omega$$

Accordingly, we would say the black box is a 100-Ω resistor generating 1 W of heat. As you can see, this is an erroneous assumption, since we have disregarded the internal workings of the black box—we didn't account for devices that perform useful work. You'll often see people treat any load (black box) as a resistor when doing circuit analysis and such. This will get you the right answer when solving for a particular variable, but it is an analysis trick and shouldn't be used to determine how much heat is being generated, unless, of course, the black box is actually a resistor.

The following example provides some insight into where power is being used and how much of it is being converted into heat.

EXAMPLE 1 (POWER LOSS IN CIRCUITS)



The total power pumped into this circuit gets converted into useful work and heat. The total "pump-in" power is:

$$P_{tot} = IV = (0.757\ \text{A})(12\ \text{V}) = 9.08\ \text{W}$$

(This is based on an open terminal battery measurement, without the rest of the circuit attached.) From there, we notice that some of the total power is wasted within the internal resistance of the battery, within the internal resistance of the wires, and within the resistance of the current-limiting resistor used by the LED. The power used to create light from the lamp and LED can be considered useful power used. However, since we can't really separate the heat from the light power for these devices, we must apply the generalized power law to them and be content with that. According to the conservation of energy (or power), all individual powers within the circuit add up to the total power.

FIGURE 2.23

Example: With an ammeter and voltmeter, you measure the current drawn by a computer to be 1.5 A and the voltage entering to be 117 V. How much power does the computer consume? Can we say how much power is lost to heat?

Answer: $P = IV$ = (1.5 A)(117 V) = 176 W. Knowing how much of the power is lost to heating is practically impossible to measure without taking the computer apart.

Example: Determine the resistance of the following four round rods of material, each 1 m long and 2 mm in diameter: copper, brass, stainless steel, and graphite. Also figure out how much power is lost to heating if a current of 0.2 A flows through each one.

Answer: Using Eq. (2.5),

$$R = \rho\frac{L}{A} = \rho\frac{L}{\pi r^2} = \rho\frac{1\ \text{m}}{\pi(0.001\ \text{m})^2} = \rho\frac{1\ \text{m}}{3.14\times10^{-6}\ \text{m}^2} = \rho(3.18\times10^5\ \text{m}^{-1})$$

Using Table 2.2:

$$\rho_{copper} = 1.72 \times 10^{-8}\ \Omega m,\ \rho_{brass} = 7.0 \times 10^{-8}\ \Omega m,\ \rho_{steel} = 7.2 \times 10^{-7}\ \Omega m,\ \rho_{graphite} = 3.5 \times 10^{-5}\ \Omega m$$

Substituting this into the resistance expression, we get:

$$R_{copper} = 5.48 \times 10^{-3}\ \Omega,\ R_{brass} = 2.23 \times 10^{-2}\ \Omega,\ R_{steel} = 2.31 \times 10^{-1}\ \Omega,\ R_{graphite} = 11.1\ \Omega$$

The power loss we get by using Eq. 2.10: $P = I^2R = (0.2\ A)^2R = (0.04\ A^2)R$:

$$P_{copper} = 2.2 \times 10^{-4}\ W,\ P_{brass} = 8.9 \times 10^{-4}\ W,\ P_{steel} = 9.2 \times 10^{-3}\ W,\ P_{graphite} = 0.44\ W$$

2.8 Thermal Heat Conduction and Thermal Resistance

How is the energy transferred in heating? Within a *gas*, heat transfer represents the transfer of energy between colliding gas molecules. Gas molecules at a hotter temperature move around more quickly—they have a high kinetic energy. When they are introduced into another, colder-temperature gas, the "hotter," fast-moving molecules impart their energy to the slower-moving molecules. Gases tend to be the worst thermal conductors, due to a low density of molecules.

In *nonmetals*, heat transfer is a result of the transfer of energy due to lattice vibrations—energetically vibrating atoms in one region of a solid (e.g., the region near a flame) transfer their energy to other regions of a solid that have less energetically vibrating atoms. The transfer of heat can be enhanced by cooperative motion in the form of propagating lattice waves, which in the quantum limit are quantized as *phonons*. The thermal conductivity of nonmetals varies, depending greatly on the lattice structure.

In terms of *metals*, heat transfer is a result of both lattice vibration effects (as seen in nonmetals) as well as kinetic energy transfer due to mobile free electrons. Recall that free electrons within a metal are moving quite fast (~10^6 m/s for most metals) at room temperature. Though quantum mechanics is required, it is possible to treat these electrons like a dense gas, capable of increasing its overall energy as heat is added, and likewise, capable of transporting this energy to regions of the metal that are lower in temperature. Notice, however, that an increase in metal temperature, as a whole, also increases the electrical resistance—the drift velocity component of the free electron goes down, due to both increased lattice vibrations and an increase in thermal velocity component. It becomes harder to influence the free electrons with an applied field. Metals are the best thermal conductors, due in part to the additional free electrons.

The energy an object has at temperature T is correlated to the internal energy—a result of its internal atomic/molecular/electron motions. It is not correct, however, to use the word *heat*, such as "object possesses heat." Heat is reserved to describe the process of energy transfer from a high-temperature object to a lower-temperature object. According to the first law of thermodynamics, which is a statement about the conservation of energy, the change in internal energy of a system ΔU is equal to the heat Q_H added to the system and the work W done by the system: $\Delta U = Q_H - W$. If we assume that no work is done (energy transferred to, say, move a piston, in the case of a gas), we say $\Delta U = Q_H$. With this assumption, we can take heat to be not a measure of internal energy of the system, but a change in internal energy. The main reason for this concept is that it is very difficult to determine the actual internal energy of a system; changes in internal energy are more meaningful and measurable.

In practice, what is most useful is the rate of heat transfer—the power loss due to heating. With the help of experimental data, the following formula can be used to determine how well certain materials transfer

heat:

$$P_{\text{heat}} = \frac{dQ_H}{dt} = -k\nabla T \tag{2.11}$$

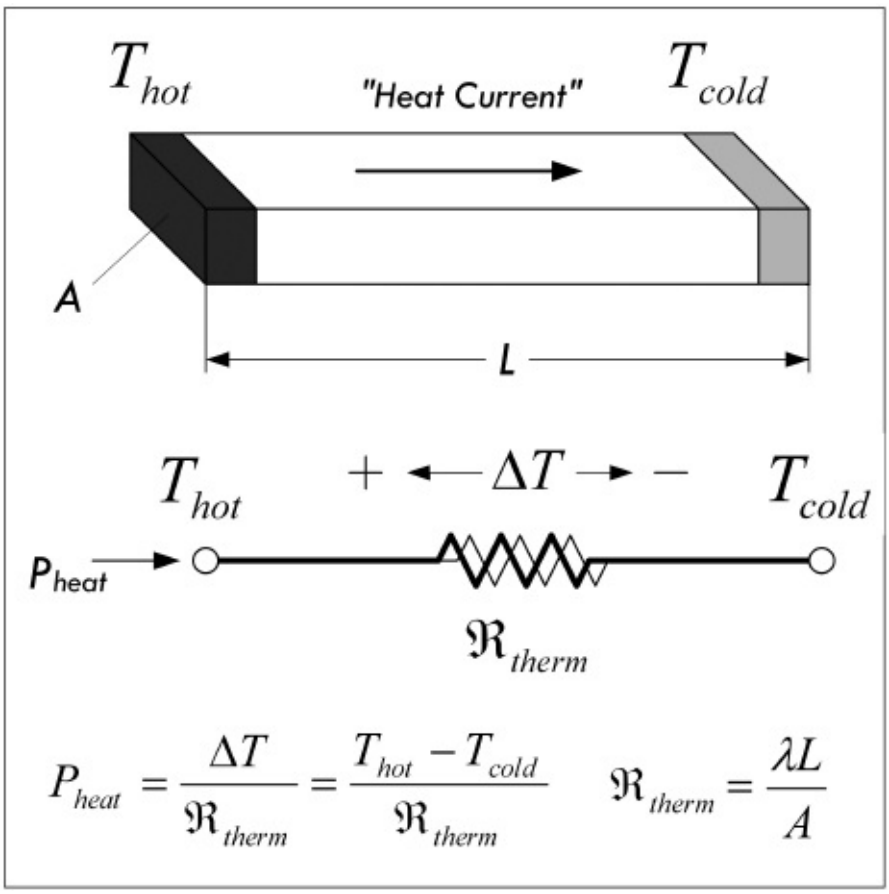
Here, k is called the *thermal conductivity* (measured in W/m °C) of the material in question, and ∇T is the temperature gradient:

$$\nabla T = \left(i\frac{\partial}{\partial t} + j\frac{\partial}{\partial t} + \frac{\partial}{\partial t} \right) T$$

Now, the gradient is probably scary to a lot of you—it is simply a way to represent temperature distributions in 3-D, with time. To keep things simple, we will stick to 2-D, and represent the gradient as acting through an area A through a thickness L, and assume steady-state conditions:

$$P_{\text{heat}} = -k\frac{A\Delta T}{L} \tag{2.12}$$

where $\Delta T = T_{\text{hot}} - T_{\text{cold}}$, measured at points across the length L of the material. The material may be steel, silicon, copper, PC board material, and so on. Figure 2.24 shows a picture of the situation.



When one end of a block of material is placed at a hot temperature, heat will be conducted through the material to the colder end. The rate of heat transfer, or power due to heating, depends on the thermal resistance of the material, which in turn is dependent on the geometry of the material and the material's thermal resistivity. A weird-looking resistor-like symbol is used here to represent the thermal resistance. Table 2.4 shows the thermal resistivities of various materials.

FIGURE 2.24

TABLE 2.4 Typical Thermal Resistivities (λ) in Units of °C·in/W

Divide λ by 39 to get value in units of °C·m/W

MATERIAL	λ	MATERIAL	λ	MATERIAL	λ
Diamond	0.06	Lead	1.14	Quartz	27.6
Silver	0.10	Indium	2.1	Glass (774)	34.8
Copper	0.11	Boron nitride	1.24	Silicon grease	46
Gold	0.13	Alumina ceramic	2.13	Water	63
Aluminum	0.23	Kovar	2.34	Mica	80
Beryllia ceramic	0.24	Silicon carbide	2.3	Polyethylene	120
Molybdenum	0.27	Steel (300)	2.4	Nylon	190
Brass	0.34	Nichrome	3.00	Silicon rubber	190
Silicon	0.47	Carbon	5.7	Teflon	190
Platinum	0.54	Ferrite	6.3	P.P.O.	205
Tin	0.60	Pyroceram	11.7	Polystyrene	380
Nickel	0.61	Epoxy (high conductivity)	24	Mylar	1040
Tin solder	0.78			Air	2280

Thermal conductivity k, like electrical conductivity, has an inverse—namely, *thermal resistivity* λ. Again, one tells you "how good" a material is at transferring heat, the other tells you "how bad" it is at doing it. The two are related by $k = 1/\lambda$.

If we consider the geometry of the material, we can create a notion of *thermal resistance* $\Re_{\text{therm}}$ (analogous to electrical resistance), which depends on the cross-sectional area A, the length of the block of material L, and the thermal conductivity k, or resistivity λ:

$$\Re_{\text{therm}} = \frac{L}{kA} \quad \text{or} \quad \Re_{\text{therm}} = \frac{\lambda L}{A} \tag{2.13}$$

Thermal resistance has units of °C/W.

Thus, putting everything together, the power transfer of heat across a block of material, from one point at one temperature to another point at a different temperature, can be expressed as:

$$P_{\text{heat}} = \frac{dQ_{\text{heat}}}{dt} = k\left(\frac{A}{L}\right)\Delta T = \frac{1}{\lambda}\left(\frac{A}{L}\right)\Delta T = \frac{\Delta T}{\Re_{\text{therm}}} \tag{2.14}$$

(k = thermal conductance) (λ = thermal resistivity) ($\Re_{\text{therm}}$ = thermal resistance)

A very useful property of Eq. 2.14 is that it is exactly analogous to Ohm's law, and therefore the same principles and methods apply to heat flow problems as to circuit problems. For example, the following correspondences hold:

Thermal conductivity k [W/(m °C)]	Electrical conductivity σ [S/m or $(\Omega m)^{-1}$]
Thermal resistivity λ [m°C/W]	Electrical resistivity ρ [Ωm]

Thermal resistance $\Re_{therm}$ [°C/W]	Electrical resistance R [Ω]
Thermal current (heat flow) P_{heat} [W]	Electrical current I [A]
Thermal potential difference ΔT [°C]	Electrical potential difference or voltage V [V]
Heat source	Current source

Example: Calculate the temperature of a 4-in (0.1 m) piece of #12 copper wire at the end that is being heated by a 25-W (input power) soldering iron, and whose other end is clamped to a large metal vise (assume an infinite heat sink), if the ambient temperature is 25°C (77°F).

Answer: First, calculate the thermal resistance of the copper wire (diameter #12 wire is 2.053 mm, cross-sectional area is 3.31×10^{-6} m^2):

$$\Re_{therm} = \frac{L}{kA} = \frac{(0.1\text{ m})}{(390\text{ W/(m°C)}(3.31\times 10^{-6}\text{ m}^2))} = 77.4\text{°C/W}$$

Then, rearranging the heat flow equation and making a realistic assumption that only around 10 W of heat actually is transferred to the wire from the 25-watt iron, we get:

$$\Delta T = P_{heat}\Re_{therm} = (10\text{ W})(77.4\text{°C/W}) = 774\text{°C}$$

So the wire temperature at the hot end is estimated at:

$$25\text{°C} + \Delta T = 799\text{°C (or 1470°F)}$$

It's important to realize that in the preceding example, we assume steady-state conditions, where the soldering iron has been held in place for a long time. Also, the assumption that only 10 W were transferred is an important point, since there is a lot of heat being radiated off as heat to the air and the iron handle, and so on. In any case, things can get very hot, even when moderate power levels are in question.

2.8.1 Importance of Heat Production

Heat production has its place in electronics (toasters, hair dryers, water heaters, etc.), but most of the time heat represents power loss that is to be minimized whenever possible, or at least taken into consideration when selecting components. All real circuit components—not just resistors but things like capacitors, transformers, transistors, and motors—contain inherent internal resistances. Though these internal resistances can often be neglected, in some situations they cannot be ignored.

Major problems arise when unintended heat generation increases the temperature of a circuit component to a critical point, causing component failure by explosion, melting, or some other catastrophic event. Less severe problems may surface as a component becomes thermally damaged, resulting in a change in characteristic properties, such as a shift in resistance that may cause undesirable effects in circuit behavior.

To avoid problems associated with heat production, it's important to use components that are rated to handle two to three or more times the maximum power they are expected to dissipate. In cases where heat presents a shift in component parameter performance, selecting a component with a lower temperature coefficient (TC) will help.

Heat dissipation (more correctly, the efficient removal of generated heat) becomes very important in

medium-to high-power circuits—power supplies, amplifier stages, transmitting circuits, and power-hungry circuits with power transistors. There are various techniques to remove heat from a circuit in order to lower the operating temperature of components to below critical levels. Passive methods include heat sinks, careful component layout, and ventilation. Heat sinks are special devices that are used to draw heat away from temperature-sensitive devices by increasing the radiating surface in air—which acts like a cooling fluid for conduction. Active methods include forced air (fans) or some sort of liquid cooling. We'll discuss these methods throughout the book.

Example: Figure 2.25 shows a thin-film resistor in an integrated circuit. How hot will it get with 2 W dissipated over its 0.1 × 0.2-in surface (100 W/in^2)? Assume the ground plane is at 80°C.

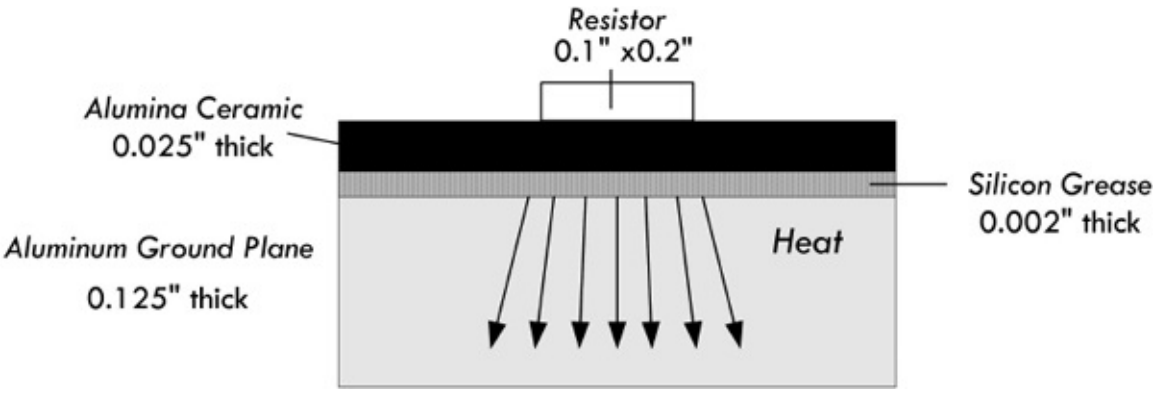


FIGURE 2.25

Answer: Since we have three different media through which this heat will be transferred, we must take into consideration thermal conduction within each. With the help of Eq. 2.14 and Table 2.4, we get the individual heat transfers through each region:

$$\Delta T_{1-2(\text{in ceramic})} = \lambda_{\text{ceramic}} \left(\frac{L}{A}\right) P_{\text{dis}} = \frac{(2.13)(0.025)(2)}{(0.1)(0.2)} = 5.3°\text{C}$$

$$\Delta T_{2-3(\text{in grease})} = \lambda_{\text{grease}} \left(\frac{L}{A}\right) P_{\text{dis}} = \frac{(46)(0.002)(2)}{(0.1)(0.2)} = 9.2°\text{C}$$

$$\Delta T_{3-4(\text{in aluminum})} = \lambda_{\text{aluminum}} \left(\frac{L}{A}\right) P_{\text{dis}} = \frac{(0.23)(0.125)(2)}{(0.1)(0.2)} = 2.9°\text{C}$$

Adding this together:

$$\Delta T_{1-4} = 5.3°\text{C} + 9.2°\text{C} + 2.9°\text{C} = 17.4°\text{C}$$

Adding this to the 80°C for the aluminum ground plane leads to an estimated 100°C for maximum resistor temperature. This is a conservative estimate, since it neglects transverse heat spreading.

2.9 Wire Gauges

In Sec. 2.5, we saw that the current density within a copper wire increased as the diameter of the wire decreased. As it turns out, a higher current density translates into a hotter wire; there are more collisions occurring between electrons and copper lattice ions. There is a point where the current density can become so large that the vibrating effect can overpower the copper-lattice binding energy, resulting in

wire meltdown (also referred to as the fusing point). To prevent this from occurring, it is important to select the appropriate wire size for anticipated current levels. Wire size is expressed in gauge number—the common standard being the American Wire Gauge (AWG)—whereby a smaller gauge number corresponds to a larger-diameter wire (high current capacity). Table 2.5 shows a short list of common AWG wires. Section 3.1, on wires and cables, provides a more in-depth list.

TABLE 2.5 Copper Wire Specifications (Bare and Enamel-Coated Wire)

WIRE SIZE (AWG)	DIAMETER (MILS)*	AREA (CM)†	FEET PER POUND BARE	OHMS PER 1000 FT, 25°C	CURRENT CAPACITY (AMPS)
4	204.3	41738.49	7.918	0.2485	59.626
8	128.5	16512.25	25.24	0.7925	18.696
10	101.9	10383.61	31.82	0.9987	14.834
12	80.8	6528.64	50.61	1.5880	9.327
14	64.1	4108.81	80.39	2.5240	5.870
18	40.3	1624.09	203.5	6.3860	2.320
20	32	1024.00	222.7	10.1280	1.463
22	25.3	640.09	516.3	16.2000	0.914
24	20.1	404.01	817.7	25.6700	0.577
28	12.6	158.76	2081	65.3100	0.227
32	8.0	64.00	5163	162.0000	0.091
40	3.1	9.61	34364	1079.0000	0.014

*** 1 mil = 0.001 in or 0.0254 mm.**
† A circular mil (CM) is a unit of area equal to that of a 1-mil-diameter circle. The CM area of a wire is the square of the mil diameter.
Diameters of wires in Fig. 2.26 are relative and not to scale.



FIGURE 2.26

Example: A load device that is known to vary in output power from 0.1 mW to 5 W is to be connected to a 12-V source that is 10 ft away from the load. Determine the minimum wire gauge, provided by Table 2.5, that can safely support any anticipated current drawn by the load.

Answer: We only care about the maximum power level, so using the generalized power law:

$$I = \frac{P}{V} = \frac{5\text{ W}}{12\text{ V}} = 0.42\text{ A}$$

Given only the selection of wire gauges provided by Table 2.5, a 22-gauge wire with a 0.914-A rating

would work, though we could be conservative and select an 18-gauge wire with a 2.32-A rating. Since the length is so short, there is no appreciable drop in voltage through the wire, so we can ignore the length.

Example: A 10-Ω heating device is powered by a 120-VAC source. How much current does it draw, and what size conductors should be used to connect to the device?

Answer: 120 VAC is an RMS value of a sinusoidal voltage—in this case, household line voltage. Though we'll discuss this later when we cover ac, it can be treated like a dc voltage in terms of power dissipated through a resistor. So,

$$P = \frac{V^2}{R} = \frac{(120\ \text{V}^2)}{10\ \Omega} = 1440\ \text{W}, \quad I\frac{P}{V} = \frac{1440\ W}{120\ \text{V}} = 12\ \text{A}$$

A 10-gauge wire would support this kind of current, though a larger 8-gauge wire would be safer.

Example: Why shouldn't you connect a wire across a voltage source? For example, if you connect a 12-gauge wire directly across a 120-V source (120-V mains outlet), what do you think will happen? What will happen when you do this to a 12-V dc supply, or to a 1.5-V battery?

Answer: In the 120-V mains case, you will likely cause a huge spark, possibly melting the wire and perhaps in the process receiving a nasty shock (if the wire isn't insulated). But more likely, your circuit breaker in the home will trip, since the wire will draw a huge current due to its low resistance—breakers trip when they sense a large level of current flowing into one of their runs. Some are rated at 10 A, others at 15 A, depending on setup. In a good dc supply, you will probably trip an internal breaker or blow a fuse, or in a bad supply, ruin the inner circuitry. In the case of a battery, there is internal resistance in the battery, which will result in heating of the battery. There will be less severe levels of current due to the internal resistance of the battery, but the battery will soon drain, possibly even destroying the battery, or in an extreme case causing the battery to rupture.

2.10 Grounds

As we left off in Sec. 2.3, we saw that understanding voltage is a relativity game. For example, to say that a point in a circuit has a voltage of 10 V is meaningless unless you have another point in the circuit with which to compare it. Often you define a point in the circuit to be a kind of 0-V reference point on which to base all other voltage measurements. This point is often called the *ground*, and is frequently represented by the symbol shown in Fig. 2.27:

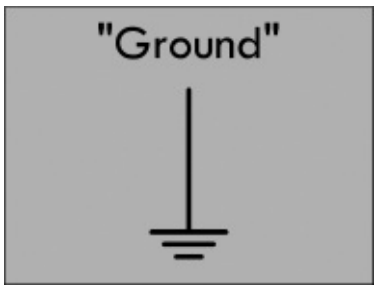


FIGURE 2.27

For example, Fig. 2.28 shows various ways in which to define voltages by selecting a ground—which in this case is simply a 0-V reference marker. The single battery provides a 1.5-V potential difference or voltage between its terminals. We can simply place a 0-V reference ground at the negative terminal and then state that the positive terminal sits at 1.5 V relative to the 0-V reference ground. The 0-V reference, or the negative terminal of the battery, is called the *return*. If a load, such as a lamp or resistor, is placed between the terminals, a load current will return to the negative terminal.

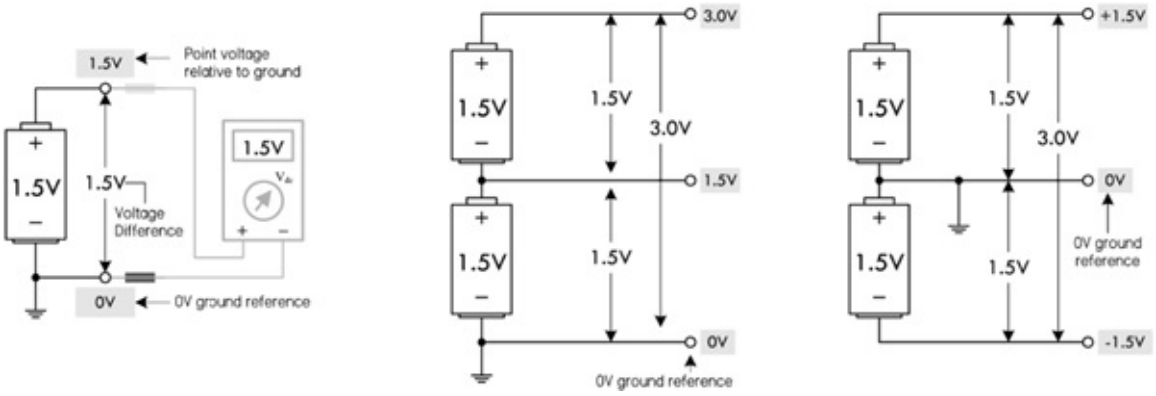


FIGURE 2.28

In the center diagram in Fig. 2.28, we have two 1.5-V batteries placed end to end. When batteries are linked this way, their voltages add, creating a combined total voltage of 3.0 V. With the 0-V reference ground at the bottom, we get 1.5-V and 3.0-V readings at the locations shown in the figure. A load placed across the two batteries (3.0-V difference) will result in a load current that returns to the lower battery's negative terminal. In this case, the return is through the 0-V reference—the lower battery's negative terminal.

Finally, it is possible to create a split supply by simply repositioning the 0-V ground reference, placing it between the batteries. This creates +1.5 V and −1.5 V leads relative to the 0-V reference. Many circuits require both positive and negative voltage relative to a 0-V ground reference. In this case, the 0-V ground reference acts as a *common return*. This is often necessary—say, in an audio circuit—where signals are sinusoidal and alternate between positive and negative voltage relative to a 0-V reference.

Now, the ground symbol shown in Fig. 2.27 as a 0-V reference, or as a return, is used all the time by various people. As it turns out, however, it really is supposed to represent a true earth ground—a physical connection to the earth through a conductive material buried in the earth. For whatever reasons, the symbol's dual meaning has survived, and this can often be a source of confusion to beginners.

2.10.1 Earth Ground

The correct definition of an *earth ground* is usually a connection terminated at a rod driven into the earth to a depth of 8 ft or more. This earth ground rod is wired directly to a mains breaker box's ground bar and sent to the various ac outlets in one's home via a green-coated or bare copper wire that is housed within the same mains cable as hot and neutral wires. The ground can then be accessed at the outlet at the ground socket. Metal piping buried in the earth is often considered an earth ground. See Fig. 2.29.

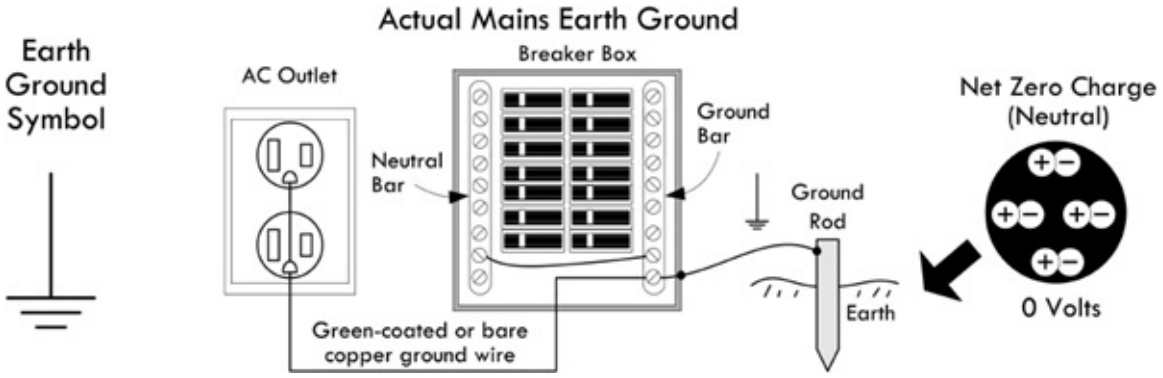


FIGURE 2.29 Earth ground.

A physical link to the earth is important because the earth provides an electrically neutral body; equal numbers of positive and negative charges are distributed through its entirety. Due to the earth's practically infinite charge neutrality, attempts at changing the earth potential, via electrical generators, batteries, static electrical mechanisms, or the like, will have essentially no measurable effect. Any introduction of new charge into the earth is quickly absorbed (the earth's moist soil is usually rather conductive). Such charge interactions occur constantly throughout the planet, and the exchanges average out to zero net charge.

For practical purposes, then, the earth is defined to be at a zero potential (relative to other things)—a potential that is practically immune to wavering. This makes the earth a convenient and useful potential on which to reference other signals. By connecting various pieces of electronics equipment to the earth ground, they can all share the earth's ground reference potential, and thus all devices share a common reference.

The actual physical connection to earth ground at a particular piece of equipment is usually through the power cord's ground wire that links to the mains ground wire network when the device is plugged in. The ground wire from the power cord is typically connected internally to the equipment chassis (frame) and, more important for our discussion, to the return portion of a channel that emanates from the interior circuitry. This is then brought out as a ground lead terminal. For example, in Fig. 2.30, an oscilloscope, function generator, and generic audiovisual device use BNC and UHF connectors for input and output channels. Internally, the outer connector body of the BNC or UHF jack is wired to the return (or source) portion of the channel, while a central conductor wire (insulated from the outer body) is wired to the source (or return) portion of the channel. The important part, now, is that the return, or outer, connector is also internally wired to the mains ground wire through the power cord cable. This sets the return to an earth ground reference. In the case of the dc power supply, a separate earth ground terminal is presented at the face in the form of a banana jack terminal. In order to ground the dc supply, a jumper wire must be connected between the negative supply terminal and the ground terminal. If no jumper is used, the supply is said to be *floating*.

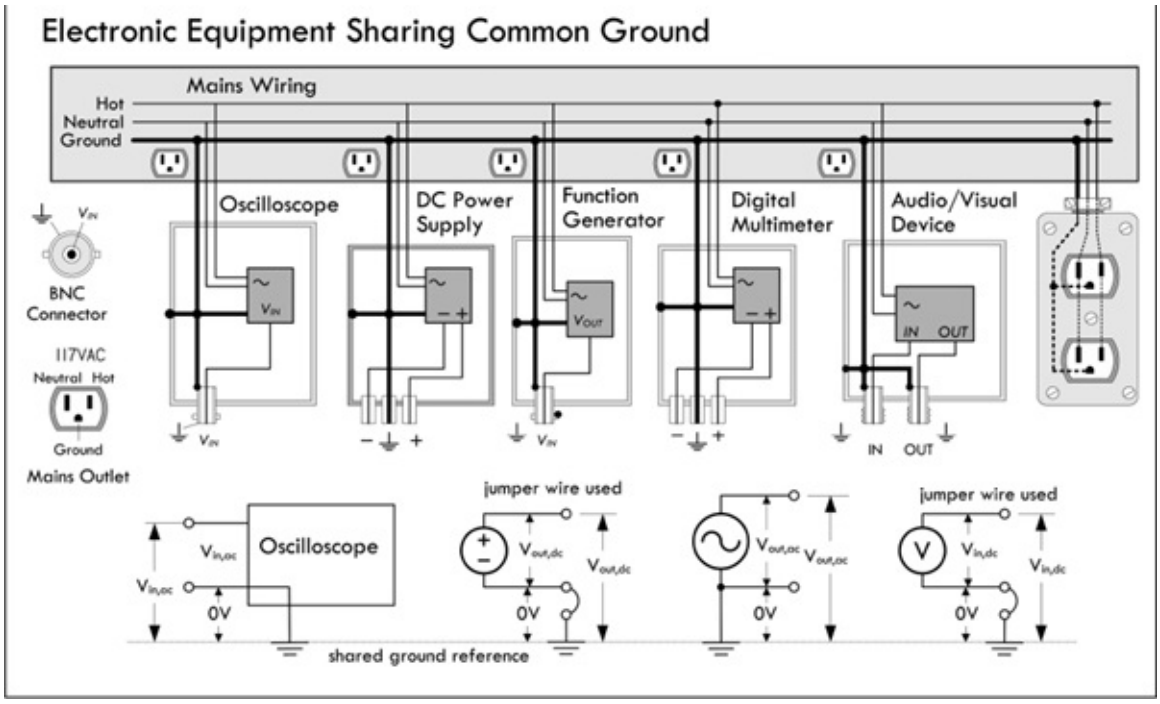


FIGURE 2.30 Illustration showing how various test instruments and an audiovisual device share a common ground connection through the mains ground wiring.

All the grounded pieces of equipment share a common ground. To prove this to yourself, try measuring the resistance between the ground terminals of any two separate pieces of test equipment in your lab. If the devices are properly grounded, you will get a measurement of 0 Ω (but a bit more for internal resistances).

Besides acting simply as a reference point, grounding reduces the possibility of electrical shock if something within a piece of equipment should fail and the chassis or cabinet becomes “hot.” If the chassis is connected to a properly grounded outlet via a three-wire electrical system ground, the path of current flow from the hot chassis will be toward ground, not through your body (which is a more resistive path). A ground system to prevent shock is generally referred to as a dc ground. We’ll discuss shock hazards and grounding later, when we cover ac.

Grounding also helps eliminate electrostatic discharge (ESD) when a statically charged body comes in contact with sensitive equipment. The charged body could be you, after a stroll you took across the carpet. Some ICs are highly vulnerable to damage from ESD. By providing a grounded work mat or using a grounded wrist strap while working with sensitive ICs, you can avoid destroying your chips by ensuring that charge is drained from your body before you touch anything.

Another big job the ground system does is provide a low-impedance path to ground for any stray RF current caused by stray radiofrequency-producing devices, such as electrical equipment, radio waves, and so on. Stray RF can cause equipment to malfunction and contributes to RFI problems. This low-impedance path is usually called RF ground. In most cases, the dc ground and the RF ground are provided by the same system.

Common Grounding Error

As previously mentioned, the ground symbol, in many cases, has been used as a generic symbol in circuit diagrams to represent the current return path, even though no physical earth ground is used. This can be confusing for beginners when they approach a three-terminal dc power supply that has a positive (+), negative (-), and ground terminal. As we have learned, the ground terminal of the supply is tied to the case of the instrument, which in turn is wired to the mains earth ground system. A common mistake for a novice to make is to attempt to power a load, such as a lamp, using the positive and ground terminals of the supply, as shown in Fig. 2.31a. This, however, doesn’t complete a current return path to the energy source

(supply), so no current will flow from the source; hence, the load current will be zero. The correct procedure, of course, is to either connect the load between the positive and negative terminals directly, thus creating a *floating load*, or, using a jumper wire between the ground and negative supply, create a *grounded load*. Obviously, many dc circuits don't need to be grounded—it will generally neither help nor hinder performance (e.g., battery-powered devices need no such connection).

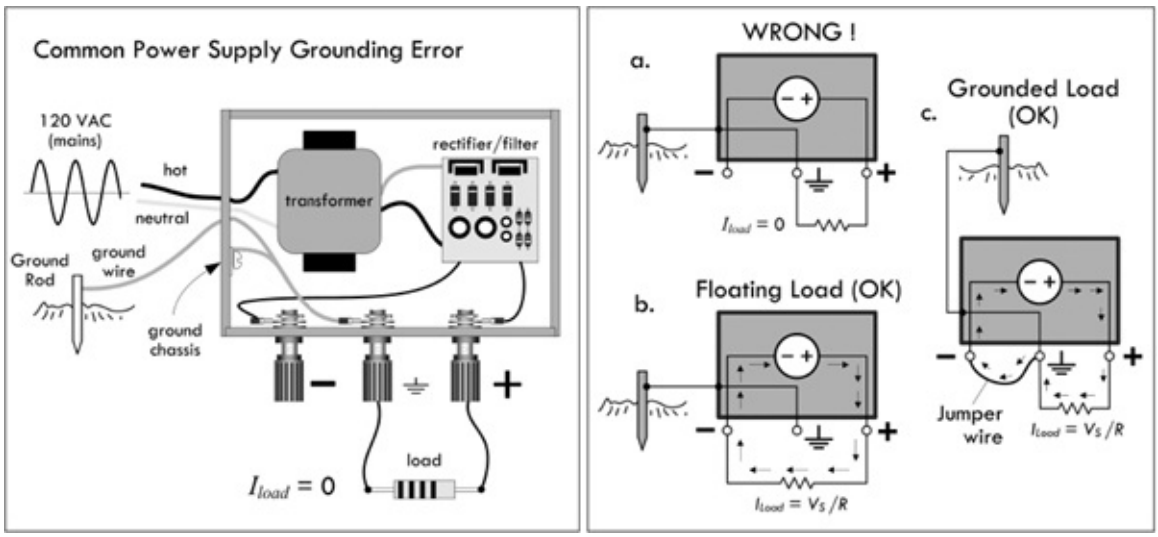


FIGURE 2.31

Circuits that require both positive and negative voltage require a power supply to provide each polarity. The supply for the positive voltage will have the negative terminal as a return, and the negative supply will have the positive terminal as the return. These two terminals are connected together, forming a common return path for load current, as shown in Fig. 2.32. The connection between the negative and the positive terminals of the supplies results in a *common* or *floating return*. The floating common may be connected to the earth ground terminal of a supply, if a particular circuit requests this. Generally, it will neither help nor hinder circuit performance.

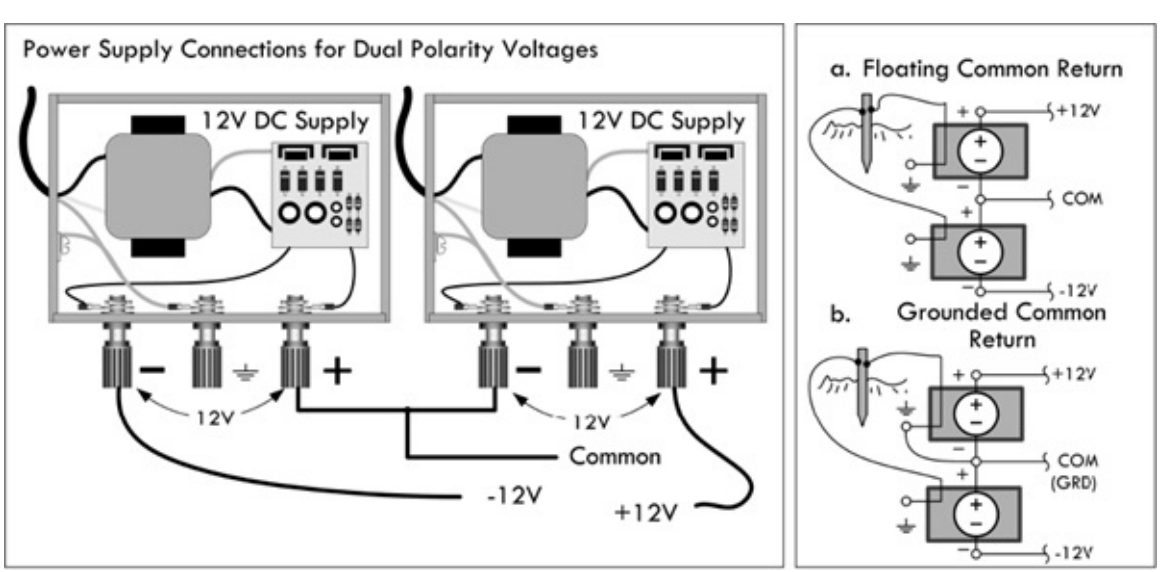


FIGURE 2.32

Unfortunately, the earth ground symbol is used a bit too loosely in electronics, often meaning different things to different people. It is used as a 0-V reference, even though no actual connection is made to Earth. Sometimes it actually means to connect a point in a circuit to earth ground. Sometimes it is used as a generic return—to eliminate the need to draw a return wire. It could be used as an actual earth ground return (using the mains ground copper wire), though this is unwise. See Fig. 2.33. To avoid

complications, alternative symbols are used, which we discuss next.

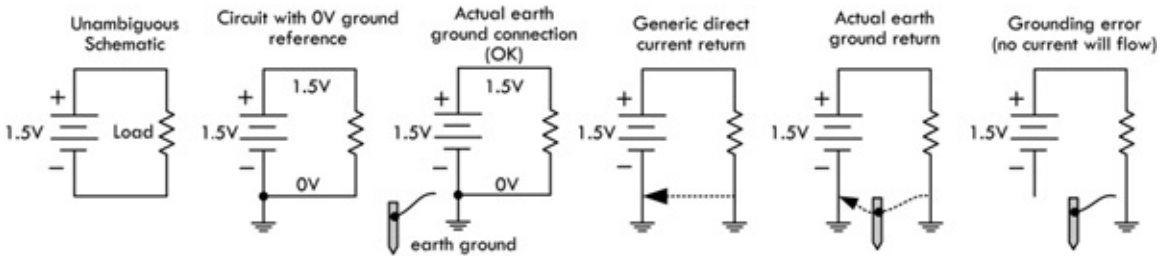


FIGURE 2.33

2.10.2 Different Types of Ground Symbols

To avoid misinterpretations regarding earth grounding, voltage references, and current return paths, less ambiguous symbols have been adopted. Figure 2.34 shows an earth ground (could be Earth or reference), a frame or chassis ground, and a digital and analog reference ground. Unfortunately, the common return for digital and analog is a bit ambiguous, too, but usually a circuit diagram will specify what the symbol is referring to.

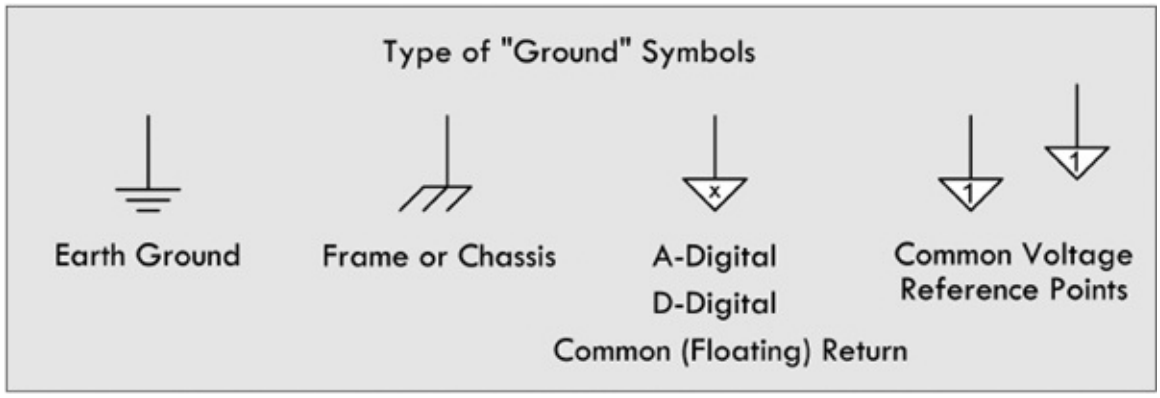
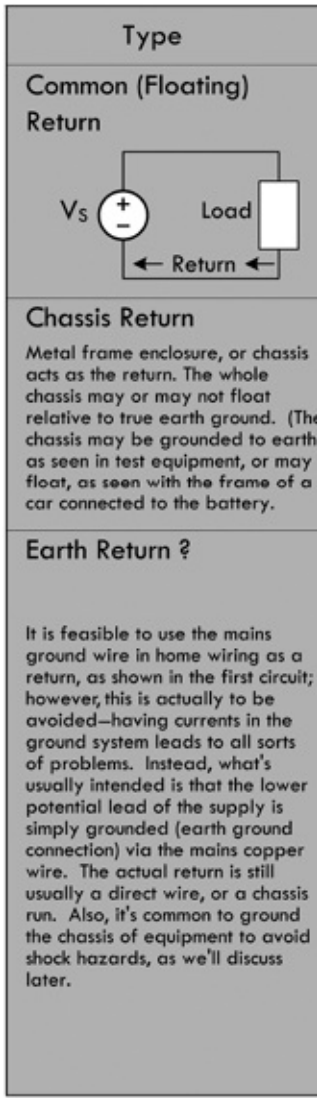


FIGURE 2.34

Table 2.6 provides a rundown on meanings behind these symbols.

TABLE 2.6 Types of Ground Symbols



2.10.3 Loose Er

There are a few loos
Fig. 2.35.

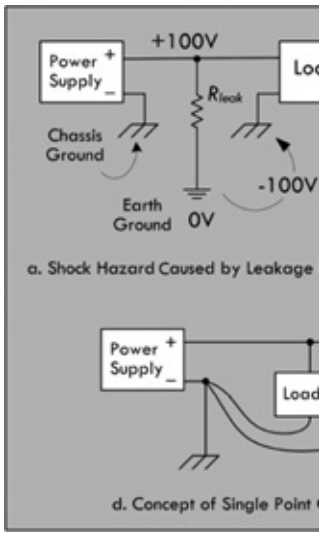


FIGURE 2.35

Shock Hazard

In instances where l

Example 2: A 47-µF capacitor is charged by the voltage sources having the following waveform. Determine the charging current. Refer to Fig. 2.100. Assume that the voltage source is ideal and has no resistance.

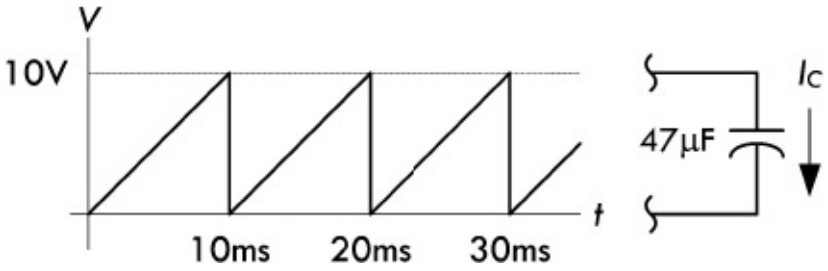


FIGURE 2.100

Answer: Since *dV*/*dt* represents the slope of the waveform, it is simply 10 V/10 ms, and the current becomes:

$$I_C = C\frac{dV_C}{dt} = (47\times10^{-6}\text{ F})\frac{10\text{ V}}{10\times10^{-3}\text{ s}} = 0.047\text{ A} = 47\text{ mA}$$

Example 3: If the voltage across a 100-µF capacitor is 5 V e^{-t}, what is the capacitor current?

Answer:

$$I_C = C\frac{dV_C}{dt} = 100\ \mu\text{F}\frac{d}{dt}[5\text{ V}e^{-t}] = -(100\ \mu\text{F})(5\text{ V})e^{-t} = -(0.0005\text{ A})e^{-t}\text{ V}$$

(Remember, these examples assume ideal capacitors. If you used real capacitors, the results would follow the same trends, but would be limited in current.)

2.23.7 Energy in a Capacitor

Finally, energy cannot be dissipated in an ideal capacitor. (This is not the case for real capacitors because of internal resistance, but since the internal resistance is so small, that energy lost to heating is often ignored.) Energy can only be stored in the electric field (or potential that exists between the plates) for later recovery. The energy stored in a capacitor is found by substituting the capacitor current into the generalized power law (*P* = *IV*), then inserting the resulting power into the definition of power (*P* = *dE*/*dt*), and solving for *E* by integration:

$$E_{cap} = \int VIdt = \int VC\frac{dV}{dt}dt = \int CVdV = \frac{1}{2}CV^2 \tag{2.42}$$

Example: How much energy is stored in a 1000-µF capacitor with an applied voltage of 5 V?

Answer:

$$E_{cap} = \frac{1}{2}CV^2 = \frac{1}{2}(1000\times10^{-6}\text{ F})(5\text{ V})^2 = 0.0125\text{ J}$$

2.23.8 RC Time Constant

When a capacitor is connected to a dc voltage source, it will charge up almost instantaneously. (In reality, a capacitor has internal resistance, as well as inductance, therefore the term "almost"—see Sec. 3.6, on capacitors, for real-life characteristics.) Likewise, a charged capacitor that is shorted with a wire will discharge almost instantaneously. However, with some resistance added, the rate of charge or discharge follows an exponential pattern, as shown in Fig. 2.101. There are numerous applications that use controlled charge and discharge rates, such as timing ICs, oscillators, waveform shapers, and low-discharge power backup circuits.

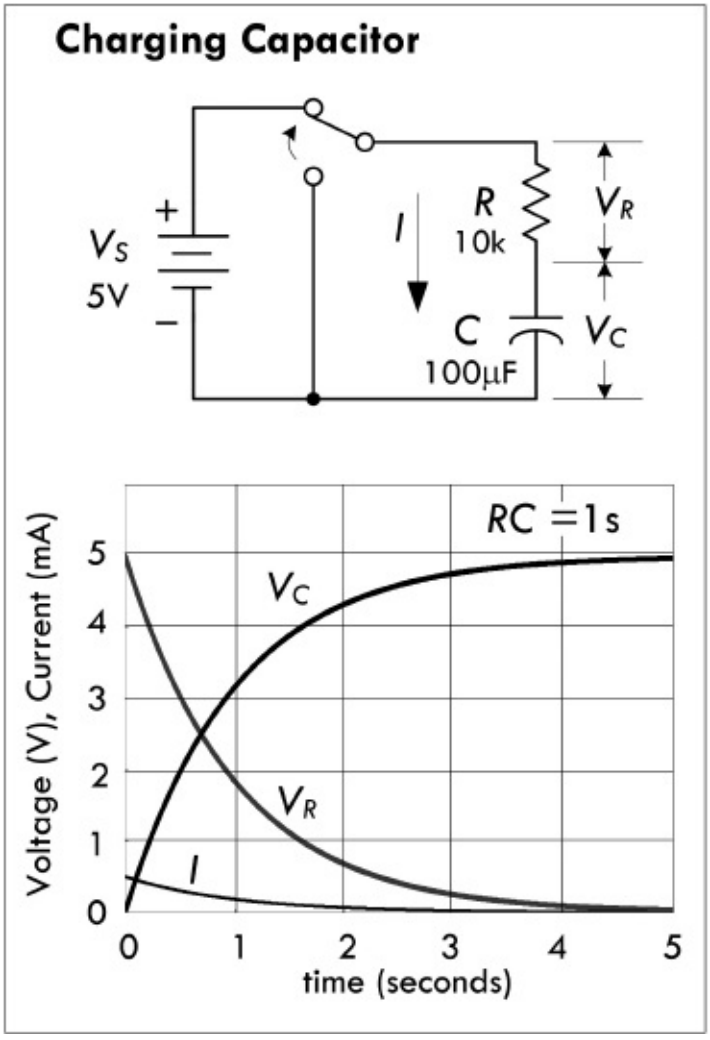


FIGURE 2.101

For a charging capacitor, the following equations are used.

Current and voltage equations for RC charging circuit

$$I = \frac{V_s}{R} e^{-t/RC} \qquad \frac{t}{RC} = -\ln\left(\frac{IR}{V_s}\right)$$

$$V_R = IR = V_s e^{-t/RC} \qquad \frac{t}{RC} = -\ln\left(\frac{V_R}{V_s}\right)$$

$$V_C = \frac{1}{C}\int I dt = V_s(1 - e^{-t/RC}) \qquad \frac{t}{RC} = -\ln\left(\frac{V_s - V_C}{V_s}\right)$$

$$\tau = RC \text{ time constant}$$

(2.43)

where I is the current in amps, V_S is the source voltage in volts, R is the resistance in ohms, C is the capacitance in

farads, t is the time in seconds after the source voltage is applied, e = 2.718, V_R is the resistor voltage in volts and V_C is the capacitor voltage in volts. Graph shown to the left is for circuit with R = 10 kΩ, and C = 100 µF. Decreasing the resistance means the capacitor charges up more quickly and the voltage across the capacitor rises more quickly.

(You can derive the preceding expressions using Kirchhoff's law, by summing the voltage around the closed loop: $V_S = RI + (1/C) \int Idt$. Differentiating and solving the differential equation, given an initial condition of current equal to V/R, the voltage across the resistor = V_S, and the voltage across the capacitor $V_C = 0$, you get the solution: $I = (V/R)e^{-t/RC}$. The voltages across the resistor and capacitor are simply found by plugging the current into the expression for the voltage across a resistor $V_R = IR$ and the voltage across a capacitor: $V_C = (1/C) \int Idt$. We'll discuss solving such circuits in detail in the section on transients in dc circuits.)

Theoretically, the charging process never really finishes, but eventually the charging current drops to an immeasurable value. A convention often used is to let $t = RC$, which makes $V(t) = 0.632$ V. The RC term is called the *time constant* of the circuit and is the time in seconds required to charge the capacitor to 63.2 percent of the supply voltage. The lowercase tau τ is often used to represent RC: $\tau = RC$. After two time constants ($t = 2RC = 2\tau$) the capacitor charges another 63.2 percent of the difference between the capacitor voltage at one time constant and the supply voltage, for a total change of 86.5 percent. After three time constants, the capacitor reaches 95 percent of the supply voltage, and so on, as illustrated in the graph in Fig. 2.101. After five time constants, a capacitor is considered fully charged, having reached 99.24 percent of the source voltage.

Example: An IC uses an external RC charging network to control its timing. The IC requires 3.4 V at V_{IN} to trigger its output to switch from high to low, at which time an internal transistor (switch) is activated, allowing the capacitor to be discharged to ground. If a 5-s timing period before the trigger point is required, what value of R is required if C = 10 µF?

Answer: Use

$$\frac{t}{RC} = -\ln\left(\frac{V_s - V_C}{V_s}\right)$$

and solve for R:

$$R = \frac{t}{-\ln\left(\frac{V_s - V_C}{V_s}\right)C} = \frac{5.0\text{ s}}{-\ln\left(\frac{5\text{V} - 3.4\text{ V}}{5\text{ V}}\right)(10 \times 10^{-6}\text{ C})}$$

$$= 4.38 \times 10^5\ \Omega$$

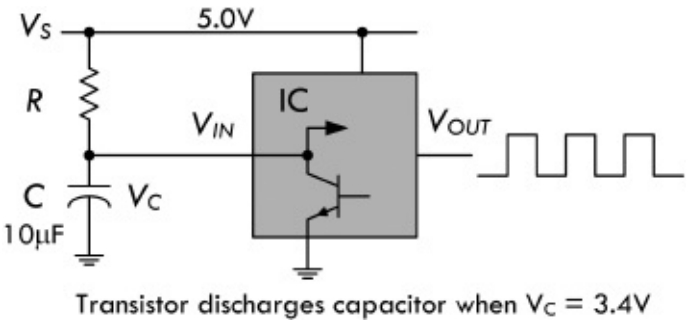


FIGURE 2.102

For a discharging capacitor, the following equations are used.

Current and voltage equations for discharging an RC circuit

$$I = \frac{V_s}{R} e^{-t/RC}, \qquad \frac{t}{RC} = -\ln\left(\frac{IR}{V_s}\right)$$

$$V_R = IR = V_s e^{-t/RC}, \qquad \frac{t}{RC} = -\ln\left(\frac{V_R}{V_s}\right)$$

$$V_C = \frac{1}{C}\int_0 I dt = V_s e^{-t/RC}, \qquad \frac{t}{RC} = -\ln\left(\frac{V_C}{V_s}\right)$$

$$\tau = RC \text{ time constant}$$

(2.44)

where I is the current in amps, V_S is the source voltage in volts, R is the resistance in ohms, C is the capacitance in farads, t is the time in seconds after the source voltage is removed, e = 2.718, V_R is the resistor voltage in volts, and V_C is the capacitor voltage in volts. Graph shown to the left in Fig. 2.103 is for circuit with R = 3 kΩ and C = 0.1 µF.

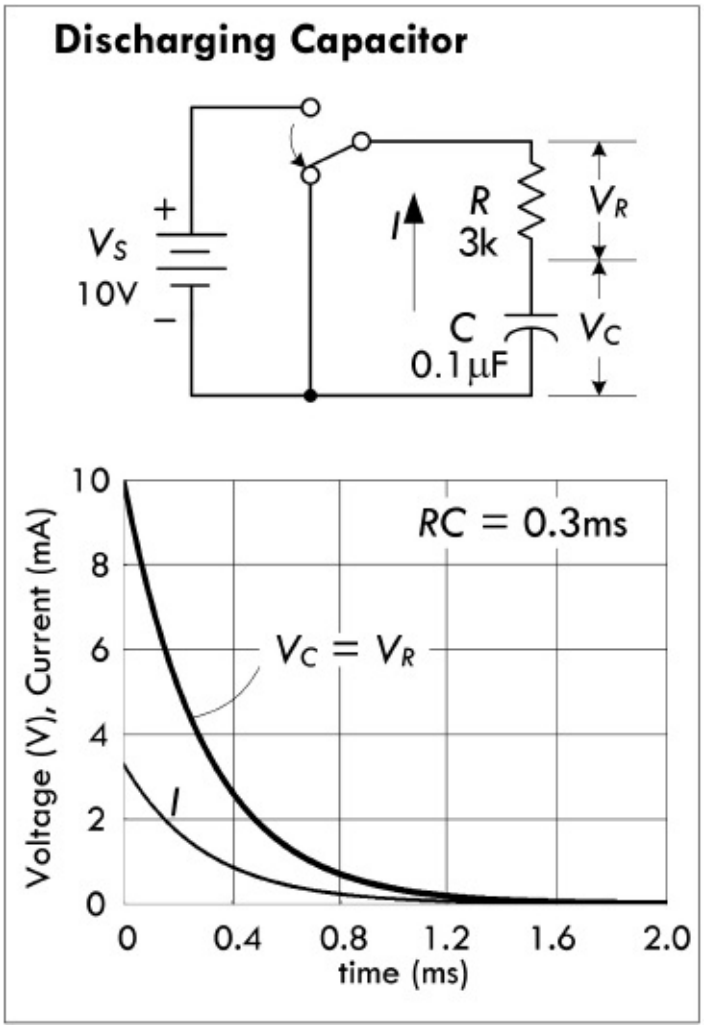


FIGURE 2.103

(You can derive this expression using Kirchhoff's law, by summing the voltage around the closed loop: $0 = RI + (1/C)\int Idt$. Differentiating and solving the differential equation, given an initial condition of current equal to 0, the voltage across the resistor = 0, and the voltage across the capacitor $V_C = V_S$, you get the solution: $I = (V/R)e^{-t/RC}$. The voltage across the resistor and capacitor is found simply by plugging the current into the expression for the voltage across a resistor $V_R = IR$ and the voltage across a capacitor: $V_C = (1/C)\int Idt$. We'll discuss solving such circuits in detail in the section on transients in dc circuits.)

The expression for a discharging capacitor is essentially the inverse of that for a charging capacitor. After one time constant, the capacitor voltage will have dropped by 63.2 percent from the supply voltage, so it will have reached 37.8 percent of the supply voltage. After five time constants, the capacitor is considered fully discharged, it will have dropped 99.24 percent, or down to 0.76 percent of the supply voltage.

Example: If a 100-µF capacitor in a high-voltage power supply is shunted by a 100k resistor, what is the minimum time before the capacitor is considered fully discharged when power is turned off?

Answer: After five time constants, a capacitor is considered discharged:

$$t = 5\tau = 5RC = (5)(100 \times 10^3\ \Omega)(100 \times 10^{-6}\ \text{F}) = 50\ \text{s}$$

2.23.9 Stray Capacitance

Capacitance doesn't exist only within capacitors. In fact, any two surfaces at different electrical potential, and that are close enough together to generate an electric field, have capacitance, and thus act like a capacitor. Such effects are often present within circuits (e.g., between conductive runs or component leads), even though they are not intended. This unintended capacitance is referred to as *stray capacitance*, and it can result in a disruption of normal current flow within a circuit. Designers of electric circuits figure out ways to minimize stray capacitance as much as possible, such as keeping capacitor leads short and grouping components in such a way as to eliminate capacitive coupling. In high-impedance circuits, stray capacitance may have a greater influence, since capacitive reactance (which we will discuss shortly) may be a greater portion of the circuit impedance. In addition, since stray capacitance usually appears in parallel with a circuit, it may bypass more of the desired signal at higher frequencies. Stray capacitance usually affects sensitive circuits more profoundly.

2.23.10 Capacitors in Parallel

When capacitors are placed in *parallel*, their capacitances add, just like resistors in series:

$$C_{\text{tot}} = C_1 + C_2 + \ldots C_n \tag{2.45}$$

(You derive this formula by applying Kirchhoff's current law at the top junction, which gives $I_{\text{tot}} = I_1 + I_2 + I_3 + \ldots I_N$. Making use of the fact that the voltage V is the same across C_1 and C_2, you can substitute the displacement currents for each capacitor into Kirchhoff's current expression as follows:

$$I = C_1\frac{dV}{dt} + C_2\frac{dV}{dt} + C_3\frac{dV}{dt} = (C_1 + C_2 + C_3)\frac{dV}{dt}$$

The term in brackets is the equivalent capacitance.)

Intuitively, you can think of capacitors in parallel representing one single capacitor with increased plate surface area. It's important to note that the largest voltage that can be applied safely to a group of capacitors in parallel is limited to the voltage rating of the capacitor with the lowest voltage rating. Both the capacitance and the voltage rating are usually included next to the capacitor symbol in schematics, but often the voltage rating is missing; you must figure out the rating based on the expected voltages present at

that point in the circuit.

2.23.11 Capacitors in Series

When two or more capacitors are connected in series, the total capacitance is less than that of the smallest capacitor in the group. The equivalent capacitance is similar to resistors in parallel:

$$\frac{1}{C_{tot}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots \frac{1}{C_N} \tag{2.46}$$

(You derive this by applying Kirchhoff's voltage law. Since the current I must flow through each capacitor, Kirchhoff's voltage expression becomes:

$$V = \frac{1}{C_1}\int Idt + \frac{1}{C_2}\int Idt + \frac{1}{C_3}\int Idt = \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots + \frac{1}{C_N}\right)\int Idt$$

The term in parentheses is called the equivalent capacitance for capacitors in series.)

Capacitors may be connected in series to enable the group to withstand a larger voltage than any individual capacitor is rated to withstand (the maximum voltage ratings add). The trade-off is a decrease in total capacitance—though that could be what you intend to do, if you can't find a capacitor or create a parallel arrangement that gives you the desired capacitance value. Notice in Fig. 2.104 that the voltage does not divide equally between capacitors. The voltage across a single capacitor—say, C_2—is a fraction of the total, expressed as $(C_{total}/C_2)V_{IN}$. There are various circuits that tap the voltage between series capacitors.

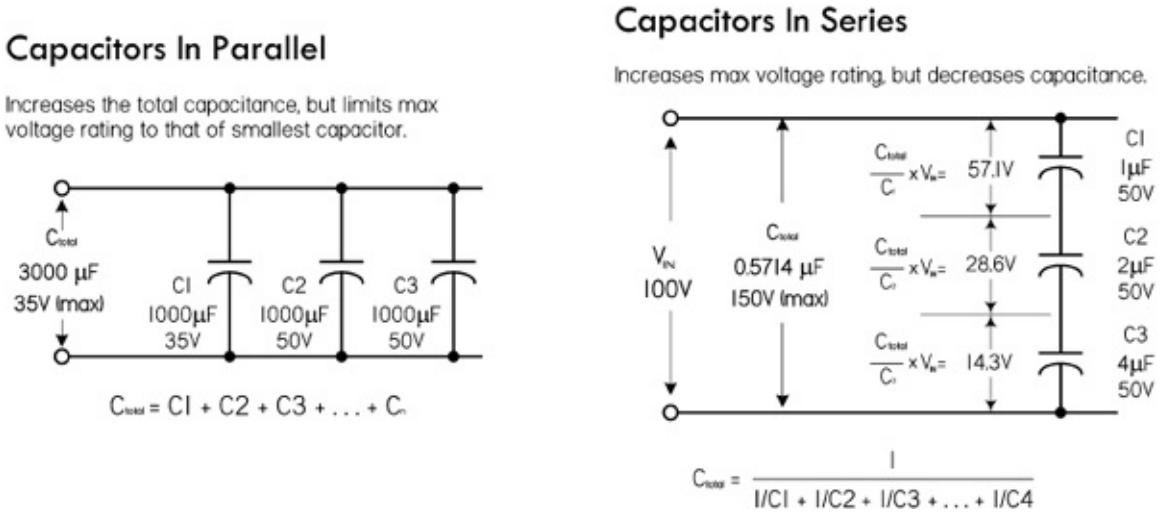
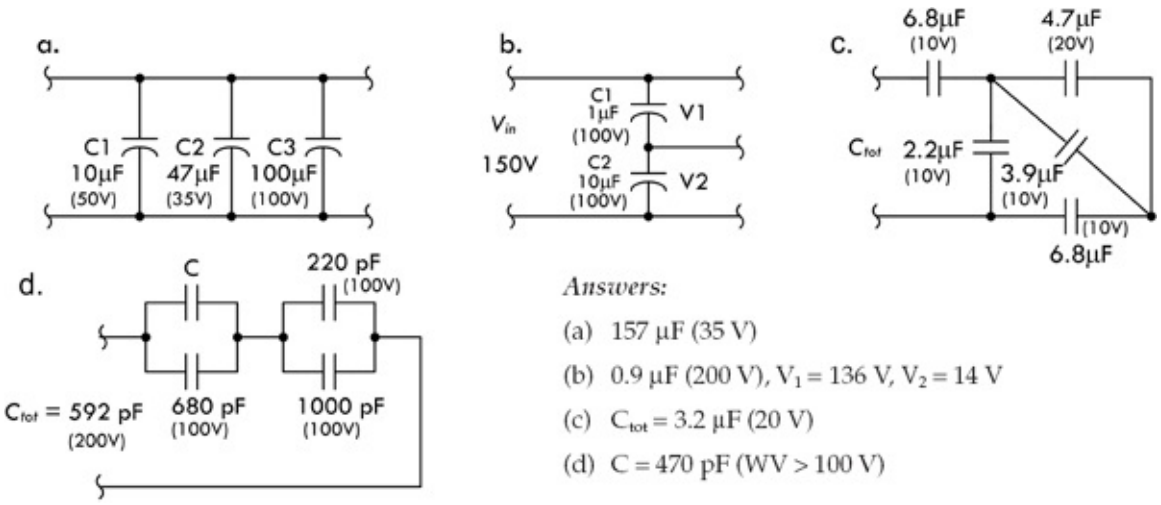


FIGURE 2.104

Use care to ensure that the voltage rating of any capacitor in the group is not exceeded. If you use capacitors in series to withstand larger voltages, it's a good idea to also connect an equalizing resistor across each capacitor. Use resistors with about 100 Ω per volt of supply voltage, and be sure they have sufficient power-handling capability. With real capacitors, the leakage resistance of the capacitor may have more effect on the voltage division than does the capacitance. A capacitor with a high parallel resistance will have the highest voltage across it. Adding equalizing resistors reduces this effect.

Example: (a) Find the total capacitance and maximum working voltage (WV) for the capacitor in the

parallel network. (b) Find the total capacitance, WV, V_1, and V_2. (c) Find the total capacitance and WV for the network of capacitors. (d) Find the value of C that gives a total capacitance of 592 pF with a total WV of 200 V. Individual capacitor WV values are in parentheses. Refer to Fig. 2.105.



Answers:

(a) 157 µF (35 V)

(b) 0.9 µF (200 V), $V_1 = 136$ V, $V_2 = 14$ V

(c) $C_{tot} = 3.2$ µF (20 V)

(d) C = 470 pF (WV > 100 V)

FIGURE 2.105

2.23.12 *Alternating Current in a Capacitor*

Everything that was discussed about capacitors in dc circuits also applies in ac circuits, with one major exception. While a capacitor in a dc circuit will block current flow (except during brief periods of charging and discharging), a capacitor in an ac circuit will either pass or limit current flow, depending on frequency. Unlike a resistor that converts current energy into heat to reduce current flow, a capacitor stores electrical energy and returns it to the circuit.

The graph in Fig. 2.106 shows the relationship between current and voltage when an ac signal is applied to a capacitor. The ac sine wave voltage has a maximum value of 100. In interval 0 to A, the applied voltage increases from 0 to 38 and the capacitor charges up to that voltage. In interval AB, the voltage increases to 71, so the capacitor has gained 33 V more. During this interval, a smaller amount of charge has been added than 0A, because the voltage rise during AB is smaller than 0A. In interval BC, the voltage rises to 21, from 71 to 92. The increase in voltage in this interval is still smaller. In interval CD, the voltage increases only 8, and therefore the increase in current is also smaller.

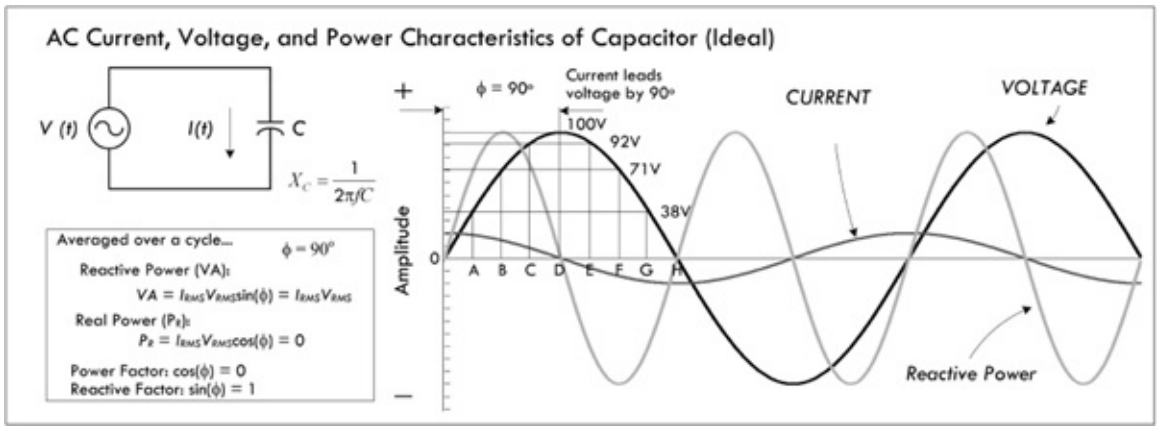


FIGURE 2.106

If you were to divide the first quarter cycle into a very large number of intervals, you'd see that the current charging the capacitor has the shape of a sine wave, just like the applied voltage. The current is largest at the start of the cycle and goes to zero at the maximum value of voltage, so there is a phase

difference of 90° between the voltage and the current.

In the second quarter cycle, from time D to H, the applied voltage decreases and the capacitor loses its charge. Using the same reasoning as before, it's apparent that the current is small during period DE and continues to increase during the other periods. The current is flowing against the applied voltage; however, since the capacitor is discharging into the circuit, the current flows in the negative direction during this quarter cycle.

The third and fourth quarter cycles repeat the events of the first and second, respectively, with one difference: the polarity of the applied voltage has reversed, and current changes to correspond. In other words, an alternating current flows in the circuit because of the alternate charging and discharging of the capacitor. As shown in the graph in Fig. 2.106, the current starts its cycle 90° before the voltage, so the current in a capacitor leads the applied voltage by 90°.

2.23.13 Capacitive Reactance

The amount of charge on a capacitor is equal to the voltage drop across the capacitor times the capacitance ($Q = CV$). Within an ac circuit, the amount of charge moves back and forth in the circuit every cycle, so the rate of movement of charge (current) is proportional to voltage, capacitance, and frequency. When the effect of capacitance and frequency are considered together, they form a quantity similar to resistance. However, since no actual heat is being generated, the effect is termed *capacitive reactance*. The unit for reactance is the ohm, just as for resistors, and the formula for calculating the reactance of a capacitor at a particular frequency is given by:

$$X_C = \frac{1}{2\pi fC} = \frac{1}{\omega C}$$

(2.47) Capacitive reactance

where X_C is the capacitive reactance in ohms, f is the frequency in Hz, C is the capacitance in farads, and $\pi = 3.1416$. Often, omega ω is used in place of $2\pi f$, where ω is called the *angular frequency*.

(You can derive this by noting that a sinusoidal voltage source placed across a capacitor will allow displaced current to flow through it because the voltage across it is changing (recall that $I = CdV/dt$ for a capacitor). For example, if the voltage source is given by $V_0 \cos(\omega t)$, you plug this voltage into V in the expression for the displacement current for a capacitor, which gives:

$$I = C\frac{dV}{dt} = -\omega C V_0 \sin(\omega t)$$

Maximum current or peak current I_0 occurs when $\sin(\omega t) = -1$, at which point $I_0 = \omega C V_0$. The ratio of peak voltage to peak current V_0/I_0 resembles a resistance in light of Ohm's law, and is given in units of ohms. However, because the physical phenomenon for "resisting" is different from that of a traditional resistor (heating), the effect is given the name capacitive reactance.)

As the frequency goes to infinity, X_C goes to 0, and the capacitor acts like a short (wire) at high frequencies; capacitors like to pass current at high frequencies. As the frequency goes to 0, X_C goes to infinity, and the capacitor acts like an open circuit; capacitors do not like to pass low frequencies.

It's important to note that even though the unit of reactance is the ohm, there is no power dissipated in reactance. The energy stored in the capacitor during one portion of the cycle is simply returned to the circuit in the next. In other words, over a complete cycle, the average power is zero. See the graph in Fig. 2.106.

Example 1: Find the reactance of a 220-pF capacitor at an applied frequency of 10 MHz.

Answer:

$$X_C = \frac{1}{2\pi \times (10 \times 10^6 \text{ Hz})(220 \times 10^{-12} \text{ F})} = 72.3\ \Omega$$

(*Note:* 1 MHz = 1 × 10^6 Hz, 1 μF = 1 × 10^{-6}, 1 nF = 1 × 10^{-9} F, 1 pF = 1 × 10^{-12} F.)

Example 2: What is the reactance of a 470-pF capacitor at 7.5 MHz and 15.0 MHz?

Answer: X_C @ 7.5 MHz = 45.2 Ω, X_C @ 15 MHz = 22.5 Ω

As you can see, the reactance decreases with an increase in frequency and/or an increase in capacitance. The left graph in Fig. 2.107 shows the reactance versus frequency of a capacitor. Real capacitors don't follow the graph and equation so precisely, a result of parasitic effects—see "Real Capacitor Models" in Fig. 2.93.

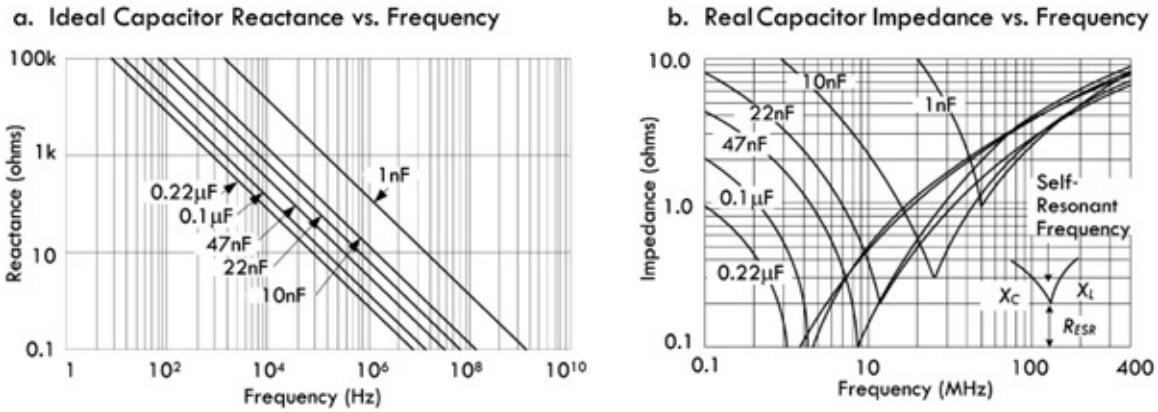
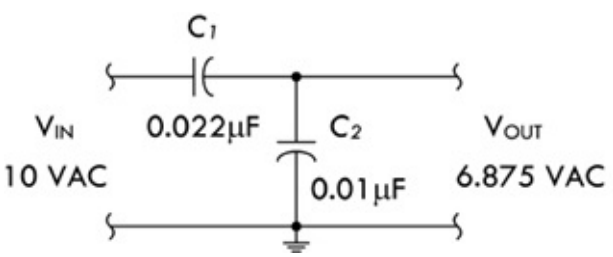


FIGURE 2.107 (a) Graph showing how reactance increases with frequency for various sizes of capacitors—all capacitors are considered ideal in nature. (b) Graph showing the frequency response for real-life capacitors, which takes into account the parasitic resistances and inductances present within a real capacitor package. The pointy dips within the graph represent self-resonance, where the capacitive and inductive reactances cancel and only internal resistance within the capacitor package is left. The frequency at which point this occurs is called the self-resonant frequency.

2.23.14 Capacitive Divider

Capacitive dividers can be used with ac input signals or even dc, since the capacitors will rapidly reach a steady state. The formula for determining the ac output voltage of a capacitive divider is different from the resistive divider, because the series element, C_1 is in the numerator, not C_2, the shunt element. See Fig. 2.108.



$$V_{out} = \frac{C_1}{C_1 + C_2} V_{in}$$

$$V_{out} = \frac{0.022\mu F}{0.032\mu F}(10\ \text{VAC}) = 6.875\ \text{VAC}$$

FIGURE 2.108

Note that the output voltage is independent of the input frequency. However, if the reactance of the

capacitors is not large at the frequency of interest (i.e., the capacitance is not large enough), the output current capability will be very low.

2.23.15 Quality Factor

Components that store energy, like a capacitor (and as we'll see, an inductor), may be compared in terms of *quality factor Q*, also known as the merit. The *Q* of any such component is the ratio of its ability to store energy to the sum total of all energy losses within the component. Since reactance is associated with stored energy and resistance is associated with energy loss, we can express the quality factor as:

$$Q = \frac{\text{Reactance}}{\text{Resistance}} = \frac{X}{R} \tag{2.48}$$

Q has no units. When considering a capacitor, the reactance (in ohms) is simply the capacitive reactance $X = X_C$. (For an inductor, we'll see that $X = X_L$, where X_L is the inductive reactance.) *R* is the sum of all resistances associated with the energy losses in the component (in ohms). The *Q* of capacitors is ordinarily high. Good-quality ceramic capacitors and mica capacitors may have *Q* values of 1200 or more. Small ceramic trimmer capacitors may have *Q* values too small to ignore in some applications. Microwave capacitors can have poor *Q* values, 10 or less at 10 GHz or higher.

2.24 Inductors

In the preceding section we saw how a capacitor stored electrical energy in the form of an electric field. Another way to store electrical energy is in a magnetic field. Circular radiating magnetic fields can be generated about a wire any time current passes through it. Increasing or decreasing current flow through the wire increases and decreases the magnetic field strength, respectively. During such changes in magnetic field strength, we encounter a phenomenon known as *inductance*. Inductance is a property of circuits somewhat analogous to resistance and capacitance; however, it is not attributed to heat production or charge storage (electric field), but rather it is associated with magnetic fields—more specifically, how changing magnetic fields influence the free electrons (current) within a circuit. Theoretically, any device capable of generating a magnetic field has inductance. Any device that has inductance is referred to as an inductor. To understand inductance requires a basic understanding of electromagnetic properties.

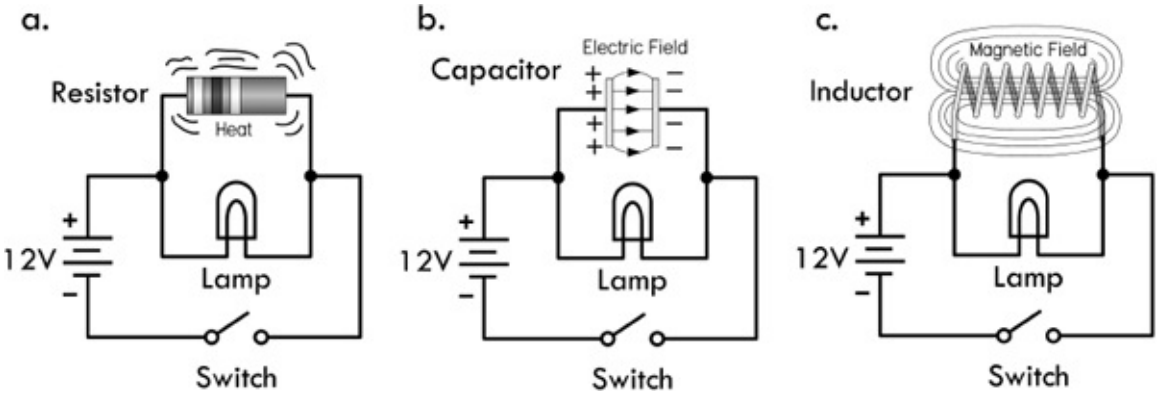


FIGURE 2.109 The three cornerstones of electronics are resistance, capacitance, and inductance. Inductance, unlike the other two, involves alternations in a circuit's current and voltage characteristics as a result of forces acting upon free electrons resulting in the creation and collapse of magnetic fields, usually concentrated in a discrete inductor device. Like a capacitor, however, inductive effects occur only during times of change, when the applied voltage/current increases or decreases with time. Resistance doesn't have a time dependency. Can you guess what will happen to the brightness of the lamp in each of the

circuits in the figure when the switch is closed? What do you think will occur when the switch is later opened? We'll discuss this a bit later.

2.24.1 Electromagnetism

According to the laws of electromagnetism, the field of a charge at rest can be represented by a uniform, radial distribution of electric field lines or lines of force (see Fig. 2.110a). For a charge moving at a constant velocity, the field lines are still radial and straight, but they are no longer uniformly distributed (see Fig. 2.110b). At the same time, the electron generates a circular magnetic field (see Fig. 2.110c). If the charge accelerates, things get a bit more complex, and a "kink" is created in the electromagnetic field, giving rise to an electromagnetic wave that radiates out (see Fig. 2.110d and e).

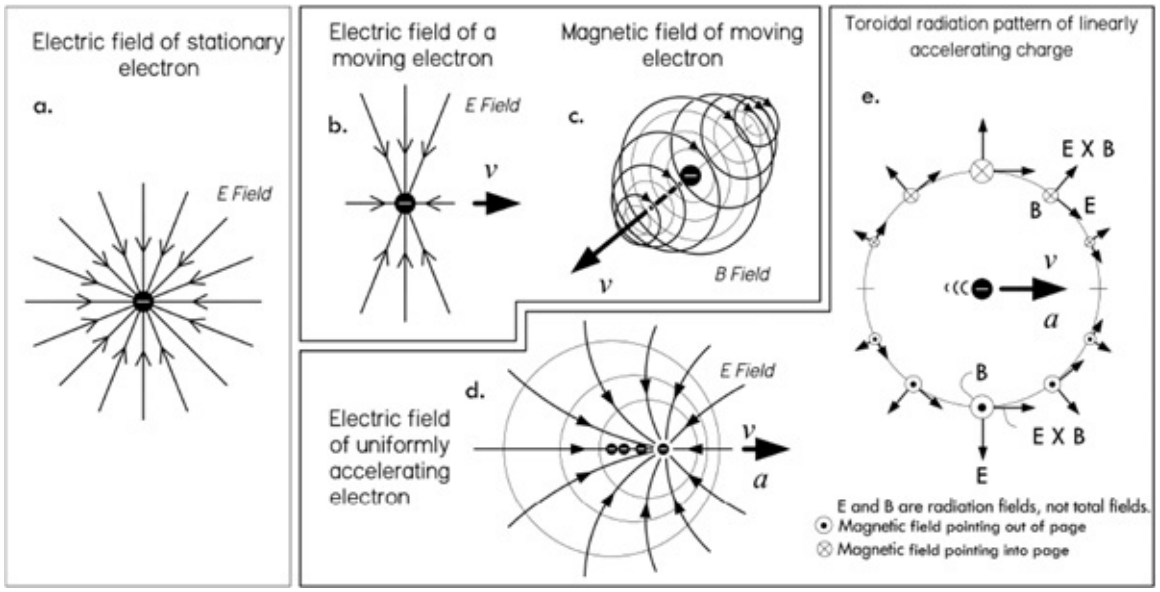


FIGURE 2.110 The electric and magnetic fields are part of the same phenomenon called electromagnetism. Magnetic fields appear whenever a charge is in motion. Interestingly, if you move along with a moving charge, the observable magnetic field would disappear—thanks to Einstein's relativity.

As depicted in Fig. 2.110c, the electric field (denoted *E*) of a moving electron—or any charge for that matter—is, in effect, partially transformed into a magnetic field (denoted *B*). Hence, it is apparent that the electric and magnetic fields are part of the same phenomenon. In fact, physics today groups electric and magnetic fields together into one fundamental field theory, referred to as *electromagnetism*. (The work of Maxwell and Einstein helped prove that the two phenomena are linked. Today, certain fields in physics paint a unique picture of field interactions using virtual photons being emitted and absorbed by charges to explain electromagnetic forces. Fortunately, in electronics you don't need to get that detailed.)

The simplest way to generate a magnetic field is to pass a current through a conductor. Microscopically, each electron within a wire should be generating a magnetic field perpendicular to its motion. However, without any potential applied across the wire, the sheer randomness of the electrons due to thermal effects, collision, and so on, cause the individual magnetic fields of all electrons to be pointing in random directions. Averaged over the whole, the magnetic field about the conductor is zero (see Fig. 2.111a.1). Now, when a voltage is applied across the conductor, free electrons gain a drift component pointing from negative to positive—conventional current in the opposite direction. In terms of electron speed, this influence is very slight, but it's enough to generate a net magnetic field (see Fig. 2.111a.2). The direction of this field is perpendicular to the direction of conventional current flow and curls in a direction described by the right-hand rule: your right thumb points in the direction of the conventional current flow; your finger curls in the direction of the magnetic field. See Fig. 2.111b. (When following electron flow instead of conventional current flow, you'd use your left hand.)

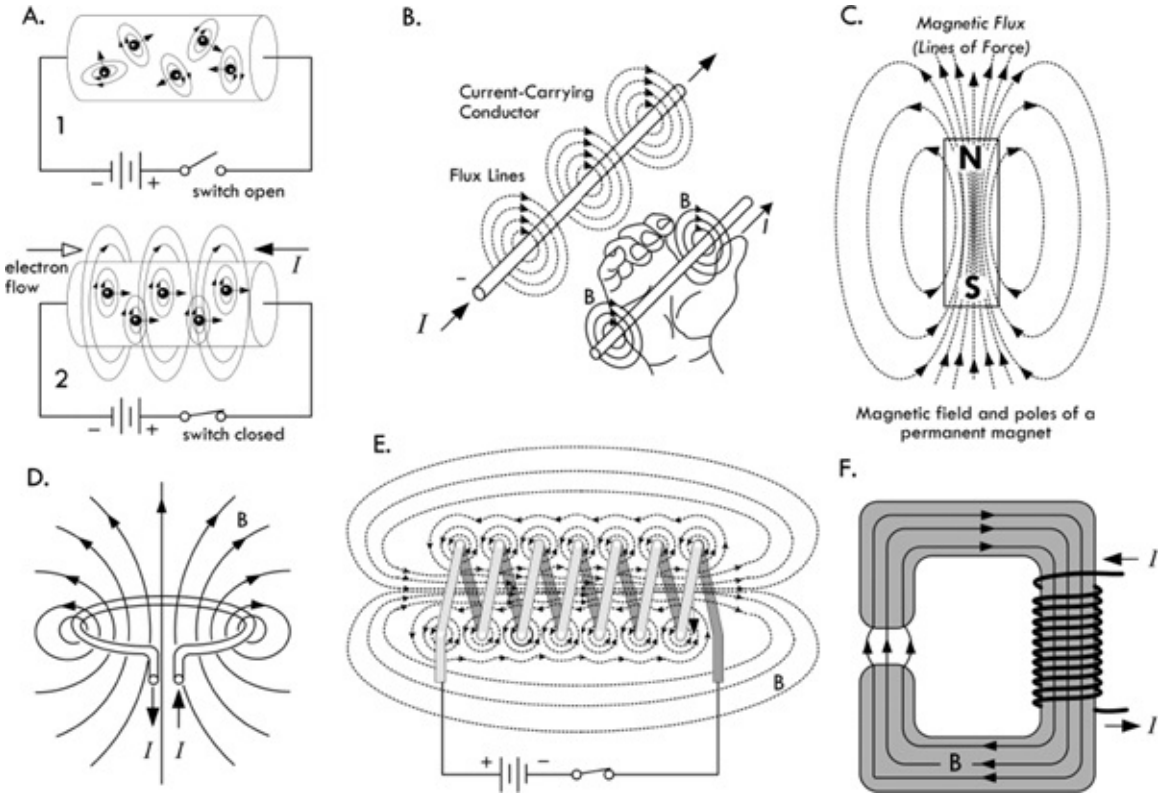


FIGURE 2.111 (a) Magnetic field generated free electrons moving in unison when voltage is applied. (b) Right-hand rule showing direction of magnetic field in relation to conventional current flow. (c) Permanent magnet. (d) Magnetic dipole radiation pattern created by current flowing through single loop of wire. (e) Energized solenoid with a magnetic field similar to permanent magnet. (f) Electromagnet, using ferromagnetic core for increased field strength.

The magnetic field created by sending current through a conductor is similar in nature to the magnetic field of a permanent magnet. (In reality, the magnetic field pattern of a permanent bar magnet, in Fig. 2.111c, is more accurately mimicked when the wire is coiled into a tight solenoid, as shown in Fig. 2.111e.) The fact that both a current-carrying wire and a permanent magnet produce magnetic fields is no coincidence. Permanent magnets made from ferromagnetic materials exhibit magnetic closed-loop fields, mainly as a result of the motion of unpaired electrons orbiting about the nucleus of an atom, as shown in Fig. 2.112, generating a dipole magnetic field. The lattice structure of the ferromagnetic material has an important role of fixing a large portion of the atomic magnetic dipoles in a fixed direction, so as to set up a net magnetic dipole pointing from north to south. This microscopic motion of the unpaired electron about the nucleus of the atom resembles the flow of current through a loop of wire, as depicted in Fig. 2.111d. (Electron spin is another source of magnetic fields, but is far weaker than that due to the electron orbital.)

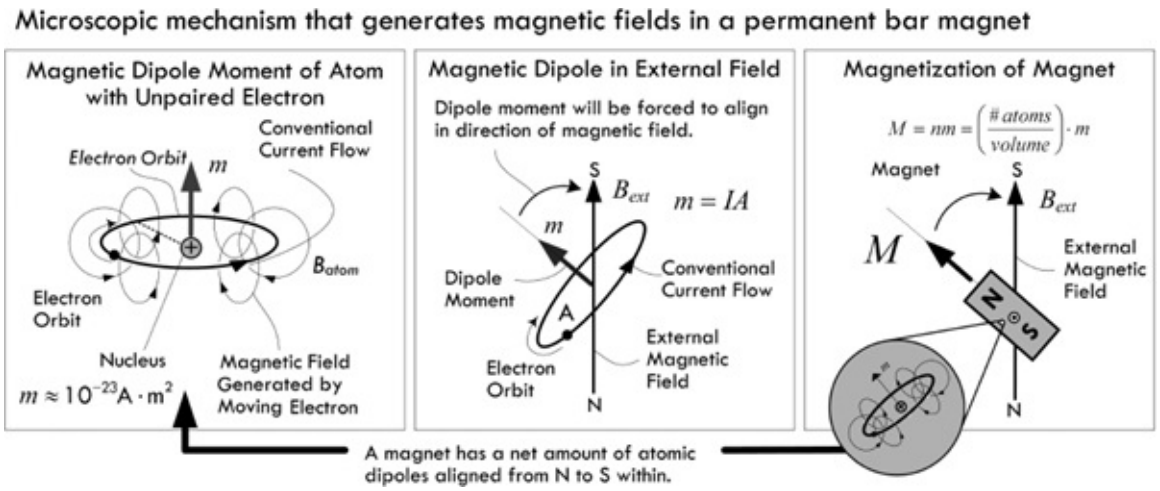


FIGURE 2.112 Microscopically, the magnetic field of a permanent magnet is a result of unpaired valence electrons fixed in a

common direction, generating a magnetic dipole. The fixing in orientation is a result of atomic bonding in the crystalline lattice structure of the magnet.

By coiling a wire into a series of loops, a solenoid is formed, as shown in Fig. 2.111e. Every loop of wire contributes constructively to make the interior field strong. In other words, the fields inside the solenoid add together to form a large field component that points to the right along the axis, as depicted in Fig. 2.111e. By placing a ferromagnetic material (one that isn't initially magnetized) within a solenoid, as shown in Fig. 2.111f, a much stronger magnetic field than would be present with the solenoid alone is created. The reason for such an increase in field strength has to do with the solenoid's field rotating a large portion of the core's atomic magnetic dipoles in the direction of the field. Thus, the total magnetic field becomes the sum of the solenoid's magnetic field and the core's temporarily induced magnetic field. Depending on material and construction, a core can magnify the total field strength by a factor of 1000.

2.24.2 Magnetic Fields and Their Influence

Magnetic fields, unlike electric fields, only act upon charges that are moving in a direction that is perpendicular (or has a perpendicular component) to the direction of the applied field. A magnetic field has no influence on a stationary charge, unless the field itself is moving. Figure 2.113a shows the force exerted upon a moving charge placed within a magnetic field. When considering a positive charge, we use our right hand to determine the direction of force upon the moving charge—the back of the hand points in the direction of the initial charge velocity, the fingers curl in the direction of the external magnetic field, and the thumb points in the direction of the force that is exerted upon the moving charge. For a negative charge, like an electron, we can use the left hand, as shown in Fig. 2.113b. If a charge moves parallel to the applied field, it experiences no force due to the magnetic field—see Fig. 2.113c.

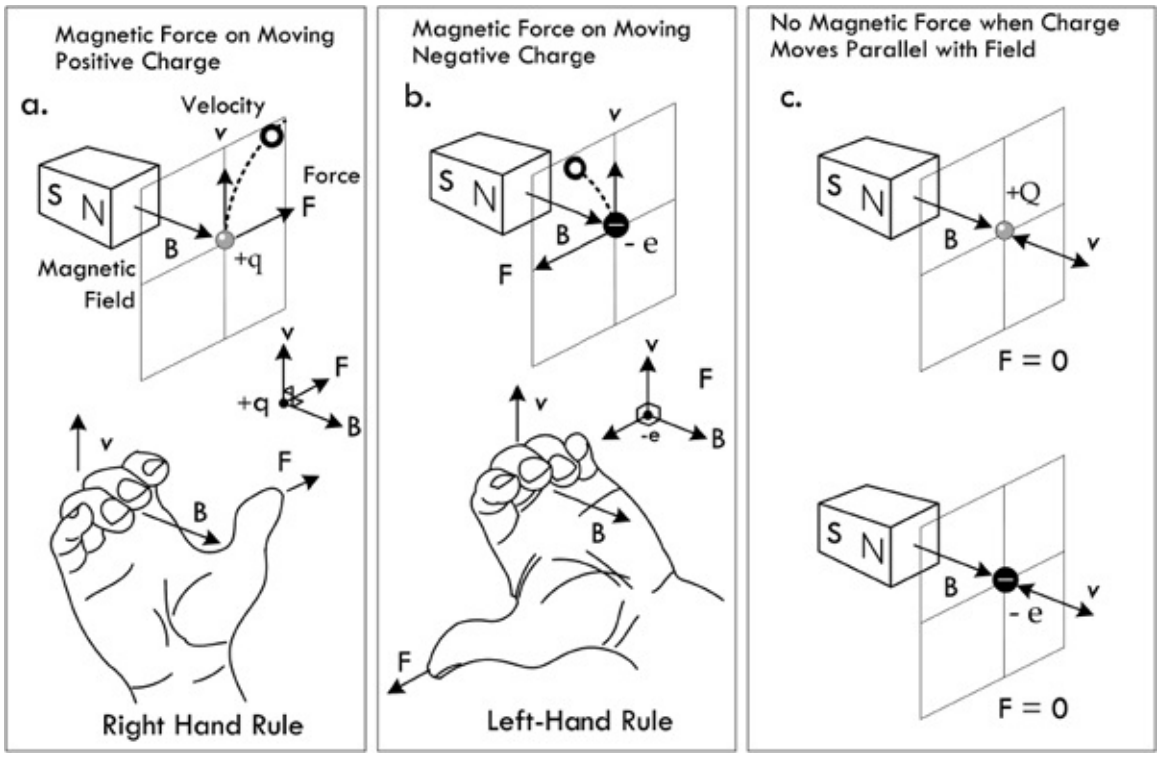


FIGURE 2.113 Illustration showing the direction of force upon a moving charge in the presence of a fixed magnetic field.

In terms of a group of moving charges, such as current through a wire, the net magnetic field of one wire will exert a force on the other wire and vice versa (provided the current is fairly large), as shown in Fig. 2.114. (The force on the wire is possible, since the electrostatic forces at the surface of the lattice structure of the wire prevent electrons from escaping from the surface.)

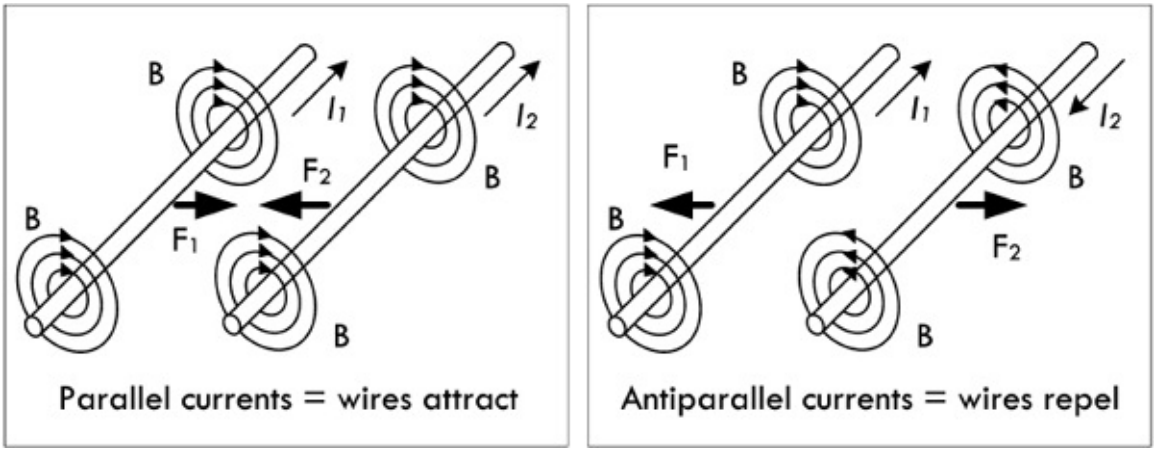


FIGURE 2.114 Forces exerted between two current-carrying wires.

Likewise, the magnetic field of a fixed magnet can exert a force on a current-carrying wire, as shown in Fig. 2.115.

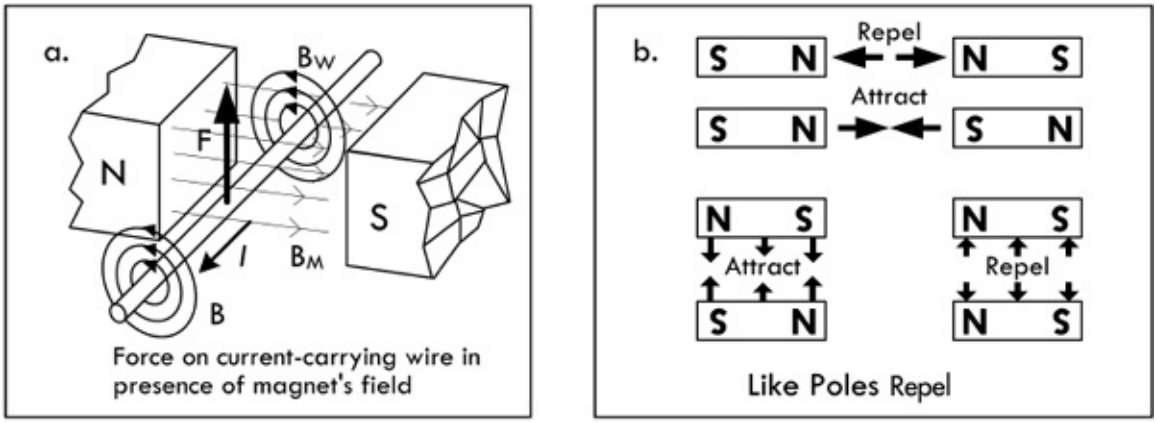


FIGURE 2.115 (a) The force a current-carrying wire experiences in the presence of a magnet's field. (b) Illustration showing how bar magnets attract and repel.

Externally, a magnet is given a north-seeking (or "north," for short) and south-seeking (or "south," for short) pole. The north pole of one magnet attracts the south pole of another, while like poles repel—see Fig. 2.115b. You may ask how two stationary magnets exert forces on each other. Isn't the requirement that a charge or field must be moving for a force to be observed? We associate the macroscopic (observed) force with the forces on the moving charges that comprise the microscopic internal magnetic dipoles which are, at the heart, electrons in motion around atoms. These orbitals tend to be fixed in a general direction called domains—a result of the lattice binding forces.

Another aspect of magnetic fields is their ability to force electrons within conductors to move in a certain direction, thus inducing current flow. The induced force is an electromotive force (EMF) being set up within the circuit. However, unlike, say, a battery's EMF, an induced EMF depends on time and also on geometry. According to Faraday's law, the EMF induced in a circuit is directly proportional to the time rate of change of the magnetic flux through the circuit:

$$\text{EMF} = -\frac{d\Phi_M}{dt}, \Phi_M = \int \vec{B} \cdot \overrightarrow{dA}, \text{EMF} = -N\frac{d\Phi_M}{dt} \text{(for a coiled wire of } N \text{ loops)} \tag{2.49}$$

where Φ_M is the magnetic flux threading through a closed-loop circuit (which is equal to the magnetic field B dotted with the direction surface area—both of which are vectors, and summed over the entire surface area—as the integral indicates). According to the law, the EMF can be induced in the circuit in

several ways: (1) the magnitude of B can vary with time; (2) the area of the circuit can change with time; (3) the angle between B and the normal of A can change with time; (4) any combination of these can occur. See Fig. 2.116.

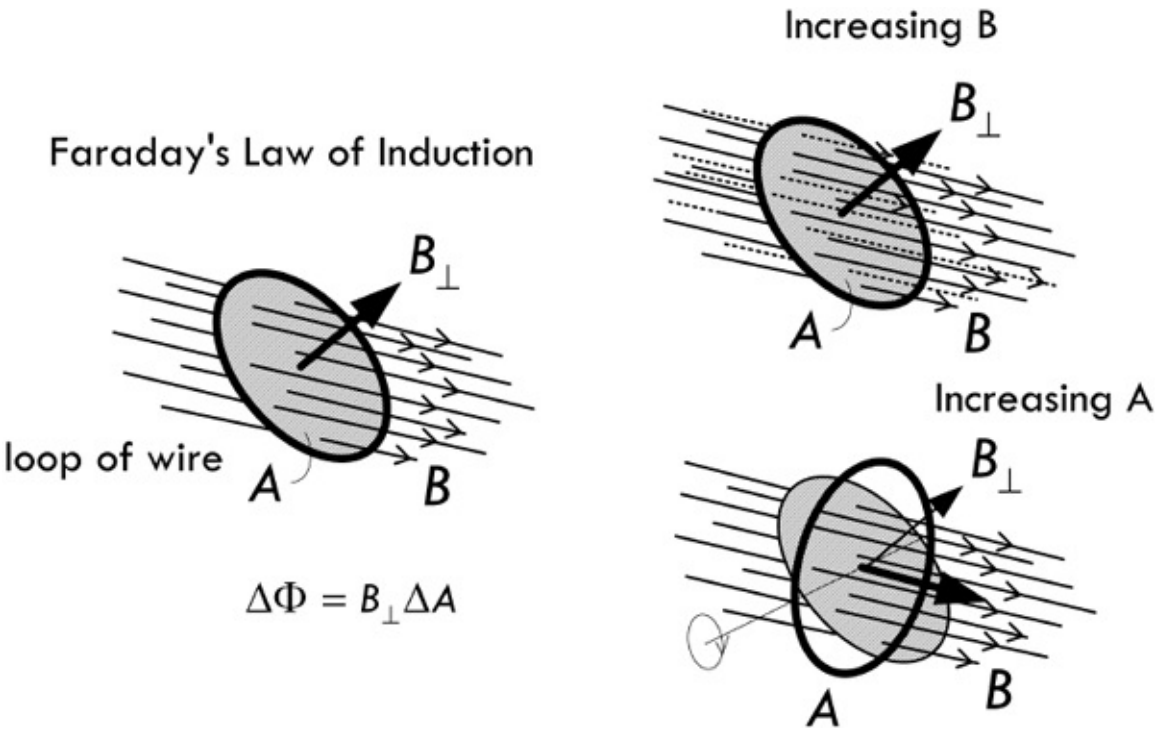


FIGURE 2.116 Illustration of Faraday's law of induction.

The simple ac generator in Fig. 2.117 shows Faraday's law in action. A simple rotating loop of wire in a constant magnetic field generates an EMF that can be used to power a circuit. As the loop rotates, the magnetic flux through it changes with time, inducing an EMF and a current in an external circuit. The ends of the loop are connected to slip rings that rotate with the loop, while the external circuit is linked to the generator by stationary brushes in contact with the slip rings.

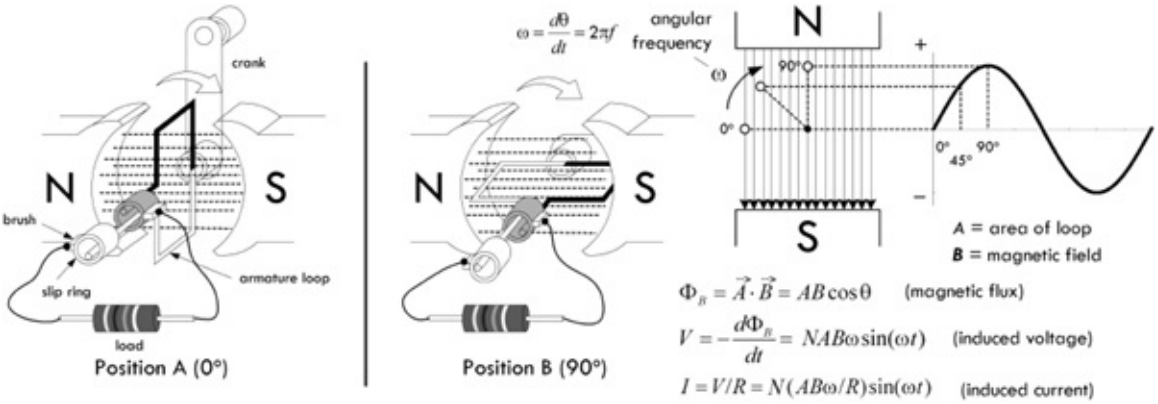


FIGURE 2.117 Basic ac generator.

A simple dc generator is essentially the same as the ac generator, except that the contacts to the rotating loop are made using a split ring or commutator. The result is a pulsating direct current, resembling the absolute value of a sine wave—there are no polarity reversals.

A motor is essentially a generator operating in reverse. Instead of generating a current by rotating a loop, a current is supplied to the loop by a battery, and the torque acting on the current-carrying loop causes it to rotate. Real ac generators and motors are much more complex than the simple ones demonstrated here. However, they still operate under the same fundamental principles of electromagnetic induction.

The circuit in Fig. 2.118 shows how it is possible to induce current within a secondary coil of wire by suddenly changing the current flow through a primary coil of wire. As the magnetic field of the primary expands, an increasing magnetic flux permeates the secondary. This induces an EMF that causes current to flow in the secondary circuit. This is the basic principle behind how transformers work; however, a real transformer's primary and secondary coils are typically wound around a common ferromagnetic core to increase magnetic coupling.

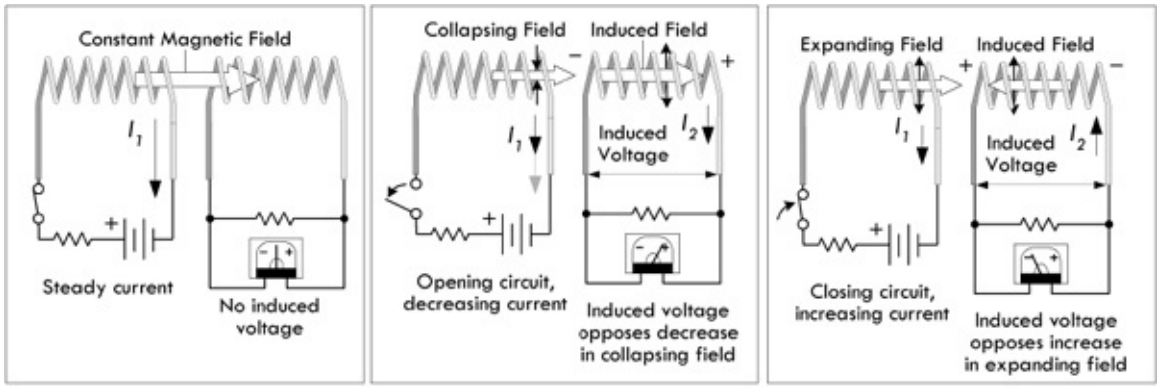


FIGURE 2.118 An induced voltage, or EMF, is generated in the secondary circuit whenever there is a sudden change in current in the primary.

2.24.3 Self-Inductance

In the previous section, we saw how an EMF could be induced in a closed-loop circuit whenever the magnetic flux through the circuit changed with time. This phenomenon of electromagnetic induction is used in a number of mechanisms, such as motors, generators, and transformers, as was pointed out. However, in each of these cases, the induced EMF was a result of an external magnetic field, such as the primary coil in relation to a secondary coil. Now, however, we will discuss a phenomenon called *self-induction*. As the name suggests, self-induction typically involves a looped wire inflicting itself with an induced EMF that is generated by the varying current that passes through it. According to Faraday's law of induction, the only time that our loop can self-inflict is when the magnetic field grows or shrinks in strength (as a result of an increase or decrease in current). Self-induction is the basis for the inductor, an important device used to store energy and release energy as current levels fluctuate in time-dependent circuits.

Consider an isolated circuit consisting of a switch, a resistor, and a voltage source, as shown in Fig. 2.119. If you close the switch, you might predict that the current flow through the circuit would jump immediately from zero to V/R, according to Ohm's law. However, according to Faraday's law of electromagnetic induction, this isn't entirely accurate. Instead, when the switch is initially closed, the current increases rapidly. As the current increases with time, the magnetic flux through the loop rises rapidly. This increasing magnetic flux then induces an EMF in the circuit that opposes the current flow, giving rise to an exponentially delayed rise in current. We call the induced EMF a *self-induced EMF*.

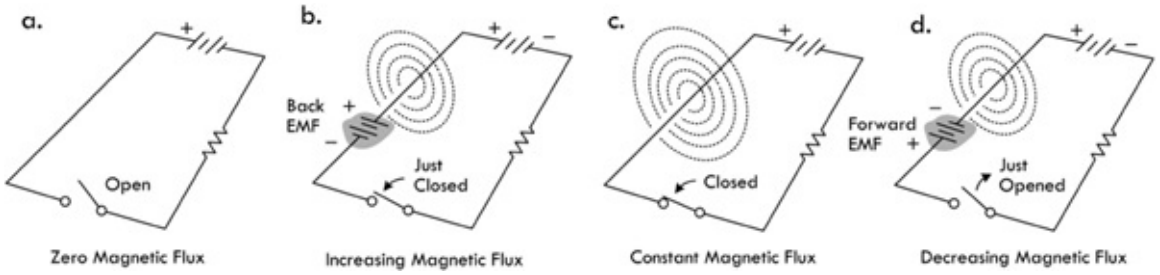


FIGURE 2.119 (a) Circuit is open and thus no current or magnetic field is generated. (b) The moment the circuit is closed, current begins to flow, but at the same time, an increasing magnetic flux is generated through the circuit loop. This increasing

flux induced a back EMF that opposes the applied or external EMF. After some time, the current levels off, the magnetic flux reaches a constant value, and the induced EMF disappears. (d) If the switch is suddenly opened, the current attempts to go to zero; however, during this transition, as the current goes to zero, the flux decreases through the loop, thus generating a forward induced voltage of the same polarity as the applied or external EMF. As we'll see later, when we incorporate large solenoid and toroidal inductors within a circuit, opening a switch such as this can yield a spark—current attempting to keep going due to a very large forward EMF.

Self-induction within a simple circuit like that shown in Fig. 2.119 is usually so small that the induced voltage has no measurable effect. However, when we start incorporating special devices that concentrate magnetic fields—namely, discrete inductors—time-varying signals can generate significant induced EMFs. For the most part, unless otherwise noted, we shall assume the self-inductance of a circuit is negligible compared with that of a discrete inductor.

2.24.4 Inductors

Inductors are discrete devices especially designed to take full advantage of the effects of electromagnetic induction. They are capable of generating large concentrations of magnetic flux, and they are likewise capable of experiencing a large amount of self-induction during times of great change in current. (Note that self-induction also exists within straight wire, but it is usually so small that it is ignored, except in special cases, e.g., VHF and above, where inductive reactance can become significant.)

The common characteristic of inductors is a looplike geometry, such as a solenoid, toroid, or even a spiral shape, as shown in Fig. 2.120. A solenoid is easily constructed by wrapping a wire around a hollow plastic form a number of times in a tight-wound fashion.

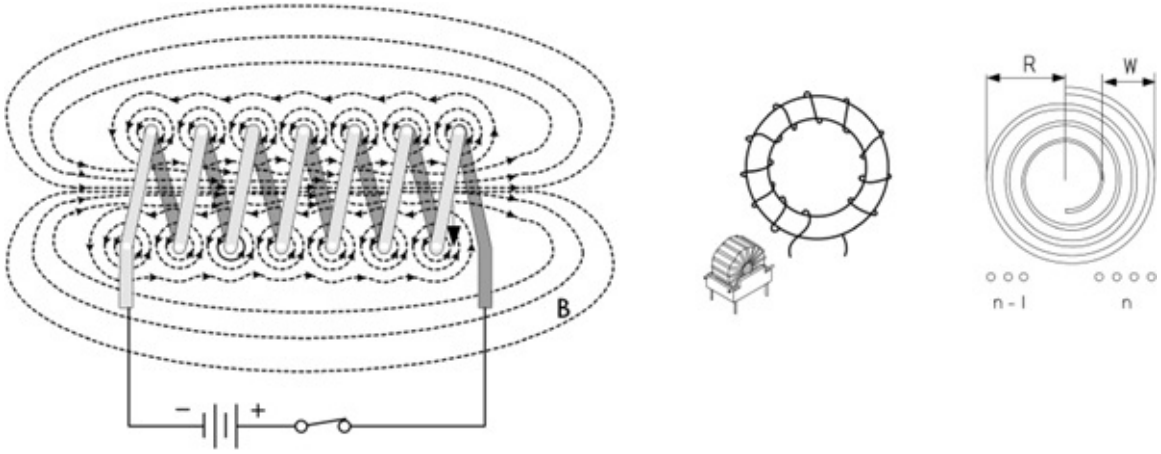


FIGURE 2.120 Various coil configurations of an inductor—solenoid, toroid, and spiral.

The basic schematic symbol of an air core inductor is given by  . Magnetic core inductors (core is either iron, iron powder, or a ferrite-type ceramic), adjustable core inductors, and a ferrite bead, along with their respective schematic symbols are shown in Fig. 2.121.

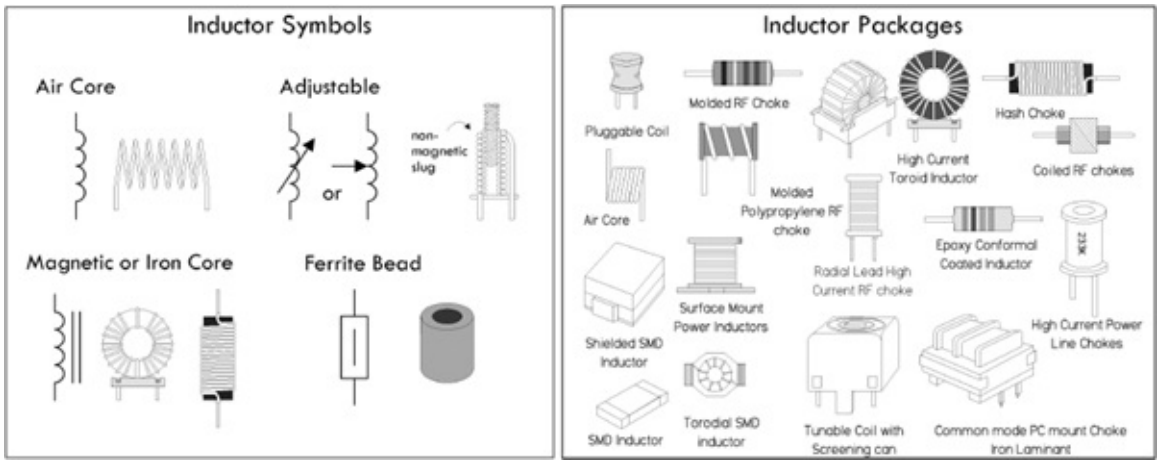


FIGURE 2.121

Magnetic core inductors are capable of generating much higher magnetic field densities than air core inductors as a result of the internal magnetization that occurs at the atomic level within the core material due to the surrounding wire coil's magnetic field. As a result, these inductors experience much greater levels of self-induction when compared to air core inductors. Likewise, it is possible to use fewer turns when using a magnetic core to achieve a desired inductance. The magnetic core material is often iron, iron powder, or a metallic oxide material (also called a ferrite, which is ceramic in nature). The choice of core material is a complex process that we will cover in a moment.

Air core inductors range from a single loop in length of wire (used at ultrahigh frequencies), through spirals in copper coating of an etched circuit board (used at very high frequencies), to large coils of insulated wire wound onto a nonmagnetic former. For radio use, inductors often have air cores to avoid losses caused by magnetic hysteresis and eddy current that occur within magnetic core–type inductors.

Adjustable inductors can be made by physically altering the effective coil length—say, by using a slider contact along the uncoated coil of wire, or more commonly by using a ferrite, powdered iron, or brass slug screwed into the center of the coil. The idea behind the slug-tuned inductor is that the inductance depends on the permeability of the material within the coil. Most materials have a relative permeability close to 1 (close to that of vacuum), while ferrite materials have a large relative permeability. Since the inductance depends on the average permeability of the volume inside the coil, the inductance will change as the slug is turned. Sometimes the slug of an adjustable inductor is made of a conducting material such as brass, which has a relative permeability near 1, in which case eddy currents flow on the outside of the slug and eliminate magnetic flux from the center of the coil, reducing its effective area.

A *ferrite bead*, also known as a *ferrite choke*, is a device akin to an inverted ferrite core inductor. Unlike a typical core inductor, the bead requires no coiling of wire (though it is possible to coil a wire around it for increased inductance, but then you've created a standard ferrite core inductor). Instead, a wire (or set of wires) is placed through the hollow bead. This effectively increases the inductance of the wire (or wires). However, unlike a standard inductor that can achieve practically any inductance based on the total number of coil turns, ferrite beads have a limited range over which they can influence inductance. Their range is typically limited to RF (radiofrequency). Ferrite beads are often slipped over cables that are known to be notorious radiators of RF (e.g., computers, dimmers, fluorescent lights, and motors). With the bead in place, the RF is no longer radiated but absorbed by the bead and converted into heat within the bead. (RF radiation can interfere with TV, radio, and audio equipment.) Ferrite beads can also be placed on cables entering receiving equipment so as to prevent external RF from entering and contaminating signals in the cable runs.

An inductor acts like a time-varying current-sensitive resistance. It only "resists" during changes in current; otherwise (under steady-state dc conditions), it passes current as if it were a wire. When the applied voltage increases, it acts like a time-dependent resistor whose resistance is greatest during times of rapid increase in current. On the other hand, when the applied voltage decreases, the inductor acts like a time-dependent voltage source (or negative resistance) attempting to keep current flowing. Maximal sourcing is greatest during times of rapid decreases in current.

In Fig. 2.122a, when an increasing voltage is applied across an inductor, resulting in an increasing current flow, the inductor acts to resist this increase by generating a reverse force on free electrons as a result of an increasing magnetic flux cutting across the coils of the solenoid (or crossing through the coil loops). This reverse force on free electrons can be viewed as an induced EMF that points in the opposite direction of the applied voltage. We deem this induced EMF a *reverse EMF*, also referred to as a *back EMF*. The result is that the inductor strongly resists during a sudden increase in current flow, but quickly loses its resistance once the current flow levels off to a constant value.

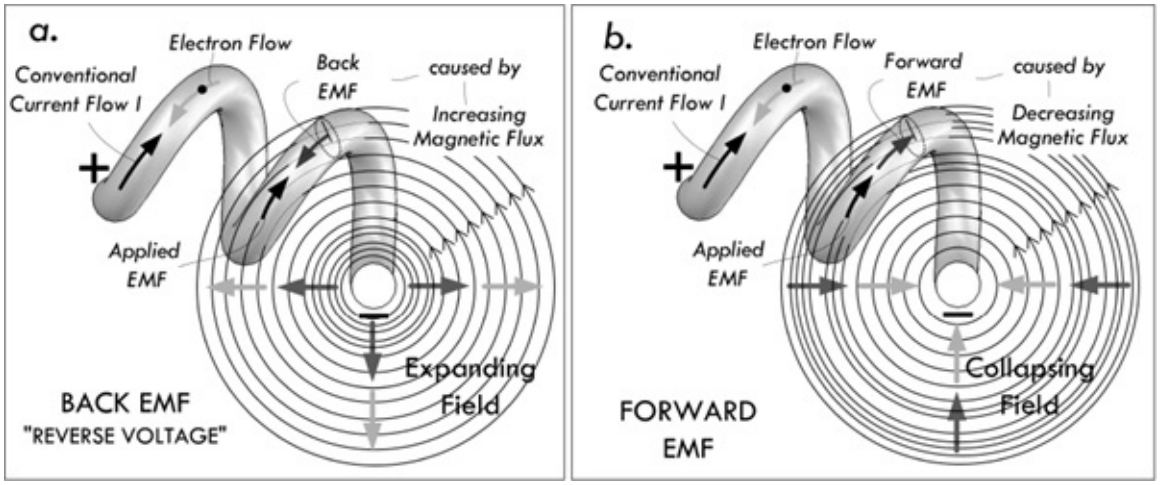


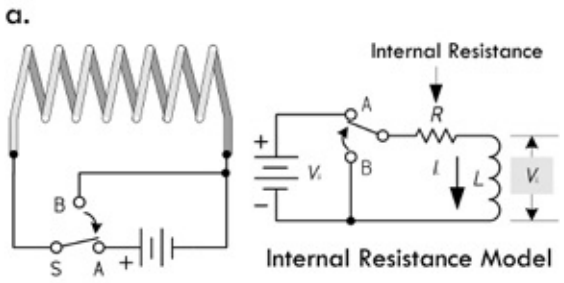
FIGURE 2.122

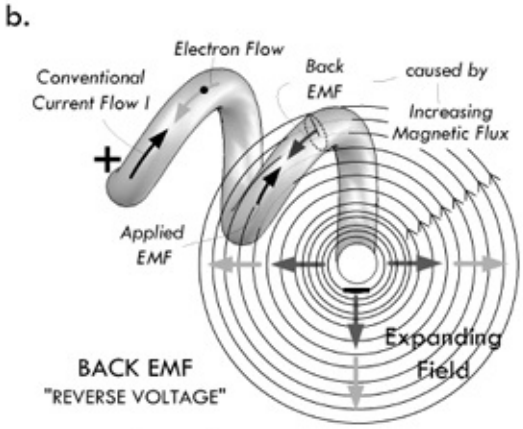
In Fig. 2.122b, when a decreasing voltage is applied across an inductor, resulting in a decreasing current flow, the inductor acts to resist this decrease by generating a forward force on free electrons as a result of a decreasing magnetic flux cutting across the coils of the solenoid (or crossing through the coil loops). This forward force on free electrons can be viewed as an induced EMF that points in the same direction as that of the applied voltage (the applied voltage that existed before the sudden change). We deem this induced EMF a *forward EMF*. The result is that the inductor acts like a voltage source during a sudden decrease in current flow, but quickly disappears once the current flow levels off to a constant value.

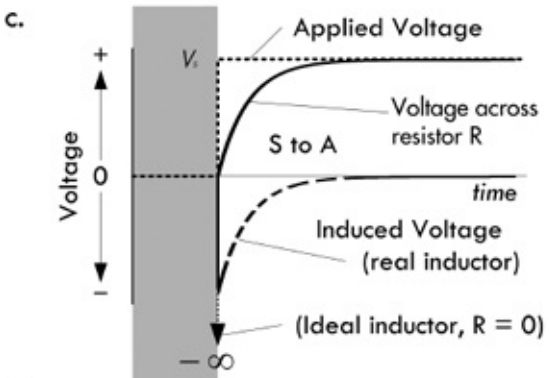
Another view of how inductors work is to consider energy transfer. The transfer of energy to the magnetic field of an inductor requires that work be performed by a voltage source connected across it. If we consider a perfect inductor where there is no resistance, the energy that goes into magnetic field energy is equal to the work performed by the voltage source. Or, in terms of power, the power is the rate at which energy is stored ($P = dW/dt$). Using the generalized power law $P = IV$, we can equate the powers and see that there must be a voltage drop across an inductor while energy is being stored in the magnetic field. This voltage drop, exclusive of any voltage drop due to resistance in a circuit, is the result of an opposing voltage induced in the circuit while the field is building up to its final value. Once the field becomes constant, the energy stored in the magnetic field is equal to the work performed by the voltage source.

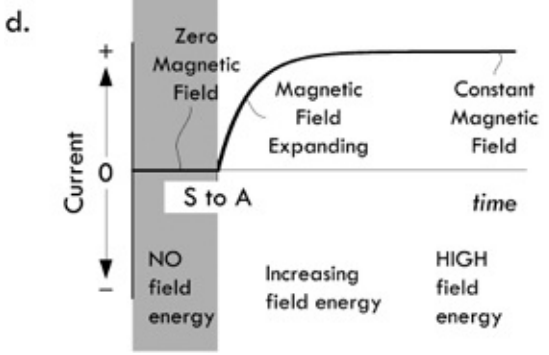
Figure 2.123 illustrates what occurs when an inductor is energized suddenly as a switch is closed.

Energizing Inductor









When an inductor is being energized, we say that electrical energy from an externally applied source is transformed into a magnetic field about the inductor. Only during times of change—say, when the switch is thrown, and the magnetic field suddenly grows in size—do we see inductive behavior that affects circuit dynamics.

In the circuit to the left, when the switch is thrown from position B to position A, a sudden change in voltage is applied across the inductor, resulting in a sudden increase in current flow. Where there was no magnetic field, there now is a rapidly growing magnetic field about the inductor. The inductor is said to be energizing. The expanding magnetic field cuts across its own inductor coils, which, by Faraday's law, exerts a force of free electrons within the coil. The force on these electrons is such as to be pointing in the opposite direction as the applied voltage. This effective reverse force and the effect it has on the free electrons it deems a reverse EMF—it's analogous to a little imaginary battery placed in the reverse direction of the applied voltage. (See Fig. 2.123b.) The result is that the inductor resists during increases in current. After a short while, the current flow through the circuit levels off, and the magnetic field stops growing and assumes a constant value. With no change in magnetic field strength, there is no increase in magnet flux through our fixed coils, so there is no more reverse EMF. The inductor acts as a simple conductor. Figure 2.123c shows the induced voltage and voltage across the resistor as a function of time.

Figure 2.123d shows the resultant current flow through the inductor as a result of the resultant voltage. Mathematically, the current flow is expressed by the following equation:

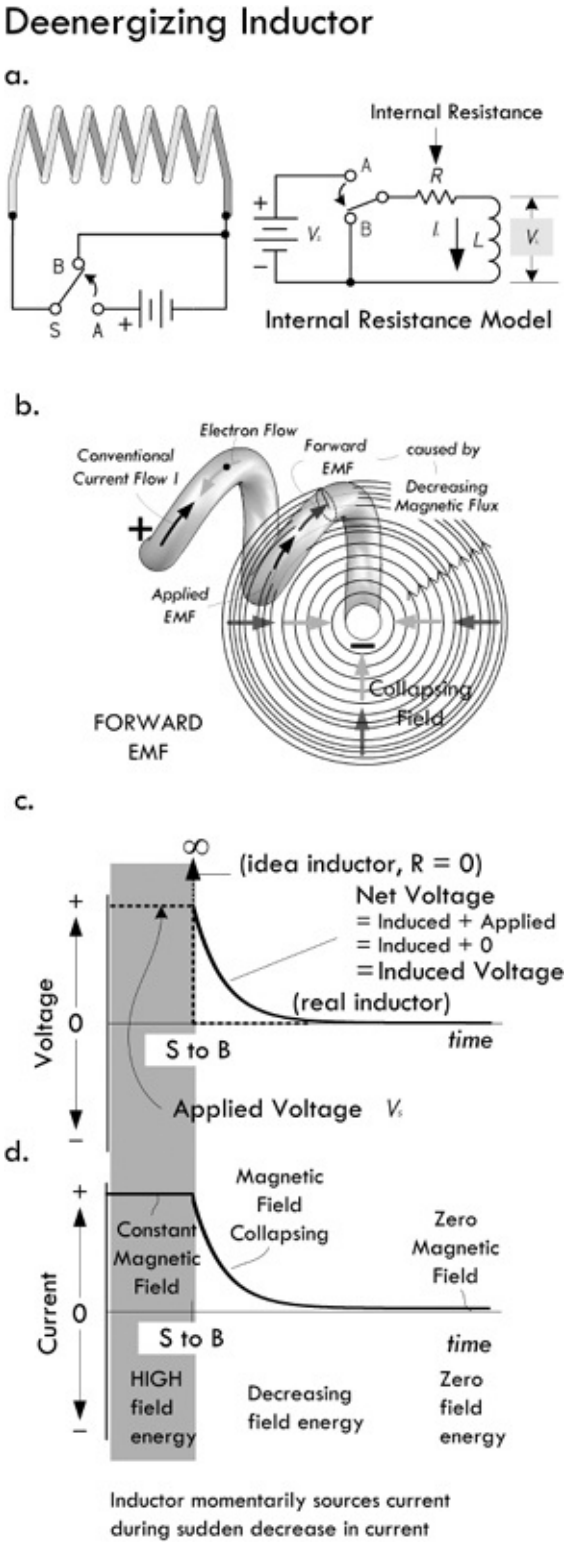
$$I = \frac{V_s}{R}(1 - e^{-t/(L/R)})$$

(You might question what happens if we assume an ideal inductor with zero internal resistance *R*. This is a very good question—one that we'll discuss when we get to defining inductance mathematically.)

In terms of energy, this can be viewed as the electrical energy being transformed into magnetic field energy. In terms of power, we see that a voltage drop occurs as energy is pumped into the magnetic field. (We associate the drop with the back EMF.) Once the current levels off, no more energy goes into the field; hence, no reverse voltage (or voltage drop) is present.

FIGURE 2.123

Figure 2.124 illustrates what occurs as an inductor is deenergized suddenly by opening a switch.



When an inductor is being deenergized, we say that the magnetic field energy of the inductor is "transformed" back into electrical energy. Again, only during times of change, which in our case involves a decreasing current flow, do we notice effects of induction.

In the circuit to the left, when the switch is thrown from position A to B, a sudden change voltage occurs across the inductor.

The inductor initially opposes this decrease by generating a collapsing magnetic field that cuts across the inductor coils. According to Faraday's law, a decreasing magnetic flux passes through the loops of the inductor, thus imparting a force upon free electrons within the coil in the same direction the applied voltage was pointing right before the switch occurred. Since the force is in the same direction, we deem the effect a forward EMF. Hence, the inductor sources current when attempts are made to decrease current flow. The energy for it to do so comes from the magnetic field, whose energy drops in proportion to the electrical energy delivered to the circuit. See (b).

Part (c) shows the resultant voltage across an inductor adding the applied voltage to the inducted voltage.

Part (d) shows the resultant current flow through the inductor as a result of the resultant voltage.

Mathematically, the current flow is expressed by the following equation:

$$I = \frac{V_s}{R} e^{-t/(L/R)}$$

(Again, you might question what happens if we assume an ideal inductor with zero internal resistance *R*. This is a very good question—one that we'll discuss when we get to defining inductance mathematically.)

***Note:* We must assume that when the switch is thrown from A to B, the transition occurs instantaneously. We'll see in a moment that when a physical break occurs, cutting current flow through an inductive circuit, the collapsing magnetic field can be large enough to generate an EMF capable of causing a spark to jump between the break points (i.e., switch contacts).**

FIGURE 2.124

2.24.5 Inductor Water Analogy

The property of inductance in electric circuits is closely analogous to mass inertia in mechanical systems. For example, the property of an inductor resisting any sudden changes in current flow (increasing or decreasing) is similar to a mass on a spinning wheel resisting any change in motion (increasing or decreasing in speed). In the following water analogy, we take this mass analogy to heart, incorporating a water turbine/flywheel device to represent a "water inductor."

To start, we consider a basic electrical inductor circuit, as shown to the left in Fig. 2.125. The field generated by a suddenly applied voltage creates a reverse induced voltage of opposite polarity that initially "resists" current flow. Quickly, depending on the inductance value, the reverse voltage disappears as the magnetic field becomes constant, at which point it has reached a maximum strength and energy. A collapsing field generated when the applied voltage is removed creates a forward induced voltage that attempts to keep current flowing. Quickly, depending on the inductance value, the forward voltage disappears and the magnetic field goes to zero.

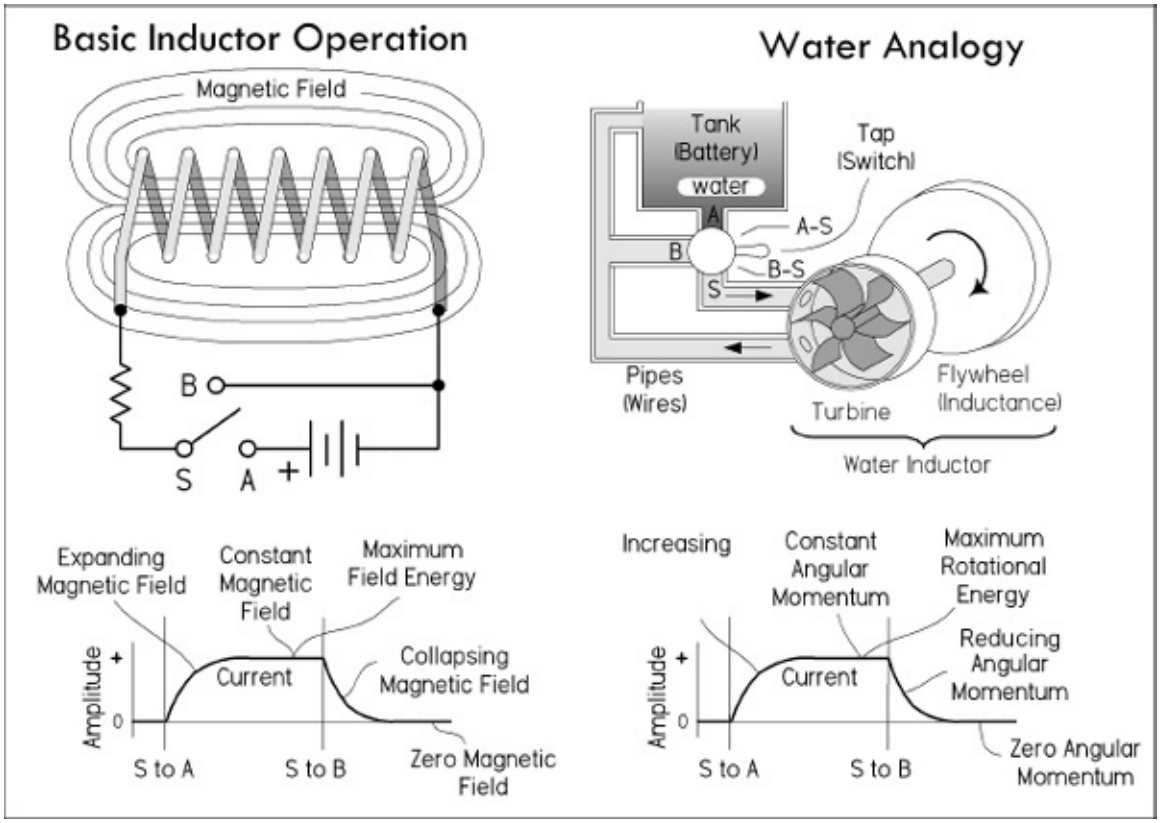


FIGURE 2.125

In the water analogy, the turbine with attached flywheel resists any sudden changes in current flow. If water pressure is suddenly applied, the turbine initially resists water flow due to its mass and that of the attached flywheel. However, the pressure exerted over the turbine blades quickly gives rise to mechanical motion. Depending on the mass of the flywheel, the time it takes for the flywheel to reach a steady angular velocity will vary—a heavier flywheel will require more time (analogous to a high-value inductor requiring more time to reach a constant current after a sudden increase in applied voltage). When the flywheel reaches this constant angular rotation, the water inductor has maximum rotational momentum and energy. This is analogous to the magnetic field strength reaching a constant maximum strength and energy when the reverse voltage disappears. If there is any sudden interruption in applied pressure—say, by turning the tap to position B-S, as shown in Fig. 2.125—the flywheel's angular momentum will attempt to keep current flowing. This is analogous to the collapsing magnetic field in an inductor inducing a forward voltage.

Example: What will happen when the switches are closed in the following three circuits? What happens when the switches are later opened? How does the size of the inductance and capacitance influence behavior? Refer to Fig. 2.126.

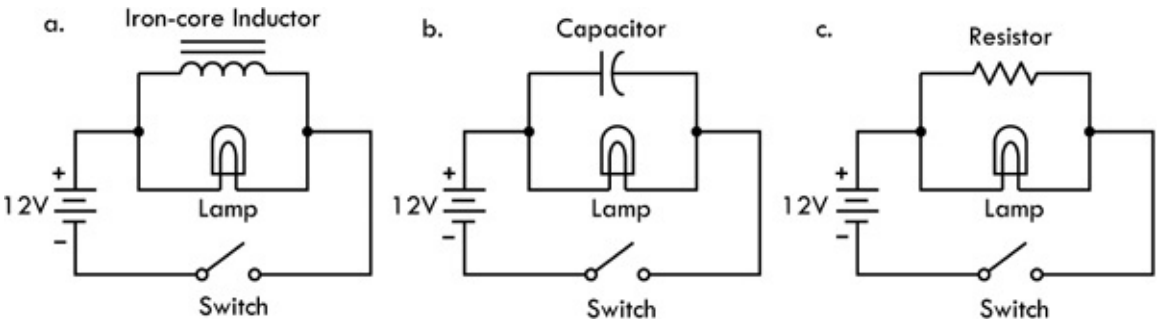


FIGURE 2.126

Answer: (a) When the switch is closed in the inductor circuit, the lamp will suddenly light up brightly, but then quickly dims out. This is because the moment the switch is closed, the inductor has very high impedance to current flow, but quickly loses its impedance as the current becomes constant (the magnetic field is no longer expanding). Once constant, the inductor acts like a short, and all current is diverted away from the lamp through the inductor. (We assume here that there is sufficient dc internal resistance in the inductor to prevent excessive current flow. Also, we assume the internal dc resistance of the inductor is much smaller than the internal resistance of the lamp.) A larger inductance value increases the time it takes for the lamp to dim out completely.

(b) When the switch is closed in the capacitor circuit, the opposite effect occurs; the lamp slowly builds up in brightness until it reaches maximum illumination. This is because when the switch is closed, the capacitor initially has very low impedance to current flow during rapid changes in applied voltage. However, as the capacitor charges up, the capacitor's impedance rises toward infinity, and consequently resembles an open circuit—hence, all current is diverted through the lamp. A larger capacitance value increases the time it takes for the lamp to reach full brightness.

(c) When the switch is closed in the resistor circuit, free-electron flow throughout the system is essentially instantaneous. Aside from small inherent inductance and capacitance built into the circuit, there are no time-dependent effects on current flow caused by discrete inductance or capacitance. Note that since the voltage source is ideal, no matter what value the parallel resistor has, there will always be 12 V across the lamp, and hence the brightness of the lamp does not change over time.

2.24.6 Inductor Equations

Conceptually, you should now understand that the amplitude of the induced voltage—be it reverse or forward induced—is proportional to the rate at which the current changes, or the rate at which the magnetic flux changes. Quantitatively, we can express this relationship by using the following equation:

$$V_L = L\frac{dI_L}{dt}$$

(2.50) Voltage across an inductor = induced EMF

If we integrate and solve for I_L, we get:

$$I_L = \frac{1}{L}\int V_L\, dt$$

(2.51) Current through an inductor

In Eq. 2.50, the proportionality constant L is called the inductance. This constant depends on a number of physical inductor parameters, such as coil shape, number of turns, and core makeup. A coil with many turns will have a higher L value than one with few turns, if both coils are otherwise physically similar. Furthermore, if an inductor is coiled around a magnetic core, such as iron or ferrite, its L value will increase in proportion to the permeability of that core (provided circuit current is below the point at which the core saturates).

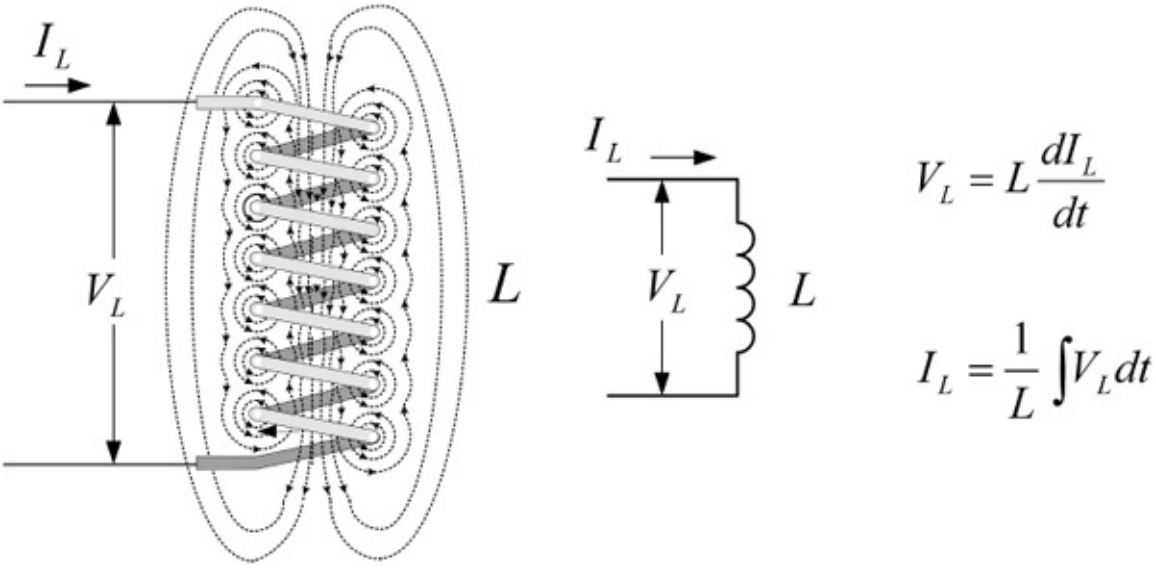


FIGURE 2.127 The voltage measured across an ideal inductor is the induced voltage, or EMF, symbolized V_L. When there is a steady-state dc current flow, there is no induced voltage (V_L is zero) and the inductor resembles a short. As we'll see, the inductor equations can yield surprisingly unrealistic values, if we don't consider internal resistance and capacitance inherently present within a real inductor.

The basic unit of inductance L is the henry, abbreviated H. One henry equals an induced voltage of 1 V when the current is varying at a rate of 1 A/s:

$$1\text{ H} = \frac{1\text{ V}}{1\text{ A/s}} \text{ (Definition of a henry)}$$

Making inductors from scratch is common in electronics, unlike the construction of capacitors, which is left almost exclusively to the manufacturers. Though we'll examine how to make inductors in a moment, it's worth taking a look at some commercial inductors. Note the inductance range, core type, current, and frequency limits listed in Table 2.7.

TABLE 2.7 Typical Characteristics of Commercial Inductors

CORE TYPE	MINIMUM H	MAXIMUM H	ADJUSTABLE?	HIGH CURRENT?	FREQUENCY LIMIT
Air core, self-supporting	20 nH	1 mH	Yes	Yes	1 GHz
Air core, on former	20 nH	100 mH	No	Yes	500 MHz
Slug tuned open winding	100 nH	1 mH	Yes	No	500 MHz
Ferrite ring	10 mH	20 mH	No	No	500 MHz
RM Ferrite Core	20 mH	0.3 H	Yes	No	1 MHz
EC or ETD Ferrite Core	50 mH	1 H	No	Yes	1 MHz
Iron	1 H	50 H	No	Yes	10 kHz

Typical values for commercial inductors range from fractions of a nanohenry to about 50 H. Inductances are most commonly expressed in terms of the following unit prefixes:

nanohenry (nH): 1 nH = 1×10^{-9} H = 0.000000001 H

microhenry (μH): 1 μH = 1×10^{-6} H = 0.000001 H

millihenry (mH): 1 mH = 1×10^{-3} H = 0.001 H

Example 1: Rewrite 0.000034 H, 1800 mH, 0.003 mH, 2000 μH, and 0.09 μH in a more suitable unit prefix format (1 ≤ numeric value < 1000).

Answer: 34 μH, 1.8 H, 3 μH, 2 mH, and 90 nH.

From another standpoint, inductance can be determined from basic physics principles. Theoretically, you can determine the inductance any time by stating that the inductance is always the ratio of the magnetic flux linkage ($N\Phi_M$) to the current:

$$L = \frac{N\Phi_M}{I} \tag{2.52}$$

For an air-filled solenoid, as shown in Fig. 1.128, if a current I flows through the coil, Ampere's law allows us to calculate the magnetic flux:

$$\Phi_M = BA = \left(\frac{\mu_0 NI}{\ell}\right) A = \mu_0 A n_{unit} I$$

where n_{unit} is the turns per unit length:

$$n_{\text{unit}} = N/\ell \tag{2.53}$$

where N is the total turns and ℓ is the length. The variable A is the cross-sectional area of the coil, and μ is the permeability of the material on which the coil is wound. For most materials (excluding iron and ferrite materials), the permeability is close to the permeability of free space:

$$\mu_0 = 4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$$

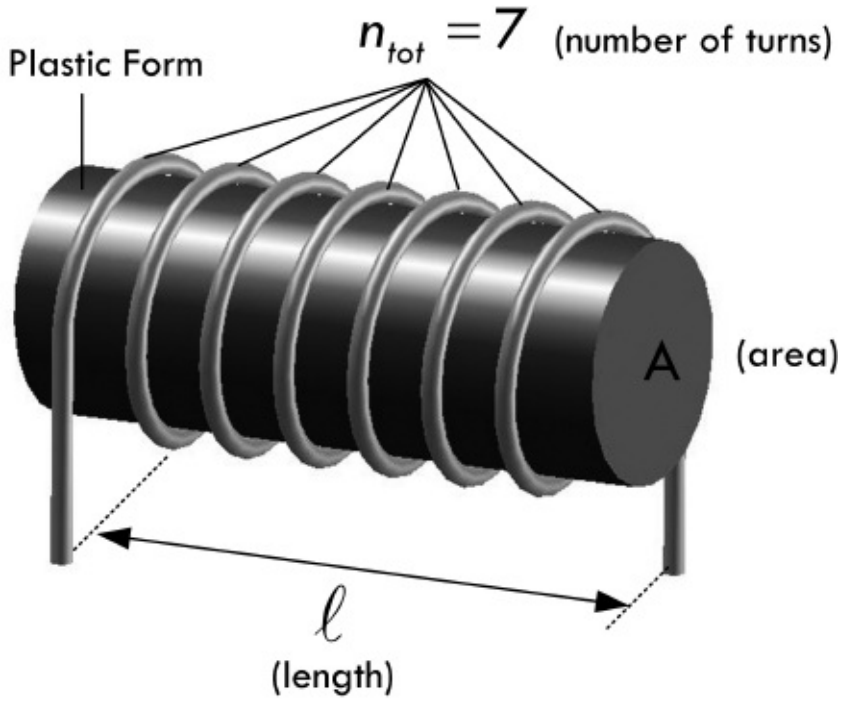


FIGURE 2.128

According to Faraday's law, an induced voltage is in each loop of the solenoid, resulting in a net induced voltage across the solenoid equal to *n* times the change in magnetic flux:

$$V_L = N\frac{d\Phi_M}{dt} = \frac{\mu N^2 A}{\ell}\frac{dI}{dt}$$

The term in front of *dI*/*dt* in the equation we call the inductance of the solenoid:

$$L_{sol} = \frac{\mu N^2 A}{\ell} \tag{2.54}$$

Inductance varies as the square of the turns. If the number of turns is doubled, the inductance is quadrupled. This relationship is inherent in the equation but is often overlooked. For example, if you want to double the inductance of a coil, you don't double the turns; rather you make the number of turns $\sqrt{2}$ (or 1.41) times the original number of turns, or 40 percent more turns.

Example 2: Find the inductance of a cylindrical coil of length 10 cm, radius 0.5 cm, having 1000 turns of wire wrapped around a hollow plastic form.

Answer:

$$L = \mu_0 N^2 A/\ell = (4\pi \times 10^{-7})\ 106(\pi \times 0.0052)/0.1$$
$$= 1 \times 10^{-3}\text{H} = 1\ \text{mH}$$

Luckily, there are some simple formulas, shown in Fig. 2.129, for air core inductors, as well as for a multiple-wound and a spiral-wound inductor. Note that the formulas' answers will not be in standard form —they assume the results are in units of microhenrys.

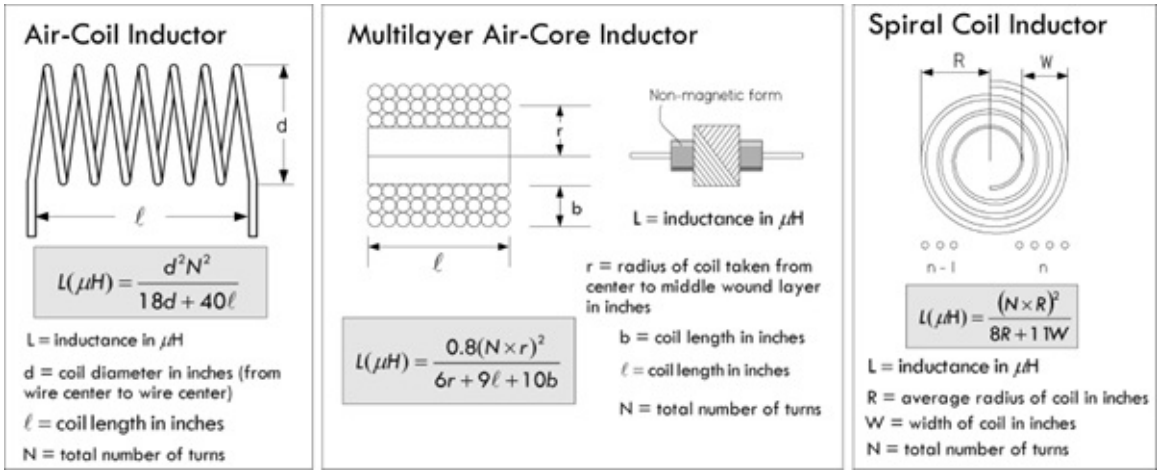


FIGURE 2.129 Practical air core inductor equations.

Example 3: What is the inductance of a coil wrapped around a 0.5-in-diameter plastic form if the coil has 38 turns wound at 22 turns per inch?

Answer: First the total length is determined:

$$\ell = \frac{N}{n_{\text{unit}}} = \frac{38 \text{ turns}}{22 \text{ turns/in}} = 1.73 \text{ in}$$

Next, using the equation for an air core inductor shown in Fig. 2.129, noting the result will be in units of microhenrys, we get:

$$L(\mu\text{H}) = \frac{d^2N^2}{18d + 40\ell} = \frac{(0.50)^2(38)^2}{18(0.50) + 40(1.73)} = \frac{361}{78} = 4.62 \ \mu\text{H}$$

Example 4: Design a solenoid inductor with an inductance of 8 μH if the form on which the coil is wound has a diameter of 1 in and a length of 0.75 in.

Answer: Rearranging the equation in the previous example:

$$N = \sqrt{\frac{L(18d + 40\ell)}{d^2}} = \sqrt{\frac{8(18(1) + 40(0.75))}{1^2}} = 19.6 \text{ turns}$$

A 20-turn coil would be close enough in practical work. Since the coil will be 0.75 in long, the number of turns per inch will be 19.6/0.75 = 26.1. A #17 enameled wire (or anything smaller) can be used. In practice, you wind the required number of turns on the form and then adjust the spacing between the turns to make a uniformly spaced coil of 0.75 in long.

On the Internet you'll find a number of free web-based inductor calculators. Some are quite advanced, allowing you to input inductance, diameter, and length values, and afterward providing you with the required number of turns, possible number of layers required, dc resistance of the wire, wire gauge to use, and so on. Check these tools out—you can save yourself the trouble of doing calculations and looking up wire dimensions and such.

2.24.7 Energy Within an Inductor

An ideal inductor, like an ideal capacitor, doesn't dissipate energy, but rather stores it in the magnetic field and later returns it to the circuit when the magnetic field collapses. The energy E_L stored in the inductor is found by using the generalized power law $P = IV$, along with the definition of power $P = dW/dt$, and the inductor equation $V = L\,dI/dt$. By equating the work W with E_L we get:

$$E_L = \int_{I=0}^{I=I\text{final}} P\,dt = \int_{I=0}^{I=I\text{final}} I\,V\,dt = \int_{I=0}^{I=I\text{final}} I\,L\frac{dI}{dt}dt = \int_{I=0}^{I=I\text{final}} L\,I\,dt = \frac{1}{2}LI^2 \tag{2.55}$$

where E_L = energy in joules, I = current in amps, and L = inductance in henrys. Note that in a real inductor a small portion of energy is lost to resistive heating through the inductor's internal resistance.

2.24.8 Inductor Cores

To conserve space and material, inductors are often wound on a magnetic core material, such as laminated iron or a special molded mix made from iron powder or ferrite material (iron oxide mixed with manganese, zinc, nickel, and other ingredients). A magnetic core increases the magnetic flux density of a coil greatly, and thus increases the inductance. This is further intensified if the magnetic core is formed into a doughnut-shaped toroid. See Fig. 2.130.

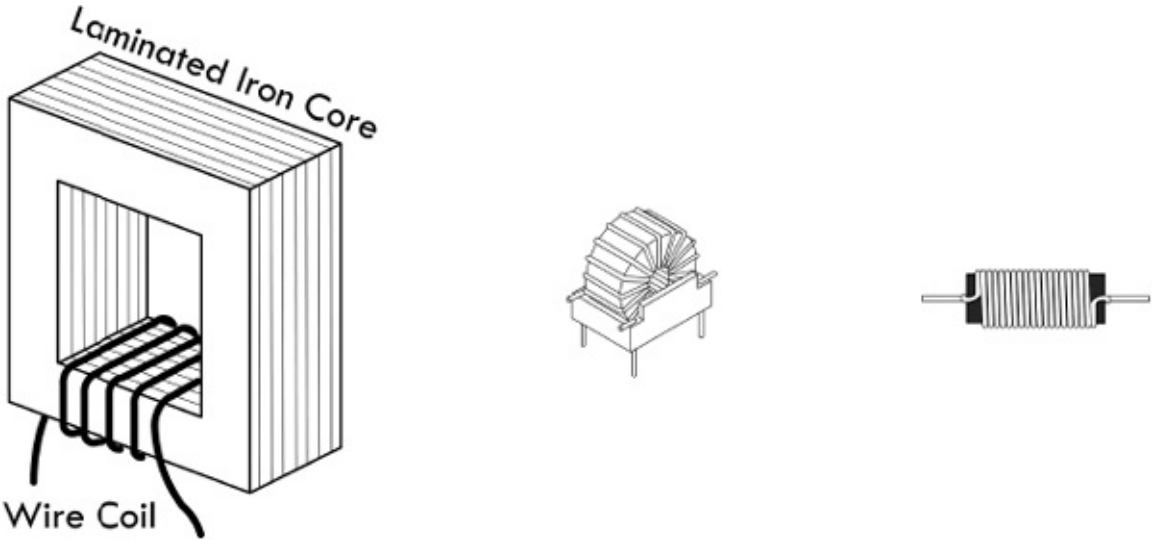


FIGURE 2.130 Magnetic core inductors.

The reason a magnetic core is so influential has to do with the magnetization that occurs within it as the outer coil passes current. When current is sent through the wire coil, a relatively weak magnetic field is set up at its center. This magnetic field, which we'll call the external magnetic field, causes the atomic rearrangement of magnetic dipoles (refer to Fig. 2.112) within the core material. This realignment is such as to rotate the dipole moments in a common direction. As more current is passed through the coil, more and more dipoles line up. The core itself is now generating a magnetic field as a result of the dipole alignment. The net magnetic field B_{total} generated by the whole inductor (coil and core) then becomes the sum of the external field (coil) and a term that is proportional to the magnetization M present in the core itself, as the external field is applied by the coil through the core. Mathematically, this is expressed:

$$B_{total} = B_{ext} + \mu_0 M \tag{1}$$

where μ_0 is the permeability of free space. The magnetic intensity H due to real current in the coil, as opposed to the magnetic intensity generated by atomic magnetization of the core, is expressed:

$$H = \frac{B_{ext}}{\mu_0} = \frac{B_{total}}{\mu_0} - M \tag{2}$$

This can be further reduced, using the notions of susceptibility and permeability (consult a physics book), to:

$$B_{total} = \mu H \tag{3}$$

where μ is the permeability of the core material.

The ratio of magnetic flux density produced by a given core material compared to the flux density produced by an air core is called the relative permeability of the material $\mu_R = \mu/\mu_0$. For example, an air core generating a flux density of 50 lines of force per square inch can be made to generate 40,000 lines of force per square inch with an iron core inserted. The ratio of these flux densities, iron core to air core, or the relative permeability is 40,000/50, or 800. Table 2.8 shows permeabilities of some popular high-permeability materials.

TABLE 2.8 Permeability of Various Materials

MATERIAL	APPROXIMATE MAXIMUM PERMEABILITY (H/M)	APPROXIMATE MAXIMUM RELATIVE PERMEABILITY	APPLICATIONS
Air	1.257×10^{-6}	1	RF
Ferrite U60	1.00×10^{-5}	8	UHF chokes
Ferrite M33	9.42×10^{-4}	750	Resonant circuits
Ferrite N41	3.77×10^{-3}	3000	Power circuits
Iron (99.8% pure)	6.28×10^{-3}	5000	
Ferrite T38	1.26×10^{-2}	10,000	Broadband transformers
45 Permalloy	3.14×10^{-2}	25,000	
Silicon 60 steel	5.03×10^{-2}	40,000	Dynamos, transformers
78 Permalloy	0.126	100,000	
Supermalloy	1.26	1,000,000	Recording heads

Problems with Magnetic Cores

When a magnetic core material is conductive (e.g., steel), there is a phenomenon known as eddy currents that arise within the core material itself when there is an applied magnetic field that is changing. For example, in Fig. 2.131a, when an increasing current is supplied through the outer coil, a changing magnetic flux passes through the core. This in turn induces a circular current flow within the core material. Eddy currents that are induced in the core represent loss in the form of resistive heating and can be a significant disadvantage in certain applications (e.g., power transformers). Eddy current losses are higher in materials with low resistivity.

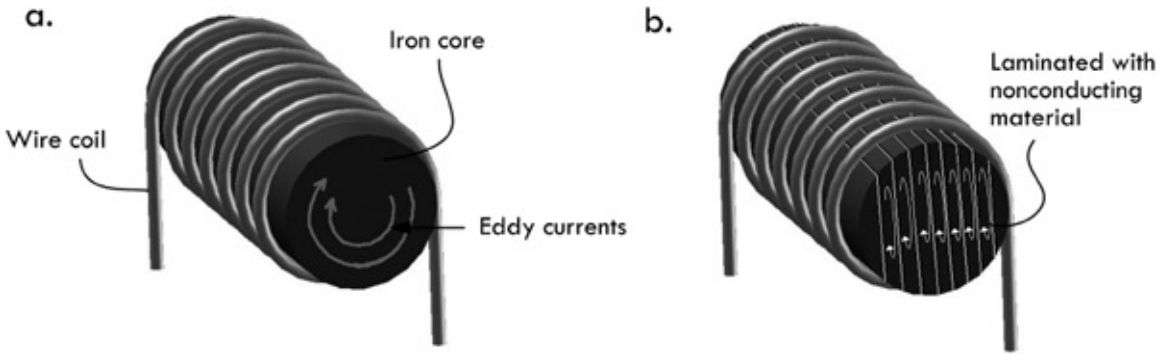


FIGURE 2.131 (a) Large eddy currents within core. (b) Eddy currents are reduced within laminated core.

To avoid eddy currents, the conductive core (steel, in this case) can be laminated together with insulating sheets of varnish or shellac. Current will still be induced in the sheets of the core, as shown in Fig. 2.131b, but because the area of the sheets is limited, so is the flux change, and therefore so are these currents.

Ferrite materials have quite high intrinsic resistivity compared to, say, steel (10 to 1,000 Ω-cm for Mn-Zn ferrites; 10^5 to 10^7 Ω-cm for Ni-Zn ferrites). Eddy current losses are therefore much less of a problem in ferrites, and this is the fundamental reason they are used in higher-frequency applications. Powdered iron cores, with their insulating compound present between iron particles, also reduce eddy currents, since the path for the eddy current is limited to the powder particle size.

A second difficulty with iron is that its permeability is not constant, but varies with the strength of the magnetic field and, hence, with the current in the windings. (It also varies with temperature.) In fact, at sufficiently high magnetic fields, the core will saturate and its relative permeability will drop to a value near unity. Not only that, but the magnetic field in the iron depends on the past history of the current in the winding. This property of remanence is essential in a permanent magnet, but in an inductor it gives rise to additional losses, called hysteresis losses. See Fig. 2.132.

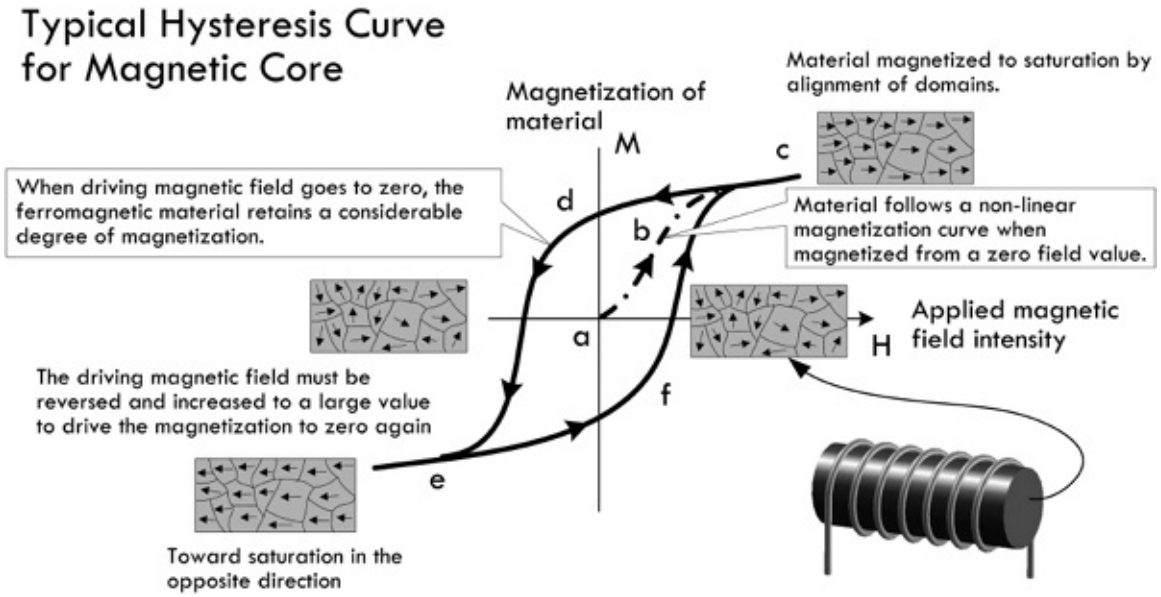


FIGURE 2.132 Hysteresis curve showing how the magnetization of the core material does not represent a reversible process. At point a, no current is applied through the coil. Along path a-b-c, coil current, and thus H (applied magnetic field or magnetizing force), increases, causing the core's magnetic dipole moments (in this case localized in domains) to rotate in proportion. As we approach point c, the core reaches saturation—increasing H causes no appreciable increase in M (magnetization or magnetic dipole density of core)—the domain's dipole moments are aligned as much as possible parallel to H. Saturation causes a rapid decrease in permeability. The saturation point of a magnetic core varies with material makeup; air and other nonmagnetic materials do not saturate—they have a permeability of 1. As H is decreased, the magnetization M does not follow the same path back down, but follows path c-d-e. Notice that when H goes to zero, the core remains magnetized. In essence, the core has become a permanent magnet. The term *retentivity* is used to describe this effect, and it presents another set of losses caused by hysteresis. In order to demagnetize the core, a reverse force is necessary to overcome the residual

magnetism retained by the core. In other words, H must go negative in the opposite direction, driving the magnetic domains back into a random orientation. At point e, the core has again reached saturation; however, the magnetic dipoles (domains) are now pointing in the opposite direction. To reach saturation in the opposite direction again, H must be applied, as shown in path e-f-c. Air cores and other materials with a permeability of 1, such as brass and aluminum, are immune to hysteresis effects and losses.

To avoid loses associated with hysteresis, it is important to not run the core inductor into saturation. This can be accomplished by running the inductor at lower current, using a larger core, altering the number of turns, using a core with lower permeability, or using a core with an air gap.

It is possible for the eddy current and hysteresis losses to be so large that the inductor behaves more like a resistor. Furthermore, there is always some capacitance between the turns of the inductor, and under some circumstances an inductor may act like a capacitor. (We'll discuss this a bit later.)

Table 2.9 shows a comparison between the various core inductors.

TABLE 2.9 Comparison of Inductor Cores

Air core	Relative permeability is equal to 1. The inductance is independent of current-flow because air does not saturate. Limited to low inductance values, but can be operated at very high frequencies (e.g., RF applications ~1 GHz).
Iron core	Permeability is a factor of 1000 greater than for air core, but inductance is highly dependent on the current flowing through the coil because the core saturates. Chiefly used in power-supply equipment. Highly prone to eddy currents (power loss) due to the core's high conductivity. Losses caused by eddy currents can be reduced by laminating the core (cutting it into thin strips separated by insulation, such as varnish or shellac. Experience significant power losses due to hysteresis. Eddy currents and hysteresis losses in iron increase rapidly as the frequency of ac increases. Limited to power-line and audio frequencies up to about 15,000 Hz. Laminated iron cores are useless at radio frequencies.
Powdered iron	Ground-up iron (powder) mixed with a binder or insulating material reduces eddy currents significantly, since the iron particles are insulated from one another. Permeability is low compared to iron core due to insulator concentration. Slug-tuned inductors, where the slug is powered iron can be made adjustable and are useful in RF work, even up to the VHF range. Manufacturers offer a wide variety of core materials, or mixes, to provide units that will perform over a desired frequency range with a reasonable permeability. Toroidal Cores are considered self-shielding. Manufacturers provide an inductance index A_L for their toroidal cores. For powdered-iron toroids, A_L provides the inductance in μH per 100 turns of wire on the core, arranged in a single layer. Formulas found in Fig. 2.133, and example problems 5 and 6 illustrate how to calculate inductance of a powdered-iron toroidal inductor.
Ferrite	Composed of nickel-zinc ferrites for lower permeability ranges and of manganese-zinc ferrites for higher permeabilities, these cores span the permeability range from 20 to above 10,000. They are used for RF chokes and wideband transformers. Ferrites are often used, since they are nonconductors and are immune to eddy currents. Like powdered-iron toroids, ferrite toroids also have an A_L value provided by the manufacturers. However, unlike powdered-iron toroids, ferrite A_L values are given in mH per 1000 turns. Formulas found in Fig. 2.133 and example problems 7 and 8 illustrate how to calculate inductance of a ferrite toroidal inductor.

Example 5: What is the inductance of a 100-turn coil on a powdered iron toroidal core with an index of inductance of 20?

Answer: To solve this you'd have to consult the manufacturer's data sheets, but, here, use a T-12-2 from the table in Fig. 2.133. Using the equation for powdered iron toroids in Fig. 2.133, and inserting $A_L = 20$ for the T-12-2, and $N = 100$:

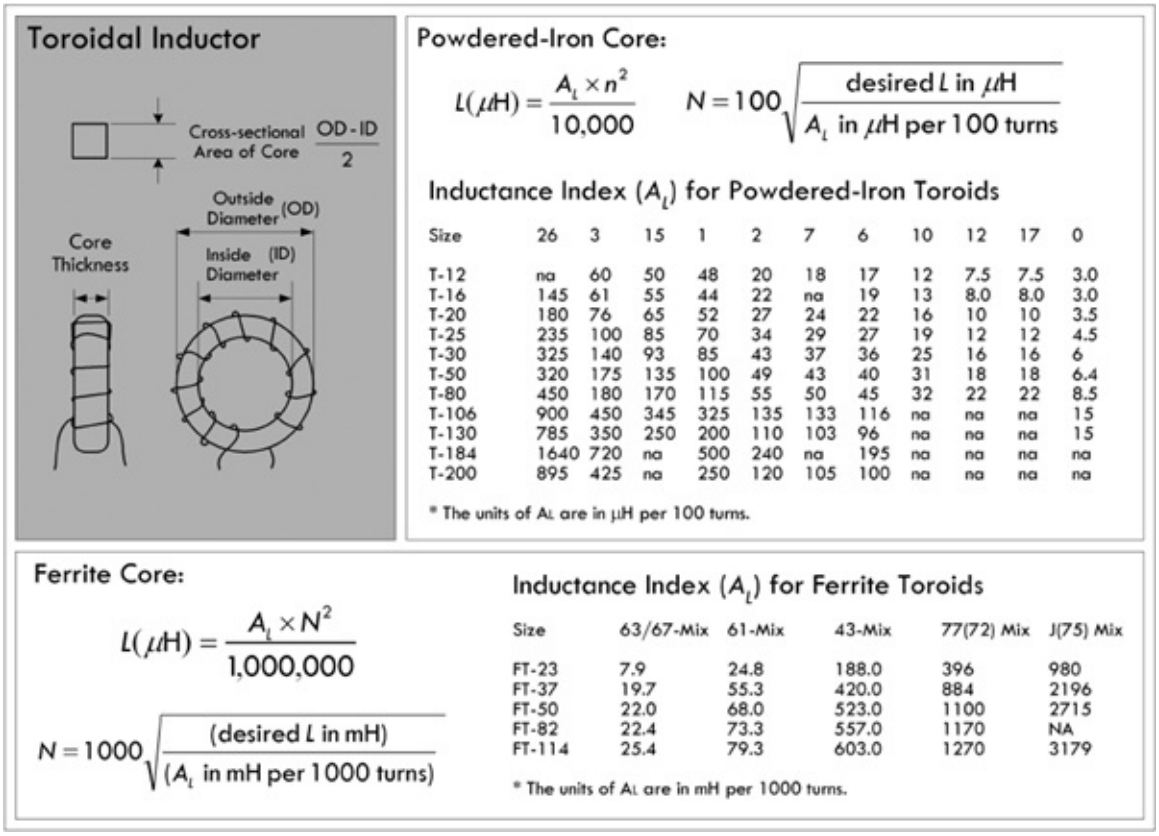


FIGURE 2.133

$$L(\mu H) = \frac{A_L \times N^2}{10{,}000} = \frac{20 \times 100^2}{10{,}000} = 200\ \mu H$$

Example 6: Calculate the number of turns needed for a 19.0-µH coil if the index of inductance for the powdered iron toroid is 36.

Answer: Use:

$$N = 100\sqrt{\frac{\text{desired } L \text{ in } \mu H}{A_L \text{ in } \mu H \text{ per 100 turns}}} = 100\sqrt{\frac{19.0}{36}} = 72.6 \text{ turns}$$

Example 7: What is the inductance of a 50-turn coil on a ferrite toroid with an index of inductance of 68?

Answer: To solve this you'd have to consult the manufacturer's data sheets, but here we'll use a FT-50 with 61-Mix, as shown in the table in Fig. 2.133. Using the equation for ferrite toroids in Fig. 2.133, and inserting A_L = 68 for the FT-50-61, and N = 50:

$$L(\mu H) = \frac{A_L \times N^2}{1{,}000{,}000} = \frac{68 \times 50^2}{1{,}000{,}000} = 0.17\ \mu H$$

Example 8: How many turns are needed for a 2.2-mH coil if the index of inductance for the ferrite toroid is 188?

Answer: Use

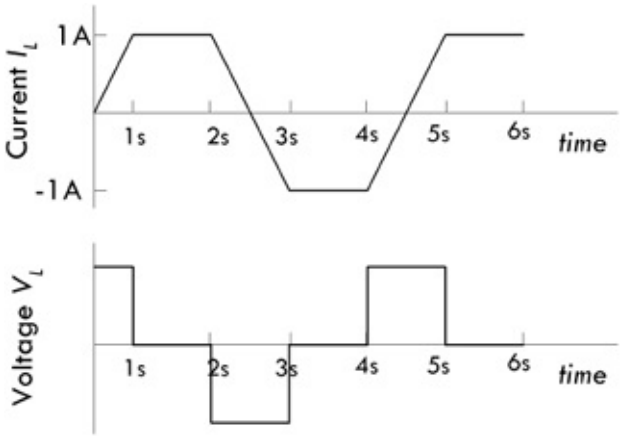
$$N = 1000\sqrt{\frac{\text{(desired } L \text{ in mH)}}{(A_L \text{ in mH per 1000 turns})}} = 1000\sqrt{\frac{2.2}{188}} = 108 \text{ turns}$$

2.24.9 Understanding the Inductor Equations

$$V_L = L\frac{dI_L}{dt}$$

The preceding inductor equation (which we derived earlier) has some curious properties that may leave you scratching your head. For starters, let's consider the dI_L/dt term. It represents the rate of change in current through the inductor with time. If there is no change in current flow through the inductor, there is no measured voltage across the inductor. For example, if we assume that a constant dc current has been flowing through an inductor for some time, then dI_L/dt is zero, making V_L zero as well. Thus, under dc conditions the inductor acts like a short—a simple wire.

However, if the current I_L is changing with time (increasing or decreasing), dI_L/dt is no longer zero, and an induced voltage V_L is present across the inductor. For example, consider the current waveform shown in Fig. 2.134. From time interval 0 to 1s, the rate of current change dI_L/dt is 1 A/s—the slope of the line. If the inductance L is 0.1 H, the induced voltage is simply (1 A/s)(0.1 H) = 0.1 V during this time interval—see the lower waveform. During interval 1s to 2s, the current is constant, making the dI_L/dt zero and, hence, the induced voltage zero. During interval 2s to 3s, dI_L/dt is −1 A/s, making the induced voltage equal to (−1 A/s)(0.1 H) = −0.1 V. The induced voltage waveform paints the rest of the picture.



From 0 - 1 and 4 - 5 sec :

$$V = L\frac{dI}{dt} = (0.1\text{H})\left(\frac{1\text{A}}{1\text{s}}\right) = 0.1\text{V}$$

From 2 - 3 sec :

$$V = L\frac{dI}{dt} = (0.1\text{H})\left(-\frac{1\text{A}}{1\text{s}}\right) = -0.1\text{V}$$

From 1 - 2, 3 - 4, 5 - 6 sec

$$V = L\frac{dI}{dt} = (0.1\text{H})\left(\frac{0}{1\text{s}}\right) = 0\text{V}$$

FIGURE 2.134

Example 9: If the current through a 1-mH inductor is given by the function $2t$ A, what is the voltage across it?

Answer:

$$V_L = L\frac{dI_L}{dt} = (1\times10^{-3}\text{ H})\frac{d}{dt}2t = (1\times10^{-3}\text{ H})\left(2\frac{I}{s}\right) = 2\times10^{-3}\text{ V} = 2\text{ mV}$$

Example 10: The current through a 4-mH inductor is given by $I_L = 3 - 2e^{-10t}$ A. What is the voltage across the inductor?

Answer:

$$V_L = L\frac{dI_L}{dt} = L\frac{d}{dt}[3-2e^{-10t}] = L(-2\times-10)e^{-10t} = (4\times10^{-3})(20)e^{-10t} = 0.08e^{-10t}\text{ V}$$

Example 11: Suppose the current flow through a 1-H inductor decreases from 0.60 A to 0.20 A during a 1-s period. Find the average voltage across the inductor during this period. See how this compared to the average induced voltage if the period is 100 ms, 10 ms, and 1 ms.

Answer: Here we ignore changes between the 1-s intervals, and take averages:

$$V_{AVE} = L\frac{\Delta I}{\Delta t} = 1\text{ H}\frac{0.20\text{ A} - 0.60\text{ A}}{1\text{ s}} = -0.40\text{ V} \qquad (1\text{ s})$$

$$V_{AVE} = L\frac{\Delta I}{\Delta t} = 1\text{ H}\frac{0.20\text{ A} - 0.60\text{ A}}{0.1\text{ s}} = -4\text{ V} \qquad (100\text{ ms})$$

$$V_{AVE} = L\frac{\Delta I}{\Delta t} = 1\text{ H}\frac{0.20\text{ A} - 0.60\text{ A}}{0.01\text{ s}} = -40\text{ V} \qquad (10\text{ ms})$$

$$V_{AVE} = L\frac{\Delta I}{\Delta t} = 1\text{ H}\frac{0.20\text{ A} - 0.60\text{ A}}{0.001\text{ s}} = -400\text{ V} \qquad (1\text{ ms})$$

$$V_{AVE} = L\frac{\Delta I}{\Delta t} = 1\text{ H}\frac{0.20\text{ A} - 0.60\text{ A}}{0\text{ s}} = -\infty \qquad (\text{Instantaneous})$$

From Example 11, notice how the induced voltage grows considerably larger as the change in current flow is more abrupt. When the change in current flow is instantaneous, the inductor equation predicts an infinite induced voltage. How can this be?

The answer to this dilemma is explained by the following example. Say you have an ideal inductor attached to a 10-V battery via a switch, as shown in Fig. 2.135. The moment the switch is closed, according to the inductor equation, *dI*/*dt* should be infinite (assuming ideal battery, wires, and coil wire). This means that the induced voltage should rise in proportion with the applied voltage, forever, and no current should flow. In other words, the reverse voltage should jump to infinity. (See Fig. 135b.) Likewise, if the switch is opened, as shown in Fig. 135c, an infinite forward voltage is predicted. The answer to this problem is rather subtle, but fundamentally important. As it turns out, these infinities we predict are never observed in the real world. There is always internal resistance within a real inductor as well as internal capacitance (as well as internal resistance and capacitance in the rest of the circuit). A realistic model of an inductor is shown in Fig. 135d, incorporating internal resistance and capacitance. It is these “imperfections” that explain why the predictions are not observed.

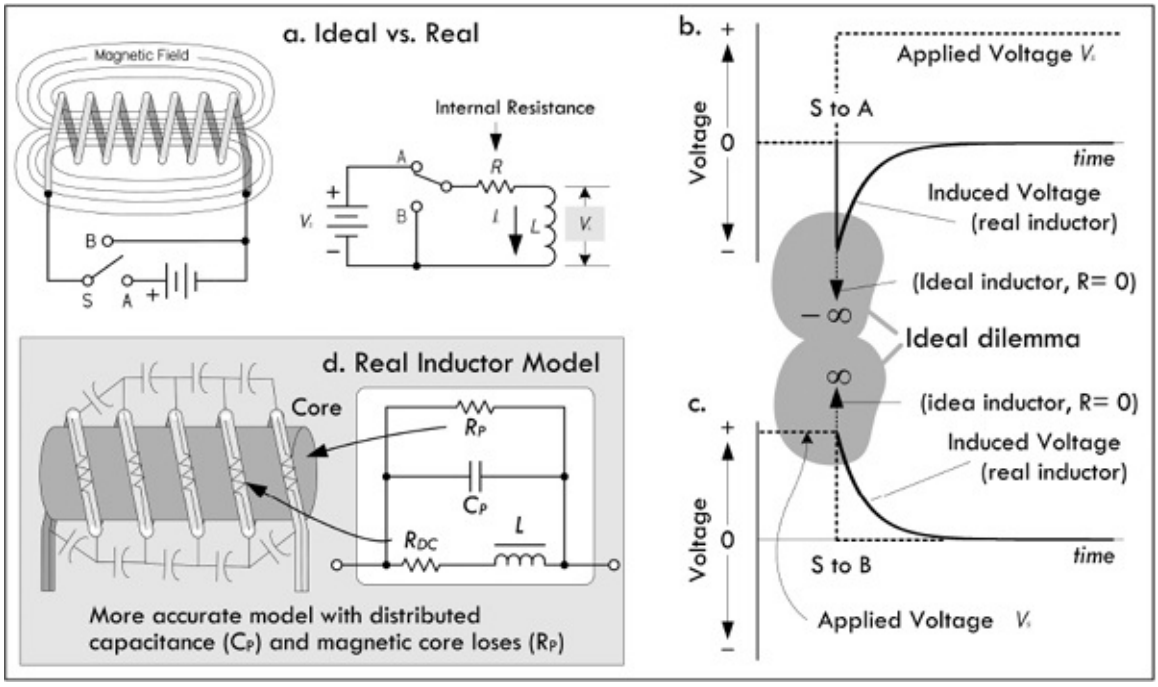


FIGURE 2.135 Real inductors never generate infinite induced voltages.

As we can see from the last example, ignoring internal resistance can cause conceptual headaches. You may be wondering why someone didn't simply create an inductor equation that automatically included an internal resistance term. The fact of the matter is that we *should*, especially when considering a simple circuit like that shown in Fig. 2.135. However, it is important to define inductance as a unique quantity and associate its effects with changing magnetic field energy alone, not with resistive heating within the coil or magnetic core losses, or with distributed capacitance between the coil loops. As it turns out, when you start analyzing more complex circuits (say RL circuits and RLC circuits), the discrete resistance present within the circuit will keep the inductor equation from freaking out. In more precise circuits, knowing the internal resistance of the inductor is critical. A standard practice is to represent an inductor as an ideal inductor in series with a resistor R_{DC}, where R_{DC} is called the dc resistance of the inductor. More accurate models throw in parallel capacitance (interloop capacitance) and parallel resistance (representing magnetic core losses), which become important in high-frequency applications.

Before we move on, it should be noted that even though we must assume there is internal resistance within an inductor, it is still possible for induced voltages to reach surprisingly high values during transient conditions. For example, turning off inductive circuits can lead to some dangerously high voltages that can cause arcing and other problems requiring special handling.

Example 12: Suppose you apply a linearly increasing voltage across an ideal 1-H inductor. The initial current through the inductor is 0.5 A, and the voltage ramps from 5 to 10 V over a 10 ms interval. Calculate the current through the inductor as a function of time.

Answer: Kirchhoff's voltage law gives:

$$V_{\text{applied}} - L\frac{dI}{dt} = 0$$

Integrating this gives:

$$\int_{t'=0}^{t} \frac{dI}{dt'}\,dt' = \int_{t'=0}^{t} V_{\text{applied}}\,dt'$$

or:

$$I(t) - I(0) = \int_{t'=0}^{t} (mt' + b)\,dt'$$

Carrying out the integration gives:

$$I(t) = I(0) + \frac{1}{2}mt^2 + bt$$

At t = 0.01 s (the end of the ramp), the current is:

$$I(0.01) = 0.5 + \frac{1}{2} \times 500 \times 0.01^2 + 5 \times 0.01 = 0.575 \text{ A}$$

The answer in the text ignores the initial current and has a numerical error in the last equality.

2.24.10 Energizing LR Circuit

When a resistor is placed in series with an inductor, the resistance controls the rate at which energy is pumped into the magnetic field of the inductor (or pumped back into the circuit when the field collapses). When we consider an RL circuit with a dc supply and a switch, as shown in Fig. 2.136, the energizing response that begins the moment the switch is closed (t = 0) is illustrated by the voltage and current response curves in Fig. 2.136, as well as the equations to the right.

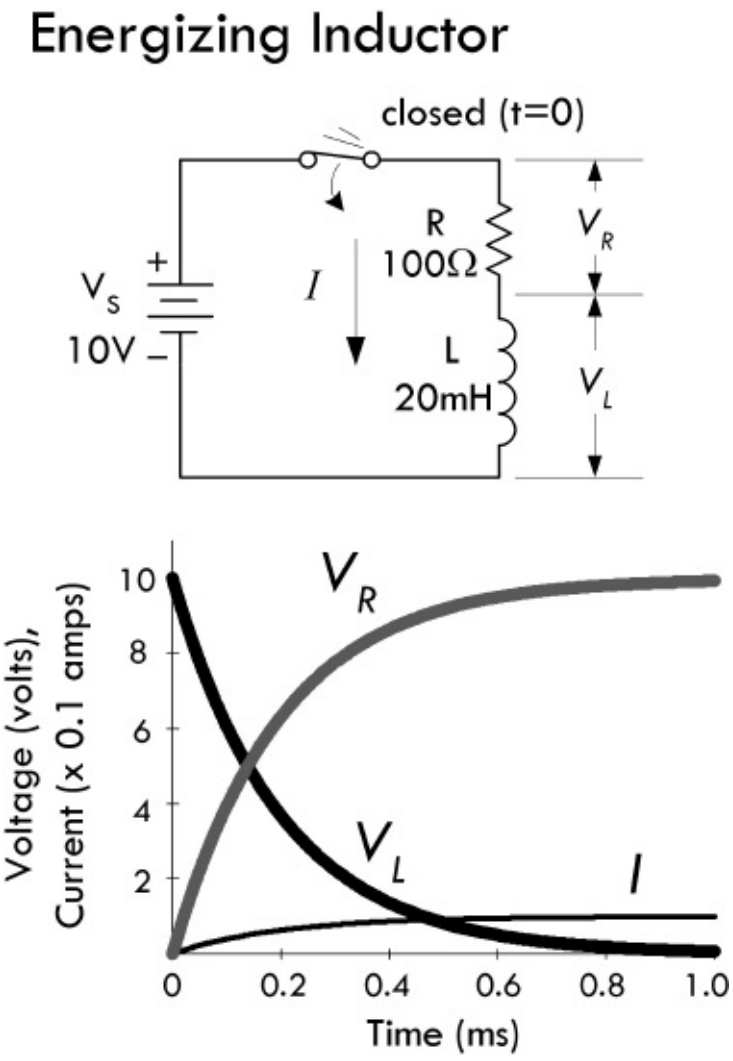


FIGURE 2.136

Current and voltage equations for RL energizing circuit:

$$I = \frac{V_s}{R}(1 - e^{-t/(L/R)}), \quad \frac{t}{(L/R)} = -\ln\left(\frac{1 - V_s/R}{V_s}\right)$$

$$V_R = IR = V_s(1 - e^{-t/(L/R)}), \quad \frac{t}{(L/R)} = -\ln\left(\frac{V_R - V_s}{V_s}\right)$$

$$V_L = L\frac{dI}{dt} = V_s e^{-t/(L/R)}, \quad \frac{t}{(L/R)} = -\ln\left(\frac{V_L}{V_s}\right)$$

$\tau = L/R$ time constant

where I is the current in amps, V_S is the source voltage in volts, R is the resistance in ohms, L is the inductance in henrys, t is the time in seconds after the source voltage is applied, e = 2.718, V_R is the resistor voltage in volts, and V_L is the inductor voltage in volts. The graph shown to the left is for a circuit with R = 100 Ω and L = 20 mH.

You can derive the expressions for the energizing response of an RL circuit by applying Kirchhoff's law, summing the voltages around the closed loop:

$$V_s = IR + L\frac{dI}{dt}$$

Rewriting in the standard form gives:

$$\frac{dI}{dt} + \frac{R}{L}I = \frac{V}{L}$$

After solving this linear, first-order nonhomogeneous differential equation, using the initial condition that the current before the switch was closed was zero I (0) = 0, the solution for the current becomes:

$$I = \frac{V_s}{R}(1 - e^{-t/(L/R)})$$

We plug this into Ohm's law to find the voltage across the resistor:

$$V_R = IR = V_S\,(1 - e^{-t/(L/R)})$$

and plug it into the expression for the inductor voltage:

$$V_L = L\frac{dI}{dt} = V_s e^{-t/(L/R)}$$

To understand what's going on within the energizing RL circuit, we first pretend the resistor value is zero. With no resistance, when the switch is closed, the current would increase forever (according to Ohm's law and assuming an ideal voltage source), always growing just fast enough to keep the self-induced voltage equal to the applied voltage.

But when there is resistance in the circuit, the current is limited—Ohm's law defines the value the current can reach. The reverse voltage generated in L must only equal the difference between the applied voltage and the drop across R. This difference becomes smaller as the current approaches the final Ohm's law value. Theoretically, the reverse voltage never quite disappears, and so the current never quite

reaches the no-inductor value. In practical terms, the differences become immeasurable after a short duration.

The time in seconds required for the current to build up to 63.2 percent of the maximum value is called the time constant, and is equal to L/R. After each time interval equal to this constant, the circuit conducts an additional 63.2 percent of the remaining current. This behavior is graphed in Fig. 2.137. As is the case with capacitors, after five time constants the current is considered to have reached its maximum value.

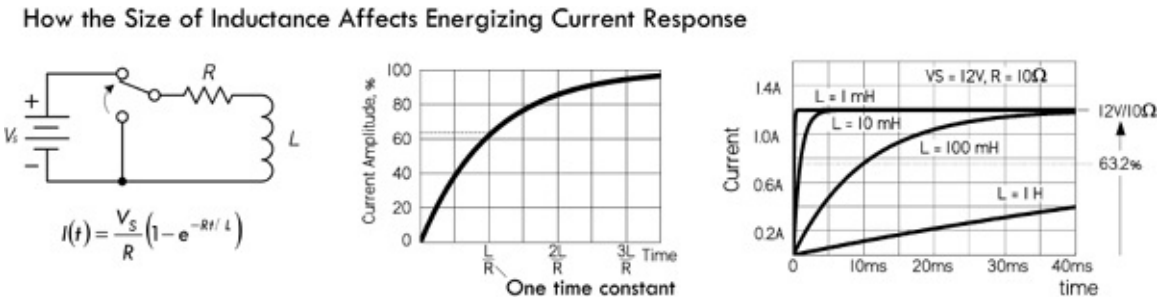


FIGURE 2.137

Example 13: If an RL circuit has an inductor of 10 mH and a series resistor of 10 Ω, how long will it take for the current in the circuit to reach full value after power is applied?

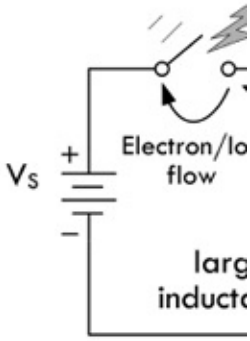
Answer: Since reaching maximum current takes approximately five time constants, $t = 5\tau = 5\ (L/R) = 5\ (10 \times 10^{-3}\ \text{H})/10\ \Omega = 5.0 \times 10^{-3}$ s or 5.0 ms.

Note that if the inductance is increased to 1.0 H, the required time increases to 0.5 s. Since the circuit resistance didn't change, the final current is the same for both cases in this example. Increasing inductance increases the time required to reach full current. Figure 2.137 shows current response curves for various inductors with the same resistance.

2.24.11 Deenergizing LR Circuit

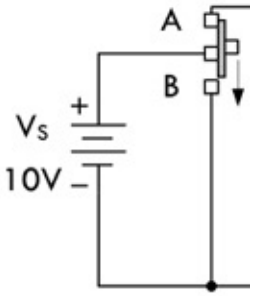
Unlike a capacitor that can store energy in the form of an electric field when an applied voltage is cut (say, by means of a switch), an inductor does not remain "charged," or energized, since its magnetic field collapses as soon as current ceases. The energy stored in the magnetic field returns to the circuit. Now, when current flow is cut (say, by means of a switch), predicting the current flow and voltage drops within an RL circuit is a bit tricky. We can say that the instant the switch is opened, a rapid collapse of the magnetic field induces a voltage that is usually many times larger than the applied voltage, since the induced voltage is proportional to the rate at which the field changes. The common result of opening the switch in such a circuit is that a spark or arc forms at the switch contacts during the instant the switch opens—see Fig. 2.138a. When the inductance is large and the current in the circuit is high, large amounts of energy are released in a very short time. It is not unusual for the switch contacts to burn or melt under such circumstances. The spark or arc at the opened switch can be reduced or suppressed by connecting a suitable capacitor and resistor in series across the contacts. Such an RC combination is called a *snubber network*. Transistor switches connected to large inductive loads, such as relays and solenoids, also require protection. In most cases, a small power diode connected in reverse across the relay coil will prevent field-collapse currents from harming the transistor—more on this in a moment.

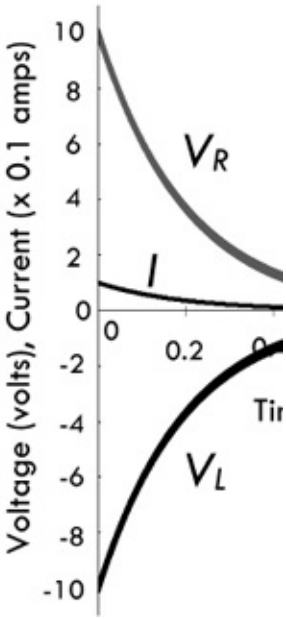
A. Break in circ



B. Assuming nc

Switched to B a





In A, a break in current generating a large forwa becomes so large that e liberated electrons make discharge across the swit

In B, if we remove th thrown to position B), we

In electronics, the bel is often used to compare electrical power levels; however, what's more common in electronics and elsewhere is to use decibels, abbreviated dB. A decibel is $\frac{1}{10}$ of a bel (similar to a millimeter being $\frac{1}{10}$ of a centimeter). It takes 10 decibels to make 1 bel. So in light of this, we can compare power levels in terms of decibels:

$$\text{dB} = 10 \log\left(\frac{P_1}{P_0}\right)$$

(2.89) Decibels in terms of power

Example: Express the gain of an amplifier (output power divided by input power) in terms of decibels, if the amplifier takes a 1-W signal and boosts it up to a 50-W signal.

Answer: Let P_0 represent the 1-W reference power, and let P_1 be the compared power:

$$\text{dB} = 10 \log\left(\frac{50\ \text{W}}{1\ \text{W}}\right) = 10 \log (50) = 17.00\ \text{dB}$$

The amplifier in this example has a gain of nearly 17.00 dB (17 decibels).

Sometimes when comparing signal levels in an electronic circuit, we know the voltage or current of the signal, but not the power. Though it's possible to calculate the power, given the circuit impedance, we take a shortcut by simply plugging ac Ohm's law into the powers in the decibel expression. Recall that $P = V^2/Z = I^2Z$. Now, this holds true only as long as the impedance of the circuit doesn't change when the voltage or current changes. As long as the impedance remains the same, we get a comparison of voltage signals and current signals in terms of decibels:

$$\text{dB} = 10 \log\left(\frac{V_1^2}{V_0^2}\right) = 20 \log\left(\frac{V_1}{V_0}\right)$$

and

$$\text{dB} = 10 \log\left(\frac{I_1^2}{I_0^2}\right) = 20 \log\left(\frac{I_1}{I_0}\right)$$

(2.90) Decibels in terms of voltage and current levels

(In the expressions above we used the laws of logarithms to remove square terms. For example:

$$10 \log (V_1^2/V_0^2) = 10(\log V_1^2 - \log V_0^2)$$

$$= 10(2 \log V_1 - 2 \log V_0) = 20(\log V_1 - \log V_2)$$

$$= 20 \log (V_1/V_0).$$

Notice that the impedance terms cancel and the final result is a factor of 2 greater—a result of the square terms in the log being pushed out (see law of logarithms). The power and voltage and current expressions are all fundamentally the same—they are all based on the ratio of powers.

There are several power ratios you should learn to recognize and be able to associate with the corresponding decibel representations.

For example, when *doubling power*, the final power is always 2 times the initial (or reference) power—it doesn't make a difference if you are going from 1 to 2 W, 40 to 80 W, or 500 to 1000 W, the ratio is always 2. In decibels, a power ratio of 2 is represented as:

$$dB = 10 \log (2) = 3.01 \text{ dB}$$

There is a 3.01-dB gain if the output power is twice the input power. Usually, people don't care about the .01 fraction and simply refer to the power doubling as a 3-dB increase in power.

When the *power is cut in half*, the ratio is always 0.5 or 1/2. Again, it doesn't matter if you're going from 1000 to 500 W, 80 to 40 W, or 2 to 1 W, the ratio is still 0.5. In decibels, a power ratio of 0.5 is represented as:

$$dB = 10 \log (0.5) = -3.01 \text{ dB}$$

A negative sign indicates a decrease in power. Again, people usually ignore the .01 fraction and simply refer to the power being cut in half as a −3-dB change in power or, more logically, a 3-dB decrease in power (the term "decrease" eliminates the need for the negative sign).

Now if you increase the power by 4, you can avoid using the decibels formula and simply associate such an increase with doubling two times: 3.01 dB + 3.01 dB = 6.02 dB or around 6 dB. Likewise, if you increase the power by 8, you, in effect, double four times, so the power ratio in decibels is 3.01 × 4 = 12.04, or around 12 dB.

The same relationship is true of power decreases. Each time you cut the power in half, there is a 3.01-dB, or around 3-dB decrease. Cutting the power by four times is akin to cutting in half twice, or 3.01 + 3.01 dB = 6.02 dB, or around a 6-dB decrease. Again, you can avoid stating "decrease" and simply say that there is a change of −6 dB.

Table 2.12 shows the relationship between several common decibel values and the power change associated with those values. The current and voltage changes are also included, but these are valid only if the impedance is the same for both values.

TABLE 2.12 Decibels and Power Ratios

COMMON DECIBEL VALUES AND POWER-RATIO EQUIVALENTS			COMMON DECIBEL VALUES AND POWER-RATIO EQUIVALENTS		
dB	P_2/P_1	V_2/V_1 or I_2/I_1	dB	P_2/P_1	V_2/V_1 or I_2/I_1
120	10^{12}	10^6	−120	10^{-12}	10^{-6}
60	10^6	10^3	−60	10^{-6}	10^{-3}
20	10^2	10.0	−20	10^{-2}	0.1000
10	10.00	3.162	−10	0.1000	0.3162
6.0206	4.0000	2.0000	−6.0206	0.2500	0.5000
3.0103	2.0000	1.4142	−3.0103	0.5000	0.7071
1	1.259	1.122	−1	0.7943	0.8913
0	1.000	1.000	0	1.000	1.000

* Voltage and current ratios hold only if the impedance remains the same.

2.31.1 Alternative Decibel Representations

It is often convenient to compare a certain power level with some standard reference. For example, suppose you measured the signal coming into a receiver from an antenna and found the power to be 2 ×

10^{-13} mW. As this signal goes through the receiver, it increases in strength until it finally produces some sound in the receiver speaker or headphones. It is convenient to describe these signal levels in terms of decibels. A common reference power is 1 mW. The decibel value of a signal compared to 1 mW is specified as "dBm" to mean decibels compared to 1 mW. In our example, the signal strength at the receiver input is:

$$\mathrm{dB}_m = 10 \log\left(\frac{2\times10^{-13}\ \mathrm{mW}}{1\ \mathrm{mW}}\right) = -127\ \mathrm{dBm}$$

There are many other reference powers used, depending upon the circuits and power levels. If you use 1 W as the reference power, then you would specify dBW. Antenna power gains are often specified in relation to a dipole (dBd) or an isotropic radiator (dBi). Anytime you see another letter following dB, you will know that some reference power is being specified. For example, to describe the magnitudes of a voltage relative to a 1-V reference, you indicate the level in decibels by placing a "V" at the end of dB, giving units of dBV (again, impedances must stay the same). In acoustics, dB, SPL is used to describe the pressure of one signal in terms of a reference pressure of 20 μPa. The term decibel is also used in the context of sound (see Section 15.1)

2.32 Input and Output Impedance

2.32.1 Input Impedance

Input impedance Z_{IN} is the impedance "seen" looking into the input of a circuit or device (see Fig. 2.186). Input impedance gives you an idea of how much current can be drawn into the input of a device. Because a complex circuit usually contains reactive components such as inductors and capacitors, the input impedance is frequency sensitive. Therefore, the input impedance may allow only a little current to enter at one frequency, while highly impeding current enters at another frequency. At low frequencies (less than 1 kHz) reactive components may have less influence, and the term *input resistance* may be used—only the real resistance is dominant. The effects of capacitance and inductance are generally more significant at high frequencies.

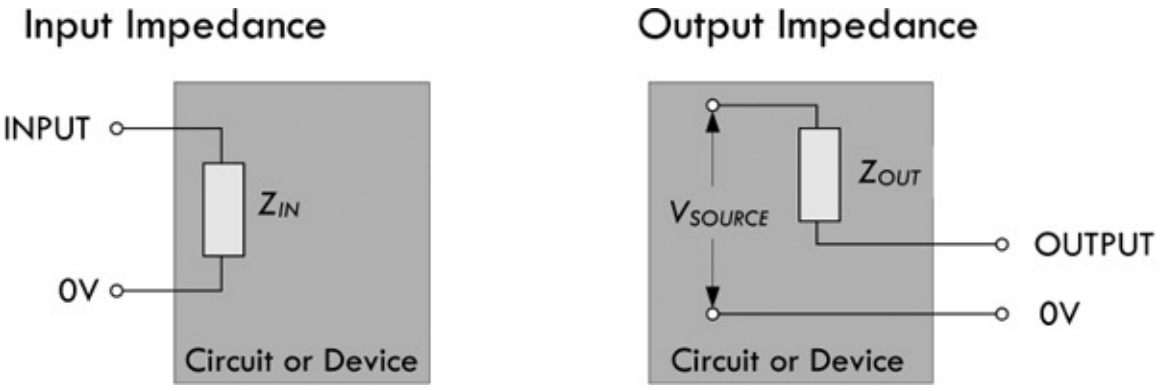


FIGURE 2.186

When the input impedance is small, a relatively large current can be drawn into the input of the device when a voltage of specific frequency is applied to the input. This typically has an effect of dropping the source voltage of the driver circuit feeding the device's input. (This is especially true if the driver circuit has a large output impedance.) A device with low input impedance is an audio speaker (typically 4 or 8 Ω), which draws lots of current to drive the voice coil.

When the input impedance is large, on the other hand, a small current is drawn into the device when a voltage of specific frequency is applied to the input, and, hence, doesn't cause a substantial drop in source voltage of the driver circuit feeding the input. An op amp is a device with a very large input impedance (1 to 10 MΩ); one of its inputs (there are two) will practically draw no current (in the nA range). In terms of audio, a hi-fi preamplifier that has a 1-MΩ radio input, a 500-kΩ CD input, and a 100-kΩ tap input all have high impedance input due to the preamp being a voltage amplifier not a current amplifier—something we will discuss later.

As a general rule of thumb, in terms of transmitting a signal, the input impedance of a device should be greater than the output impedance of the circuit supplying the signal to the input. Generally, the value should be 10 times as great to ensure that the input will not overload the source of the signal and reduce the strength by a substantial amount. In terms of calculations and such, the input impedance is by definition equal to:

$$Z_{in} = \frac{V_{in}}{I_{in}}$$

For example, Figure 2.187 shows how to determine what the input impedance is for two resistor circuits. Notice that in example 2, when a load is attached to the output, the input impedance must be recalculated, taking R_2 and R_{load} in parallel.

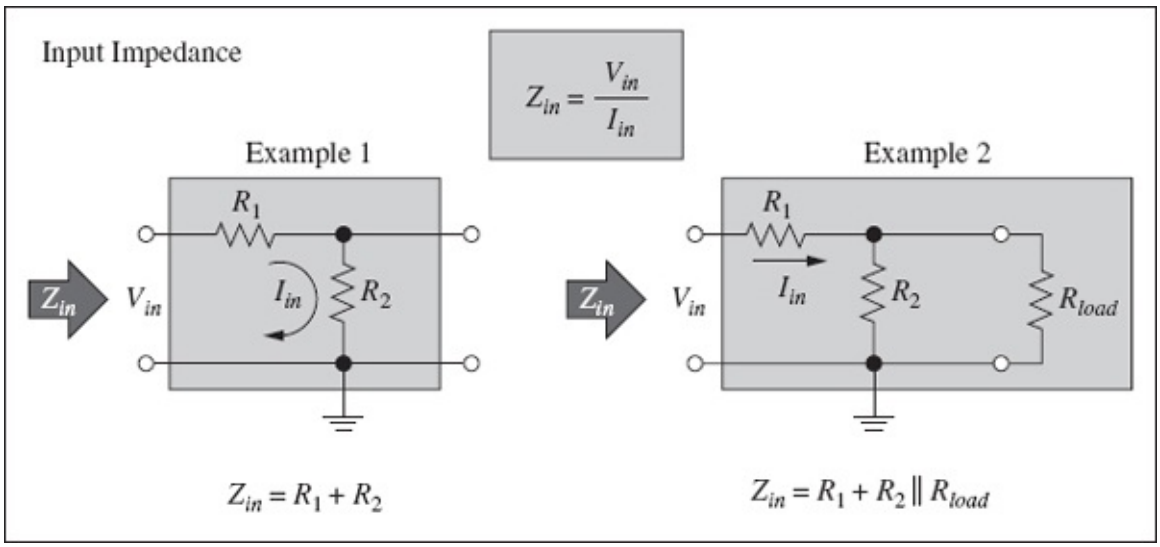


FIGURE 2.187

In Sec. 2.33.1, which covers filter circuits, you'll see how the input impedance becomes dependent on frequency.

2.32.2 Output Impedance

Output impedance Z_{OUT} refers to the impedance looking back into the output of a device. The output of any circuit or device is equivalent to an output impedance Z_{OUT} in series with an ideal voltage source V_{SOURCE}. Figure 2.186 shows the equivalent circuit; it represents the combined effect of all the voltage sources and the effective total impedance (resistances, capacitance, and inductance) connected to the output side of the circuit. You can think of the equivalent circuit as a Thevenin equivalent circuit, in which case it should be clear that the V_{SOURCE} present in Fig. 2.186 isn't necessarily the actual supply voltage of the circuit but the Thevenin equivalent voltage. As with input impedance, output impedance can be frequency dependent. The term *output resistance* is used in cases where there is little reactance within the circuit, or when the

frequency of operation is low (say, less than 1 kHz) and reactive effects are of little consequence; the effects of capacitance and inductance are generally most significant at high frequencies.

When the output impedance is small, a relatively large output current can be drawn from the device's output without significant drop in output voltage. A source with an output impedance much lower than the input impedance of a load to which it is attached will suffer little voltage loss driving current through its output impedance. For example, a lab dc power supply can be viewed as an ideal voltage source in series with a small internal resistance. A decent supply will have an output impedance in the milliohm (mΩ) range, meaning it can supply considerable current to a load without much drop in supply voltage. A battery typically has a higher internal resistance and tends to suffer a more substantial drop in supply voltage as the current demands increase. In general, a small output impedance (or resistance) is considered a good thing, since it means that little power is lost to resistive heating in the impedance and a larger current can be sourced. Op amps, which, we saw, have large input impedance and tend to have low output impedance.

When the output impedance is large, on the other hand, a relatively small output current can be drawn from the output of a device before the voltage at the output drops substantially. If a source with a large output impedance attempts to drive a load that has a much smaller input impedance, only a small portion appears across the load; most is lost driving the output current through the output impedance.

Again, the rule of thumb for efficient signal transfer is to have an output impedance that is at least $^1/_{10}$ that of the load's input impedance to which it is attached.

In terms of calculations, the output impedance of a circuit is simply its Thevenin equivalent resistance R_{th}. The output impedance is sometimes called the source impedance. In terms of determining the output impedance in circuit analysis, it amounts to "killing" the source (shorting it) and finding the equivalent impedance between the output terminals—the Thevenin impedance.

For example, in Figure 2.188, to determine the output impedance of this circuit, we effectively find the Thevinin resistance by "shorting the source" and removing the load, and determining the impedance between the output terminals. In this case, the output impedance is simply R_1 and R_2 in parallel.

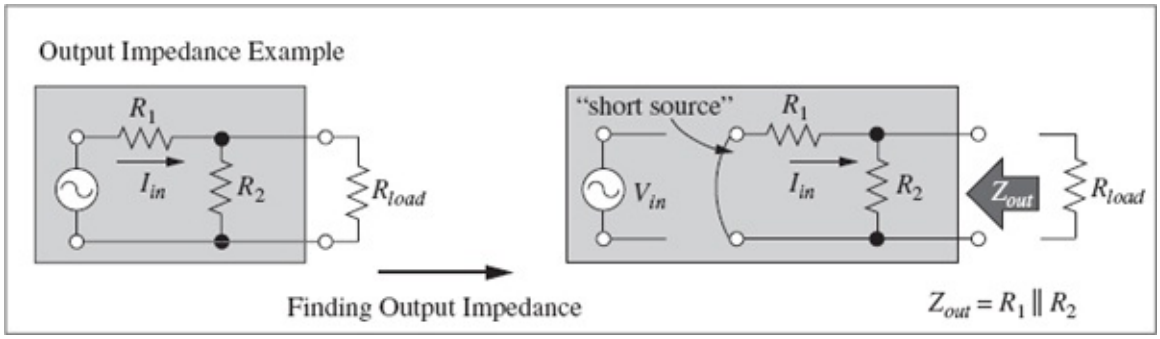


FIGURE 2.188

In Sec. 2.3.1, when we cover filter circuits, you'll see how the output impedance becomes dependent on frequency.

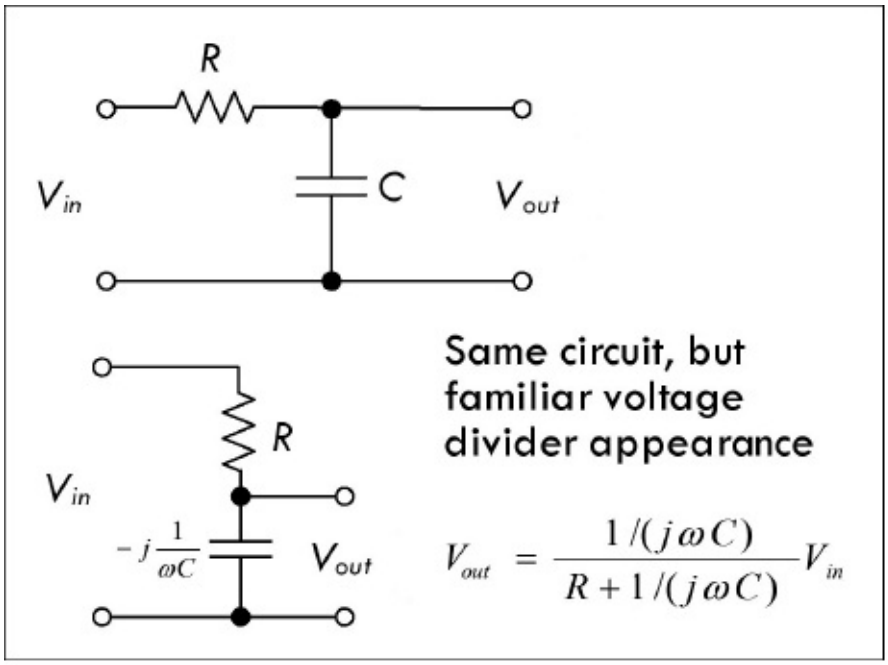
2.33 Two-Port Networks and Filters

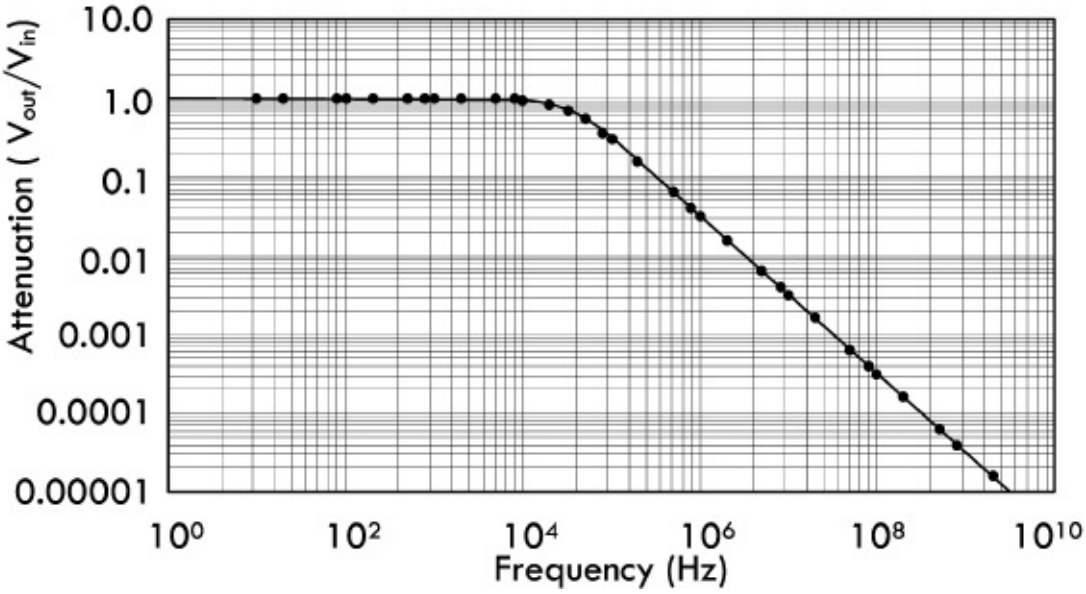
2.33.1 Filters

By combining resistors, capacitors, and inductors in special ways, you can design networks that are capable of passing certain frequencies of signals while rejecting others. This section examines four basic kinds of filters: low-pass, high-pass, bandpass, and notch filters.

Low-Pass Filters

The simple RC filter shown in Fig. 2.189 acts as a low-pass filter—it passes low frequencies but rejects high frequencies.





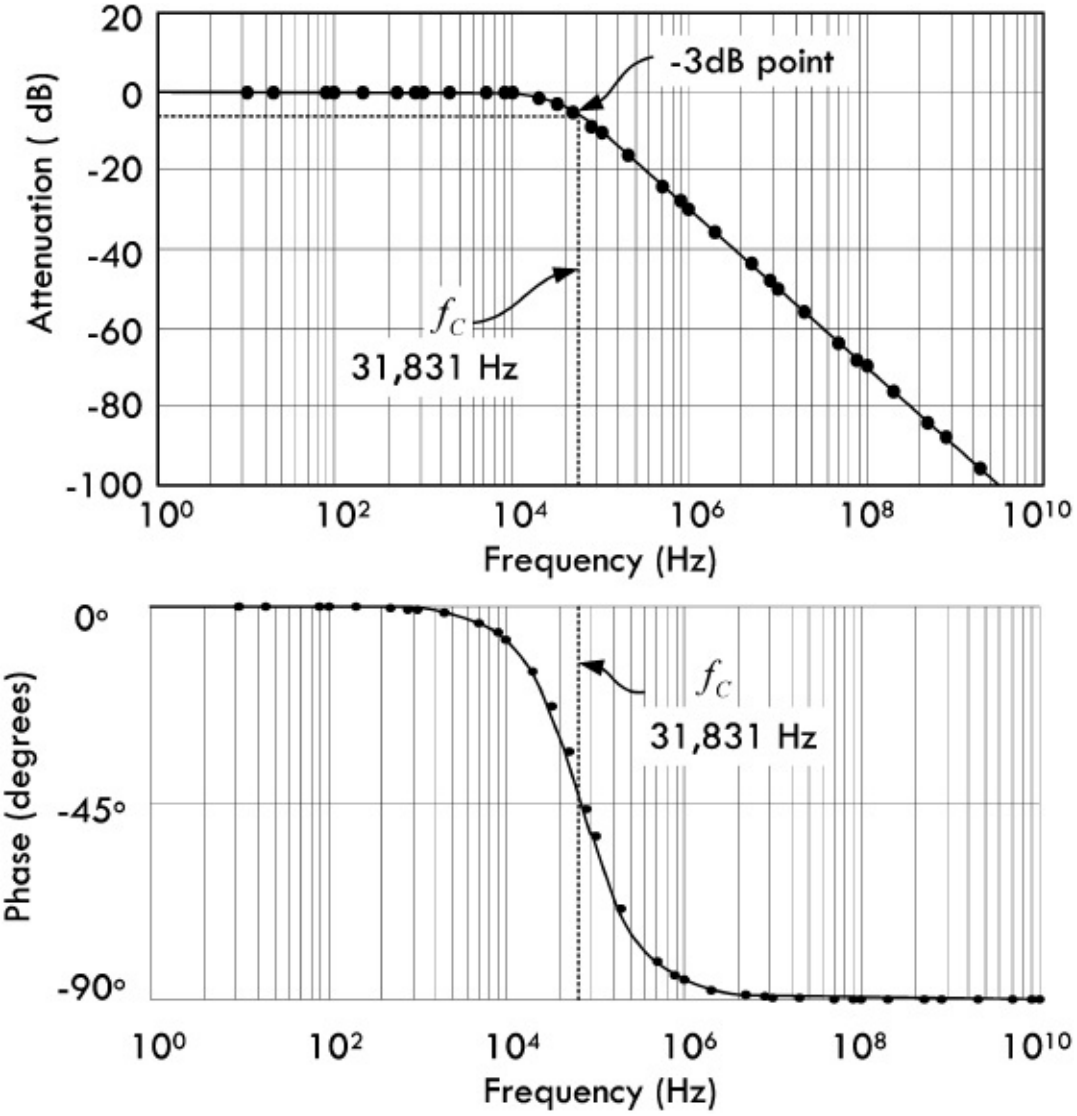


FIGURE 2.189

Example: To figure out how this network works, we find the transfer function. We begin by using the voltage divider to find V_{out} in terms of V_{in}, and consider there is no load (open output or $R_L = \infty$):

$$V_{out} = \frac{1/(j\omega C)}{R + 1/(j\omega C)} V_{in} = \frac{1}{1 + j\omega RC} V_{in}$$

The transfer function is then found by rearranging the equation:

$$H = \frac{V_{out}}{V_{in}} = \frac{1}{1 + j\omega RC} V_{in}$$

$$H = \frac{V_{out}}{V_{in}} = \frac{1}{1 + j\tau\omega}, \quad \tau = RC$$

$$H = \frac{V_{out}}{V_{in}} = \frac{1}{1 + j\omega/\omega_C}, \quad \omega_C = \frac{1}{RC}$$

The magnitude and phase of H are:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{1}{\sqrt{1+\tau^2\omega^2}}$$

$$\arg(H) = \varphi = -\tan^{-1}\left(\frac{\omega}{\omega_C}\right)$$

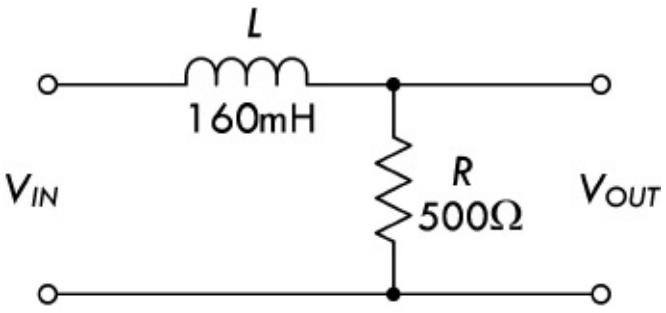
Here, τ is called the time constant, and ω_C is called the angular cutoff frequency of the circuit—related to the standard cutoff frequency by $\omega_C = 2\pi f_C$. The cutoff frequency represents the frequency at which the output voltage is attenuated by a factor of $1/\sqrt{2}$, the equivalent of half power. The cutoff frequency in this example is:

$$f_C = \frac{\omega_C}{2\pi} = \frac{1}{2\pi RC} = \frac{1}{2\pi(50\,\Omega)(0.1\times10^{-6}\,\text{F})}$$
$$= 31{,}831\ \text{Hz}$$

Intuitively, we imagine that when the input voltage is very low in frequency, the capacitor's reactance is high, so it draws little current, thus keeping the output amplitude near the input amplitude. However, as the frequency of the input signal increases, the capacitor's reactance decreases and the capacitor draws more current, which in turn causes the output voltage to drop. Figure 2.189 shows attenuation versus frequency graphs—one graph expresses the attenuation in decibels.

The capacitor produces a delay, as shown in the phase plot in Fig. 2.189. At very low frequency, the output voltage follows the input—they have similar phases. As the frequency rises, the output starts lagging the input. At the cutoff frequency, the output voltage lags by 45°. As the frequency goes to infinity, the phase lag approaches 90°.

Figure 2.190 shows an RL low-pass filter that uses inductive reactance as the frequency-sensitive element instead of capacitive reactance, as was the case in the RC filter.



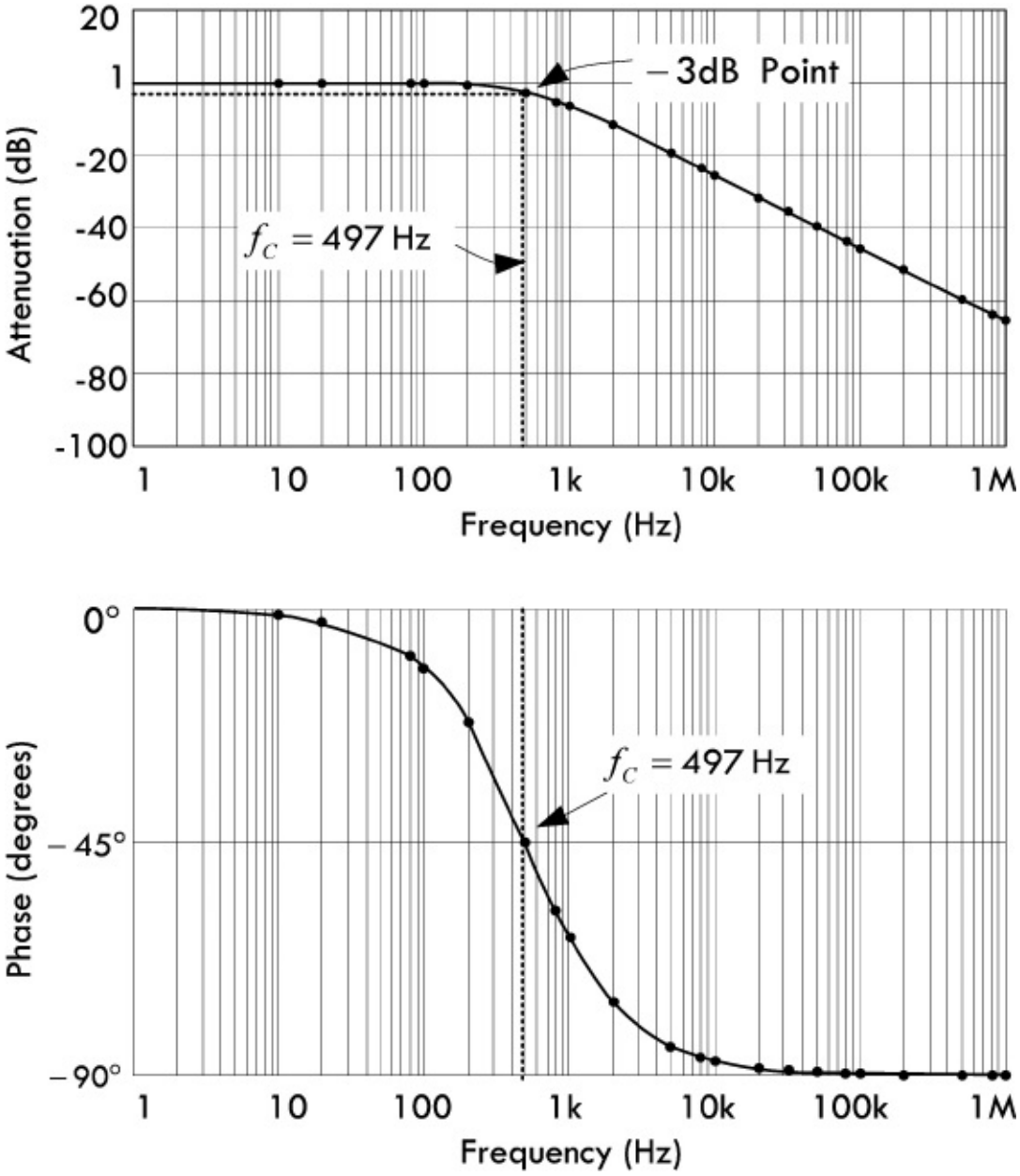


FIGURE 2.190

The input impedance can be found by definition $Z_{in} = \frac{V_{in}}{I_{in}}$, while the output impedance can be found by "killing the source" (see Fig. 2.191):

$$Z_{in} = R + \frac{1}{j\omega C} \text{ and } Z_{in}|_{min} = R$$

$$Z_{out} = R \,||\, \frac{1}{j\omega C} \text{ and } Z_{out}|_{max} = R$$

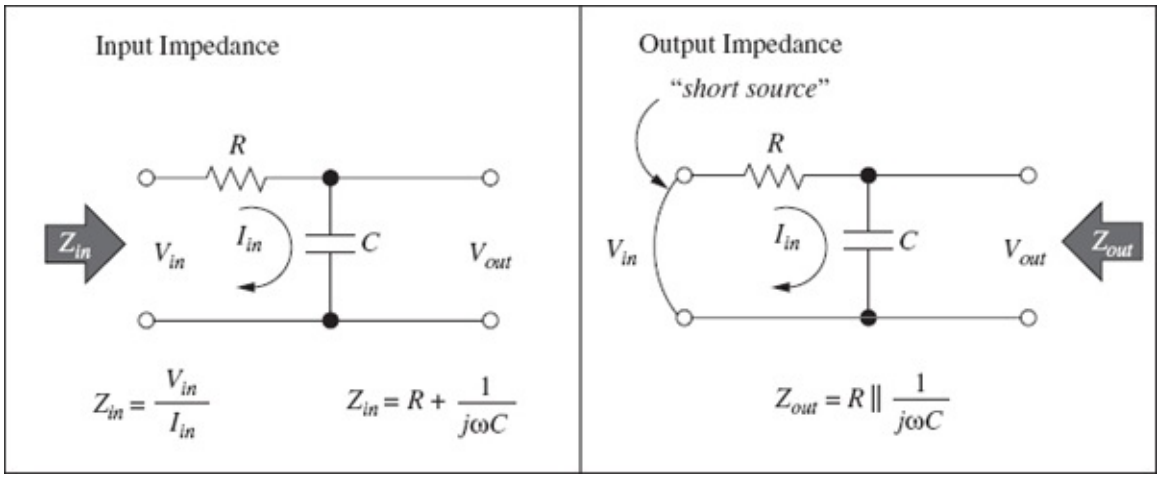


FIGURE 2.191

Now what happens when we put a finite load resistance R_L on the output? Doing the voltage divider stuff and preparing the voltage transfer function, we get:

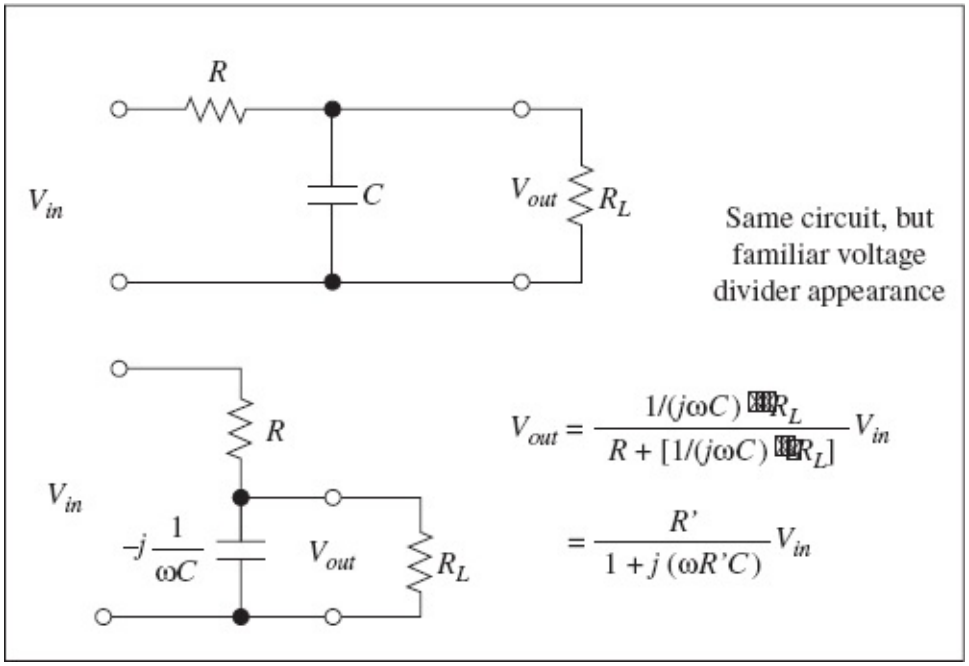


FIGURE 2.192

$$H = \frac{V_{out}}{V_{in}} = \frac{1/(j\omega C) \,||\, R_L}{R + \left[1/(j\omega C) \,||\, R_L\right]} = \frac{R'}{1 + j(\omega R'C)}$$

$$\text{where } R' = R \,||\, R_L$$

This is similar to the transfer function for the unterminated RC filter, but with resistance R being replaced by R'. Therefore:

$$\omega = \frac{1}{R'C} = \frac{1}{(R \,||\, R_L)C} \qquad \text{and} \qquad H = \frac{R'/R}{1 + j\left(\omega/\omega_C\right)}$$

As you can see, the load has the effect of reducing the filter gain ($K = R'/R < 1$) and shifting the cutoff frequency to a higher frequency as ($R' = R \| R_L < R$).

The input and output impedance with load resistance become:

$$Z_{in} = R + \frac{1}{j\omega C} || R_L \qquad \text{and} \qquad Z_{in}|_{min} = R$$

$$Z_{out} = R || \frac{1}{j\omega C} \qquad \text{and} \qquad Z_{out}|_{max} = R$$

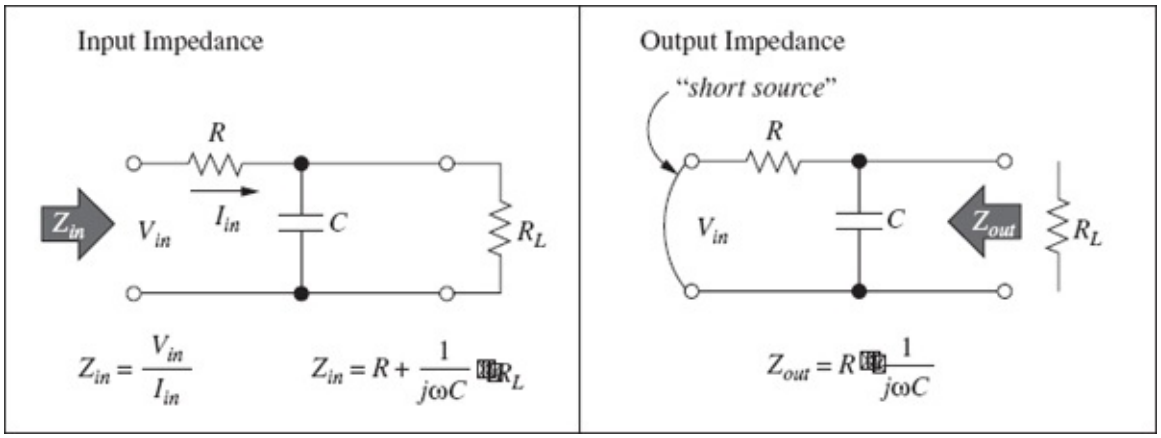


FIGURE 2.193

As long as $R_L >> Z_{out}$ or $R >> Z_{out}|_{max} = R$ (condition for good voltage coupling), $R' \approx R$ and the terminated RC filter will look exactly like an unterminated filter. The filter gain is one, the shift in cutoff frequency disappears, and the input and output resistance become the same as before.

Example: To find the transfer function or attenuation of the RL circuit with no load, we find V_{out} in terms of V_{in} using the voltage divider:

$$V_{out} = \frac{R}{R + j\omega L} V_{in} = \frac{1}{1 + j(\omega L/R)} V_{in}$$

$$H = \frac{V_{out}}{V_{in}} = \frac{1}{1 + j(\omega L/R)}$$

$$H = \frac{V_{out}}{V_{in}} = \frac{1}{1 + j(\omega/\omega_C)}, \quad \omega_C = \frac{R}{L}$$

The magnitude and phase become:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{1}{\sqrt{1 + (\omega L/R)^2}}$$

$$\arg(H) = \varphi = \tan^{-1}\left(\frac{\omega L}{R}\right) = \tan^{-1}\left(\frac{\omega}{\omega_C}\right)$$

Here, ω_C is called the angular cutoff frequency of the circuit—related to the standard cutoff frequency by $\omega_C = 2\pi f_C$. The cutoff frequency represents the frequency at which the output voltage is attenuated by a factor of $1/\sqrt{2}$, the equivalent of half power. The cutoff frequency in this example is:

$$f_C = \frac{\omega_C}{2\pi} = \frac{R}{2\pi L} = \frac{500\Omega}{2\pi(160 \times 10^{-3}\ \text{H})} = 497\ \text{Hz}$$

Intuitively we imagine that when the input voltage is very low in frequency, the inductor doesn't have a

hard time passing current to the output. However, as the frequency gets big, the inductor's reactance increases, and the signal becomes more attenuated at the output. Figure 2.190 shows attenuation versus frequency graphs—one graph expresses the attenuation in decibels.

The inductor produces a delay, as shown in the phase plot in Fig. 2.190. At very low frequency, the output voltage follows the input—they have similar phases. As the frequency rises, the output starts lagging the input. At the cutoff frequency, the output voltage lags by 45°. As the frequency goes to infinity, the phase lag approaches 90°.

The input impedance can be found using the definition of the input impedance:

$$Z_{\text{in}} = \frac{V_{\text{in}}}{I_{\text{in}}} = j\omega L + R$$

The value of the input impedance depends on the frequency ω. For good voltage coupling, the input impedance of this filter should be much larger than the output impedance of the previous stage. The minimum value of Z_{in} is an important number, and its value is minimum when the impedance of the inductor is zero (ω → 0):

$$Z_{\text{in}}\big|_{\text{min}} = R$$

The output impedance can be found by sorting the source and finding the equivalent impedance between output terminals:

$$Z_{\text{out}} = j\omega L \,||\, R$$

where the source resistance is ignored. The output impedance also depends on the frequency ω. For good voltage coupling, the output impedance of this filter should be much smaller than the input impedance of the next stage. The maximum value of Z_{out} is also an important number, and it is maximum when the impedance of the inductor is infinity (ω → ∞):

$$Z_{\text{out}}\big|_{\text{max}} = R$$

When the RL low-pass filter is terminated with a load resistance R_L, the voltage transfer function changes to:

$$H = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{1 + j\omega/\omega_C} \text{ where } \omega_C = (R \,||\, R_L)/L$$

The input impedance becomes:

$$Z_{\text{in}} = j\omega L = R \,||\, R_L,\ Z_{\text{in}}\big|_{\text{min}} = R \,||\, R_L$$

The output impedance becomes:

$$Z_{\text{out}} = (j\omega L) \,||\, R,\ Z_{\text{out}}\big|_{\text{max}} = R$$

The effect of the load is to shift the cutoff frequency to a lower value. Filter gain is not affected. Again, for $R_L >> Z_{\text{out}}$ or $R_L >> Z_{\text{out}}|_{\text{max}} = R$ (condition for good voltage coupling), the shift in cutoff frequency disappears, and the filter will look exactly like an unterminated filter.

High-Pass Filters

Example: To figure out how this network works, we find the transfer function by using the voltage

divider equation and solving in terms of V_{out} and V_{in}:

$$H = \frac{V_{out}}{V_{in}} = \frac{R}{R + 1/(j\omega C)} = \frac{1}{1 - j(1/\omega RC)} = \frac{j\omega\tau}{1 + j\omega\tau}, \qquad \tau = RC$$

or

$$H = \frac{V_{out}}{V_{in}} = \frac{j(\omega/\omega_C)}{1 + j(\omega/\omega_C)} = \frac{1}{1 - j\omega_C/\omega}, \qquad \omega_C = \frac{1}{RC}$$

The magnitude and phase of H are:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{\tau\omega}{\sqrt{1 + \tau^2\omega^2}}$$

$$\arg(H) = \varphi = \tan^{-1}\left(\frac{\omega_C}{\omega}\right)$$

Here τ is called the time constant and ω_C is called the angular cutoff frequency of the circuit—related to the standard cutoff frequency by $\omega_C = 2\pi f_C$. The cutoff frequency represents the frequency at which the output voltage is attenuated by a factor of $1/\sqrt{2}$, the equivalent of half power. The cutoff frequency in this example is:

$$f_0 = \frac{\omega_0}{2\pi} = \frac{1}{2\pi RC}$$

$$= \frac{1}{2\pi(10{,}000\ \Omega)(0.1 \times 10^{-6}\ \text{F})} = 159\ \text{Hz}$$

Intuitively, we imagine that when the input voltage is very low in frequency, the capacitor's reactance is very high, and hardly any signal is passed to the output. However, as the frequency rises, the capacitor's reactance decreases, and there is little attenuation at the output. Figure 2.194 shows attenuation versus frequency graphs—one graph expresses the attenuation in decibels.

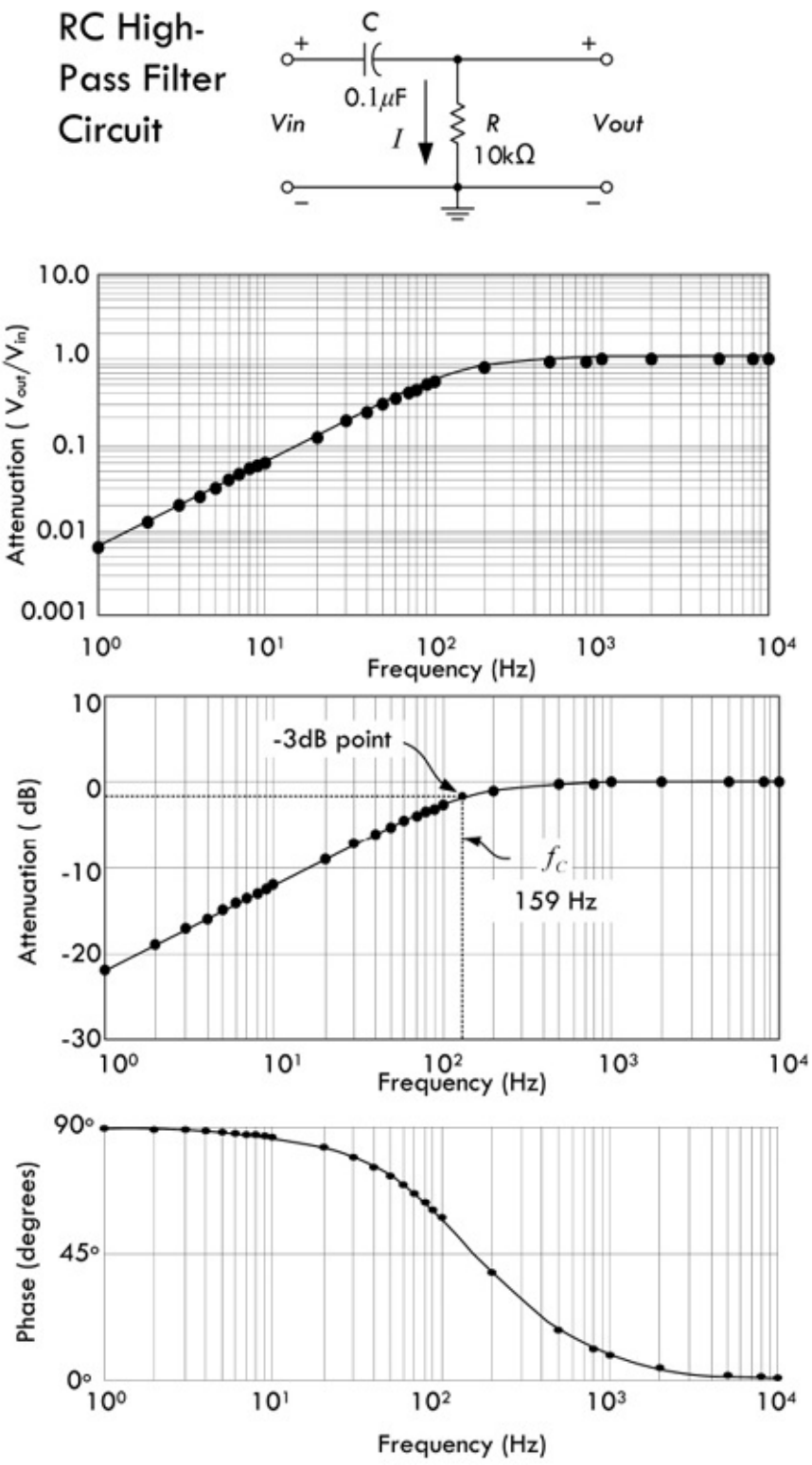


FIGURE 2.194

In terms of phase, at very low frequency the output leads the input in phase by 90°. As the frequency rises to the cutoff frequency, the output leads by 45°. When the frequency goes toward infinity, the phase approaches 0, the point where the capacitor acts like a short.

Input and output impedances of this filter can be found in a way similar to finding these impedances for low-pass filters:

$$Z_{in} = R + \frac{1}{j\omega C} \quad \text{and} \quad Z_{in}|_{min} = R$$

$$Z_{out} = R \,||\, \frac{1}{j\omega C}, \; Z_{out}|_{max} = R$$

With a terminated load resistance, the voltage transfer function becomes:

$$H = \frac{V_{out}}{V_{in}} = \frac{R||R_L}{R||R_L + 1/(j\omega C)} = \frac{1}{1 - j(1/\omega R'C)} \text{ where } R' = R||R_L$$

This is similar to the transfer function for the unterminated RC filter, but with resistance R being replaced by R':

$$\omega_C = \frac{1}{R'C} = \frac{1}{(R||R_L)C} \text{ and } H = \frac{1}{1 - j\omega_C/\omega}$$

The load has the effect of shifting the cutoff frequency to a higher frequency ($R' = R||R_L < R$).

The input and output impedances are:

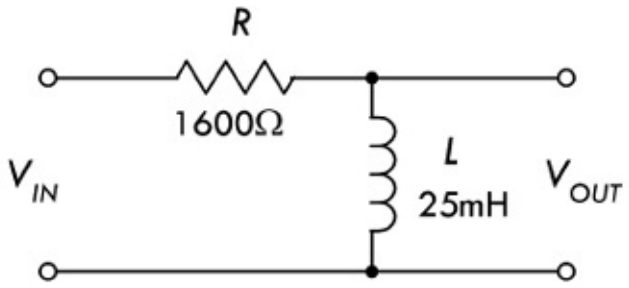
$$Z_{in} = \frac{1}{j\omega C} + R||R_L,\ Z_{in}|_{min} = R||R_L$$

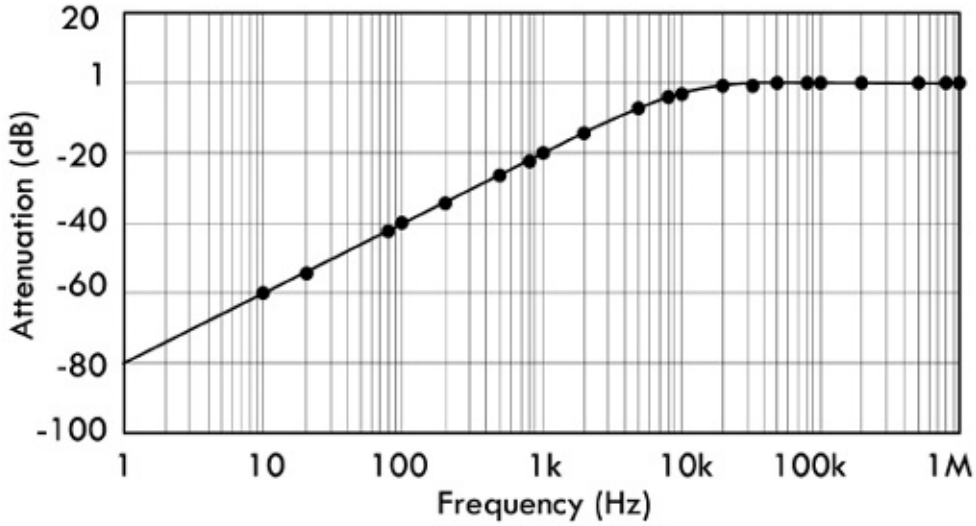
$$Z_{out} = R||\frac{1}{j\omega C},\ Z_{out}|_{max} = R$$

As long as $R_L >> Z_{out}$ or $R_L >> Z_{out}|_{max} = R$ (condition for good voltage coupling), $R' \approx R$ and the terminated RC filter will look like an unterminated filter. The shift in cutoff frequency disappears, and input and output resistance become the same as before.

RL High-Pass Filter

Figure 2.195 shows an RL high-pass filter that uses inductive reactance as the frequency-sensitive element instead of capacitive reactance, as was the case in the RC filter.





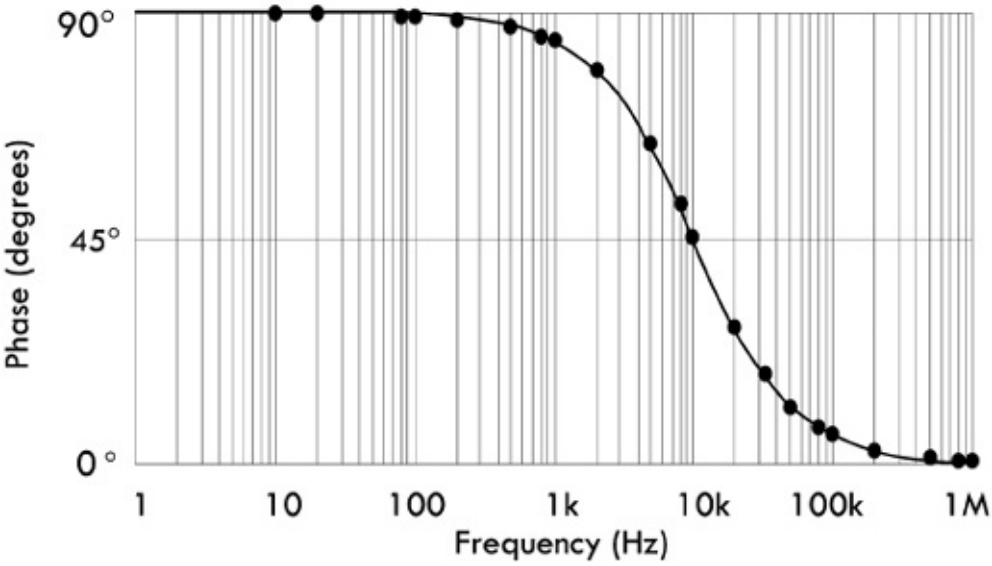


FIGURE 2.195

Example: To find the transfer function or attenuation of the RL circuit, we again use the voltage divider equation and solve for the transfer function or attenuation of the RL circuit in terms of V_{out} and V_{in}:

$$H = \frac{V_{out}}{V_{in}} = \frac{j\omega L}{R + j\omega L}$$

$$= \frac{\omega L \angle 90^{\circ}}{\sqrt{R^2 + (\omega L)^2} \angle \tan^{-1}(\omega L / R)}$$

or

$$H = \frac{1}{1 - j\omega_C/\omega}, \qquad \omega_C = \frac{R}{L}$$

The magnitude and phase of H are:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{\omega L}{\sqrt{R^2 + (\omega L)^2}} = \frac{\omega/\omega_C}{\sqrt{1 + (\omega/\omega_C)^2}}$$

$$\arg(H) = \varphi = 90^\circ - \tan^{-1}\left(\frac{\omega L}{R}\right)$$

Here ω_C is called the angular cutoff frequency of the circuit—related to the standard cutoff frequency by $\omega_C = 2\pi f_C$. The cutoff frequency represents the frequency at which the output voltage is attenuated by a factor of $1\sqrt{2}$, the equivalent of half power. The cutoff frequency in this example is:

$$f_C = \frac{\omega_0}{2\pi} = \frac{R}{2\pi L} = \frac{1600\ \Omega}{2\pi(25 \times 10^{-3}\ \text{H})} = 10{,}186\ \text{Hz}$$

Intuitively, we imagine that when the input voltage is very low in frequency, the inductor's reactance is very low, so most of the current is diverted to ground—the signal is greatly attenuated at the output. However, as the frequency rises, the inductor's reactance increases and less current is passed to ground—the attenuation decreases. Figure 2.195 shows attenuation versus frequency graphs—one graph expresses the attenuation in decibels.

In terms of phase, at very low frequency the output leads the input in phase by 90°. As the frequency rises to the cutoff frequency, the output leads by 45°. When the frequency goes toward infinity, the phase approaches 0, the point where the inductor acts like an open circuit.

The input and output impedances are:

$$Z_{in} = R + j\omega L,\ Z_{in}|_{min} = R$$

$$Z_{out} = R\,||\,j\omega L,\ Z_{out}|_{max} = R$$

For a terminated RL high-pass filter with load resistance, we do a similar calculation as we did with the RC high-pass filter, replacing the resistance with R':

$$H = \frac{V_{out}}{V_{in}} = \frac{R'/R}{1 - j\omega_C/\omega}, \qquad \omega_C = \frac{R'}{L}, \qquad R' = R\,||\,R_L$$

The input and output impedances are:

$$Z_{in} = R + j\omega L\,||\,R_L,\ Z_{in}|_{min} = R$$

$$Z_{out} = R\,||\,j\omega L,\ Z_{out}|_{max} = R$$

The load has the effect of lowering the gain, $K = R'/R < 1$, and it shifts the cutoff frequency to a lower value. As long as $R_L >> Z_{out}$ or $R_L >> Z_{out}|_{max} = R$ (condition for good voltage coupling), $R' \approx R$ and the terminated RC filter will look like an unterminated filter.

Bandpass Filter

The RLC bandpass filter in Fig. 2.196 acts to pass a narrow range of frequencies (band) while attenuating or rejecting all other frequencies.

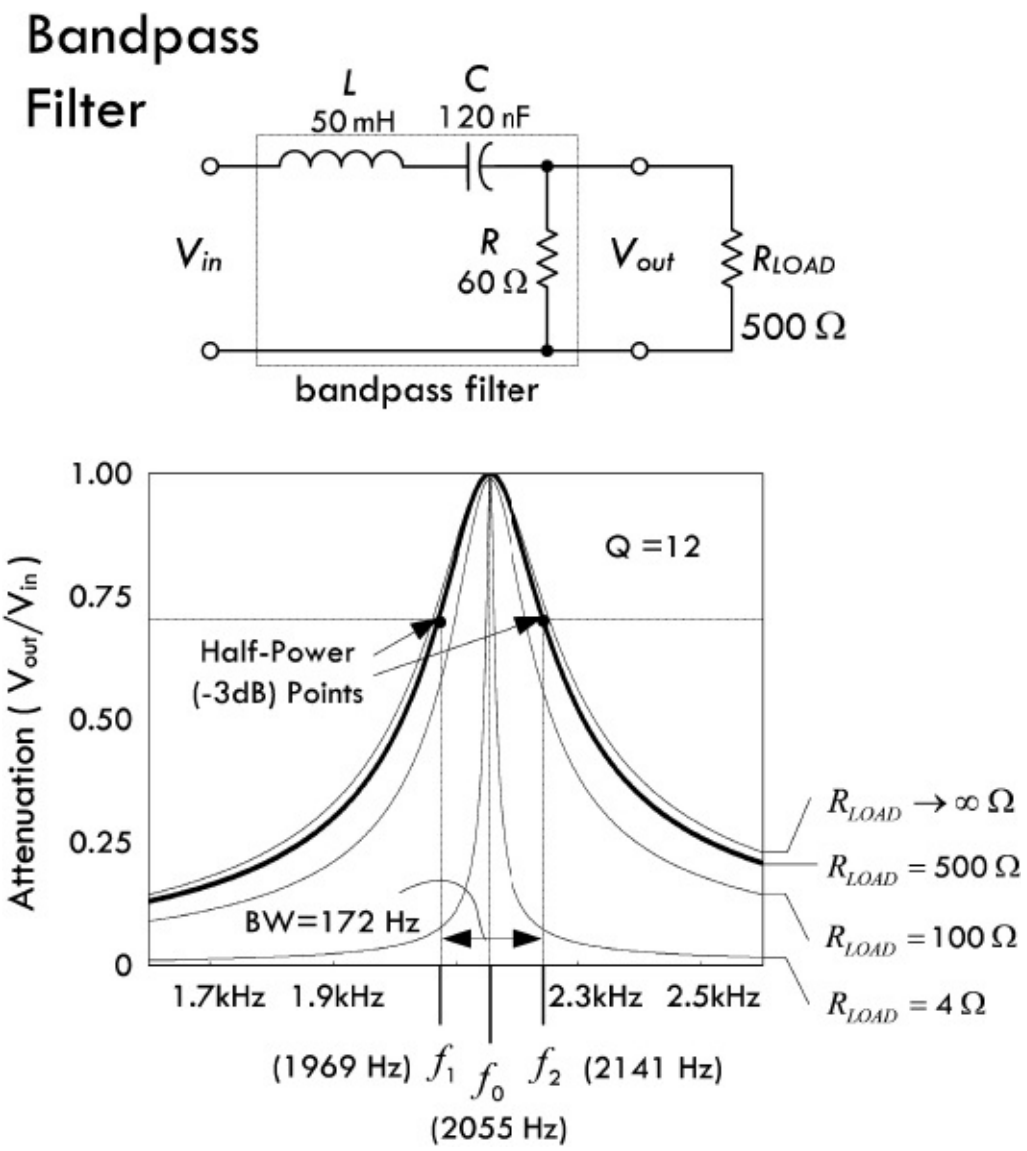


FIGURE 2.196 The parallel bandpass filter shown here yields characteristics similar to those of the previous bandpass filter. However, unlike the previous filter, as you approach the resonant frequency of the tuned circuit, the LC (RL coil) section's impedance gets large, not allowing current to be diverted away from the load. On either side of resonance, the impedance goes down, diverting current away from load.

Example: To find the transfer function or attenuation of the unloaded RLC circuit, we set up equations for V_{in} and V_{out}:

$$V_{\text{in}} = \left(j\omega L - j\frac{1}{\omega C} + R \right) \times I$$

$$V_{\text{out}} = R \times I$$

The transfer function becomes:

$$H = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{R}{R + j(\omega L - 1/\omega C)}$$

This is the transfer function for an unloaded output. However, now we get more realistic and have a load resistance attached to the output. In this case we must replace R with R_T, which is the parallel resistance of R and R_{LOAD}:

$$R_T = \frac{R \times R_{LOAD}}{R + R_{LOAD}} = \frac{500(60)}{500 + 60} = 54\ \Omega$$

Placing this in the unloaded transfer function and solving for the magnitude, we get:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{R_T}{\sqrt{R_T^2 + [\omega L - 1/(\omega C)]^2}}$$

Plugging in all component values and setting $\omega = 2nf$, we get:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right|$$

$$= \frac{54}{\sqrt{54^2 + [0.314f - 1/(7.54 \times 10^{-7} f)]^2}}$$

The attenuation versus frequency graph based on this equation is shown in Fig. 2.196. (Three other families of curves are provided for loads of 4 Ω, 100 Ω, and an infinite resistance load.)

The resonant frequency, *Q*, bandwidth, and upper and lower cutoff frequencies are given by:

$$f_0 = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{2\pi\sqrt{(50 \times 10^{-3})(120 \times 10^{-9})}} = 2055\ \text{Hz}$$

$$Q = \frac{X_{L,0}}{R_T} = \frac{2\pi f_0 L}{R_T} = \frac{2\pi(2055)(50 \times 10^{-3})}{54} = 12$$

$$\text{BW} = \frac{f_0}{Q} = \frac{2055}{12} = 172\ \text{Hz}$$

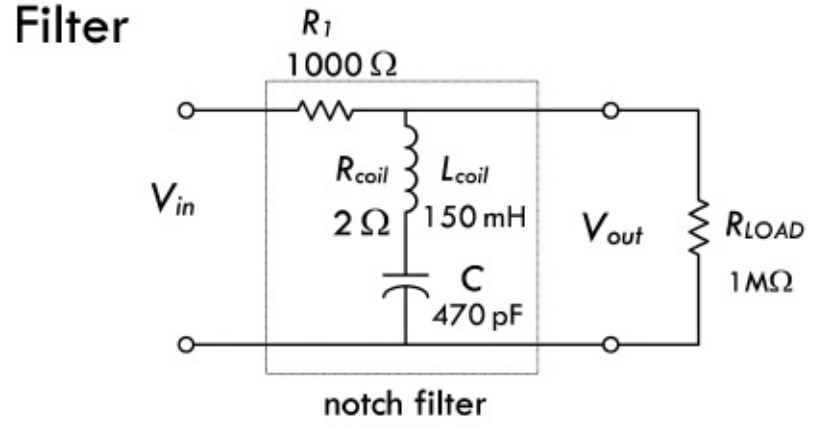
$$f_1 = f_0 - \text{BW}/2 = 2055 - 172/2 = 1969\ \text{Hz}$$

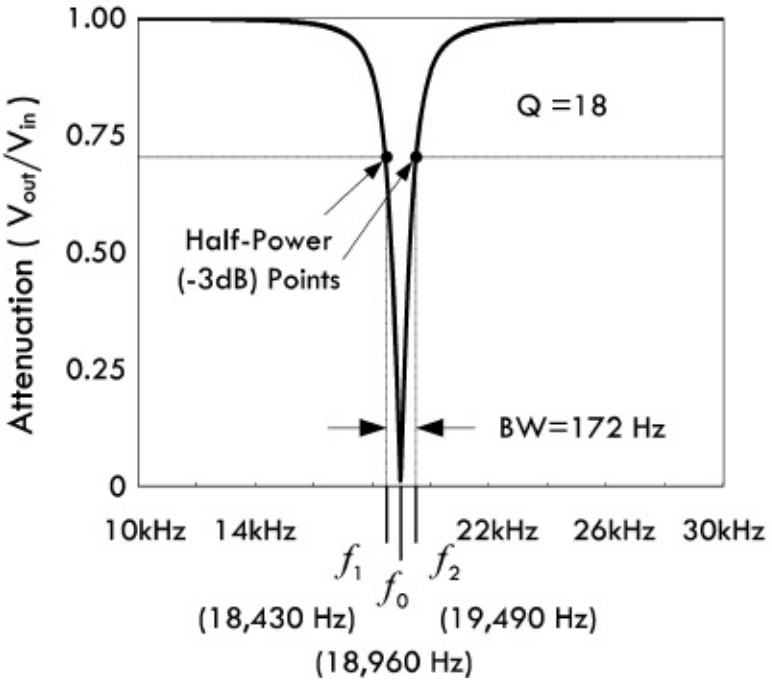
$$f_2 = f_0 + \text{BW}/2 = 2055 + 172/2 = 2141\ \text{Hz}$$

Notch Filter

The notch filter in Fig. 2.197 acts to pass a wide range of frequencies, while attenuating (rejecting) a narrow band of frequencies.

Notch Filter





Parallel Notch Filter

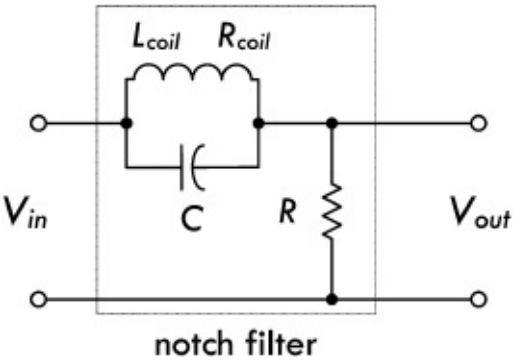


FIGURE 2.197 The notch filter shown here yields stop-band characteristics similar to those of the previous notch filter. However, unlike the previous filter, as you approach the resonant frequency of the tuned circuit, the LC (RL coil) section's impedance gets large, not allowing current to be sent to the load. On either side of resonance, the impedance goes down, allowing current to reach load.

Example: To find the transfer function or attenuation of the unloaded RLC circuit, we set up equations for V_{in} and V_{out}:

$$V_{in} = \left(R_1 + R_{coil} + j\omega L - j\frac{1}{\omega C}\right) \times I$$

$$V_{out} = \left(R_{coil} + j\omega L - j\frac{1}{\omega C}\right) \times I$$

The transfer function becomes:

$$H = \frac{V_{out}}{V_{in}} = \frac{R_{coil} + j[\omega L - 1/\omega C]}{(R_1 + R_{coil}) + j[\omega L - 1/(\omega C)]}$$

This is the transfer function for an unloaded output. Now we get more realistic and have a load resistance attached to the output. However, in this case, the load resistance is so large that we can assume it draws inconsequential current, so we don't need to place it into the equation:

The magnitude of the transfer function is:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{\sqrt{R_{coil^2} + [\omega L - 1/\omega C]^2}}{\sqrt{(R_1 + R_{coil})^2 + j[\omega L - 1/(\omega C)]^2}}$$

Plugging in all component values and setting $\omega = 2\pi f$, we get:

$$|H| = \left|\frac{V_{out}}{V_{in}}\right| = \frac{\sqrt{4 + (0.94f - 3.38 \times 10^8/f)^2}}{\sqrt{1.00 \times 10^6 + (0.94f - 3.38 \times 10^8/f)^2}}$$

The attenuation versus frequency graph based on this equation is shown in Fig. 2.197.

The resonant frequency, Q, bandwidth, and upper and lower cutoff frequencies are given by:

$$f_0 = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{2\pi\sqrt{(150 \times 10^{-3})(470 \times 10^{-12})}} = 18,960 \text{ Hz}$$

$$Q = \frac{X_{L,0}}{R_T} = \frac{2\pi f_0 L}{R_1} = \frac{2\pi(18,960)(150 \times 10^{-3})}{1000} = 18$$

$$\text{BW} = \frac{f_0}{Q} = \frac{18960}{18} = 1053 \text{ Hz}$$

$$f_1 = f_0 - \text{BW}/2 = 18,960 - 1053/2 = 18,430 \text{ Hz}$$

$$f_2 = f_0 + \text{BW}/2 = 18,960 + 1053/2 = 19,490 \text{ Hz}$$

2.33.2 Attenuators

Often it is desirable to attenuate a sinusoidal voltage by an amount that is independent of frequency. We can do this by using a voltage divider, since its output is independent of frequency. Figure 2.198 shows a simple voltage divider attenuator network inserted between a source and a load circuit to decrease the source signal's magnitude before it reaches the load. As we go through this example, you will pick up additional insight into input and output impedances.

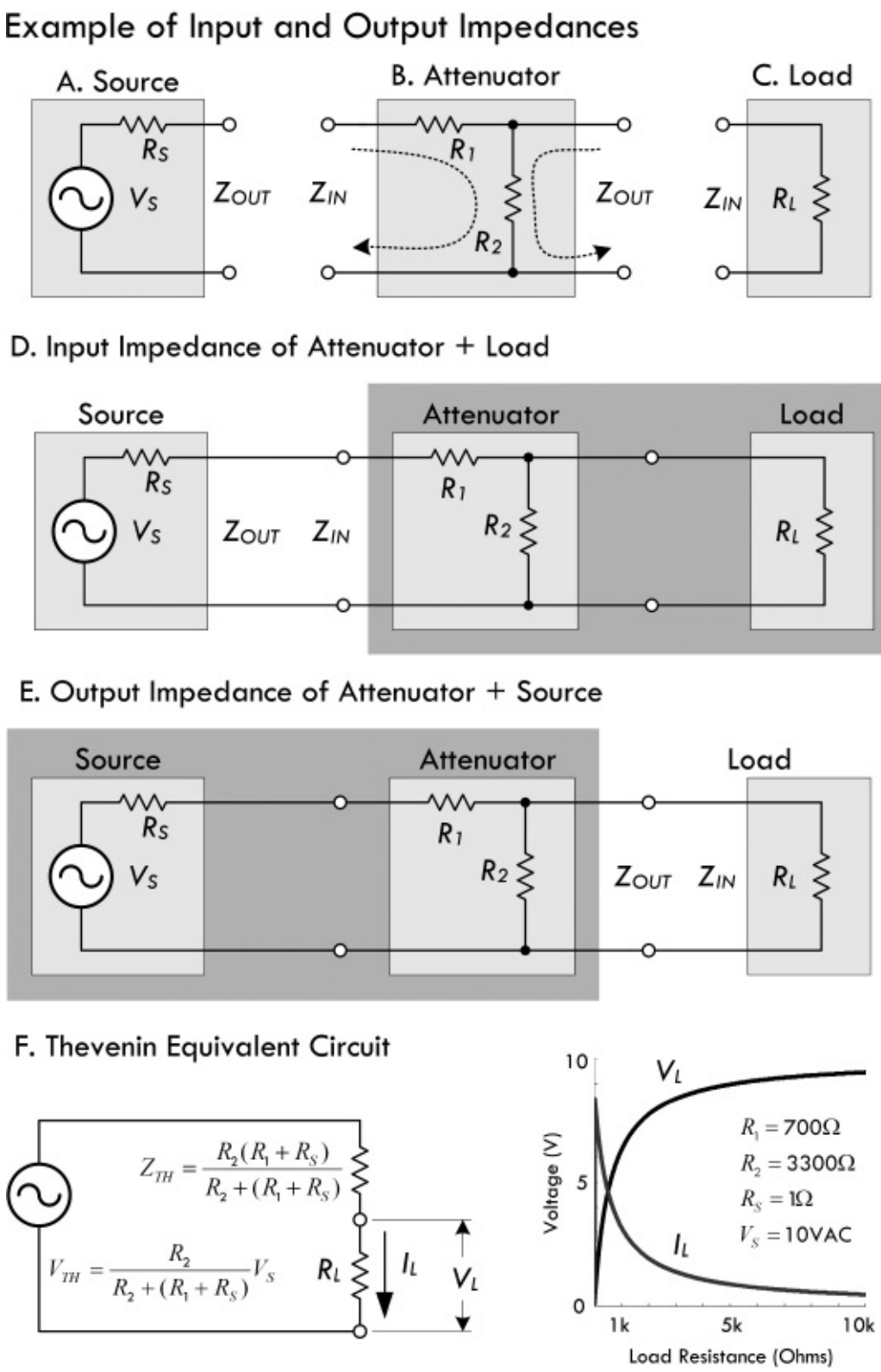


FIGURE 2.198

Example: In (a), the source has an output impedance equal to the internal resistance of the source:

$$Z_{out} = R_S$$

In (b), the attenuator network has input and output impedances of:

$$Z_{in} = R_1 + R_2$$

$$Z_{out} = R_2$$

We can come up with a transfer function for this, taking the same current to flow through R_2:

$$H = \frac{V_{out}}{V_{in}} = \frac{I \times Z_{out}}{I \times Z_{in}} = \frac{Z_{out}}{Z_{in}} = \frac{R_2}{R_1 + R_2}$$

If you rearrange terms, you can see this is simply a voltage divider.

In (c), the load has an input impedance of:

$$Z_{in} = R_L$$

However, when we assemble the circuit, the useful input and output impedances change. Now, looking at things from the point of view of the source, as shown in (d), the source sees an input impedance of the attenuator combined with the load impedance, which is R_1 in series with the parallel combination of R_2 and R_L:

$$Z_{in} = \frac{R_1(R_2 + R_L)}{R_1 + (R_{21} + R_L)}$$

But now, from the point of view of the load, as shown in (e), the output impedance of the attenuator and source combined is R_2 in parallel with the series combination of R_1 and R_S:

$$Z_{out} = \frac{R_2(R_1 + R_S)}{R_2 + (R_1 + R_S)}$$

(The input impedance to the load is still R_L.)

This output impedance is equivalent to the Thevinen equivalent impedance Z_{TH}, as shown in (f). If we substitute R_1 = 100 Ω, R_2 = 3300 Ω, R_S = 1 Ω, and V_S = 10 VAC, we get the graph shown in (f). If we set the load R_L = 330 Ω, using the Thevenin circuit, we find:

$$V_L = I_L \times R_L = \frac{V_{TH}}{Z_{TH}} R_L = \frac{8.25 \text{ VAC}}{4001\ \Omega}(330\ \Omega) = 6.8 \text{ VAC}$$

Compensated Attenuator

We just saw how a voltage divider could be used to attenuate a signal in a manner that is independent of frequency. However, in practice, there is always stray capacitance in a real circuit, and eventually a frequency is reached at which the voltage divider behaves like either a low-or a high-pass filter. This problem can be overcome by using a compensated attenuator circuit, as shown in Fig. 2.199.

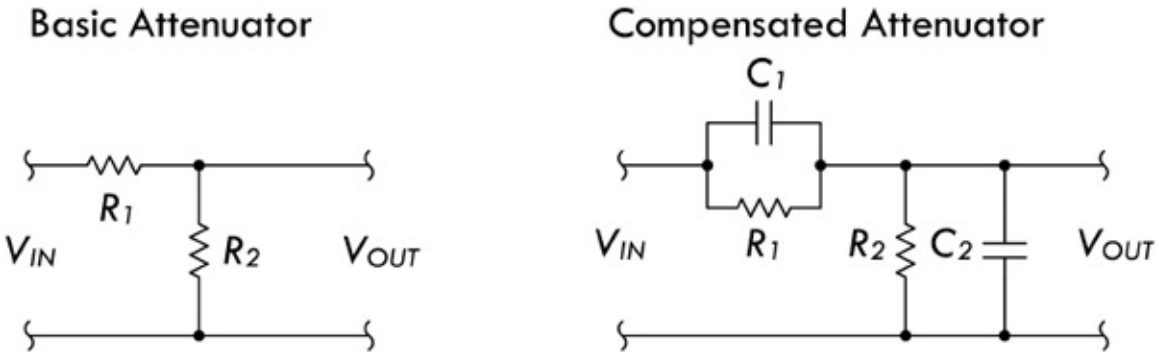


FIGURE 2.199

At low frequencies the circuit behaves like an ordinary resistive divider, but at high frequencies the capacitive reactance dominates, and the circuit behaves like a capacitive voltage divider. The attenuation is independent of frequency, provided that:

$$R_1C_1 = R_2C_2 \tag{2.91}$$

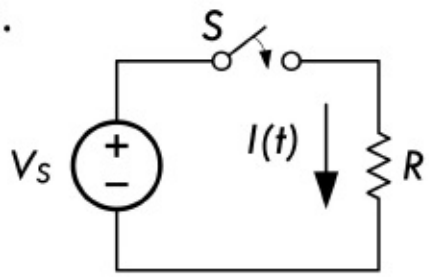
In practice, one of the capacitors is usually variable, so that the attenuator can be adjusted to compensate for any stray capacitance.

Such compensated attenuators are often used at the input of an oscilloscope to raise the input resistance and lower the input capacitance so as to make the oscilloscope into a more nearly ideal voltmeter. However, this results in a decrease in sensitivity of the scope to input voltage.

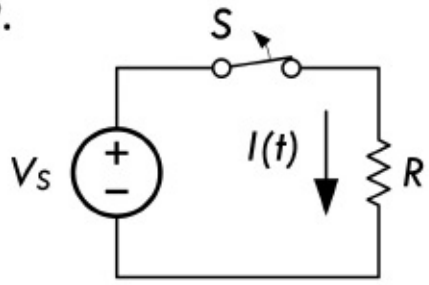
2.34 Transient Circuits

Transients within circuits remove any steady-state condition. They represent a sudden change in voltage introduced by an external agent, such as a switch being thrown or a transistor switching states. During a transient, the voltages and currents throughout the circuit readjust to a new dc value in a brief but nonnegligible time interval immediately following the transient event. The initial condition is a dc circuit; the final condition is a different dc circuit; but the interval in between—while the circuit is readjusting to the new conditions—may exhibit complex behavior. The introduction of a transient into a circuit containing reactive components usually requires solving differential equations, since the response is time dependent. The following simple example is an exception to the rule (no reactances—no differential equations), but is a good illustration of transient behavior, nevertheless. Refer to Fig. 2.200.

1.



2.



3.

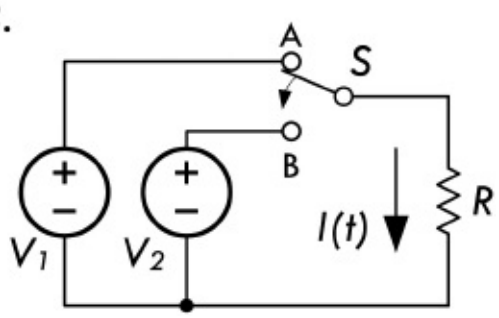


FIGURE 2.200

Example:

1. Initially the switch S is open. The instant it is closed ($t = 0$), V_S is applied across R and current flows immediately, according to Ohm's law. If the switch remains closed thereafter ($t > 0$), the current remains the same as it was at the instant S was closed:

$$I(t) = \begin{cases} 0 & t < 0 \quad \text{(before } S \text{ closed)} \\ V_s/R & t = 0 \quad (S \text{ just closed)} \\ V_s/R & t > 0 \quad \text{(time after } S \text{ closed)} \end{cases}$$

2. Initially S is closed. The instant it is open ($t = 0$), the voltage across and the current through the resistor go to zero. The voltage and current remain zero thereafter ($t > 0$), until the switch is opened again.

$$I(t) = \begin{cases} V_s/R & t < 0 \quad \text{(before } S \text{ opened)} \\ 0 & t = 0 \quad (S \text{ just opened)} \\ 0 & t > 0 \quad \text{(time after } S \text{ opened)} \end{cases}$$

3. Initially S is in position A ($t < 0$), and the voltage across R is V_1, thus making the current $I = V_1/R$. Neglecting the time delay for the switch element to move to position B, the instant S is thrown to position B ($t = 0$), the voltage immediately changes and the voltage across R is V_2, thus making $I = V_2/R$.

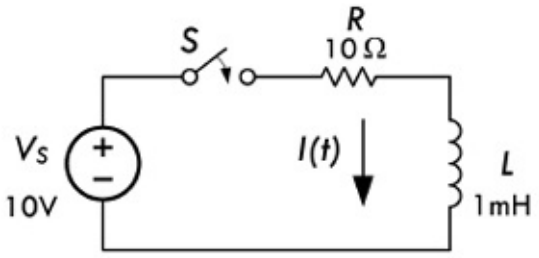
$$I(t) = \begin{cases} V_1/R & t < 0 \quad (S \text{ in position A)} \\ V_2/R & t = 0 \quad (S \text{ just switched to B)} \\ V_2/R & t > 0 \quad \text{(time after } S \text{ to B)} \end{cases}$$

The preceding example may have seemed like child's play, but there is an important thing to notice from it. Just before the transient event of switching, the voltage across and the current through the resistor were both zero (or a constant value). Immediately after the switching event, the voltage and current jumped to new levels instantaneously. The resistor's natural response after the event was time independent as indicated by Ohm's law, $V = IR$. In other words, under a forced response (source is now in the picture), the resistor voltage can change instantly to a new steady state; under a forced response, the resistor current can change instantly to a new steady state.

When we consider capacitors and inductors, a forced response (applying or removing source influence) does not result in voltages and currents instantly jumping to a new steady state. Instead, there is a natural response after the forced response where the voltage and currents vary with time. With the help of Kirchhoff's laws, transient circuits can be modeled; the resulting differential equations take into account transient events by applying initial conditions. The following two examples, which were actually covered earlier in the capacitor and inductor sections, show what happens when voltage is suddenly applied to an RL and an RC circuit.

Example: The circuit in Fig. 2.201 can be modeled using Kirchhoff's voltage equation:

$$V_s - L\frac{dI}{dt} - RI = 0 \qquad \text{or} \qquad L\frac{dI}{dt} + RI = V_s$$



$$I(t)=\begin{cases}0 & t<0 \text{ (before S closed)}\\ 0 & t=0 \text{ (S just closed)}\\ (V_S/R)(1-e^{-(R/L)t}) & t>0 \text{ (after S closed)}\\ V_S/R & t\to\infty \text{ (steady state)}\end{cases}$$

$$V_R(t)=\begin{cases}0 & t<0 \text{ (before S closed)}\\ 0 & t=0 \text{ (S just closed)}\\ V_S(1-e^{-(R/L)t}) & t>0 \text{ (after S closed)}\\ V_S & t\to\infty \text{ (steady state)}\end{cases}$$

$$V_L(t)=\begin{cases}0 & t<0 \text{ (before S closed)}\\ V_S & t=0 \text{ (S just closed)}\\ V_S e^{-(R/L)t} & t>0 \text{ (after S closed)}\\ 0 & t\to\infty \text{ (steady state)}\end{cases}$$

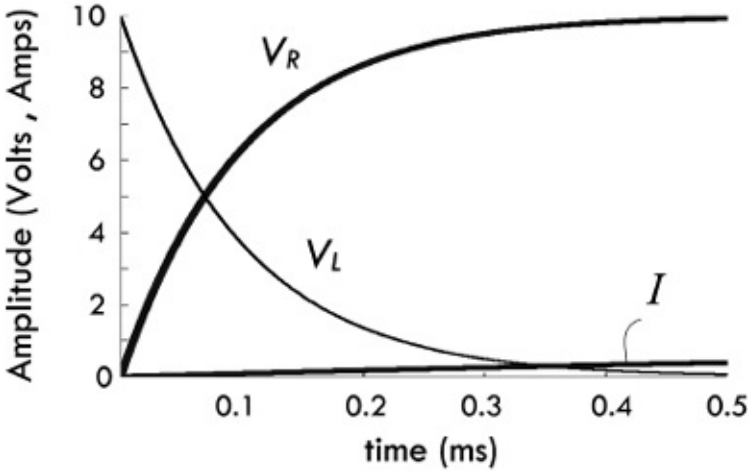


FIGURE 2.201

This is a first-order nonhomogeneous differential equation. To solve this equation, we separate the variables and integrate:

$$\int \frac{L}{V_S - RI}\,dI = \int dt$$

The solution of integration:

$$-\frac{L}{R}\ln(V_S - RI) = t + C$$

Using the initial conditions $t(0) = 0$, $I(0) = 0$, we can determine the constant of integration:

$$C = -\frac{L}{R} \ln V_S$$

Substituting this back into the solution, we get:

$$I = \frac{V_S}{R}(1 - e^{-Rt/L})$$

Using the component values shown in Fig. 2.201, this equation becomes:

$$I = \frac{10\ V}{10\ \Omega}(1 - e^{-10t/0.001}) = 1.0\ (1 - e^{-10{,}000t})\text{A}$$

Once the current is known, the voltage across the resistor and inductor can be easily determined:

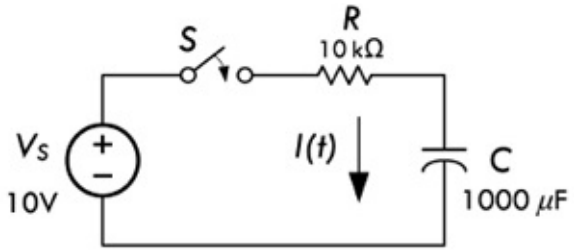
$$V_R = IR = V_S\ (1 - e^{-Rt/L}) = 10\ \text{V}\ (1 - e^{10{,}000t})$$

$$V_L = L\frac{dI_L}{dt} = V_S e^{-R_t/L} = 10\ \text{V} e^{-10{,}000t}$$

The graph in Fig. 2.201 shows how these voltages change with time. The section on inductors explains some important details of the RL circuit not mentioned here and also explains how the RL deenergizing circuit works.

Example: The circuit in Fig. 2.202 can be modeled using Kirchhoff's voltage equation:

$$V_S = RI + \frac{1}{C}\int_0^t I(t')\,dt'$$



$$I(t) = \begin{cases} 0 & t < 0 \text{ (before S closed)} \\ V_S / R & t = 0 \text{ (S just closed)} \\ (V_S / R)e^{-t/RC} & t > 0 \text{ (after S closed)} \\ 0 & t \to \infty \text{ (steady state)} \end{cases}$$

$$V_R(t) = \begin{cases} 0 & t < 0 \text{ (before S closed)} \\ V_S & t = 0 \text{ (S just closed)} \\ V_S e^{-t/RC} & t > 0 \text{ (after S closed)} \\ 0 & t \to \infty \text{ (steady state)} \end{cases}$$

$$V_C(t) = \begin{cases} 0 & t < 0 \text{ (before S closed)} \\ 0 & t = 0 \text{ (S just closed)} \\ V_S(1 - e^{-t/RC}) & t > 0 \text{ (after S closed)} \\ V_S & t \to \infty \text{ (steady state)} \end{cases}$$

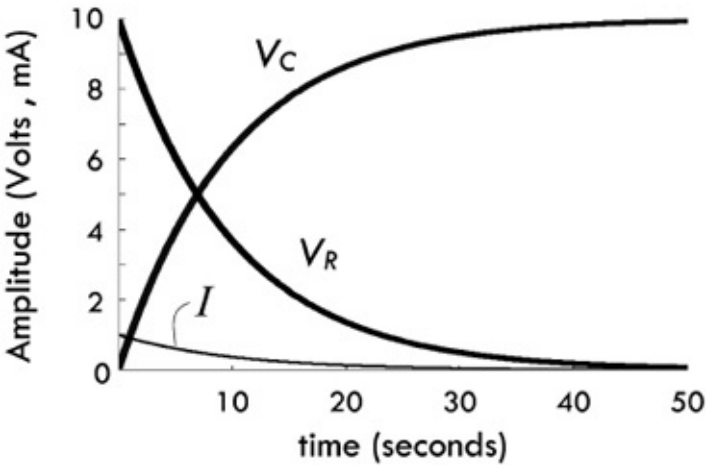


FIGURE 2.202

The first step in solving such an equation is to eliminate the integral by differentiating each term:

$$0 = R\frac{dI}{dt} + \frac{1}{C}I$$

This is an example of a linear first-order homogeneous differential equation. It is linear because the unknown appears only once to the first power in each term. It is first-order because the highest derivative is the first, and it is homogeneous because the right-hand side, which would contain any terms not dependent on the unknown *I*, is zero. The solution to all linear first-order homogeneous differential equations is of the form:

$$I = I_0 e^{\alpha t}$$

where the constant α is determined by substituting the solution back into the differential equation and solving the resulting algebraic equation:

$$\alpha + \frac{1}{RC} = 0$$

In this case the solution is:

$$\alpha = -\frac{1}{RC}$$

The constant I_0 is determined from the initial condition at t = 0, at which time the voltage across the capacitor cannot change abruptly, and thus if the capacitor has zero voltage across it before the switch is closed, it will also have zero voltage immediately after the switch is closed. The capacitor initially behaves like a short circuit, and the initial current is:

$$I(0) = I_0 = \frac{V_S}{R}$$

Therefore, the complete solution for the transient series RC circuit for an initially discharged capacitor is:

$$I = \frac{V_S}{R} e^{-t/RC}$$

Using the component values shown in Fig. 2.202, this equation becomes:

$$I = 0.001 \text{ A } e^{-0.1t}$$

Once the current is known, the voltage across the resistor and capacitor can be easily determined:

$$V_R = IR = V_S e^{-t/RC} = 10 \text{ V } e^{-0.1t}$$

$$V_C = \frac{1}{C}\int_0^t I(t')dt' = V_S(1 - e^{-t/RC}) = 10 \text{ V}(1 - e^{-0.1t})$$

The graph in Fig. 2.202 shows how these voltages change with time. The section on capacitors explains some important details of the RC circuit not mentioned here and also explains how the discharging RC circuit works.

Once you've memorized the response equations for a charging and discharging RC as well as the equations for an energizing and a deenergizing RL circuit (see following equations), it's often possible to incorporate these equations into transient circuits without starting from scratch and doing the differential equations.

RC charging	RC discharging	LC energizing	LC deenergizing
$I = \frac{V_S}{R} e^{-t/RC}$	$I = \frac{V_S}{R} e^{-t/RC}$	$I = \frac{V_S}{R}(1 - e^{-Rt/L})$	$I = \frac{V_S}{R} e^{-Rt/L}$
$V_R = V_S e^{-t/RC}$	$V_R = V_S e^{-t/RC}$	$V_R = V_S\,(1 - e^{-Rt/L})$	$V_R = V_S e^{-Rt/L}$
$V_C = V_S\,(1 - e^{-t/RC})$	$V_C = V_S e^{-t/RC}$	$V_L = V_S e^{-Rt/L}$	$V_L = -V_S e^{-Rt/L}$

The following example problems illustrate how this is done.

Example 1: The circuit in Fig. 2.203 was under steady state before the switch was opened. Determine the current flow I_2 the instant the switch is opened, and then determine both I_2 and the voltage across R_2 1 ms after that.

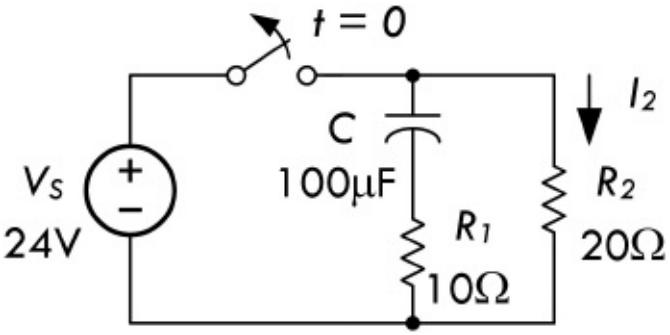


FIGURE 2.203

Answer: Unlike a resistor's voltage, a capacitor's voltage cannot change instantaneously (instead, it takes time). Because of this, the instant the switch is opened, the capacitor voltage remains as it was prior to the event:

$$V_C(0^+) = V_C(0^-) = 24 \text{ V}$$

where 0^+ means the instant after the switch is flipped, and 0^- means the instant before the switch is flipped.

The instant after the switch is opened ($t = 0^+$), the source is no longer part of the circuit, and we are left with the capacitor and two resistors all in series. Using Kirchhoff's law for this new circuit:

$$V_C + V_{R1} + V_{R2} = 0$$

$$24 \text{ V} + I(10\ \Omega + 20\ \Omega) = 0$$

$$I = -24 \text{ V}/30\ \Omega = 0.800 \text{ A}$$

This is the same current through I_2 the instant the switch is opened. To determine the current 1 ms ($t = 0.001$ s) after the switch is opened, simply treat the circuit as an RC discharge circuit (see Eq. 2.44), taking $R = R_1 + R_2$:

$$I = \frac{V_C}{R} e^{-t/RC} = \frac{24 \text{ V}}{(30\ \Omega)} e^{-0.001/(0.003)} = 0.573 \text{ A}$$

The voltage across R_2 at this time is:

$$V_{R2} = IR_2 = (0.573 \text{ A})(20\ \Omega) = 11.46 \text{ V}$$

Example 2: In the circuit in Fig. 2.204, determine the current I when the switch is opened ($t = 0$). Also determine the voltage across the resistor and inductor at $t = 0.1$ s.

Answer: When the switch is opened, the inductor voltage cannot change instantaneously, so it resembles a short, which gives rise to the forced response:

$$I_f = \frac{24\ \text{V}}{(10\ \Omega + 20\ \Omega)} = 0.80\ \text{A}$$

where the inductor's resistance is in series with *R*.

For any time after $t = 0$, the natural, or source-free, response is simply a deenergizing RL circuit, where *R* is the combined resistance of the fixed resistor and the resistance of the inductor:

$$I_n = C_e^{-Rt/L} = C_e^{-30t/5} = C_e^{-6t}$$

The total current is the sum of the forced and natural responses:

$$I = I_f + I_n = 0.80\ \text{A} + C_e^{-6t}$$

The trick then is to find *C* by finding the current $I(0^+)$:

$$I(0^+) = V_S/R_L = 24\ \text{V}/10\ \Omega = 2.4\ \text{A}$$

$$C = \frac{2.4\ \text{A} - 0.80\ \text{A}}{e^{-6(0)}} = 1.60\ \text{A}$$

Thus:

$$I(t) = (0.8 + 1.60\ e^{-6t})\ \text{A}$$

The voltage across the resistor and the inductor at $t = 0.1$ s can be found by first calculating the current at this time:

$$I(0.1) = (0.8 + 1.6e^{(-6 \times 0.1)} = 1.68\ \text{A}$$

Then:

$$V_R = IR = 1.68\ \text{A} \times 20\ \Omega = 33.6\ \text{V}$$

$$V_L = 24\ \text{V} - V_R = 24\ \text{V} - 33.6\ \text{V} = -9.6\ \text{V}$$

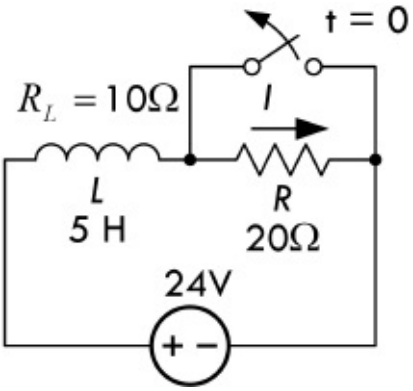


FIGURE 2.204

Example 3: Calculate the current I_L in the circuit in Fig. 2.205 at $t = 0.3$ s.

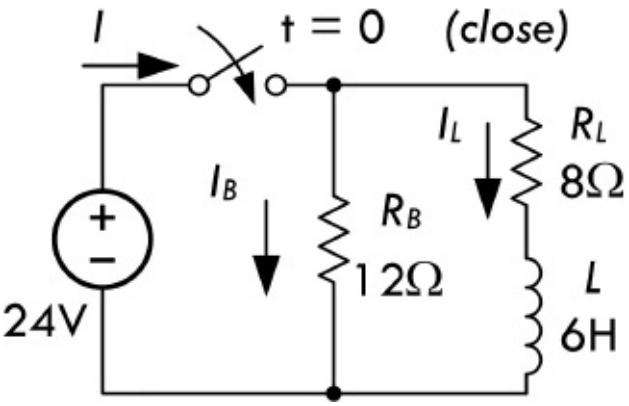


FIGURE 2.205

Answer: Notice that the 12-Ω resistor has no effect on the current I_L. Therefore, you get a simple RL energizing circuit with R_L in series with L:

$$I_L = \frac{V_S}{R_L}(1 - e^{-(R_L/L)t}) = \frac{24\text{ V}}{8\,\Omega}(1 - e^{-1.3t})$$

$$= (3\text{ A})(1 - e^{-1.3t})$$

So at $t = 0.3$ s:

$$I_L = (3\text{ A})(1 - e^{-1.3(0.3\text{ s})}) = 0.99\text{ A}$$

Here are some important things to notice during a forced response in regard to resistors, capacitors, and inductors:

Resistor: Under a forced response, a voltage is instantly placed across a resistor and a current immediately flows. There is no delay in voltage or current response (ideally).

Capacitor: Under a forced response, the voltage across a capacitor cannot change instantly, so at the instant a transition occurs it acts like an open circuit or constant voltage source. The voltage at instant $t = 0^-$ or $t = 0^+$ is a constant—the voltage that was present before the event. Also, at the instant $t = 0^-$ or $t = 0^+$ the current is zero, since no time transpires for charge to accumulate. However, after $t = 0^+$, the capacitor voltage and current have a natural response that is a function of time.

Inductor: Under a forced response, an inductor voltage cannot change instantaneously, so it acts like a short, meaning there is no voltage across it at $t = 0^-$ or $t = 0^+$. The current, however, at $t = 0^-$ or $t = 0^+$ will be a constant—the value of the current prior to the transient event. However, after $t = 0^+$, the inductor voltage and current have a natural response that is a function of time.

(Recall that $t = 0^-$ means the instant prior to the transient event, and $t = 0^+$ is the instant immediately after the transient event.)

Sometimes determining the voltage and currents within a transient circuit is a bit tricky, and requires a different approach than we encountered in the last three examples. The following few examples provide a good illustration of this.

Example 4: In the circuit in Fig. 2.206, determine the current through the inductor and the voltage across

the capacitor the instant before the switch is closed ($t = 0^-$) and the instant after the switch is closed ($t = 0^+$). Find I_C and I_L when the switch has been closed for $t = 0.5$ s.

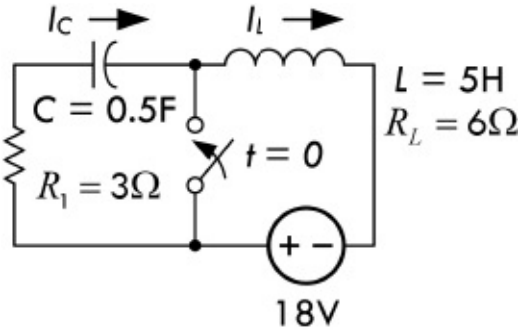


FIGURE 2.206

Answer: Before the switch is closed, the capacitor acts as an open circuit, preventing current from flowing. The instant after the switch is closed, the capacitor cannot change in voltage instantaneously, so the current is still zero. The instant before and the instant after the switch is closed, the voltage across the capacitor is equal to the supply voltage. All this can be expressed by:

$$I_L(0^-) = I_L(0^+) = 0$$

$$V_C(0^-) = V_C(0^+) = 18 \text{ V}$$

When the switch is closed, we have a source-free RC circuit for which:

$$I_C = Be^{-t/RC} = Be^{-t/(1.5)}$$

At $t = 0^+$,

$$I_C(0^+)R + V_C(0^+) = 18 \text{ V}$$

So,

$$I_C = (0^+) = -\frac{V_C(0^+)}{R} = -\frac{18 \text{ V}}{3\ \Omega} = -6 \text{ A}$$

Plugging this back in to find A we get:

$$B = \frac{I_C(0^+)}{e^{-0/(1.5)}} = \frac{-6 \text{ A}}{1} = -6 \text{ A}$$

So the complete expression for I_C is:

$$I_C = (-6 \text{ A})e^{-t/(1.5)}$$

At $t = 0.5$ s,

$$I_C = (-6 \text{ A})e^{-0.5/(1.5)} = -4.3 \text{ A}$$

To find I_L at t = 0.5 s, we consider an RL circuit excited by an 18-V source (forced response) plus the natural response of an RL circuit:

$$I_L = I_f + I_n = \frac{V_S}{R_I} + Ae^{-6t/5}$$

Example 5: State all the initial conditions for the circuit in Fig. 2.207, which is under steady state for t < 0, and the switch is opened at t = 0. Determine the current in the circuit 0.6 s after the switch is opened. What is the induced voltage in the inductor at t = 0.4 s?

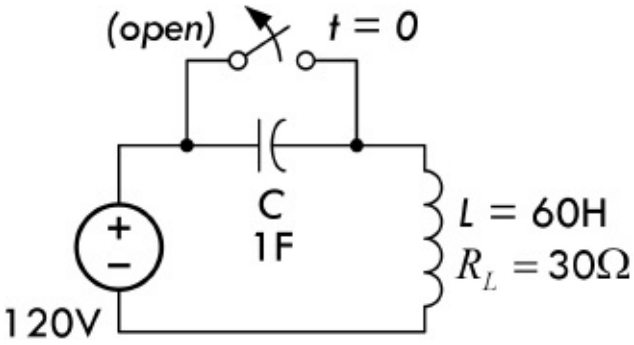


FIGURE 2.207

Answer: Since before the switch is initially closed, the capacitor is shorted out, there is no voltage across it, and the current is simply equal to the supply voltage divided by the resistance of the inductor:

$$I_L(0^-) = I_L(0^+) = \frac{120\text{ V}}{30\ \Omega} = 4\text{ A}$$

$$V_C(0^-) = V_C(0^+) = 0\text{ V}$$

By Kirchhoff's voltage law, we get the following equation for the circuit:

$$L\frac{dI}{dt} + R_L I + \frac{1}{C}\int Idt = 120\text{ V}$$

$$(60\text{ H})\frac{dI}{dt} + (30\ \Omega)I + \frac{1}{1\text{ F}}\int Idt = 120\text{ V}$$

from which the characteristic equation is:

$$60\,p^2 + 20\,p + 1 = 0$$

The characteristic roots are:

$$p = -0.46, -0.04$$

The complete current response is the sum of the forced and natural responses:

$$I = I_f + I_n = 0 + A_1e^{-0.46t} + A_2e^{-0.04t}$$

$I(0^+)$ = 4 implies that $A_1 + A_2 = 4$.

At t = 0,

$$(60\text{ H})\frac{dI}{dt}(0^+) + (30\ \Omega)/(0^+) + V_C(0^+) = 120\text{ V}$$

$$\frac{dI}{dt}(0^+) = 0 = 0.46A_1 + 0.04A_2$$

Solving for A_1 and A_2 gives $A_1 = -0.38$, $A_2 = 4.38$. Thus:

$$I = -0.38e^{-0.46t} + 4.38e^{-0.04t}$$

The induced voltage in the inductor at $t = 0.4$ s is simply found by plugging I into the definition of voltage for an inductor:

$$V_L = L\frac{dI}{dt} = 60[(-0.38)(-0.46)e^{-0.46t} + (4.38)(-0.04)e^{-0.04t}]$$

At $t = 0.4$ s,

$$V_L = 60(0.145 - 0.172) = -1.62\text{ V}$$

Example 6: In the circuit in Fig. 2.208, the switch is moved from 1 to 2 at $t = 0$. Determine I as a function of time thereafter.

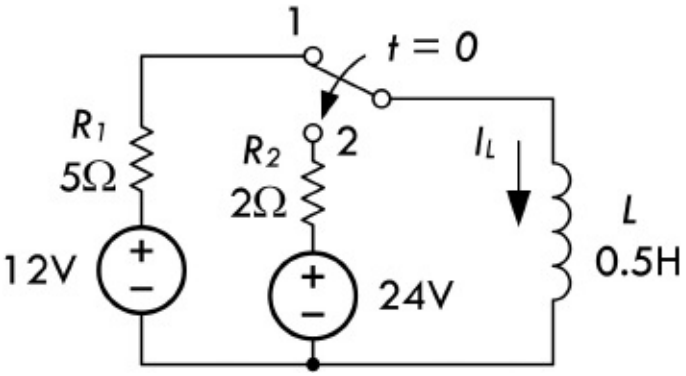


FIGURE 2.208

Answer: The complete current response when the switch is thrown is the sum of the forced response and the natural response:

$$I = I_f + I_n = \frac{V_{24\text{ V}}}{R_2} + I_L e^{-Rt/L}$$

$$= \frac{24\text{ V}}{2\ \Omega} + \left(\frac{12\text{ V}}{5\ \Omega} - \frac{24\text{ V}}{2\ \Omega}\right)e^{-2t/0.5}$$

$$= 12\text{ A} - (9.6\text{A})e^{-4t}$$

2.34.1 Series RLC Circuit

There's another transient example, which is a bit heavy in the math but is a classic analog of many other phenomena found in science and engineering. See Fig. 2.209.

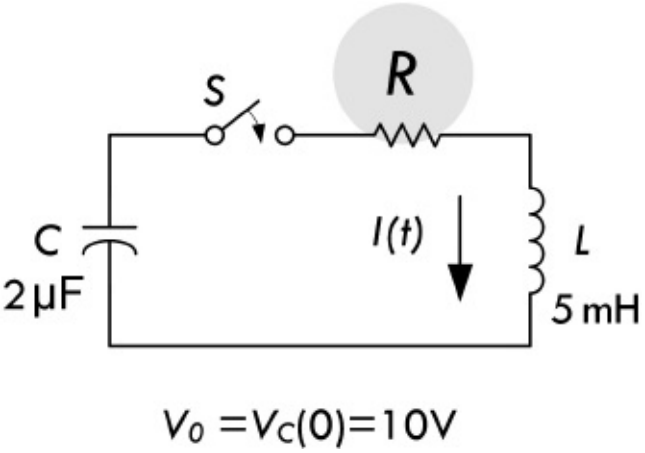


FIGURE 2.209

Assume the capacitor is charged to a voltage V_0, and then at $t = 0$, the switch is closed. Kirchhoff's voltage law for $t \geq 0$:

$$\frac{1}{C}\int I dt + IR + L\frac{dI}{dt} = 0$$

Rewriting in the standard form gives:

$$\frac{d^2I}{dt^2} + \frac{R}{L}\frac{dI}{dt} + \frac{1}{LC}I = 0$$

This is an example of a linear second-order homogeneous differential equation. It is reasonable to guess that the solution is of the same form as for the first-order homogeneous differential equation encountered earlier:

$$I = I_0 e^{\alpha t}$$

Substituting this into the differential equation gives:

$$\alpha^2 + \frac{R}{L}\alpha + \frac{1}{LC} = 0$$

Note that a solution of the form $e^{\alpha t}$ always reduces a linear homogeneous differential equation to an algebraic equation in which first derivates are replaced by α and second derivates by α^2, and so forth. A linear second-order homogeneous differential equation then becomes a quadratic algebraic equation, and so on. This particular algebraic equation has the following solutions:

$$\alpha_1 = -\frac{R}{2L} + \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$

$$\alpha_2 = -\frac{R}{2L} - \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} \qquad (2.92)$$

Since either value of α represents a solution to the original differential equation, the most general solution is one in which the two possible solutions are multiplied by arbitrary constants and added together:

$$I = I_1 e^{\alpha_1 t} + I_2 e^{\alpha_2 t}$$

The constants I_1 and I_2 must be determined from the initial conditions. An nth-order differential equation will generally have n constants that must be determined from initial conditions. In this case the constants can be evaluated from a knowledge of $I(0)$ and $dI/dt(0)$. Since the current in the inductor was zero for $t < 0$, and since it cannot change abruptly, we know that:

$$I(0) = 0$$

The initial voltage across the inductor is the same as across the capacitor, so that:

$$I_1 = -I_2 = \frac{V_0}{(\alpha_1 - \alpha_2)L}$$

The solution for the current in the series RLC circuit is thus:

$$I = \frac{V_0}{(\alpha_1 - \alpha_2)L}(e^{\alpha_1 t} - e^{\alpha_2 t}) \tag{2.93}$$

where α_1 and α_2 are given in Eq. 2.92.

The solution to the preceding equation has a unique character, depending on whether the quantity under the square root in Eq. 2.92 is positive, zero, or negative. We consider these three cases.

Case 1: Overdamped

For $R^2 > 4L/C$, the quantity under the square root is positive, and both values of α are negative with $|\alpha_2| > |\alpha_1|$. The solution is the sum of a slowly decaying positive term and a more rapidly decaying negative term of equal initial magnitude. The solution is sketched in Fig. 2.210. An important limiting case is the one in which $R^2 >> 4L/C$. In that limit, the square root can be approximated as:

$$\sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} = \frac{R}{2L}\sqrt{1 - \frac{4L}{R^2C}} \approx \frac{R}{2L} - \frac{1}{RC}$$

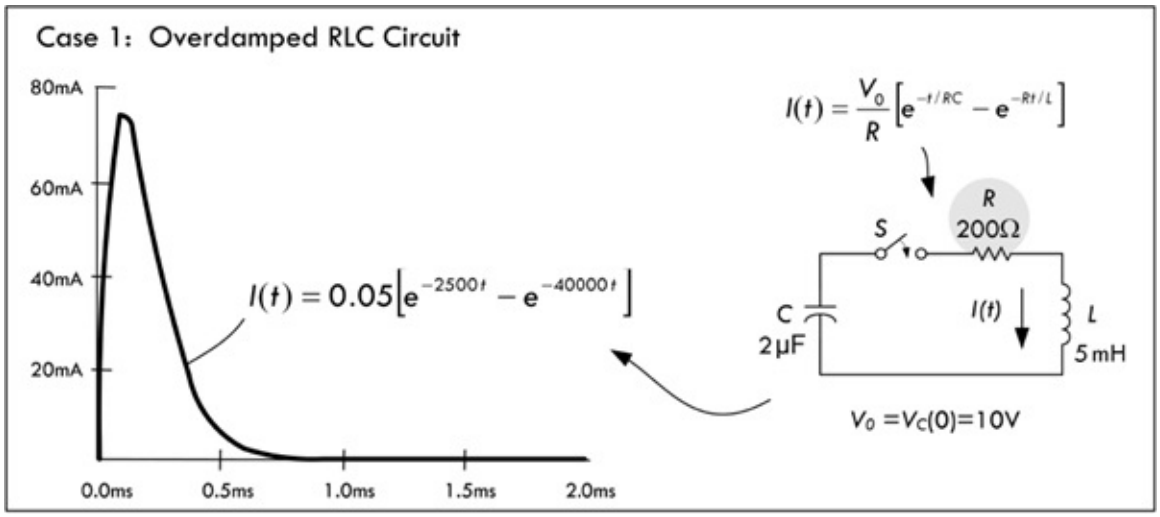


FIGURE 2.210

and the corresponding values of α are:

$$\alpha_1 = -\frac{1}{RC} \quad \text{and} \quad \alpha_2 = -\frac{R}{L}$$

Then the current in Eq. 2.93 is:

$$I \approx \frac{V_0}{R}(e^{-t/RC} - e^{-Rt/L}) \tag{2.94}$$

In this limit, the current rises very rapidly (in a time ~L/R) to a value near V_0/R and then decays very slowly back to zero. The overdamped response curve is shown in Fig. 2.210.

Case 2: Critically Damped

For $R^2 = 4L/C$, the quantity under the square root is zero, and $\alpha_1 = \alpha_2$. Equation 2.93 is then zero divided by zero, which is undefined. Therefore, the method of solution outlined here fails. A more productive approach is to let

$$\varepsilon = \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}$$

and take the limit of Eq. 2.93 as $\varepsilon \to 0$. Then:

$$\alpha_1 = -\frac{R}{2L} + \varepsilon \quad \text{and} \quad \alpha_2 = -\frac{R}{2L} - \varepsilon$$

and Eq. 2.93 becomes:

$$I = \frac{V_0}{2\pi L} e^{-Rt/2L}(e^{\varepsilon t} - e^{-\varepsilon t})$$

Using the expansion $e^x \approx 1 + x$, for $|x| << 1$, the preceding equation becomes:

$$I = \frac{V_0}{L} e^{-Rt/2L} \tag{2.95}$$

The shape of the curve is not very different from the overdamped case, except that it approaches zero as fast as possible without overshooting the t axis and going negative. The critically damped response curve is shown in Fig. 2.211.

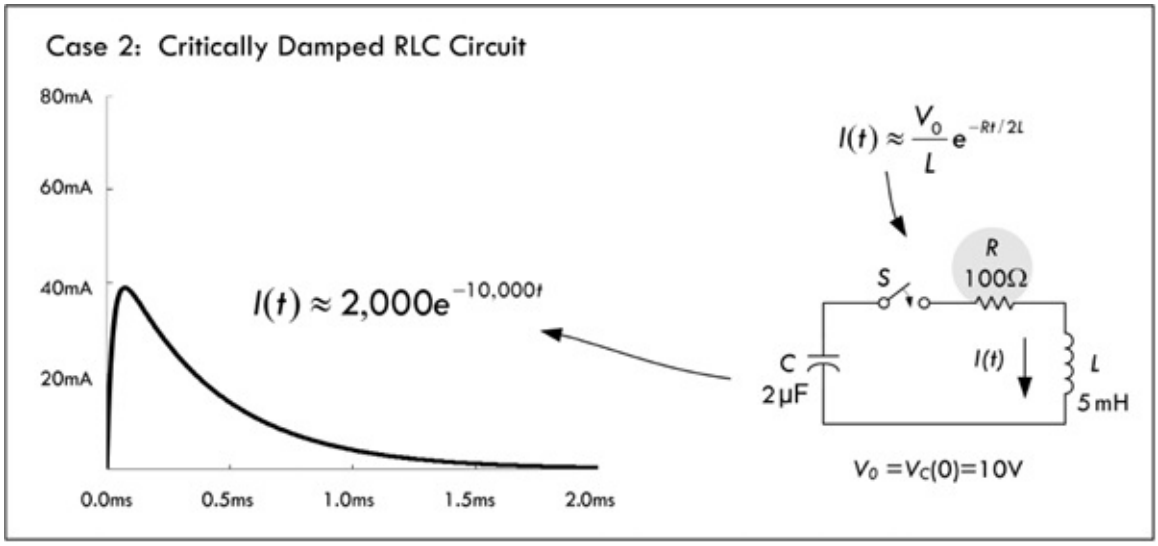


FIGURE 2.211

Critical damping is difficult to achieve: only a small change in R moves from this point. Small temperature change will cause that to occur. Energy transfer from C to L is now smaller than loss in R.

Case 3: Underdamped

For $R^2 < 4L/C$, the quantity under the square root is negative, and α can be written as:

$$\alpha = -\frac{R}{2L} \pm \frac{j}{\sqrt{LC}}\sqrt{1-\frac{R^2C}{4L}}$$

where $j = \sqrt{-1}$

It is useful to define another symbol, ω, which we call the angular frequency:

$$\omega = \frac{1}{\sqrt{LC}}\sqrt{1-\frac{R^2C}{4L}} \tag{2.96}$$

When $R^2 << 4L/C$, the angular frequency is:

$$\omega \approx \frac{1}{\sqrt{LC}}$$

and this approximation will usually suffice for most cases of interest. With these substitutions, Eq. 2.93 becomes:

$$I = \frac{V_0}{2j\omega L}e^{-Rt/2L}(e^{j\omega t} - e^{-j\omega t})$$

We now make use of Euler's equation $e^{j\theta} = \cos\theta + j\sin\theta$ to express the current as follows:

$$I = \frac{V_0}{\omega L}e^{-Rt/2L}\sin(\omega t) \tag{2.97}$$

Notice that the solu
amplitude decaying e
frequency, note that i
units of cycles per se

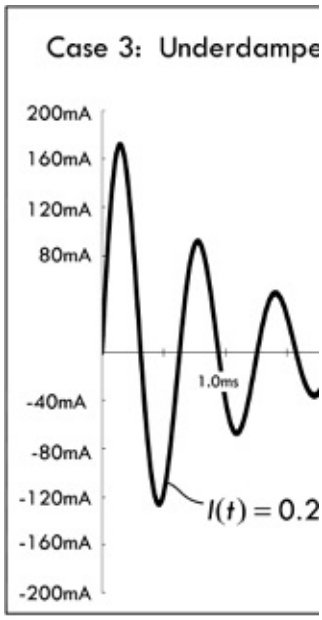


FIGURE 2.212

The underdamped
current increases, er
cycle has passed, at
the inductor decrease
time all the energy e
slosh back and forth
resistance would osc

This type of diffe
science and enginee
automobile, for exar
damped.

2.35 Circuits

Suppose that you are
that is used to drive
circuit? The circuit i
cannot directly apply

If all else fails, yc
before going any furt
is, even if you could
term. For example, l
expression for a pe
pretend that you can
into Kirchhoff's law
then, because things

CHAPTER 3

Basic Electronic Circuit Components

3.1 Wires, Cables, and Connectors

Wires and cables provide low-resistance pathways for electric currents. Most electrical wires are made from copper or silver and typically are protected by an insulating coating of plastic, rubber, or lacquer. Cables consist of a number of individually insulated wires bound together to form a multiconductor transmission line. Connectors, such as plugs, jacks, and adapters, are used as mating fasteners to join wires and cable with other electrical devices.

3.1.1 Wires

A wire's diameter is expressed in terms of a *gauge number*. The gauge system, as it turns out, goes against common sense. In the gauge system, as a wire's diameter increases, the gauge number decreases. At the same time, the resistance of the wire decreases. When currents are expected to be large, smaller-gauge wires (large-diameter wires) should be used. If too much current is sent through a large-gauge wire (small-diameter wire), the wire could become hot enough to melt. Table 3.1 shows various characteristics for B&S-gauged copper wire at 20°C. For rubber-insulated wire, the allowable current should be reduced by 30 percent.

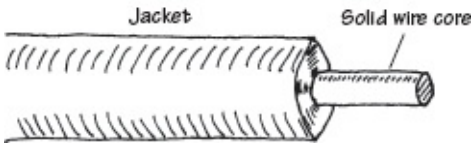
TABLE 3.1 Copper Wire Specifications (Bare and Enamel-Coated Wire)

WIRE SIZE (AWG)	DIAMETER (MILS)*	DIAMETER (MM)	OHMS PER 1000 FT	OHMS PER KM	CURRENT CARRY-ING CAPACITY (A)	NEAREST BRIT-ISH SWG NO.
1	289.3	7.35	0.1239	0.41	119.564	1
2	257.6	6.54	0.1563	0.51	94.797	2
3	229.4	5.83	0.1971	0.65	75.178	4
4	204.3	5.19	0.2485	0.82	59.626	5
5	181.9	4.62	0.3134	1.03	47.268	6
6	162.0	4.12	0.3952	1.30	37.491	7
7	144.3	3.67	0.4981	1.63	29.746	8
8	128.5	3.26	0.6281	2.06	23.589	9
9	114.4	2.91	0.7925	2.60	18.696	11
10	101.9	2.59	0.9987	3.28	14.834	12
11	90.7	2.31	1.2610	4.13	11.752	13
12	80.8	2.05	1.5880	5.21	9.327	13
13	72.0	1.83	2.0010	6.57	7.406	15
14	64.1	1.63	2.5240	8.29	5.870	15
15	57.1	1.45	3.1810	10.45	4.658	16
16	50.8	1.29	4.0180	13.17	3.687	17
17	45.3	1.15	5.0540	16.61	2.932	18
18	40.3	1.02	6.3860	20.95	2.320	19
19	35.9	0.91	8.0460	26.42	1.841	20
20	32.0	0.81	10.1280	33.31	1.463	21
21	28.5	0.72	12.7700	42.00	1.160	22
22	25.3	0.64	16.2000	52.96	0.914	22
23	22.6	0.57	20.3000	66.79	0.730	24
24	20.1	0.51	25.6700	84.22	0.577	24
25	17.9	0.46	32.3700	106.20	0.458	26
26	15.9	0.41	41.0200	133.90	0.361	27
27	14.2	0.36	51.4400	168.90	0.288	28
28	12.6	0.32	65.3100	212.90	0.227	29
29	11.3	0.29	81.2100	268.50	0.182	31
30	10.0	0.26	103.7100	338.60	0.143	33
31	8.9	0.23	130.9000	426.90	0.113	34
32	8.0	0.20	162.0000	538.30	0.091	35
33	7.1	0.18	205.7000	678.80	0.072	36
34	6.3	0.16	261.3000	856.00	0.057	37
35	5.6	0.14	330.7000	1079.00	0.045	38
36	5.0	0.13	414.8000	1361.00	0.036	39
37	4.5	0.11	512.1000	1716.00	0.029	40

* 1 Mil = 2.54×10^{-5} m

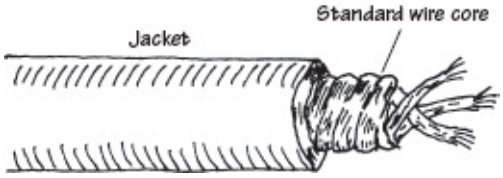
Wire comes in solid core, stranded, or braided forms.

Solid Core



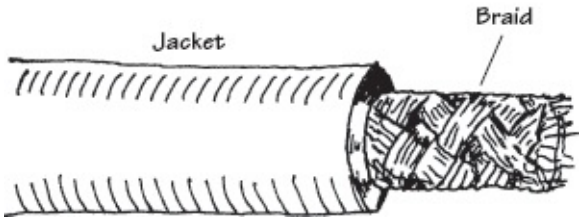
This wire is useful for wiring breadboards; the solid-core ends slip easily into breadboard sockets and will not fray in the process. These wires have the tendency to snap after a number of flexes.

Stranded Wire



The main conductor is comprised of a number of individual strands of copper. Stranded wire tends to be a better conductor than solid-core wire because the individual wires together comprise a greater surface area. Stranded wire will not break easily when flexed.

Braided Wire

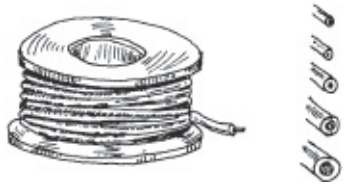


A braided wire is made up of a number of individual strands of wire braided together. Like stranded wires, these wires are better conductors than solid-core wires, and they will not break easily when flexed. Braided wires are frequently used as an electromagnetic shield in noise-reduction cables and also may act as a wire conductor within the cable (e.g., coaxial cable).

FIGURE 3.1

Kinds of Wires

Pretinned Solid Bus Wire

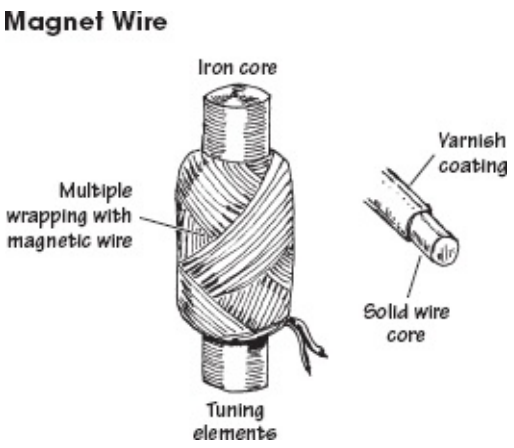


This wire is often referred to as hookup wire. It includes a tin-lead alloy to enhance solderability and is usually insulated with polyvinyl chloride (PVC), polyethylene, or Teflon. Used for hobby projects, preparing printed circuit boards, and other applications where small bare-ended wires are needed.

Speaker Wire



This wire is stranded to increase surface area for current flow. It has a high copper content for better conduction.



This wire is used for building coils and electromagnets or anything that requires a large number of loops, say, a tuning element in a radio receiver. It is built of solid-core wire and insulated by a varnish coating. Typical wire sizes run from 22 to 30 gauge.

FIGURE 3.2

3.1.2 Cables

A cable consists of a multiple number of independent conductive wires. The wires within cables may be solid core, stranded, braided, or some combination in between. Typical wire configurations within cables include the following:

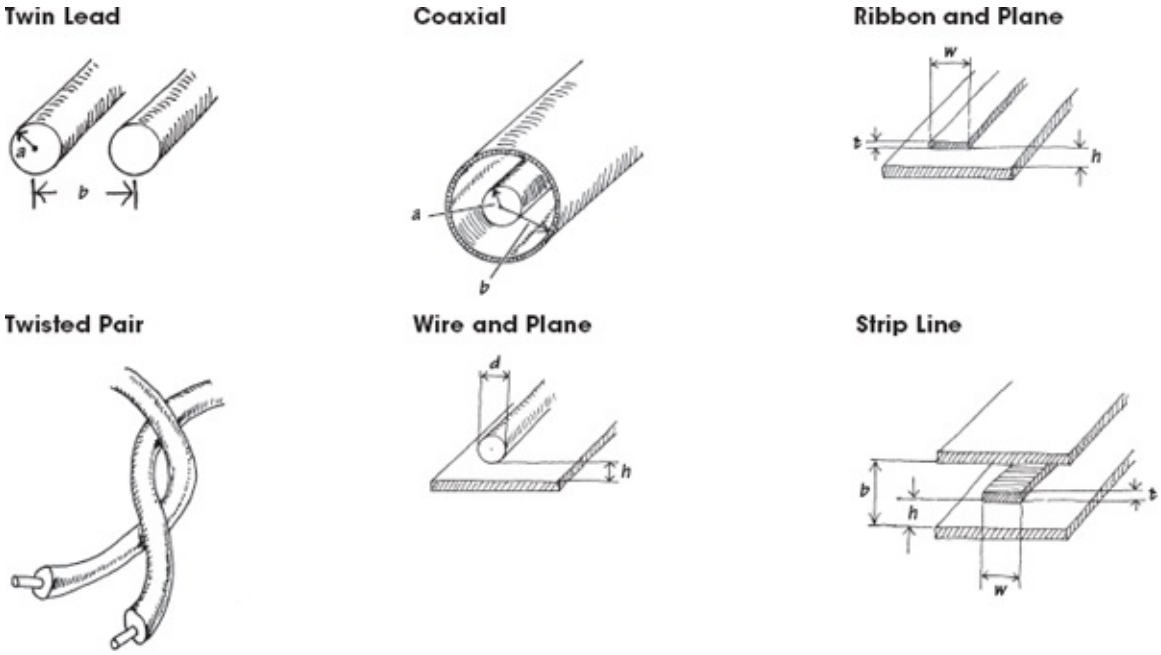
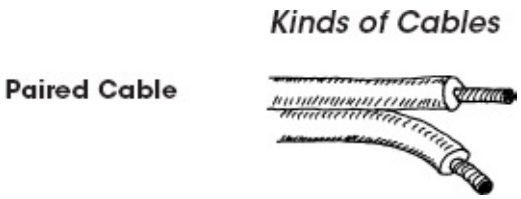


FIGURE 3.3

3.1.3 Connectors

The following is a list of common plug and jack combinations used to fasten wires and cables to electrical devices. Connectors consist of plugs (male-ended) and jacks (female-ended). To join dissimilar connectors together, an adapter can be used.



This cable is made from two individually insulated conductors. Often it is used in dc or low-frequency ac applications.

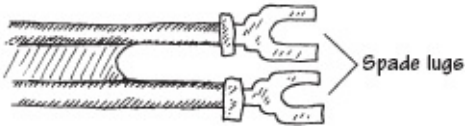
Twisted Pair



This cable is composed of two interwound insulated wires. It is similar to a paired cable, but the wires are held together by a twist.

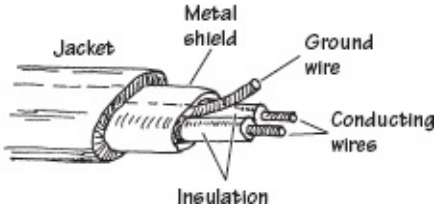
The common CAT5 cable used for Ethernet, among other things, is based on a set of four twisted pairs.

Twin Lead



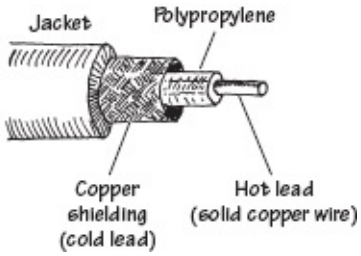
This cable is a flat two-wire line, often referred to as 300-Ω line. The line maintains an impedance of 300 Ω. It is used primarily as a transmission line between an antenna and a receiver (e.g., TV, radio). Each wire within the cable is stranded to reduce skin effects.

Shielded Twin Lead



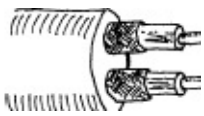
This cable is similar to paired cable, but the inner wires are surrounded by a metal-foil wrapping that's connected to a ground wire. The metal foil is designed to shield the inner wires from external magnetic fields—potential forces that can create noisy signals within the inner wires.

Unbalanced Coaxial



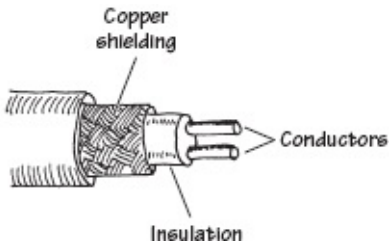
This cable typically is used to transport high-frequency signals (e.g., radio frequencies). The cable's geometry limits inductive and capacitive effects and also limits external magnetic interference. The center wire is made of solid-core copper or aluminium wire and acts as the hot lead. An insulative material, such as polyethylene, surrounds the center wire and acts to separate the center wire from a surrounding braided wire. The braided wire, or copper shielding, acts as the cold lead or ground lead. Characteristic impedances range from about 50 to 100 Ω.

Dual Coaxial



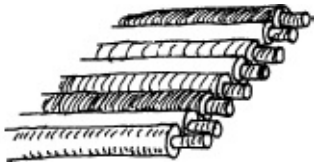
This cable consists of two unbalanced coaxial cables in one. It is used when two signals must be transferred independently.

Balanced Coaxial



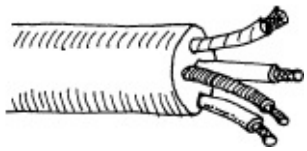
This cable consists of two solid wires insulated from one another by a plastic insulator. Like unbalanced coaxial cable, it too has a copper shielding to reduce noise pickup. Unlike unbalanced coaxial cable, the shielding does not act as one of the conductive paths; it only acts as a shield against external magnetic interference.

Ribbon



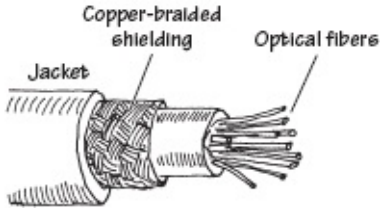
This type of cable is used in applications where many wires are needed. It tends to flex easily. It is designed to handle low-level voltages and often is found in digital systems, such as computers, to transmit parallel bits of information from one digital device to another.

Multiple Conductor



This type of cable consists of a number of individually wrapped, color-coded wires. It is used when a number of signals must be sent through one cable.

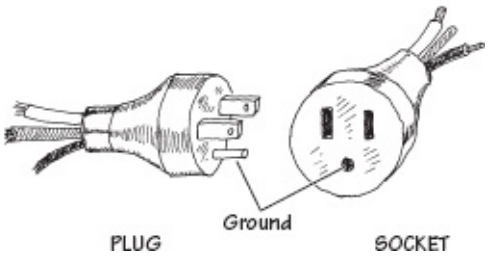
Fiberoptic



Fiberoptic cable is used in the transport of electromagnetic signals, such as light. The conducting-core medium is made from a glass material surrounded by a fiberoptic cladding (a glass material with a higher index of refraction than the core). An electromagnetic signal propagates down the cable by multiple total internal reflections. It is used in direct transmission of images and illumination and as waveguides for modulated signals used in telecommunications. One cable typically consists of a number of individual fibers.

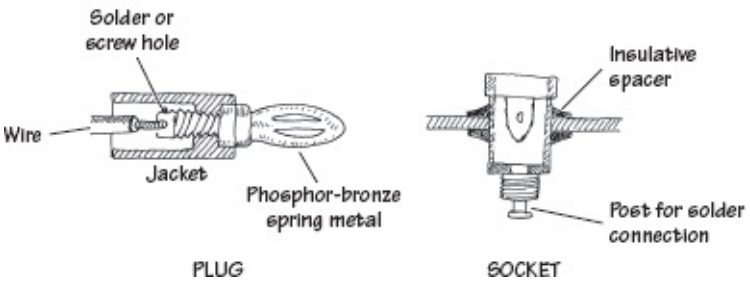
FIGURE 3.4

117-Volt



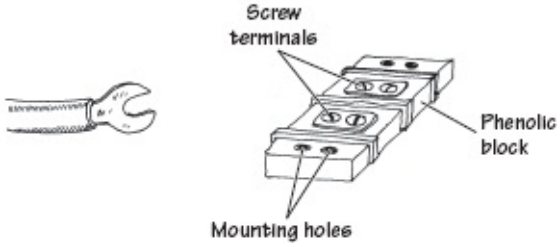
This is a typical home appliance connector. It comes in unpolarized and polarized forms. Both forms may come with or without a ground wire.

Banana



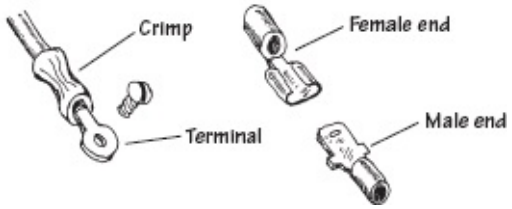
This is used for connecting single wires to electrical equipment. It is frequently used with testing equipment. The plug is made from a four-leafed spring tip that snaps into the jack.

Spade Lug/Barrier Strip



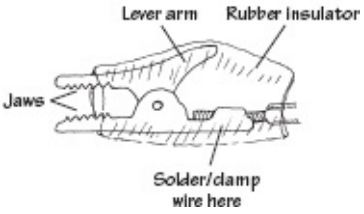
This is a simple connector that uses a screw to fasten a metal spade to a terminal. A barrier strip often acts as the receiver of the spade lugs.

Crimp



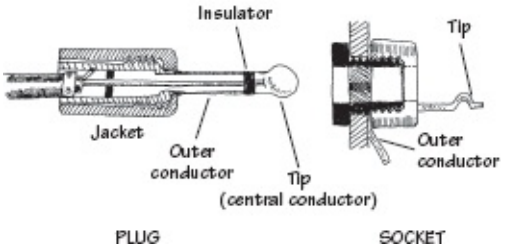
Crimp connectors are color-coded according to the wire size they can accommodate. They are useful as quick, friction-type connections in dc applications where connections are broken repeatedly. A crimping tool is used to fasten the wire to the connector.

Alligator



Alligator connectors are used primarily as temporary test leads.

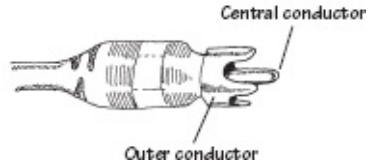
Phone



These connectors accept shielded braid, but they are larger in size. They come in two-or three-element types and have a barrel that is 1¼ in (31.8 mm) long. They are used as connectors in microphone cables and for other low-voltage, low-current applications.

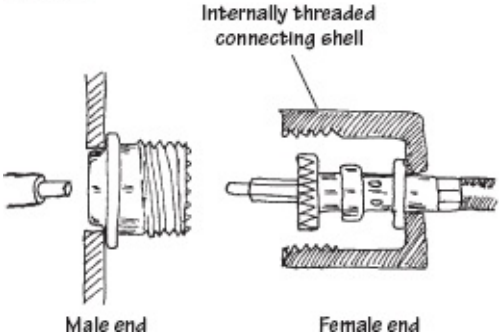
3.5 mm and even 2.5 mm versions of these connectors are also commonly used.

Phono



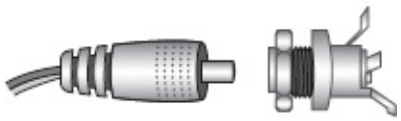
Phono connectors are often referred to as RCA plugs or pin plugs. They are used primarily in audio connections.

F-Type



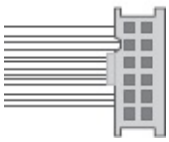
F-type connectors are used with a variety of unbalanced coaxial cables. They are commonly used to interconnect video components. F-type connectors are either threaded or friction-fit together.

Tip



These connectors are commonly used to supply low voltage dc between 3 and 15V.

Mini



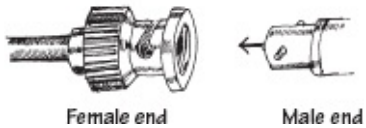
IDC connectors are often found in computers. The plug is attached to ribbon cable using v-shaped teeth that are squeezed into the cable insulation to make a solderless contact.

PL-259



These are often referred to as UHF plugs. They are used with RG-59/U coaxial cable. Such connectors may be threaded or friction-fit together.

BNC



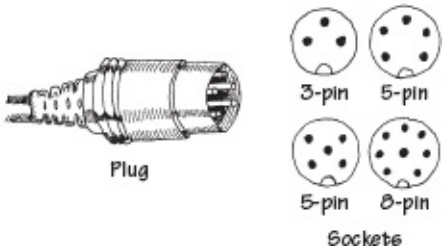
BNC connectors are used with coaxial cables. Unlike the F-type plug, BNC connectors use a twist-on bayonet-like locking mechanism. This feature allows for quick connections

T-Connector



T-connectors consist of two plug ends and one central jack end. They are used when a connection must be made somewhere along a coaxial cable.

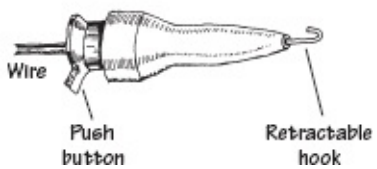
DIN Connector



These connectors are used with multiple conductor wires. They are often used for interconnecting audio and computer accessories.

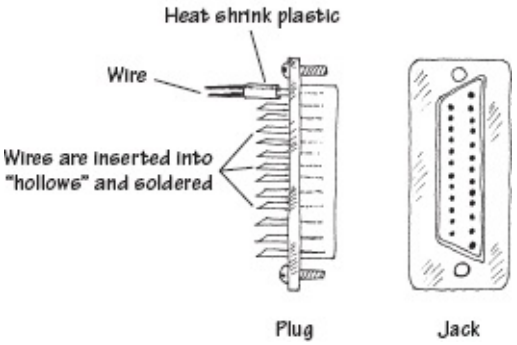
Smaller versions of these connectors (mini-DIN) are also widely used.

Meat Hook



These connectors are used as test probes. The spring-loaded hook opens and closes with the push of a button. The hook can be clamped onto wires and component leads.

D-Connector



D-connectors are used with ribbon cable. Each connector may have as many as 50 contacts. The connection of each individual wire to each individual plug pin or jack socket is made by sliding the wire in a hollow metal collar at the backside of each connector. The wire is then soldered into place.

FIGURE 3.5

3.1.4 Wiring and Connector Symbols

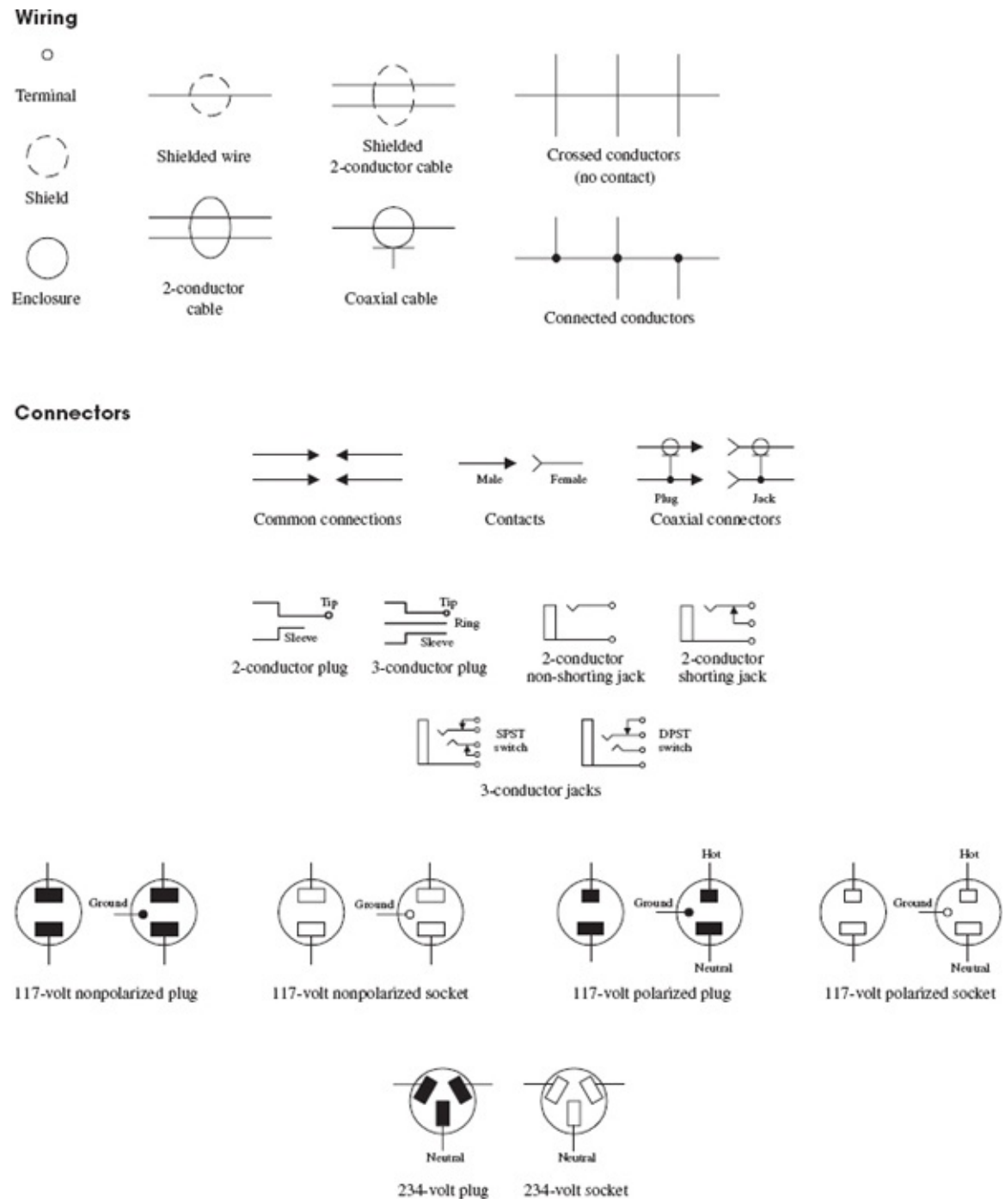


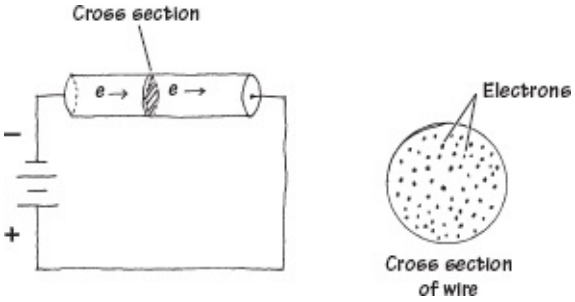
FIGURE 3.6

3.1.5 High-Frequency Effects Within Wires and Cables

Weird Behavior in Wires (Skin Effect)

When dealing with low-current dc hobby projects, wires and cables are straightforward—they act as simple conductors with essentially zero resistance. However, when you replace dc currents with very high-frequency ac currents, weird things begin to take place within wires. As you will see, these "weird things" will not allow you to treat wires as perfect conductors.

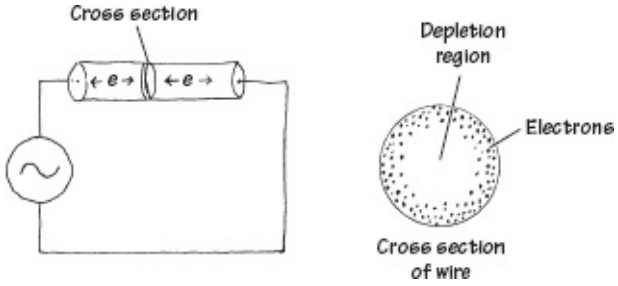
First, let's take a look at what is going on in a wire when a dc current is flowing through it.



A wire that is connected to a dc source will cause electrons to flow through the wire in a manner similar to the way water flows through a pipe. This means that the path of any one electron essentially can be anywhere within the volume of the wire (e.g., center, middle radius, surface).

FIGURE 3.7

Now, let's take a look at what happens when a high-frequency ac current is sent through a wire.



An ac voltage applied across a wire will cause electrons to vibrate back and forth. In the vibrating process, the electrons will generate magnetic fields. By applying some physical principles (finding the forces on every electron that result from summing up the individual magnetic forces produced by each electron), you find that electrons are pushed toward the surface of the wire. As the frequency of the applied signal increases, the electrons are pushed further away from the center and toward the surface. In the process, the center region of the wire becomes devoid of conducting electrons.

FIGURE 3.8

The movement of electrons toward the surface of a wire under high-frequency conditions is called the *skin effect*. At low frequencies, the skin effect does not have a large effect on the conductivity (or resistance) of the wire. However, as the frequency increases, the resistance of the wire may become an influential factor. Table 3.2 shows just how influential skin effect can be as the frequency of the signal increases (the table uses the ratio of ac resistance to dc resistance as a function of frequency).

TABLE 3.2 ac/dc Resistance Ratio as a Function of Frequency

	R_{AC}/R_{DC}			
WIRE GAUGE	10^6 HZ	10^7 HZ	10^8 HZ	10^9 HZ
22	6.9	21.7	68.6	217
18	10.9	34.5	109	345
14	17.6	55.7	176	557
10	27.6	87.3	276	873

One thing that can be done to reduce the resistance caused by skin effects is to use stranded wire—the combined surface area of all the individual wires within the conductor is greater than the surface area for a solid-core wire of the same diameter.

Weird Behavior in Cables (Lecture on Transmission Lines)

Like wires, cables also exhibit skin effects. In addition, cables exhibit inductive and capacitive effects that result from the existence of magnetic and electrical fields within the cable. A magnetic field produced

by the current through one wire will induce a current in another. Likewise, if two wires within a cable have a net difference in charge between them, an electrical field will exist, thus giving rise to a capacitive effect.

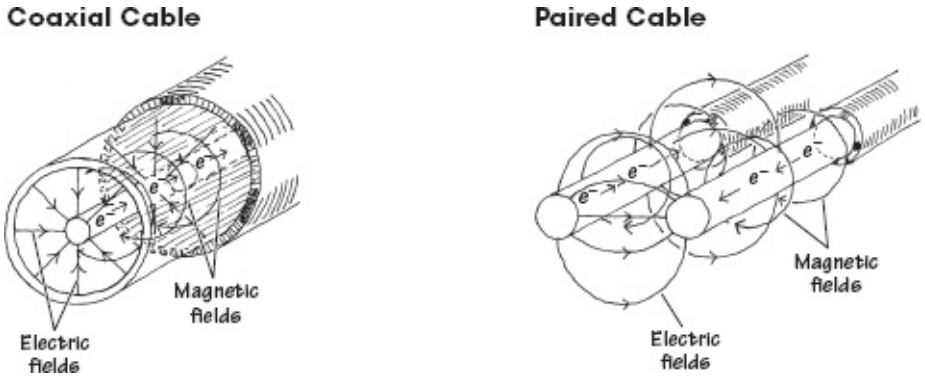
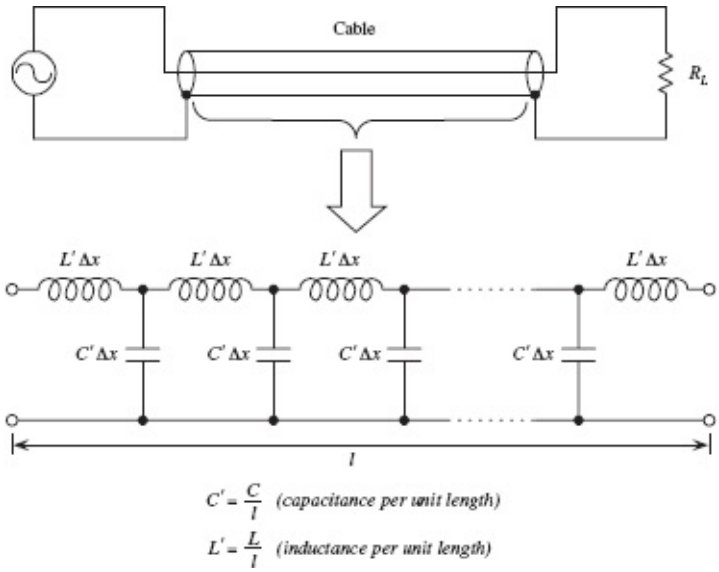


FIGURE 3.9 Illustration of the electrical and magnetic fields within a coaxial and paired cable.

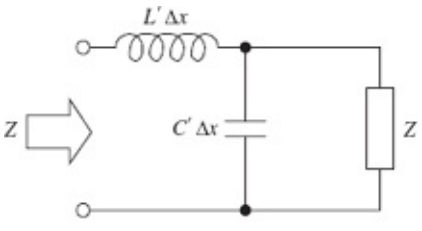
Taking note of both inductive and capacitive effects, it is possible to treat a cable as if it were made from a number of small inductors and capacitors connected together. An equivalent inductor-capacitor network used to model a cable is shown in Fig. 3.10.



The impedance of a cable can be modeled by treating it as a network of inductors and capacitors.

FIGURE 3.10

To simplify this circuit, we apply a reduction trick; we treat the line as an infinite ladder and then assume that adding one "rung" to the ladder (one inductor-capacitor section to the system) will not change the overall impedance Z of the cable. What this means—mathematically speaking—is we can set up an equation such that $Z = Z + (LC \text{ section})$. This equation can then be solved for Z. After that, we find the limit as Δx goes to zero. The mathematical trick and the simplified circuit are shown below.



$$Z = j\omega L'\Delta x + \frac{Z/j\omega C'\Delta x}{Z + 1/j\omega C'\Delta x} = j\omega L'\Delta x + \frac{Z}{1 + j\omega C'Z\Delta x}$$

When $\Delta x \rightarrow$ small,

$$Z = \sqrt{L'/C'} = \sqrt{\frac{Ll}{Cl}} = \sqrt{LC}$$

FIGURE 3.11

By convention, the impedance of a cable is called the *characteristic impedance* (symbolized Z_0). Notice that the characteristic impedance Z_0 is a real number. This means that the line behaves like a resistor despite the fact that we assumed the cable had only inductance and capacitance built in.

The question remains, however, what are L and C? Well, figuring out what L and C should be depends on the particular geometry of the wires within a cable and on the type of dielectrics used to insulate the wires. You could find L and C by applying some physics principles, but instead, let's cheat and look at the answers. The following are the expressions for L and C and Z_0 for both a coaxial and parallel-wire cable:

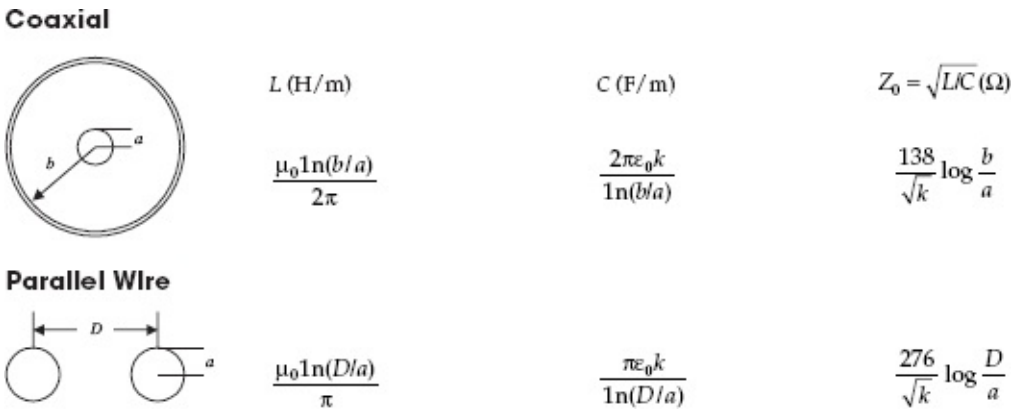


FIGURE 3.12 Inductance, capacitance, and characteristic impedance formulas for coaxial and parallel wires.

Here, k is the dielectric constant of the insulator, $\mu_0 = 1.256 \times 10^{-6}$ H/m is the permeability of free space, and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m is the permitivity of free space. Table 3.3 provides some common dielectric materials with their corresponding constants.

TABLE 3.3 Common Dielectrics and Their Constants

MATERIAL	DIELECTRIC CONSTANT (k)
Air	1.0
Pyrex glass	4.8
Mica	5.4
Paper	3.0
Polyethylene	2.3
Polystyrene	5.1–5.9
Quartz	3.8
Teflon	2.1

Often, cable manufacturers supply capacitance per foot and inductance per foot values for their cables. In this case, you can simply plug the given manufacturer's values into $Z_0 = \sqrt{L/C}$ to find the characteristic impedance of the cable. Table 3.4 shows capacitance per foot and inductance per foot values for some common cable types.

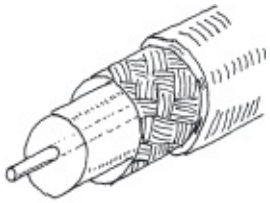
TABLE 3.4 Capacitance and Inductance per Foot for Some Common Transmission-Line Types

CABLE TYPE	CAPACITANCE/FT (pF)	INDUCTANCE/FT (μH)
RG-8A/U	29.5	0.083
RG-11A/U	20.5	0.115
RG-59A/U	21.0	0.112
214-023	20.0	0.107
214-076	3.9	0.351

Sample Problems (Finding the Characteristic Impedance of a Cable)

EXAMPLE 1

An RG-11AU cable has a capacitance of 21.0 pF/ft and an inductance of 0.112 μH/ft. What is the characteristic impedance of the cable?



You are given the capacitance and inductance per unit length: *C′* = *C*/ft, *L′* = *L*/ft. Using $Z_0 = \sqrt{L/C}$ and substituting *C* and *L* into it, you get

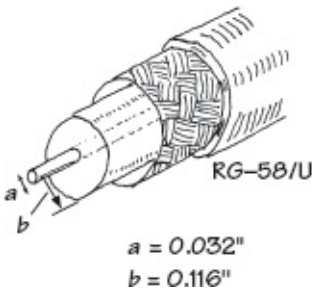
$$Z_0 = \sqrt{L/C}$$

$$Z_0 = \sqrt{\frac{0.112 \times 10^{-6}}{21.0 \times 10^{-12}}} = 73\ \Omega$$

FIGURE 3.13

EXAMPLE 2

What is the characteristic impedance of the RG-58/U coaxial cable with polyethylene dielectric (k = 2.3) shown below?



a = 0.032"
b = 0.116"

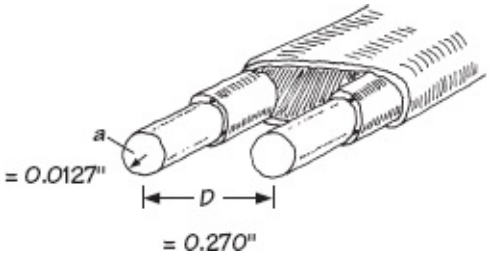
$$Z_0 = \frac{138}{\sqrt{k}} \log \frac{b}{a}$$

$$Z_0 = \frac{138}{\sqrt{2.3}} \log \left(\frac{0.116}{0.032} \right) = 91 \times 0.056 = 51\Omega$$

FIGURE 3.14

EXAMPLE 3

Find the characteristic impedance of the parallel-wire cable insulated with polyethylene (k = 2.3) shown below.



$$Z_0 = \frac{276}{\sqrt{k}} \log \frac{D}{a}$$

$$Z_0 = \frac{276}{\sqrt{2.3}} \log \frac{0.270}{0.0127} = 242\Omega$$

FIGURE 3.15

Impedance Matching

Since a transmission line has impedance built in, the natural question to ask is, How does the impedance affect signals that are relayed through a transmission line from one device to another? The answer to this question ultimately depends on the impedances of the devices to which the transmission line is attached. If the impedance of the transmission line is not the same as the impedance of, say, a load connected to it, the signals propagating through the line will only be partially absorbed by the load. The rest of the signal will be reflected back in the direction it came. Reflected signals are generally bad things in electronics. They represent an inefficient power transfer between two electrical devices. How do you get rid of the reflections? You apply a technique called *impedance matching*. The goal of impedance matching is to make the impedances of two devices—that are to be joined—equal. The impedance-matching techniques make use of special matching networks that are inserted between the devices.

Before looking at the specific methods used to match impedances, let's first take a look at an analogy that should shed some light on why unmatched impedances result in reflected signals and inefficient power transfers. In this analogy, pretend that the transmission line is a rope that has a density that is analogous to the transmission line's characteristic impedance Z_0. Pretend also that the load is a rope that has a density that is analogous to the load's impedance Z_L. The rest of the analogy is carried out below.

Unmatched Impedances ($Z_0 < Z_L$)

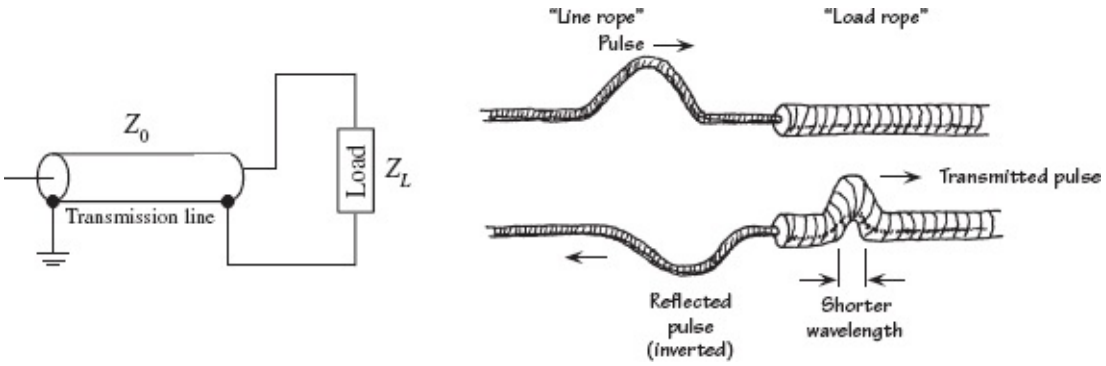


FIGURE 3.16a

A low-impedance transmission line that is connected to a high-impedance load is analogous to a low-density rope connected to a high-density rope. In the rope analogy, if you impart a pulse at the left end of the low-density rope (analogous to sending an electrical signal through a line to a load), the pulse will travel along without problems until it reaches the high-density rope (load). According to the laws of physics, when the wave reaches the high-density rope, it will do two things. First, it will induce a smaller-wavelength pulse within the high-density rope, and second, it will induce a similar-wavelength but inverted and diminished pulse that rebounds back toward the left end of the low-density rope. From the analogy, notice that only part of the signal energy from the low-density rope is transmitted to the high-density rope. From this analogy, you can infer that in the electrical case similar effects will occur—only now you are dealing with voltage and currents and transmission lines and loads.

Unmatched Impedances ($Z_0 > Z_L$)

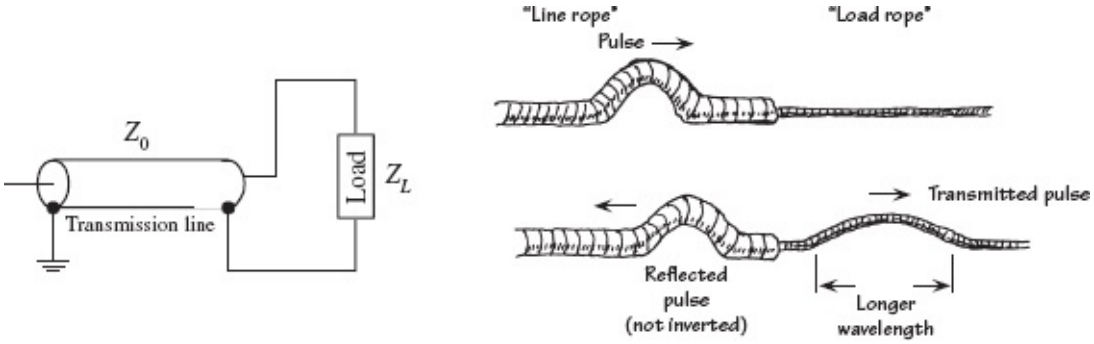


FIGURE 3.16b

A high-impedance transmission line that is connected to a low-impedance load is analogous to a high-density rope connected to a low-density rope. If you impart a pulse at the left end of the high-density rope (analogous to sending an electrical signal through a line to a load), the pulse will travel along the rope without problems until it reaches the low-density rope (load). At that time, the pulse will induce a longer-wavelength pulse within the low-density rope and will induce a similar-wavelength but inverted and diminished pulse that rebounds back toward the left end of the high-density rope. From this analogy, again you can see that only part of the signal energy from the high-density rope is transmitted to the low-density rope.

Matched Impedances ($Z_0 = Z_L$)

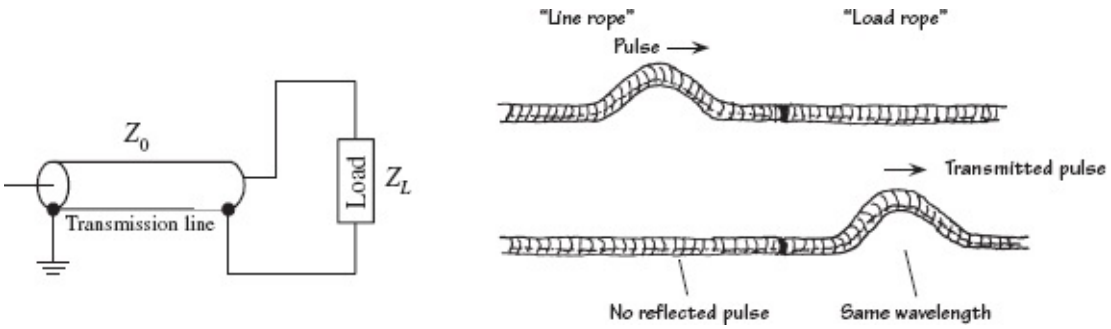


FIGURE 3.16c

Connecting a transmission line and load of equal impedances together is analogous to connecting two ropes of similar densities together. When you impart a pulse in the “transmission line” rope, the pulse will travel along without problems. However, unlike the first two analogies, when the pulse meets the load rope, it will continue on through the load rope. In the process, there will be no reflection, wavelength change, or amplitude change. From this analogy, you can infer that if the impedance of a transmission line matches the impedance of the load, power transfer will be smooth and efficient.

Standing Waves

Let’s now consider what happens to an improperly matched line and load when the signal source is producing a continuous series of sine waves. You can, of course, expect reflections as before, but you also will notice that a superimposed standing-wave pattern is created within the line. The standing-wave pattern results from the interaction of forward-going and reflected signals. Figure 3.17 shows a typical resulting standing-wave pattern for an improperly matched transmission line attached between a sinusoidal transmitter and a load. The standing-wave pattern is graphed in terms of amplitude (expressed in terms of V_{rms}) versus position along the transmission line.

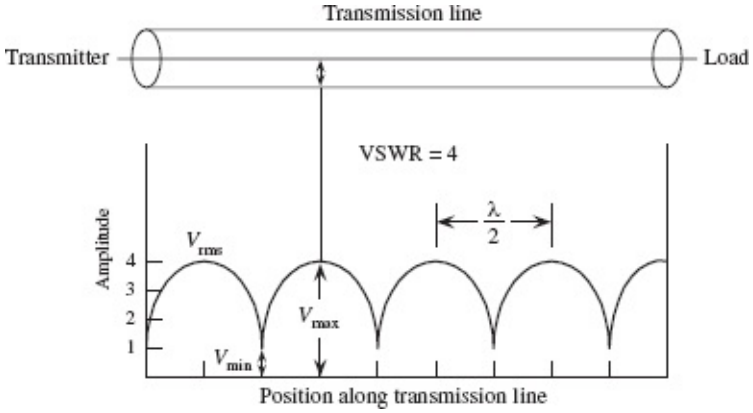


FIGURE 3.17 Standing waves on an improperly terminated transmission line. The VSWR is equal to V_{max}/V_{min}.

A term used to describe the standing-wave pattern is the *voltage standing-wave ratio* (VSWR). The VSWR is the ratio between the maximum and minimum rms voltages along a transmission line and is expressed as

$$VSWR = \frac{V_{rms,max}}{V_{rms,min}}$$

The standing-wave pattern shown in Fig. 3.17 has VSWR of 4/1, or 4.

Assuming that the standing waves are due entirely to a mismatch between load impedance and characteristic impedance of the line, the VSWR is simply given by either

$$VSWR = \frac{Z_0}{R_L} \quad \text{or} \quad VSWR = \frac{R_L}{Z_0}$$

whichever produces a result that is greater than 1.

A VSWR equal to 1 means that the line is properly terminated, and there will be no reflected waves. However, if the VSWR is large, this means that the line is not properly terminated (e.g., a line with little or no impedance attached to either a short or open circuit), and hence there will be major reflections.

The VSWR also can be expressed in terms of forward and reflected waves by the following formula:

$$VSWR = \frac{V_F + V_R}{V_F - V_R}$$

To make this expression meaningful, you can convert it into an expression in terms of forward and reflected power. In the conversion, you use $P = IV = V^2/R$. Taking P to be proportional to V^2, you can rewrite the VSWR in terms of forward and reflected power as follows:

$$VSWR = \frac{\sqrt{P_F} + \sqrt{P_R}}{\sqrt{P_F} - \sqrt{P_R}}$$

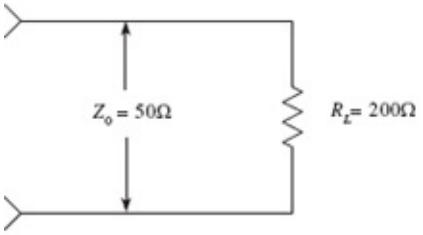
Rearranging this equation, you get the percentage of reflected power and percentage of absorbed power in terms of VSWR:

$$\%\ \text{reflected power} = \left[\frac{\text{VSWR} - 1}{\text{VSWR} + 1}\right]^2 \times 100\%$$

$$\%\ \text{absorbed power} = 100\% - \%\ \text{reflected power}$$

EXAMPLE (VSWR)

Find the standing-wave ratio (VSWR) of a 50-Ω line used to feed a 200-Ω load. Also find the percentage of power that is reflected at the load and the percentage of power absorbed by the load.



$$\text{VSWR} = \frac{Z_0}{R_L} = \frac{200}{50} = 4$$

VSWR is 4:1

$$\%\ \text{reflected power} = \frac{\text{VSWR} - 1^2}{\text{VSWR} + 1} \times 100\%$$

$$= \frac{4 - 1^2}{4 + 1} \times 100\% = 36\%$$

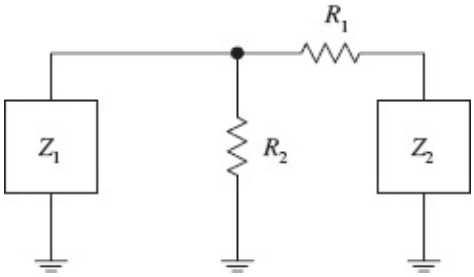
$$\%\ \text{absorbed power} = 100\% - \%\ \text{reflected power} = 64\%$$

FIGURE 3.18

Techniques for Matching Impedances

This section looks at a few impedance-matching techniques. As a rule of thumb, with most low-frequency applications where the signal's wavelength is much larger than the cable length, there is no need to match line impedances. Matching impedances is usually reserved for high-frequency applications. Moreover, most electrical equipment, such as oscilloscopes, video equipment, etc., has input and output impedances that match the characteristic impedances of coaxial cables (typically 50 Ω). Other devices, such as television antenna inputs, have characteristic input impedances that match the characteristic impedance of twin-lead cables (300 Ω). In such cases, the impedance matching is already taken care of.

IMPEDANCE-MATCHING NETWORK



A general method used to match impedance makes use of the impedance-matching network shown here. To match impedances, choose

$$R_1 = \sqrt{Z_2(Z_2 - Z_1)}$$

$$R_2 = Z_1\sqrt{Z_2/(Z_2 - Z_1)}$$

The attenuation seen from the Z_1 end will be $A_1 = R_1/Z_2 + 1$. The attenuation seen from the Z_2 end will be $A_2 = R_1/R_2 + R_1/Z_1 + 1$.

For example, if $Z_1 = 50$ W, and $Z_2 = 125$ W, then R_1, R_2, A_1, and A_2 are

$$R_1 = \sqrt{Z_2(Z_2 - Z_1)} = \sqrt{125(125 - 50)} = 97\Omega$$

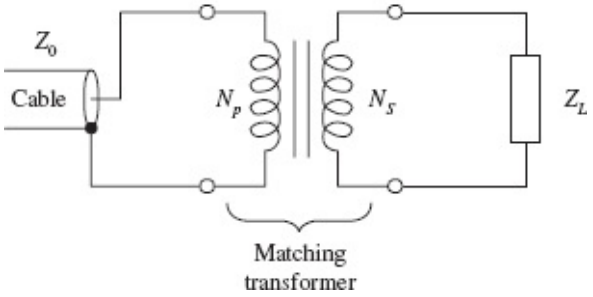
$$R_2 = Z_1\sqrt{\frac{Z_2}{Z_2 - Z_1}} = 50\sqrt{\frac{125}{125 - 50}} = 65\Omega$$

$$A_1 = \frac{R_1}{Z_2} + 1 = \frac{96.8}{125} + 1 = 1.77$$

$$A_2 = \frac{R_1}{R_2} + \frac{R_1}{Z_1} + 1 = \frac{96.8}{64.6} + \frac{96.8}{50} + 1 = 4.43$$

FIGURE 3.19

IMPEDANCE TRANSFORMER



Here, a transformer is used to match the characteristic impedance of a cable with the impedance of a load. By using the formula

$$N_p/N_s = \sqrt{Z_0/Z_L}$$

you can match impedances by choosing appropriate values for N_P and N_S so that the ratio N_P/N_S is equal to $\sqrt{Z_0/Z_L}$.

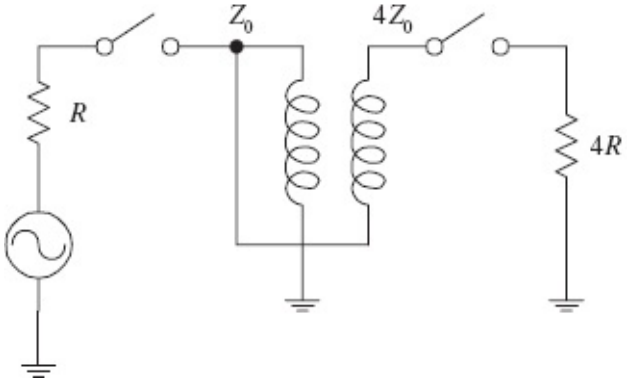
For example, if you wish to match an 800-Ω impedance line with an 8-Ω load, you first calculate

$$\sqrt{Z_0/Z_L} = \sqrt{800/8} = 10$$

To match impedances, you select N_P (number of coils in the primary) and N_S (number of coils in the secondary) in such a way that N_P/N_S = 10. One way of doing this would be to set N_P equal to 10 and N_S equal to 1. You also could choose N_P equal to 20 and N_S equal to 2 and you would get the same result.

FIGURE 3.20

BROADBAND TRANSMISSION-LINE TRANSFORMER



A broadband transmission-line transformer is a simple device that consists of a few turns of miniature coaxial cable or twisted-pair cable wound about a ferrite core. Unlike conventional transformers, this device can more readily handle high-frequency matching (its geometry eliminates capacitive and inductive resonance behavior). These devices can handle various impedance transformations and can do so with incredibly good broadband performance (less than 1 dB loss from 0.1 to 500 MHz).

FIGURE 3.21

QUARTER-WAVE SECTION

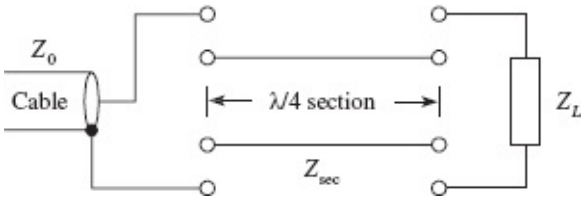


FIGURE 3.22

A transmission line with characteristic impedance Z_0 can be matched with a load with impedance Z_L by inserting a wire segment that has a length equal to one quarter of the transmitted signal's wavelength (λ/4) and which has an impedance equal to

$$Z_{sec} = \sqrt{Z_0 Z_L}$$

To calculate the segment's length, you must use the formula λ = *v*/*f*, where *v* is the velocity of propagation of a signal along the cable and *f* is the frequency of the signal. To find *v*, use

$$\upsilon = c/\sqrt{k}$$

where $c = 3.0 \times 10^8$ m/s, and *k* is the dielectric constant of the cable's insulation.

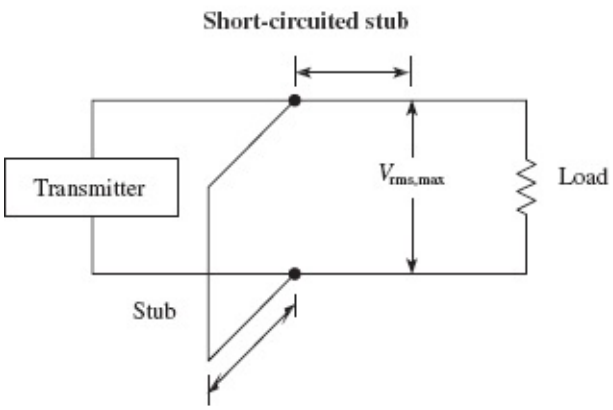
For example, say you wish to match a 50-Ω cable that has a dielectric constant of 1 with a 200-Ω load. If you assume the signal's frequency is 100 MHz, the wavelength then becomes

$$\lambda = \frac{\upsilon}{f} = \frac{c/\sqrt{k}}{f} = \frac{3\times10^8/1}{100\times10^6} = 3\text{ m}$$

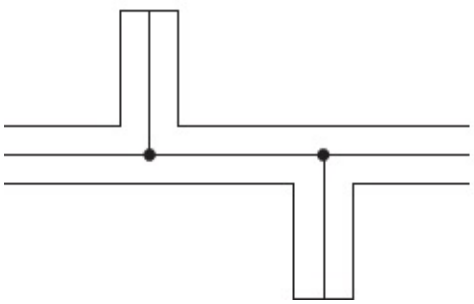
To find the segment length, you plug λ into λ/4. Hence the segment should be 0.75 m long. The wire segment also must have an impedance equal to

$$Z_{sec} = \sqrt{(50)(200)} = 100\ \Omega$$

STUBS



Pair of open-ended matching stubs



A short length of transmission line that is open ended or short-circuit terminated possesses the property of having an impedance that is reactive. By properly choosing a segment of open-circuit or short-circuit line and placing it in shunt with the original transmission line at an appropriate position along the line, standing waves can be eliminated. The short segment of wire is referred to as a *stub*. Stubs are made from the same type of cable found in the transmission line. Figuring out the length of a

stub and where it should be placed is fairly tricky. In practice, graphs and a few formulas are required. A detailed handbook on electronics is the best place to learn more about using stubs.

FIGURE 3.23

3.2 Batteries

A battery is made up of a number of *cells*. Each cell contains a positive terminal, or *cathode*, and a negative terminal, or *anode*. (Note that most other devices treat *anodes* as positive terminals and *cathodes* as negative terminals.)

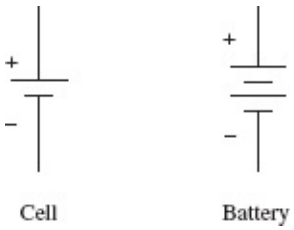


FIGURE 3.24

When a load is placed between a cell's terminals, a conductive bridge is formed that initiates chemical reactions within the cell. These reactions produce electrons in the anode material and remove electrons from the cathode material. As a result, a potential is created across the terminals of the cell, and electrons from the anode flow through the load (doing work in the process) and into the cathode.

A typical cell maintains about 1.5 V across its terminals and is capable of delivering a specific amount of current that depends on the size and chemical makeup of the cell. If more voltage or power is needed, a number of cells can be added together in either series or parallel configurations. By adding cells in series, a larger-voltage battery can be made, whereas adding cells in parallel results in a battery with a higher current-output capacity. Figure 3.25 shows a few cell arrangements.

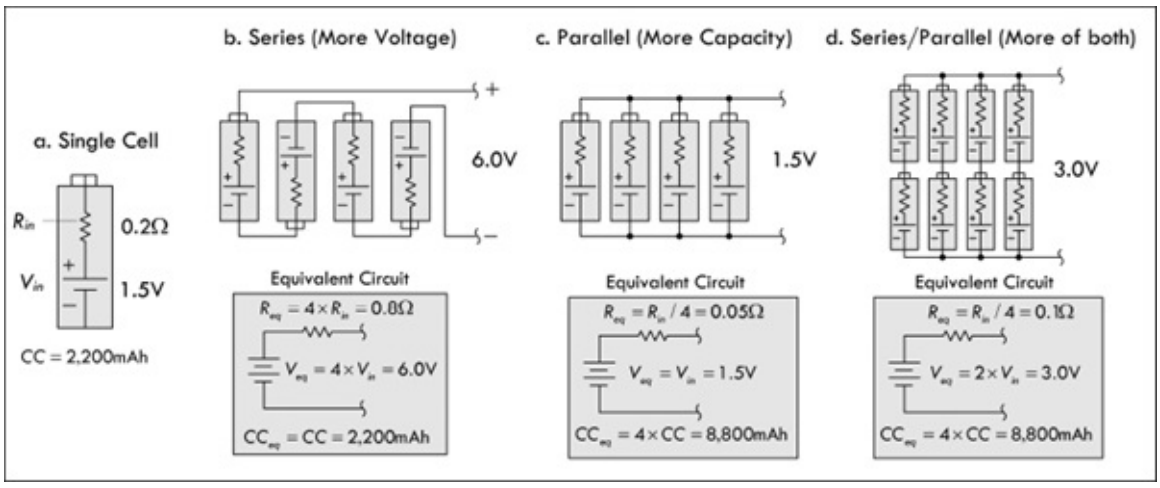


FIGURE 3.25

Battery cells are made from a number of different chemical ingredients. The use of a particular set of ingredients has practical consequences on the battery's overall performance. For example, some cells are designed to provide high open-circuit voltages, whereas others are designed to provide large current capacities. Certain kinds of cells are designed for light-current, intermittent applications, whereas others are designed for heavy-current, continuous-use applications. Some cells are designed for pulsing applications, where a large burst of current is needed for a short period of time. Some cells have good shelf lives; others have poor shelf lives. Batteries that are designed for one-time use, such as carbon-zinc

and alkaline batteries, are called *primary batteries*. Batteries that can be recharged a number of times, such as nickel metal hydride and lead-acid batteries, are referred to as *secondary batteries*.

3.2.1 How a Cell Works

A cell converts chemical energy into electrical energy by going through what are called *oxidation-reduction reactions* (reactions that involve the exchange of electrons). The three fundamental ingredients of a cell used to initiate these reactions include two chemically dissimilar metals (positive and negative electrodes) and an electrolyte (typically a liquid or pastelike material that contains freely floating ions). The following is a little lecture on how a simple lead-acid battery works.

For a lead-acid cell, one of the electrodes is made from pure lead (Pb); the other electrode is made from lead oxide (PbO_2); and the electrolyte is made from a sulfuric acid solution ($H_2O + H_2SO_4 \rightarrow 3H^+ + SO_4^{2-} + OH^-$).

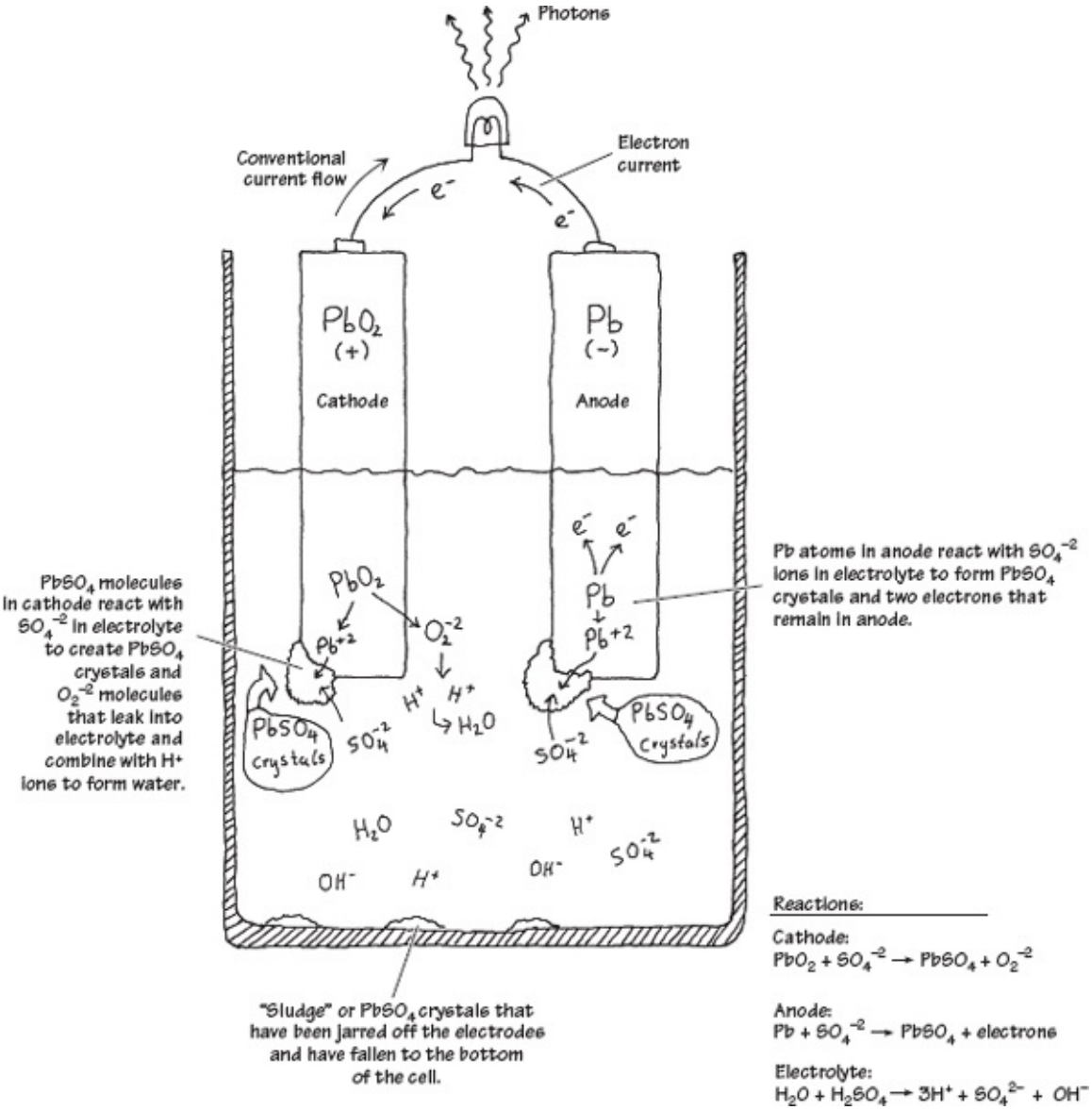


FIGURE 3.26

When the two chemically dissimilar electrodes are placed in the sulfuric acid solution, the electrodes react with the acid (SO_4^{-2}, H^+ ions), causing the pure lead electrode to slowly transform into $PbSO_4$ crystals. During this transformation reaction, two electrons are liberated within the lead electrode. Now, if you examine the lead oxide electrode, you also will see that it too is converted into $PbSO_4$ crystals.

However, instead of releasing electrons during its transformation, it releases O_2^{2-} ions. These ions leak out into the electrolyte solution and combine with the hydrogen ions to form H_2O (water). By placing a load element, say, a lightbulb, across the electrodes, electrons will flow from the electron-abundant lead electrode, through the bulb's filament, and into the electron-deficient lead oxide electrode.

As time passes, the ingredients for the chemical reactions run out (the battery is drained). To get energy back into the cell, a reverse voltage can be applied across the cell's terminals, thus forcing the reactions backward. In theory, a lead-acid battery can be drained and recharged indefinitely. However, over time, chunks of crystals will break off from the electrodes and fall to the bottom of the container, where they are not recoverable. Other problems arise from loss of electrolyte due to gasing during electrolysis (a result of overcharging) and due to evaporation.

3.2.2 Primary Batteries

Primary batteries are one-shot deals—once they are drained, it is all over. Common primary batteries include carbon-zinc batteries, alkaline batteries, mercury batteries, silver oxide batteries, zinc air batteries, and silver-zinc batteries. Here are some common battery packages:

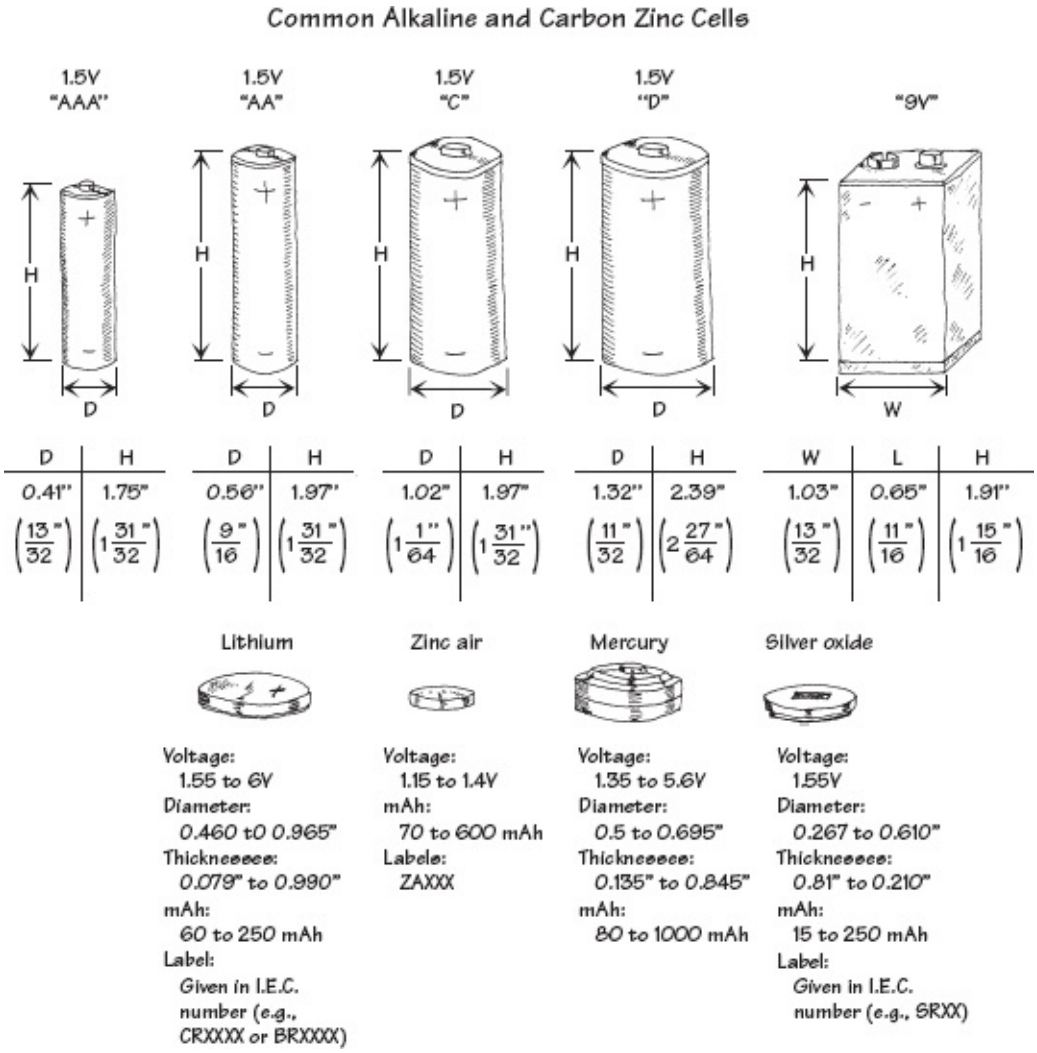


FIGURE 3.27

3.2.3 Comparing Primary Batteries

Carbon-Zinc Batteries

Carbon-zinc batteries ("standard-duty") are general-purpose primary-type batteries that were popular

back in the 1970s, but have become obsolete with the advent of alkaline batteries. These batteries are not suitable for continuous use (only for intermittent use) and are susceptible to leakage. The nominal voltage of a carbon-zinc cell is about 1.5 V, but this value gradually drops during service. Shelf life also tends to be poor, especially at elevated temperatures. The only really positive aspect of these batteries is their low cost and wide size range. They are best suited to low-power applications with intermittent use, such as in radios, toys, and general-purpose low-cost devices. Don't use these cells in expensive equipment or leave them in equipment for long periods of time—there is a good chance they will leak. Standby applications and applications that require a wide temperature range should also be avoided. In all, these batteries are to be avoided, if you can even find them.

Zinc-Chloride Batteries

Zinc-chloride batteries ("heavy-duty") are a heavy-duty version of the carbon-zinc battery, designed to deliver more current and provide about 50 percent more capacity. Like the carbon-zinc battery, zinc-chloride batteries are essentially obsolete as compared to alkaline batteries. The terminal voltage of a zinc-chloride cell is initially about 1.5 V, but drops as chemicals are consumed. Unlike carbon-zinc batteries, zinc-chloride batteries perform better at low temperatures, and slightly better at higher temperatures, too. The shelf life is also longer. They tend to have lower internal resistance and higher capacities than carbon-zinc, allowing higher currents to be drawn for longer periods. These batteries are suited to moderate, intermittent use. However, an alkaline battery will provide better performance in similar applications.

Alkaline Batteries

Alkaline batteries are the most common type of household battery you can buy—they have practically replaced the carbon-zinc and zinc-chloride batteries. They are relatively powerful and inexpensive. The nominal voltage of an alkaline cell is again 1.5 V, but doesn't drop as much during discharge as the previous two battery types. The internal resistance is also considerably lower, and remains so until near the end of the battery's life cycle. They have very long shelf lives and better high- and low-temperature performance, too. General-purpose alkaline batteries don't work particularly well on high-drain devices like digital cameras, since the internal resistance limits output current flow. They will still work in your device, but the battery life will be greatly reduced. They are well suited to most general-purpose applications, such as toys, flashlights, portable audio equipment, flashes, digital cameras, and so on. Note that there is a rechargeable version of an alkaline battery, as well.

Lithium Batteries

Lithium batteries use a lithium anode, one of a number of different kinds of cathodes, and an organic electrolyte. They have a nominal voltage of 3.0 V—twice that of most other primary cells—that remains almost flat during the discharge cycle. They also have a very low self-discharge rate, giving them an excellent shelf life—as much as 10 years. The internal resistance is also quite low, and remains so during discharge. It performs well in both low and high temperatures, and advanced versions of this battery are used on satellites, on space vehicles, and in military applications. They are ideal for low-drain applications such as smoke detectors, data-retention devices, pacemakers, watches, and calculators.

Lithium-Iron Disulfide Batteries

Unlike other lithium cells that have chemistries geared to obtaining the greatest capacity in a given package, lithium-iron disulfide cells are a compromise. To match existing equipment and circuits, their chemistry has been tailored to the standard nominal 1.5-V output (whereas other lithium technologies produce double that). These cells are consequently sometimes termed *voltage-compatible lithium* batteries. Unlike other lithium technologies, lithium-iron disulfide cells are not rechargeable. Internally,

the lithium-iron disulfide cell is a sandwich of a lithium anode, a separator, and an iron disulfide cathode with an aluminum cathode collector. The cells are sealed but vented. Compared to the alkaline cells—with which they are meant to compete—lithium-iron disulfide cells are lighter (weighing about 66 percent of same-size alkaline cells) and higher in capacity, and they have a much longer shelf life—even after 10 years on the shelf, lithium disulfide cells still retain most of their capacity. Lithium iron-disulfide cells operate best under heavier loads. In high-current applications, they can supply power for a duration exceeding 260 percent of the time that a similar-sized alkaline cell can supply. This advantage diminishes at lower loads, however, and at very light loads may disappear or even reverse. For example, under a 20-mA load, a certain manufacturer rates its AA-size lithium-disulfide cells to provide power for about 122 hours while its alkaline cells will last for 135 hours. However, under a heavy load of 1 A, the lithium disulfide cell overshadows the alkaline counterpart by lasting 2.1 hours versus only 0.8 hours for the alkaline battery.

Mercury Cells

Zinc-mercuric oxide, or “mercury,” cells take advantage of the high electrode potential of mercury to offer a very high energy density combined with a very flat discharge curve. Mercuric oxide forms the positive electrode, sometimes mixed with manganese dioxide. The nominal terminal voltage of a mercury cell is 1.35 V, and this remains almost constant over the life of the cell. They have an internal resistance that is fairly constant. Although made only in small button sizes, mercury cells are capable of reasonably high-pulsed discharge current and are thus suitable for applications such as quartz analog watches and hearing aids as well as voltage references in instruments, and the like.

Silver Oxide Batteries

The silver oxide battery is the predominate miniature battery found on the market today. Silver oxide cells are made only in small button sizes of modest capacity but have good pulsed discharge capability. They are typically used in watches, calculators, hearing aids, and electronic instruments. This battery’s general characteristics include higher voltage than comparable mercury batteries, flatter discharge curve than alkaline batteries, good low-temperature performance, good resistance to shock and vibration, essentially constant internal resistance, excellent service maintenance, and long shelf life—exceeding 90 percent charge after storage for five years. The nominal terminal voltage of a silver oxide cell is slightly over 1.5 V and remains almost flat over the life of the cell. Batteries built from cells range from 1.5 V to 6.0 V and come in a variety of sizes. Silver oxide hearing aid batteries are designed to produce greater volumetric energy density at higher discharge rates than silver oxide watch or photographic batteries. Silver oxide photo batteries are designed to provide constant voltage or periodic high-drain pulses with or without a low drain background current. Silver oxide watch batteries, using a sodium hydroxide (NaOH) electrolyte system are designed primarily for low-drain continuous use over long periods of time—typically five years. Silver oxide watch batteries using potassium hydroxide (KOH) electrolyte systems are designed primarily for continuous low drains with periodic high-drain pulse demands, over a period of about two years.

Zinc Air Batteries

Zinc air cells offer very high energy density and a flat discharge curve, but have relatively short working lives. The negative electrode is formed of powdered zinc, mixed with the potassium hydroxide electrolyte to form a paste. This is retained inside a small metal can by a separator membrane that is porous to ions, and on the other side of the membrane is simply air to provide the oxygen (which acts as the positive electrode). The air/oxygen is inside an outer can of nickel-plated steel that also forms the cell’s positive connection, lined with another membrane to distribute the oxygen over the largest area. Actually there is

no oxygen or air in the zinc-oxygen cell when it's made. Instead, the outer can has a small entry hole with a covering seal, which is removed to admit air and activate the cell. The zinc is consumed as the cell supplies energy, which is typically for around 60 days. The nominal terminal voltage of a zinc-oxygen cell is 1.45 V, and the discharge curve is relatively flat. The internal resistance is only moderately low, and they are not suitable for heavy or pulsed discharging. They are found mainly in button and pill packages, and are commonly used in hearing aids and pagers. Miniature zinc air batteries are designed primarily to provide power to hearing aids. In most hearing aid applications, zinc air batteries can be directly substituted for silver oxide or mercuric oxide batteries and will typically give the longest hearing aid service of any common battery system. Notable characteristics include high capacity-to-volume ratio for a miniature battery, more stable voltage at high currents when compared to mercury or silver oxide batteries, and essentially constant internal resistance. They are activated by removing the covering (adhesive tab) from the air access hole, and they are most effective in applications that consume battery capacity in a few weeks.

TABLE 3.5 Primary Battery Comparison Chart

TYPE (CHEMISTRY)	COMMON NAME(S)	NOMINAL CELL VOLTS	INTERNAL RESISTANCE	MAXIMUM DISCHARGE RATE	COST	PROS AND CONS	TYPICAL APPLICATIONS
Carbon-zinc	Standard-duty	1.5	Medium	Medium	Low	Low cost, various sizes, but terminal voltage drops steadily during cell life	Radios, toys, and general-purpose electrical equipment
Zinc-chloride	Heavy-duty	1.5	Low	Medium to high	Low to medium	Low cost at higher discharge rates and at lower temperature; terminal voltage still drops	Motor-driven portable devices, clocks, remote controls
Alkaline zinc-manganese dioxide	Alkaline	1.5	Very low	High	Medium to high	Better for high continuous or pulsed loads and at low temperatures, but terminal voltage drops	Photoflash units, battery shavers, digital cameras, handheld transceivers, portable CD players, etc.
Lithium-manganese dioxide	Lithium	1.5	Low	Medium to high	High	High energy density, very low self-discharge rate (excellent shelf-life), good temperature tolerance	Watches, calculators, cameras (digital and film), DMMs, and other test instruments
Zinc-mercuric oxide	Mercury cell	1.35	Low	Low	High	High energy density (compact), very flat discharge curve, good at higher temperatures	Calculators, pagers hearing aids, watches, test intruments
Zinc-silver oxide	Silver oxide cell	1.5	Low	Low	High	Very high energy density (very compact), very flat discharge curve, reasonable at lower temperatures	Calculators, pagers, hearing aids, watches test instruments
Zinc-oxygen	Zinc air cell	1.45	Medium	Low	Medium	High energy density, very lightweight, flat discharge curve, but must have access to air	Hearing aids and pagers

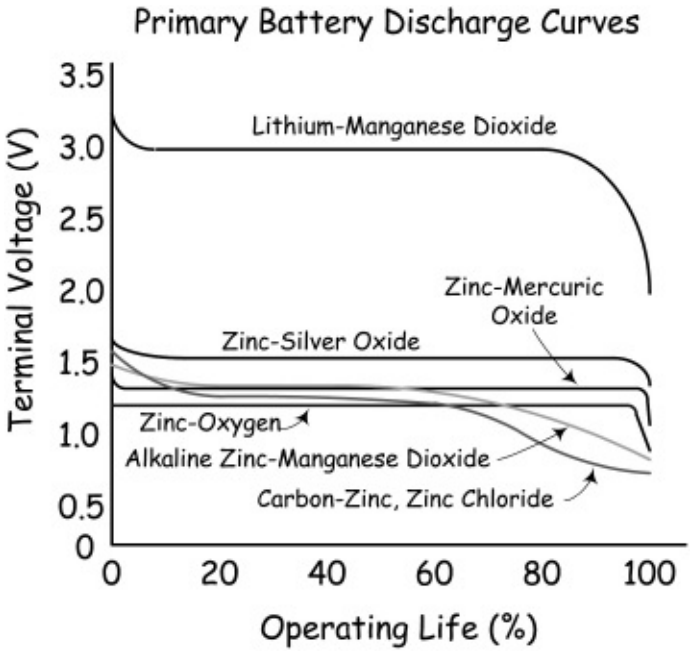


FIGURE 3.28

3.2.4 Secondary Batteries

Secondary batteries, unlike primary batteries, are rechargeable by nature. The actual discharge characteristics for secondary batteries are similar to those of primary batteries, but in terms of design, secondary batteries are made for long-term, high-power-level discharges, whereas primary batteries are designed for short discharges at low power levels. Most secondary batteries come in packages similar to those of primary batteries, with the exception of, say, lead-acid batteries and special-purpose batteries. Secondary batteries are used to power such devices as laptop computers, portable power tools, electric vehicles, emergency lighting systems, and engine starting systems.

Here are some common packages for secondary batteries:

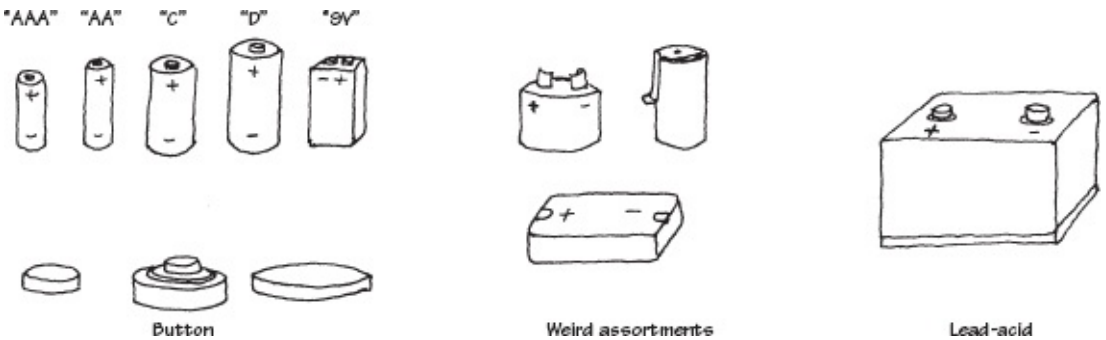


FIGURE 3.29

Comparing Secondary (Rechargeable) Batteries

LEAD-ACID BATTERIES

Lead-acid batteries are typically used for high-power applications, such as motorized vehicle power and battery backup applications. There are basically three types of lead-acid batteries: flooded lead-acid, valve-regulated lead-acid (VRLA), and sealed lead-acid (SLA). The flooded types must be stood upright and tend to lose electrolytes while producing gas over time. The SLA and VRLA are designed for a low overvoltage potential to prohibit the battery from reaching its gas-generating potential during discharge. However, SLA and VRLA can never be charged to their full potential. VRLA is generally used for stationary applications, while the SLA can be used in various positions. Lead-acid batteries typically come in 2-V, 4-V, 6-V, 8-V, and 12-V versions, with capacities ranging from 1 to several thousand amp-

hours. The flooded lead-acid battery is used in automobiles, forklifts, wheelchairs, and UPS devices.

An SLA battery uses a gel-type electrolyte rather than a liquid to allow it to be used in any position. However, to prevent gas generation, it must be operated at a lower potential—meaning it's never fully charged. This means that it has a relatively poor energy density—the lowest for all sealed secondary batteries. However, they're the cheapest secondary, making them best suited for applications where low-cost, stationary power storage is the main concern. SLA batteries have the lowest self-discharge rate of any of the secondary batteries (about 5 percent per month). They do not suffer from memory effect (as displayed in NiCad batteries), and they perform well with shallow cycling; in fact, they tend to prefer it to deep cycling, although they perform well with intermittent heavy current demands, too. SLA batteries aren't designed for fast charging—typically 8 to 16 hours for full recharge. They must also always be stored in a charged state. Leaving them in a discharged state can lead to sulfation, a condition that makes the batteries difficult, if not impossible, to recharge. Also, SLA batteries have an environmentally unfriendly electrolyte.

The basic technique for recharging lead-acid batteries, be they flooded, sealed, or valve-regulated, is to read the technical directions that come with them. If you don't know what you're doing—say, trying to make your own battery recharger—you may run into a serious problem, such as blowing up batteries with too much pressure, melting them, or destroying the chemistry. (The procedure for charging lead-acid batteries is different from that for NiCad and NiMH batteries in that voltage limiting is used instead of current limiting.)

NICKEL-CADMIUM (NiCAD) BATTERIES

Nickel-cadmium batteries are made using nickel hydroxide as the positive electrode and cadmium hydroxide as the negative electrode, with potassium hydroxide as the electrolyte. Nickel-cadmium batteries have been a very popular rechargeable battery over the years; however, with the introduction of NiMH batteries, they have seen a decline in use. In practical terms, NiCad batteries don't last very long before needing a recharge. They put out less voltage (per cell) than a standard alkaline (1.2 V versus 1.5 V for alkaline). This means that applications that require four or more alkaline batteries might not work at all with comparable-sized NiCad batteries. During discharge, the average voltage of a sealed NiCad cell is about 1.2 V per cell. At nominal discharge rates, the characteristic is very nearly flat until the cell approaches full discharge. The battery provides most of its energy above 1.0 V per cell. The self-discharge rate of a NiCad is not great, either—around two to three months. However, like SLAs, sealed NiCads can be used in virtually any position. NiCads have a higher energy density than SLAs (about twice as much), and with a relatively low cost, they are popular for powering compact portable equipment: cordless power tools, model boats and cars, and appliances such as flashlights and vacuum cleaners. NiCads suffer from memory effect and are therefore not really suitable for applications that involve shallow cycling or spending most of their time on a float charger. They perform best in situations where they're deeply cycled. They have a high number of charge/discharge cycles—around 1000.

Use a recommended charger—a constant current–type charger with due regard for heat dissipation and wattage rating. Improper charging can cause heat damage or even high-pressure rupture. Observe proper charging polarity. The safe charge rate for sealed NiCad cells for extended charge periods is 10 hours, or C/10 rate.

NICKEL METAL HYDRIDE (NiMH) BATTERIES

NiMH batteries are very popular secondary batteries, replacing NiCad batteries in many applications. NiMH batteries use a nickel/nickel hydroxide positive electrode, a hydrogen-storage alloy (such as lanthanum-nickel or zirconium-nickel) as a negative electrode, and potassium hydroxide as the electrolyte. They have a higher energy density than standard NiCad batteries (about 30 to 40 percent

higher) and don't require special disposal requirements, either. The nominal voltage of a NiMH battery is 1.2 V per cell, which must be taken into consideration when substituting them into devices that use standard 1.5-V cells such as alkaline cells. They self-discharge in about two to three months and do display some memory effect, but not as bad as NiCad batteries. They are not as happy with a deep discharge cycle as a NiCad battery, and they tend to have a shorter work life. Best results are achieved with load currents of 0.2-C to 0.5-C (one-fifth to one-half of the rated capacity). Typical applications include remote-control vehicles and power tools (although NiMH batteries are rapidly being superceded by Li-ion and LiPo batteries).

Recharging NiMH batteries is a bit complex due to significant heat generation; the charge uses a special algorithm that requires trickle charging and temperature sensing. Unlike NiCad batteries, NiMH batteries have little memory effect. The batteries require regular full discharge to prevent crystalline formation.

Li-ION BATTERIES

Lithium is the lightest of all metals and has the highest electrochemical potential, giving it the possibility of an extremely high energy density. However, the metal itself is highly reactive. While this isn't a problem with primary cells, it poses an explosion risk with rechargeable batteries. For these to be made safe, lithium-ion technology had to be developed; the technology uses lithium ions from chemicals such as lithium-cobalt dioxide, instead of the metal itself. Typical Li-ion batteries have a negative electrode of aluminum coated with a lithium compound such as lithium-cobalt dioxide, lithium-nickel dioxide, or lithium-manganese dioxide. The positive electrode is generally of copper, coated with carbon (generally either graphite or coke), while the electrolyte is a lithium salt such as lithium-phosphorous hexafluoride, dissolved in an organic solvent such as a mixture of ethylene carbonate and dimethyl carbonate. Li-ion batteries have roughly twice the energy density of NiCads, making them the most compact rechargeable yet in terms of energy storage. Unlike NiCad or NiMH batteries, they are not subject to memory effect, and have a relatively low self-discharge rate—about 6 percent per month, less than half that of NiCads. They are also capable of moderately deep discharging, although not as deep as NiCads, as they have a higher internal resistance. On the other hand, Li-ion batteries cannot be charged as rapidly as NiCads, and they cannot be trickle or float charged, either. They also are significantly more costly than either NiCads or NiMH batteries, making them the most expensive rechargeables of all. Part of this is that they must be provided with built-in protection against both excessive discharging and overcharging (both of which pose a safety risk). Most Li-ion batteries are therefore supplied in self-contained battery packs, complete with "smart" protective circuitry. They are subject to aging, even if not used, and have moderate discharge currents. The main applications for Li-ion batteries are in places where as much energy as possible needs to be stored in the smallest possible space, and with as little weight as possible. They are found in laptop computers, PDAs, camcorders, and cell phones.

Li-ion batteries require special voltage-limiting recharging devices. Commercial Li-ion battery packs contain a protection circuit that prevents the cell voltage from going too high while charging. The typical safety threshold is set to 4.30 V/cell. In addition, temperature sensing disconnects the charging device if the internal temperature approaches 90°C (194°F). Most cells feature a mechanical pressure switch that permanently interrupts the current path if a safe pressure threshold is exceeded. The charge time of all Li-ion batteries, when charged at a 1-C initial current, is about three hours. The battery remains cool during charge. Full charge is attained after the voltage has reached the upper voltage threshold and the current has dropped and leveled off at about 3 percent of the nominal charge current. Increasing the charge current on a Li-ion charger does not shorten the charge time by much. Although the voltage peak is reached more quickly with higher current, the topping charge will take longer.

LITHIUM POLYMER (Li-POLYMER) BATTERIES

The lithium polymer batteries are a lower-cost version of the Li-ion batteries. Their chemistry is similar to that of the Li-ion in terms of energy density, but uses a dry solid polymer electrolyte only. This electrolyte resembles a plastic-like film that does not conduct electricity but allows an exchange of ions (electrically charged atoms or groups of atoms). The dry polymer is more cost effective during fabrication, and the overall design is rugged, safe, and thin. With a cell thickness measuring as little as 1 mm, it is possible to use this battery in thin compact devices where space is an issue. It is possible to create designs which form part of a protective housing, are in the shape of a mat that can be rolled up, or are even embedded into a carrying case or piece of clothing. Such innovative batteries are still a few years away, especially for the commercial market.

Unfortunately, the dry Li-polymer suffers from poor ion conductivity, due to high internal resistance; it cannot deliver the current bursts needed for modern communication devices. However, it tends to increase in conductivity as the temperature rises, a characteristic suitable for hot climates. To make a small Li-polymer battery more conductive, some gelled electrolyte may be added. Most of the commercial Li-polymer batteries used today for mobile phones are hybrids and contain gelled electrolytes.

The charge process of a Li-polymer battery is similar to that of the Li-ion battery. The typical charge time is around one to three hours. Li-polymer batteries with gelled electrolyte, on the other hand, are almost identical to Li-ion batteries. In fact, the same charge algorithm can be applied.

NICKEL-ZINC (NiZn) BATTERIES

Nickel-zinc batteries are commonly used in light electric vehicles. They are considered the next generation of batteries used for high-drain applications, and are expected to replace sealed lead-acid batteries due to their higher energy densities (up to 70 percent lighter for the same power). They are also relatively cheap compared to NiCad batteries.

NiZn batteries are chemically very similar to NiCad batteries; both use an alkaline electrolyte and a nickel electrode, but they differ significantly in their voltage. The NiZn cell delivers more than 0.4 V of additional voltage both at open circuit and under load. With the additional 0.4 V per cell, multicell batteries can be constructed in smaller packages. For example, a 19.2-V pack can replace a 14.4-V NiCad pack, representing a 25 percent lower cell space and delivering higher power and a 45 percent lower impedance. They are also less expensive than most rechargeables. They are safe (abuse-tolerant). The life cycle is a bit better than for NiCad batteries for typical applications. They have superior shelf life when compared to lead-acid. Also, they are considered environmentally green—both nickel and zinc are nontoxic and easily recycled.

In terms of recharge times, it takes less than two hours to achieve full recharge; there is an 80 percent charge in one hour. This feature makes them useful in cordless power tools. Their high energy density and high discharge rate make them suitable for applications that demand large amounts of power in small, lightweight packages. They are found in cordless power tools, UPS systems, electric scooters, high-intensity dc lighting and the like.

NICKEL-IRON (NiFe) BATTERIES

Nickel-iron batteries, also called nickel alkaline or NiFe batteries, were introduced in 1900 by Thomas Edison. These are very robust batteries that are tolerant of abuse and can have very long life spans (30 years or more). The open-circuit voltage of these cells is 1.4 V, and the discharge voltage is about 1.2 V. They withstand overcharge and over-discharge. They accept high depth of discharge (deep cycling) and can remain discharged for long periods without damage, unlike lead-acid batteries that need to be stored in a charged state. They are, however, very heavy and bulky. Also, the low reactivity of the active components limits high-discharge performance. The cells take a charge slowly, give it up slowly, and

have a steep voltage dropoff with state of charge. Furthermore, they have a low energy density compared to other secondary batteries, and a high self-discharge rate. NiFe batteries are used in applications similar to those for lead-acid batteries, but oriented toward a necessity of longevity. (A typical lead-acid battery will last around five years, compared to around 30 to 80 years for a NiFe battery.)

TABLE 3.6 Rechargeable Battery Comparison Chart

TYPE (CHEMISTRY)	NOMINAL CELL VOLTS (APPROX.)	ENERGY DENSITY (WH/KG)	CYCLE LIFE	CHARGE TIME	MAX. DISCHARGE RATE	COST	PROS AND CONS	TYPICAL APPLICATIONS
Sealed lead-acid	2.0	Low (30)	Long (shallow cycles)	8–16 h	Medium (0.2 C)	Low	Low cost, low self-discharge, happy float charging, but prefers shallow charging	Emergency lighting, alarm systems solar power systems, wheelchairs, etc.
Rechargeable alkaline-manganese	1.5	High (75 initial)	Short to medium	2–6 h (pulsed)	Medium (0.3 C)	Low	Low cost, low self-discharge, prefer shallow cycling, no memory effect but short cycle life	Portable emergency lighting, toys, portable radios, CD players, test instruments, etc.
NiCad	1.2	Medium (40–60)	Long (deep cycles)	14–16 h (0.1 C) or <2 h with care (1 C)	High (>2 C)	Medium	Prefer deep cycling, good pulse capacity, but have memory effect, fairly high self-discharge rate, environmentally unfriendly	Portable tools and appliances, model cars and boats, data loggers, camcorders, portable transceivers, and test equipment
NiMH	1.2	High (60–80)	Medium	2–4 h	Medium (0.2–0.5 C)	Medium	Very compact energy source, but have some memory effect, high self-discharge rate	Remote control vehicles, cordless mobile phones, personal DVD and CD players, power tools
NiZn	1.65	High (>170)	Medium to high	1–2 h	—	Medium	Low cost, environmentally green, twice energy density of NiCad	Exceptional performance, no memory, long shelf-life
NiFe	1.4	High (>200)	Extremely long	Long	—	Low	High cycle life, incredibly long life up to 80 years, environmentally friendly	Forklifts and other, similar SLA-like applications, but where longevity is important
Li-ion/LiPo	3.6	Very high (>100)	Medium	3–4 h (1 C–0.03 C)	Med/high (<1 C)	High	Very compact, low maintenance, low self-discharge, but needs great care with charging	Compact cell phones and notebook PCs, digital cameras, and similar very small portable device

RECHARGEABLE ALKALINE-MANGANESE (RAM) BATTERIES

Rechargeable alkaline-manganese, or RAM, batteries are the rechargeable version of primary alkaline batteries. Like the primary technology, they use a manganese dioxide positive electrode and potassium hydroxide electrode, but the negative electrode is now a special porous zinc gel designed to absorb hydrogen during the charging process. The separator is also laminated to prevent it being pierced by zinc dendrites. These are often considered a poor substitute for a rechargeable, as compared to a NiCad or NiMH battery. RAM batteries have a tendency to plummet in capacity over few recharge cycles. It is feasible for a RAM battery to lose 50 percent of its capacity after only eight cycles. On the positive side, they are inexpensive and readily available. They can usually be used as a direct replacement for non-rechargeable batteries, but they usually have a lower nominal voltage, making them unsuitable for some devices, except in high-drain devices like digital cameras. They have a low self-discharge rate and can be stored on standby for up to 10 years. Also, they are environmentally friendly (no toxic metals are used) and maintenance-free; there is no need for cycling or worrying about memory effect. On the short side, they have limited current-handling capability and are limited to light-duty applications such as flashlights and other low-cost portable electronic devices that require shallow cycling. Recharging a RAM battery requires a special recharger; if you charge them in a standard charger, they may explode.

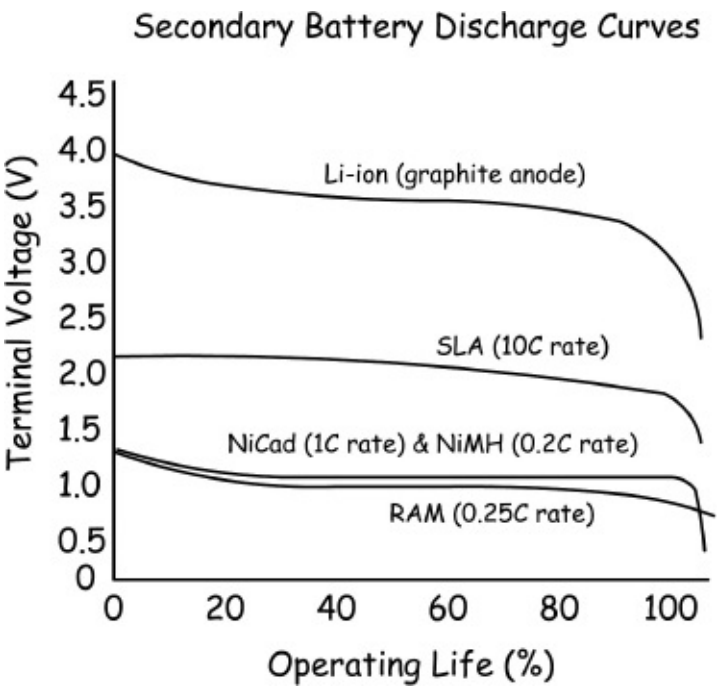


FIGURE 3.30

THE SUPERCAPACITOR

The supercapacitor isn't really a battery but a cross between a capacitor and a battery. It resembles a regular capacitor, but uses special electrodes and some electrolytes. There are three kinds of electrode material found in a supercapacitor: high-surface-area-activated carbons, metal oxide, and conducting polymers. The one using high-surface-area-activated carbons is the most economical to manufacture. This system is also called double layer capacitor (DLC) because the energy is stored in the double layer formed near the carbon electrode surface. The electrolyte may be aqueous or organic. The aqueous electrolyte offers low internal resistance but limits the voltage to 1 V. In contrast, the organic electrolyte allows 2 and 3 V of charge, but the internal resistance is higher.

To make the supercapacitor practical for use in electronic circuits, higher voltages are needed. Connecting the cells in series accomplishes this task. If more than three or four capacitors are connected in series, voltage balancing must be used to prevent any cell from reaching overvoltage.

Supercapacitors have values from 0.22 F upwards to several F. They have higher energy storage capacity than electrolytic capacitors, but a lower capacity than a battery (approximately $^{1}/_{10}$ that of a NiMH battery). Unlike electrochemical batteries that deliver a fairly steady voltage, the voltage of a supercapacitor drops from full voltage to zero volts without the customary flat voltage curve characteristic of most batteries. For this reason, supercapacitors are unable to deliver the full charge. The percentage of charge that is available depends on the voltage requirements of the applications. For example, a 6-V battery is allowed to discharge to 4.5 V before the equipment cuts off; the supercapacitor reaches that threshold with the first quarter of the discharge. The remaining energy slips into an unusable voltage range.

The self-discharge of the supercapacitor is substantially higher than that of the electrochemical battery. Typically, the voltage of the supercapacitor with an organic electrolyte drops from full charge to the 30 percent level in as little as 10 hours. Other supercapacitors can retain the charged energy longer. With these designs, the capacity drops from full charge to 85 percent in 10 days. In 30 days, the voltage drops to roughly 65 percent, and to 40 percent after 60 days.

The most common supercapacitor applications are memory backup and standby power for real-time clock ICs. Only in special applications can the supercapacitor be used as a direct replacement for a chemical battery. Often the supercapacitor is used in tandem with a battery (placed across its terminals, with a provision in place to limit high influx of current when equipment is turned on) to improve the

current handling of the battery: during low load current the battery charges the supercapacitor; the stored energy of the supercapacitor kicks in when a high load current is requested. In this way the supercapacitor acts to filter and smooth pulsed load currents. This enhances the battery's performance, prolongs the runtime, and even extends the longevity of the battery.

Limitations include an inability to use the full energy spectrum—depending on the application, not all energy is available. A supercapacitor has low energy density, typically holding 1/5 to 1/10 the energy of an electrochemical battery. Cells have low voltages—serial connections are needed to obtain higher voltages. Voltage balancing is required if more than three capacitors are connected in series. Furthermore, the self-discharge is considerably higher than that of an electrochemical battery.

Advantages include a virtually unlimited life cycle—supercapacitors are not subject to the wear and aging experienced by electrochemical batteries. Also, low impedance can enhance pulsed current demands on a battery when placed in parallel with the battery. Supercapacitors experience rapid charging—with low-impedance versions reaching full charge within seconds. The charge method is simple—the voltage-limiting circuit compensates for self-discharge.

Selecting the Right Battery (Comparison Chart)

● Yes, or common
● Yes, but not the best choice
○ Borderline
No

	Carbon Zinc	Zinc Chloride	Alkaline	Lithium	Zinc air	Silver Oxide	Mercury	RAM	Lead Acid (SLA)	NiCad	NiMH	Lithium Ion	Lithium Polymer	NiZn	Supercapacitor
Characteristics of single cell															
Obsolete (not recommended)	●	●						●							
Rechargeable								●	●	●	●	●	●	●	●
Stable voltage				●	●	●	●				●	●	●	●	
High energy density (Wh/kg)		○	●	●	●	●	●	●			○	●	●	●	
High capacity rating (mAh)														●	
High peak load-current rating	○	●	●	●				●	●	●	●	○	○	●	d
High pulsed discharge current		●	●	●		●	●		●	●		●	●	●	●
Low self-discharge rate			●	●				●	●	○		●	●	●	
Good at high temperatures		○	●	●			●								
Good at low temperatures		●	●	●		●									
High cycle life									●	●	●	●	●	●	●
Miniature				●	●	●	●							●	
Memory effect										●	○				
Expensive (cost)				●	●	●	●			●	●	●	●	○	
Environmentally unfriendly									●	●					
Applications															
Very small portable devices				●	●	●	●					●	●		
Pagers, hearing aids (h), watches (w)				●	h	w	●								
Radios, toys, general purpose	●	●	●					●							
Small motorized items		●	●					○			●				
Camcorders, digital cameras, test equipment			●							●	●				
Remotes, clocks, calculators		●	●	●				●							
Cellphones, mobile phones, laptops			○							○	●	●	●		
Low-self discharge (smoke detectors, data loggers)				●			●								
Power tools, model cars, electric toothbrush, etc.										●	●			●	
Motorized vehicles (electric bike, scooters, trolling motors, power tools, mower)									●	●	●			●	
High-power battery backup (emergency lighting, solar power storage, UPS)									●						
Power backup for short-term primary power outages (e.g. CMOS RAM, motor start, etc.)			●	●											●
Standard cell and battery packages															
AAA	●	●	●	●				●		●	●				
AA	●	●	●	●				●		●	●				
C	●	●	●	●				●		●	●				
D	●	●	●	●				●		●	●				
9V	●	●	●	●						●	●				
6V Lantern	●	●	●												
Button (coin)			●	●	●	●	●								●
Special battery pack			●	●					●	●	●	●	●	●	
Plastic Box									●					●	
PCB mount				●											●

FIGURE 3.31

3.2.5 Battery Capacity

Batteries are given a capacity rating that indicates how much electrical energy they are capable of delivering over a period of time. The capacity rating is specified in terms of ampere-hours (Ah) and millampere-hours (mAh). Knowing the battery capacity, it is possible to estimate how long the battery will last before being considered dead. The following example illustrates this.

Example: A battery with a capacity of 1800 mAh is to be used in a device that draws 120 mA continuously. Ignoring possible loss in capacity as a result of load current magnitude, how long should the battery be able to deliver power?

Answer: Ideally, this would be:

$$t = \frac{1800 \text{ mAh}}{120 \text{ mA}} = 15 \text{ h}$$

Note: In reality, you must consult the battery manufacturer's data sheets and analyze their discharge graphs (voltage as a function of time and of load current) to get an accurate determination of actual discharge time. As the load current increases, there is an apparent loss in battery capacity caused by internal resistance.

Typical capacity ratings for AAA, AA, C, D, and 9-V NiMH batteries are 1000 mAh (AAA), 2300 mAh (AA), 5000 mAh (C), 8500 mAh (D), 250 mAh (9).

C Rating

The charge and discharge currents of a battery are measured in capacity rating or C rating. The capacity represents the efficiency of a battery to store energy and its ability to transfer this energy to a load. Most portable batteries, with the exception of lead-acid, are rated at 1 C. A discharge rate of 1 C draws a current equal to the rated capacity that takes one hour (h). For example, a battery rated at 1000 mAh provides 1000 mA for 1 hour if discharged at 1 C rate. The same discharge at 0.5 C provides 500 mA for 2 hours. At 2 C, the same battery delivers 2000 mA for 30 minutes. 1 C is often referred to as a 1-hour discharge; 0.5 C would be 2 hours, and 0.1C would be a 10-hour discharge. The discrepancy in C rates between different batteries is largely dependent on the internal resistance.

Example: Determine the discharge time and average current output of a battery with a capacity rating of 1000 mAh if it is discharged at 1 C. How long would it take to discharge at 5 C, 2 C, 0.5 C, 0.2 C, and 0.05 C?

Answer: At 1 C, the battery is attached to a load drawing 1000 mA (rated capacity/hour), so the discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/1 C} = 1 \text{ h}$$

At 5 C, the battery is attached to a load drawing 5000 mA (five times rated capacity/hour), so the discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/5 C} = 0.2 \text{ h}$$

At 2 C, the battery is attached to a load drawing 2000 mA (two times rated capacity/hour), so the

discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/2 C} = 0.5 \text{ h}$$

At 0.5 C, the battery is attached to a load drawing 500 mA (half the rated capacity/hour), so the discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/0.5 C} = 2 \text{ h}$$

At 0.2 C, the battery is attached to a load drawing 200 mA (20 percent rated capacity/hour) so the discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/0.2 C} = 5 \text{ h}$$

At 0.05 C, the battery is attached to a load drawing 50 mA (5 percent rated capacity/hour) so the discharge time is:

$$t = 1 \text{ hC/C rating} = 1 \text{ hC/0.05 C} = 20 \text{ h}$$

Again, note that these values are estimates. When load currents increase (especially when C values get large), the capacity level drops below nominal values—due to nonideal internal characteristics such as internal resistance—and must be determined using manufacturer's discharge curves and Peurkert's equation. Do a search of the Internet, using "Peurkert's equation" as a keyword, to learn more.

3.2.6 Note on Internal Voltage Drop of a Battery

Batteries have an internal resistance that is a result of the imperfect conducting elements that make up the battery (resistance in electrodes and electrolytes). Though the internal resistance may appear low (around 0.1 Ω for an AA alkaline battery, or 1 to 2 Ω for a 9-V alkaline battery), it can cause a noticeable drop in output voltage if a low-resistance (high-current) load is attached to it. Without a load, we can measure the open-circuit voltage of a battery, as shown in Fig. 3.32a. This voltage is essentially equal to the battery's rated nominal voltage—the voltmeter has such a high input resistance that it draws practically no current, so there is no appreciable voltage drop. However, if we attach a load to the battery, as shown in Fig. 3.32, the output terminal voltage of the battery drops. By treating the internal resistance R_{in} and the load resistance R_{load} as a voltage divider, you can calculate the true output voltage present across the load—see the equation in Fig. 3.32b.

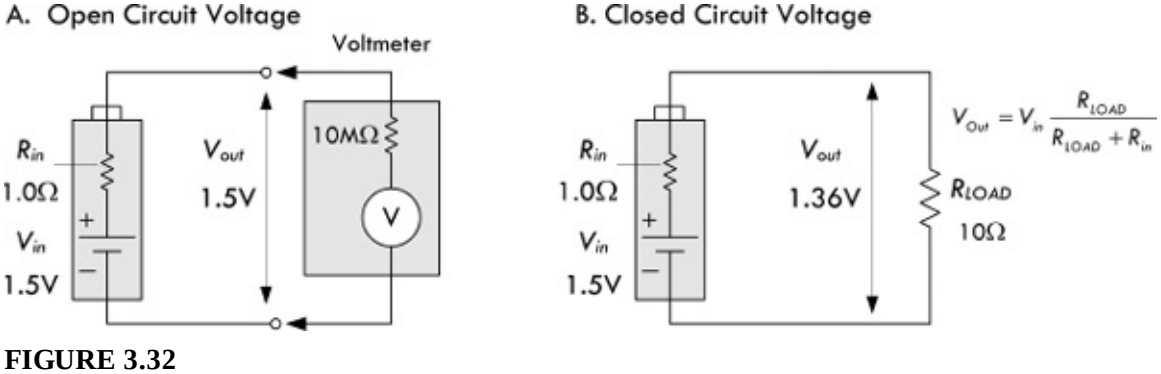


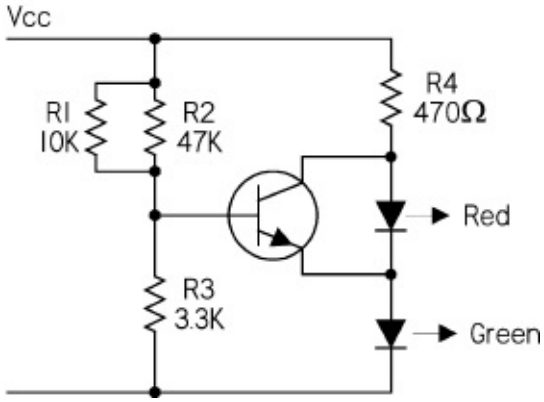
FIGURE 3.32

Batteries with large internal resistances show poor performance in supplying high current pulses.

(Consult the battery comparison section and tables to determine which batteries are best suited for high-current, high-pulse applications.) Internal resistance also increases as the battery discharges. For example, a typical alkaline AA battery may start out with an internal resistance of 0.15 Ω when fresh, but may increase to 0.75 Ω when 90 percent discharged. The following list shows typical internal resistance for various batteries found in catalogs. The values listed should not be assumed to be universal—you must check the specs for your particular batteries.

9-V zinc carbon	35 Ω
9-V lithium	16 to 18 Ω
9-V alkaline	1 to 2 Ω
AA alkaline	0.15 Ω (0.30 Ω at 50 percent discharge)
AA NiMH	0.02 Ω (0.04 Ω at 50 percent discharge)
D Alkaline	0.1 Ω
D NiCad	0.009 Ω
D SLA	0.006 Ω
AC13 zinc air	5 Ω
76 silver	10 Ω
675 mercury	10 Ω

Battery OK / LOW Indicator



Here a green LED is used to indicate that the battery is okay. This stays on all the time to indicate that the battery is live, and the red LED comes on when the battery voltage falls below the set threshold. A green LED has around 2.0 V on when it is illuminated. This value varies a bit with different manufacturers, but is pretty well matched within any batch. Add the base emitter voltage, and you need 2.6 V on the base of the right transistor (i.e., across the 3k3) to turn on the transistor. 2.6 V across 3k3 needs 9.1 across the supply rail. Below this threshold voltage, the transistor is off and the red LED is on. Above this voltage, the red LED is off. By adjusting the values of the three resistors, you can alter the threshold level. We'll discuss transistors and LEDs later on in this book.

FIGURE 3.33

3.3 Switches

A *switch* is a mechanical device that interrupts or diverts electric current flow within a circuit.

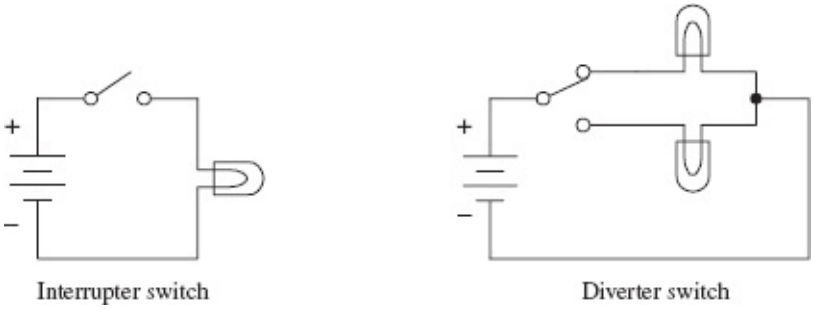
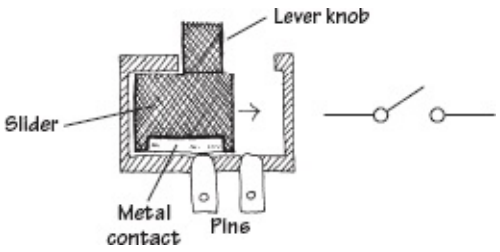


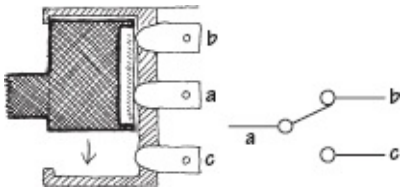
FIGURE 3.34

3.3.1 How a Switch Works

Two slider-type switches are shown in Fig. 3.35. The switch in Fig. 3.35a acts as an interrupter, whereas the switch in Fig. 3.35b acts as a diverter.



(*a*) When the lever is pushed to the right, the metal strip bridges the gap between the two contacts of the switch, thus allowing current to flow. When the lever is pushed to the left, the bridge is broken, and current will not flow.



(*b*) When the lever is pushed upward, a conductive bridge is made between contacts *a* and *b*. When the lever is pushed downward, the conductive bridge is relocated to a position where current can flow between contact *a* and *c*.

FIGURE 3.35

Other kinds of switches, such as push-button switches, rocker switches, magnetic-reed switches, etc., work a bit differently than slider switches. For example, a magnetic-reed switch uses two thin pieces of leaflike metal contacts that can be forced together by a magnetic field. This switch, as well as a number of other unique switches, will be discussed later on in this section.

3.3.2 Describing a Switch

A switch is characterized by its number of *poles* and by its number of *throws*. *A* pole represents, say, contact a in Fig. 3.35b. A throw, on the other hand, represents the particular contact-to-contact connection, say, the connection between contact a and contact *b* or the connection between contact *a* and contact *c* in Fig. 3.35b. In terms of describing a switch, the following format is used: (number of poles) "*P*" and (number of throws) "*T*." The letter *P* symbolizes "pole," and the letter T symbolizes "throw." When specifying the number of poles and the number of throws, a convention must be followed: When the number of poles or number of throws equals 1, the letter S, which stands for "single," is used. When the number of poles or number of throws equals 2, the letter *D*, which stands for "double," is used. When the number of poles or number of throws exceeds 2, integers such as 3, 4, or 5 are used. Here are a few examples: SPST, SPDT, DPST, DPDT, DP3T, and 3P6T. The switch shown in Fig. 3.35a represents a

single-pole single-throw switch (SPST), whereas the switch in Fig. 3.35b represents a single-pole double-throw switch (SPDT).

Two important features to note about switches include whether a switch has momentary contact action and whether the switch has a center-off position. Momentary-contact switches, which include mainly pushbutton switches, are used when it is necessary to only briefly open or close a connection. Momentary-contact switches come in either normally closed (NC) or normally open (NO) forms. A normally closed pushbutton switch acts as a closed circuit (passes current) when left untouched. A normally open pushbutton switch acts as an open circuit (broken circuit) when left untouched. Center-off position switches, which are seen in diverter switches, have an additional "off" position located between the two "on" positions. It is important to note that not all switches have center-off or momentary-contact features—these features must be specified.

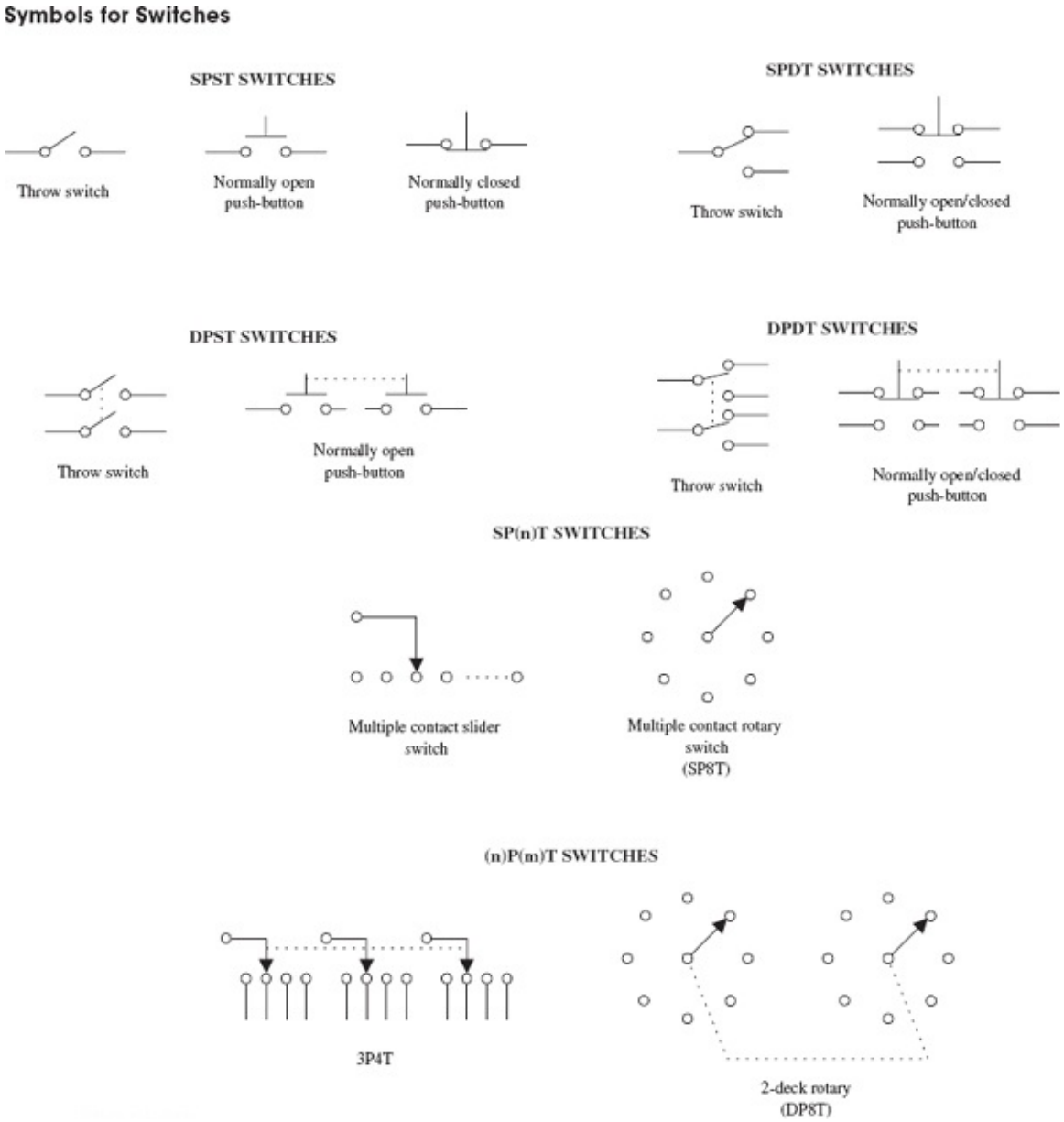
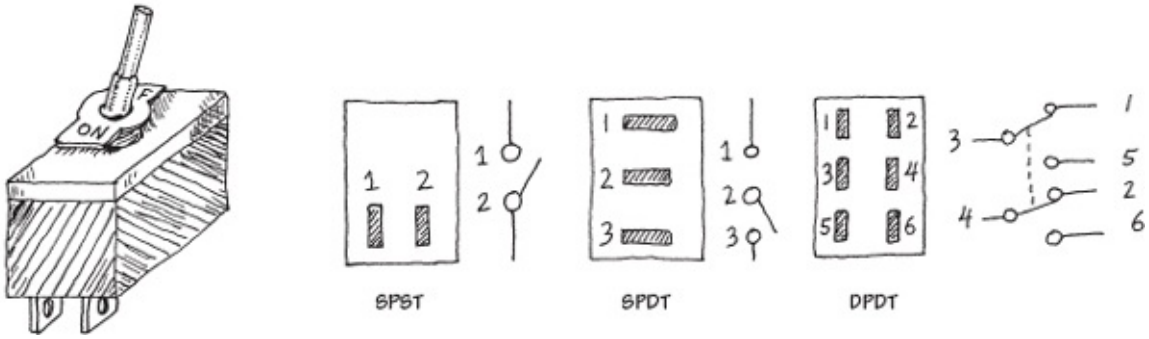


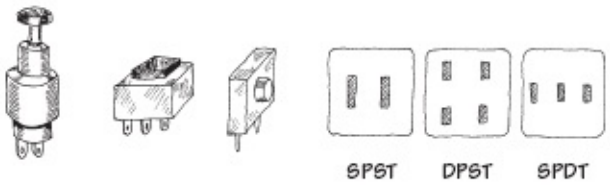
FIGURE 3.36

3.3.3 Kinds of Switches

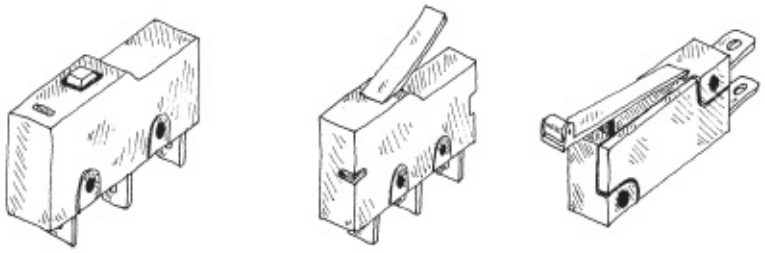
Toggle Switch



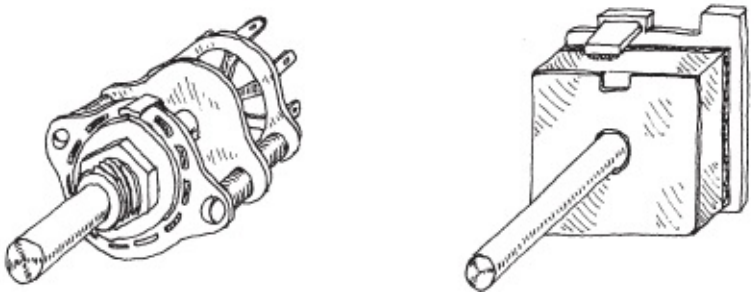
Pushbutton Switch



Snap Switch (Microswitch)



Rotary Switch



Magnetic Reed Switch

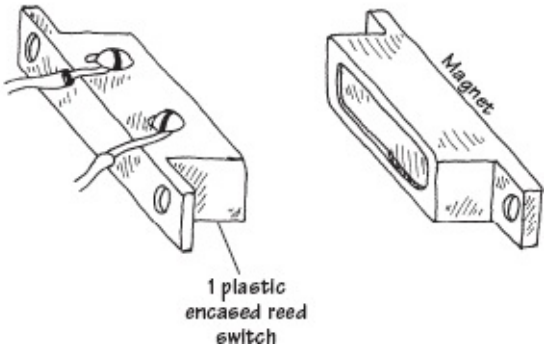
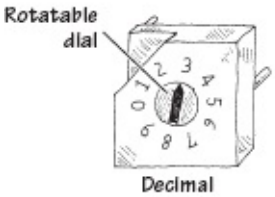


FIGURE 3.37

A reed switch consists of two closely spaced leaflike contacts that are enclosed in an air-tight container. When a magnetic field is brought nearby, the two contacts will come together (if it is a normally open reed switch) or will push apart (if it is a normally closed reed switch).

Binary-Coded Switch



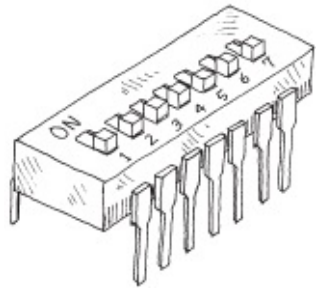


These switches are usec between the switch pairs binary/hexadecimal and c

True Binary/Hexadecimal

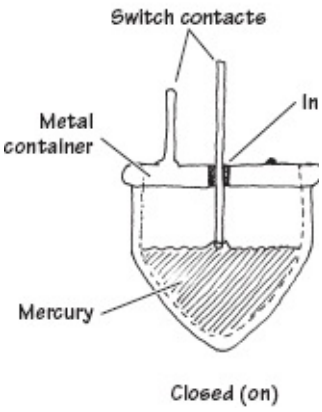
Type		Position	Code 1	2	4
Hexadecimal	Decimal	0			
		1	•		
		2		•	
		3	•	•	
		4			•
		5	•		•
		6		•	•
		7	•	•	•
		8			
		9	•		
		A		•	
		B	•	•	
		C			•
		D	•		•
		E		•	•
		F	•	•	•

DIP Switch



***DIP* stands for "dual-inli can be wired directly into**

Mercury Tilt-Over



This type of switch is us

Problem 3: If a 100-μF capacitor in a high-voltage power supply is shunted by a 100-K resistor, what is the minimum time before the capacitor is considered fully discharged when power is turned off?

Answer: 50 s. Hint: After five time constants, a capacitor is considered discharged.

Problem 4: An IC uses an external RC charging network to control its timing. To match its internal circuitry, the IC requires 0.667 of the supply voltage (5 V), or 3.335 V. If the capacitor's value is 10 μF, what value resistor is required in the RC network to obtain a 5.0-s timing period?

Answer: 500 kΩ. Hint: Use

$$V(t) = V(1 - e^{-(t/RC)}) = 0.667 \text{ V}$$

Problem 5: What is the reactance of a 470-pF capacitor at *7.5* MHz and 15.0 MHz?

Answer: X_C(7.5 MHz) = 45.2 Ω, X_C(15 MHz) = 22.5 Ω

Problem 6: Find the $f_{-3dB,}$ the attenuation in decibels A_{dB}, and phase shift ϕ for the given frequencies f_1, f_2, f_3, and f_4 of the filter circuits shown in Fig. 3.81.

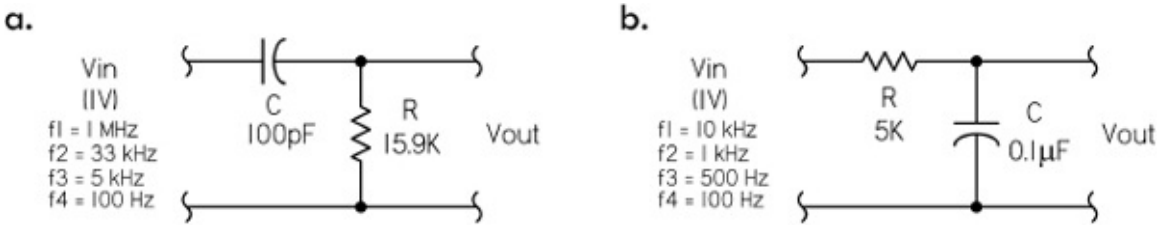


FIGURE 3.81

Answer: (a) f_{-3dB} = 100,097 Hz; f_1 = 1 MHz, A_{dB} = −0.043 dB, ϕ = 5.71°; f_2 = 33 kHz, A_{dB} = −10.08 dB, ϕ = 71.7°; f_3 = 5 kHz, A_{dB} = −26.04 dB, ϕ = 87.1°; f_4 = 100 Hz, A_{dB} = −60.00 dB, ϕ = 89.9°.

(b) f_{-3dB} = 318 Hz; f_1 = 10 kHz, A_{dB} = −29.95 dB, ϕ = −88.2°; f_2 = 1 kHz, A_{dB} = −10.36 dB, ϕ = −72.3°; f_3 = 500 Hz, A_{dB} = −5.40 dB, ϕ = −57.5°; f_4 = 100 Hz, A_{dB} = −0.41 dB, ϕ = −17.4°.

Problem 7: For the half-wave rectifier supply shown in Fig. 3.82, find the average dc output voltage, the peak-to-peak ripple voltage, the rms ripple voltage, the ripple factor, and percent ripple.

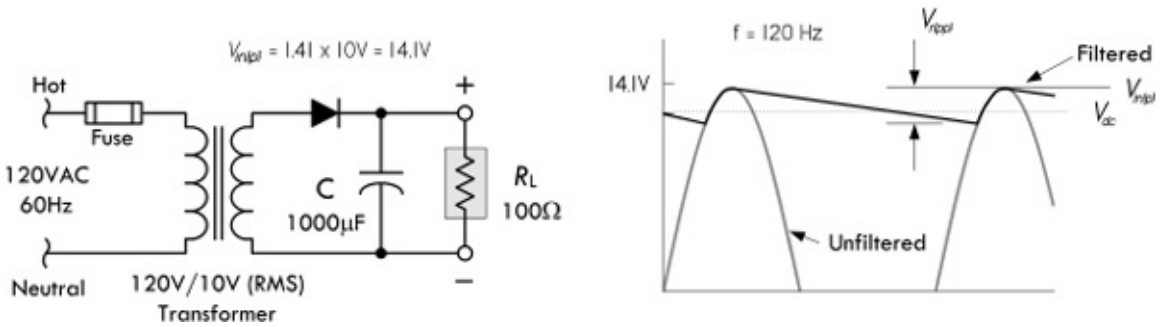


FIGURE 3.82

Answer: V_{dc} = 12.92 V, $V_{r(pp)}$ = 2.35 V, $V_{r(rms)}$ = 0.68 VAC, ripple factor = 0.0526, percent ripple = 5.26%. See Fig. 3.74.

3.7 Inductors

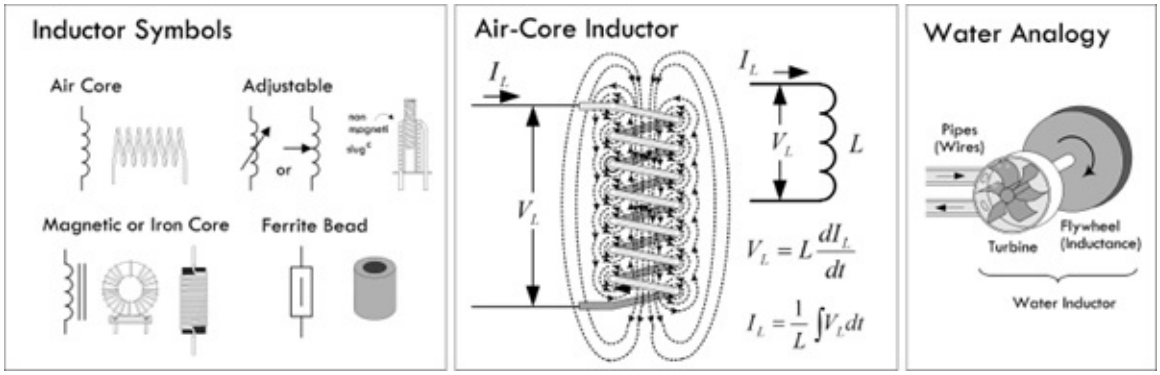


FIGURE 3.83

The basic role of an inductor is to prevent any sudden changes in current from flowing through it. (Refer to Sec. 2.24 for the details of how this works.) Under ac conditions, an inductor's impedance (reactance) increases with frequency; an inductor acts to block high-frequency signals while allowing low-frequency signals to pass through it. By selecting the proper inductance value, it is possible to create high-frequency chokes (e.g., RF/EMI chokes) that, when placed in series with power or signal paths, will prevent RF (radiofrequency) or EMI (electromagnetic interference) from entering into the main circuit, where they could introduce undesired hum and false triggering effects.

Filter networks can be created when inductors are used alongside other components such as resistors and capacitors. For example, Fig. 3.84a and b show low-pass and high-pass filters using an inductor as the reactive element. In the low-pass filter, the inductor "chokes" out the high-frequency components, while in the high-pass filter the inductor passes the low-frequency components to ground—high-frequency components are prevented from taking the same path and follow the signal path. Series-resonant (bandpass) and parallel-resonant (notch) filters are shown in Figs. 3.84c and d. Parallel-resonant filters are used in oscillator circuits to eliminate, from an amplifier, any input any frequencies significantly different from the resonant frequency of the LC filter. Such circuits are often used to generate radiofrequency carrier signals for transmitters. Resonant filters also act as tuned circuits used in radio reception.

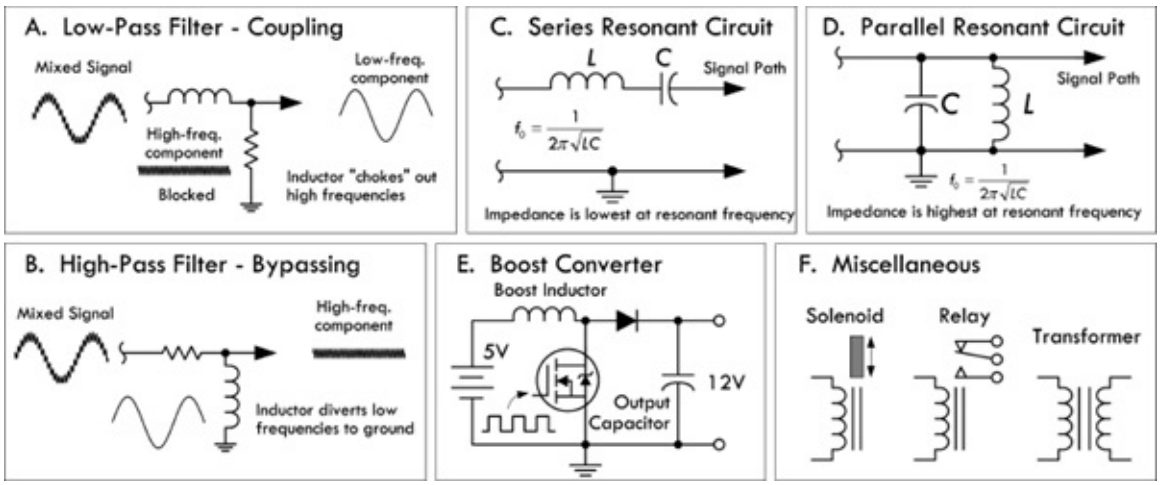


FIGURE 3.84

The energy storage nature of an inductor can be utilized in switching power supplies. For example, in Fig. 3.84e, a step-up switching regulator or "boost converter" is used to increase a 5-V input voltage to a 12-V output voltage. When the control element (transistor) is on, energy is stored in the inductor. The load, isolated by the diode, is supplied by the charge stored in the capacitor. When the control element is off, the energy stored in the inductor is added to the input voltage. At this time, the inductor supplies load current as well as restoring charge to the capacitor. Other converter configurations include step-down and inverting switching regulators.

A heavily wound, iron core inductor can be used as an electromagnet, capable of attracting steel and other ferromagnetic materials. A solenoid is an electromagnet that has a mechanical mechanism that is pulled when the solenoid is energized by current. The movement mechanism may involve latching or unlatching a door, or opening or closing a valve (i.e., a solenoid valve), or making or breaking switch contacts (as in an electrical relay).

Coupled inductors that share magnetic flux linkage are used to create transformers—devices that utilize mutual inductance to step up or step down ac voltages and currents.

3.7.1 Inductance

The voltage across an inductor is proportional to the rate at which the current changes. The relationship between voltage and current are described by the following equations:

$$V_L = L\frac{dI_L}{dt} \quad \text{(Voltage across inductor = induced voltage)}$$

$$I_L = \frac{1}{L}\int V_L dt \quad \text{(Current through inductor)}$$

The proportionality constant L is the inductance, which depends on a number of physical parameters, such as coil shape, number of turns, and core makeup. The basic unit of inductance is the henry, abbreviated H. One henry equals an induced voltage of 1 V when the current is varying at a rate of 1 A/s:

$$1\,\text{H} = \frac{1\,\text{V}}{1\,\text{A/s}}$$

Typical values of inductance for commercial inductors vary from a few nanohenrys for small air core inductors to 50 H for large iron core inductors.

3.7.2 Constructing Inductors

Although constructing a capacitor from scratch is rare and usually unnecessary, constructing an inductor from scratch is quite common. Often it is necessary, since specific inductance values may be difficult to

find or overly expensive in commercial units. Even stock inductors often require trimming to match an accurately desired value. Figure 3.85 shows some formulas used to construct inductors. Refer to Sec. 2.24 to see example calculations.

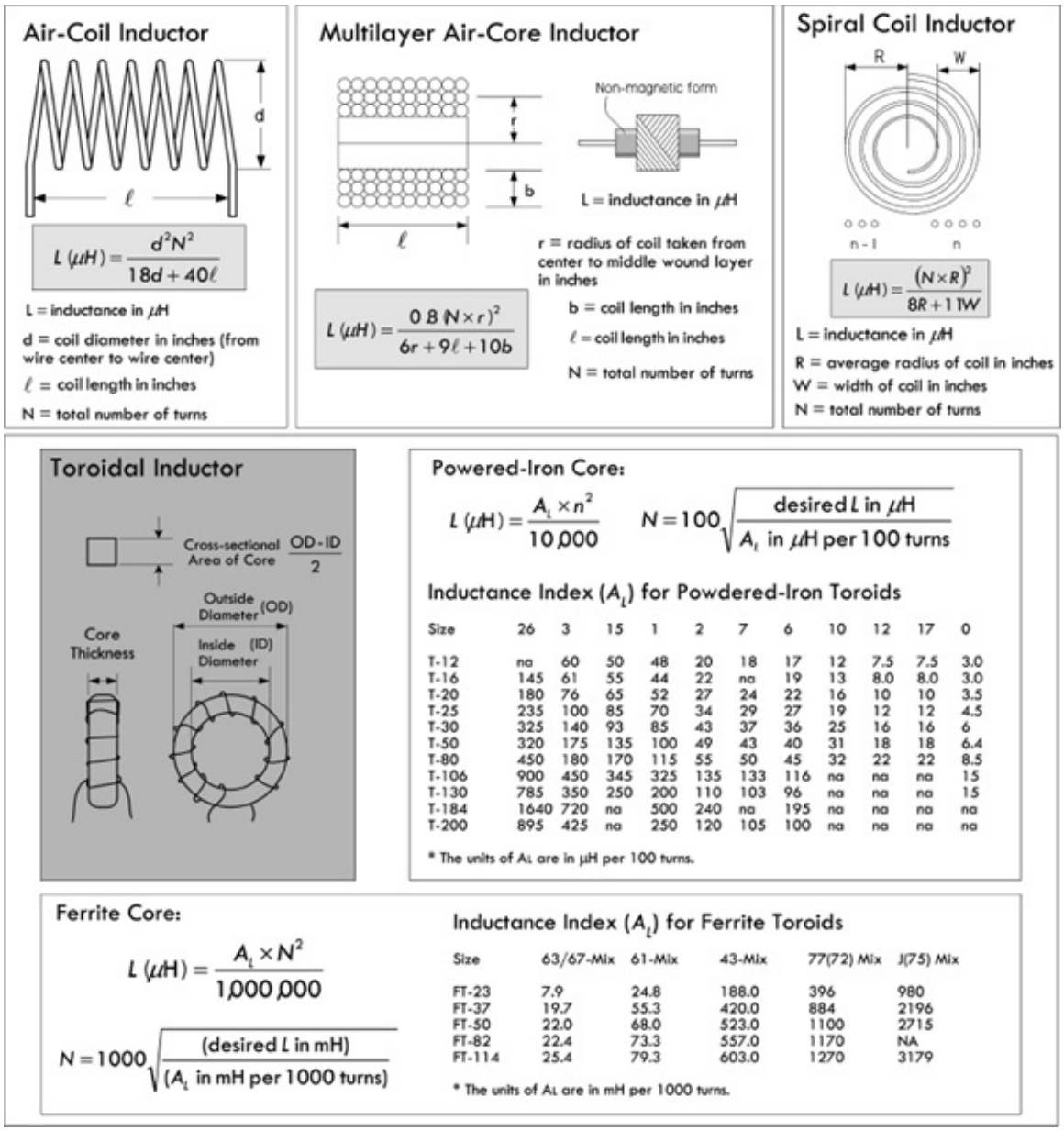


FIGURE 3.85

3.7.3 Inductors in Series and Parallel

When two or more inductors are connected in series, the total inductance is equal to the sum of the individual inductances, provided the coils are sufficiently separated so that coils are not in the magnetic field of one another:

$$L_{TOT} = L_1 + L_2 + L_3 + \cdots + L_N \quad \text{(Inductor in series)}$$

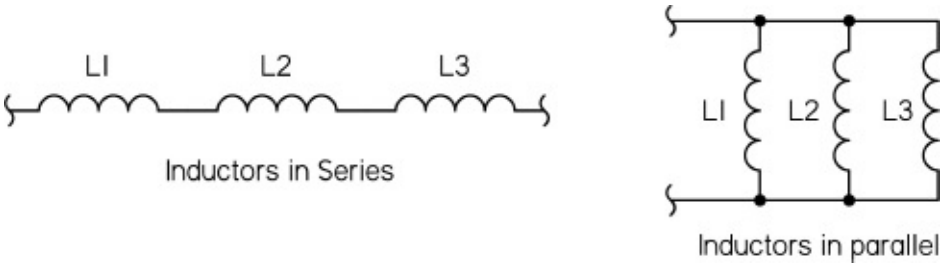


FIGURE 3.86

If inductors are connected in parallel, and if the coils are separated sufficiently, the total inductance is given by:

$$\frac{1}{L_{TOT}} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3} + \cdots \frac{1}{L_N} \quad \text{(Inductor in parallel)}$$

When only two inductors are in parallel, the formula simplifies to $L_{TOT} = (L_1 \times L_2)/(L_1 + L_2)$. Refer to Sec. 2.24 to see how these formulas were derived.

3.7.4 RL Time Constant

When a resistor is placed in series with an inductor, the resistance controls the rate at which energy is pumped into the magnetic field of an inductor (or pumped back into the circuit when the field collapses). Figure 3.87 shows energizing and deenergizing RL circuits with corresponding current-versus-time response curves.

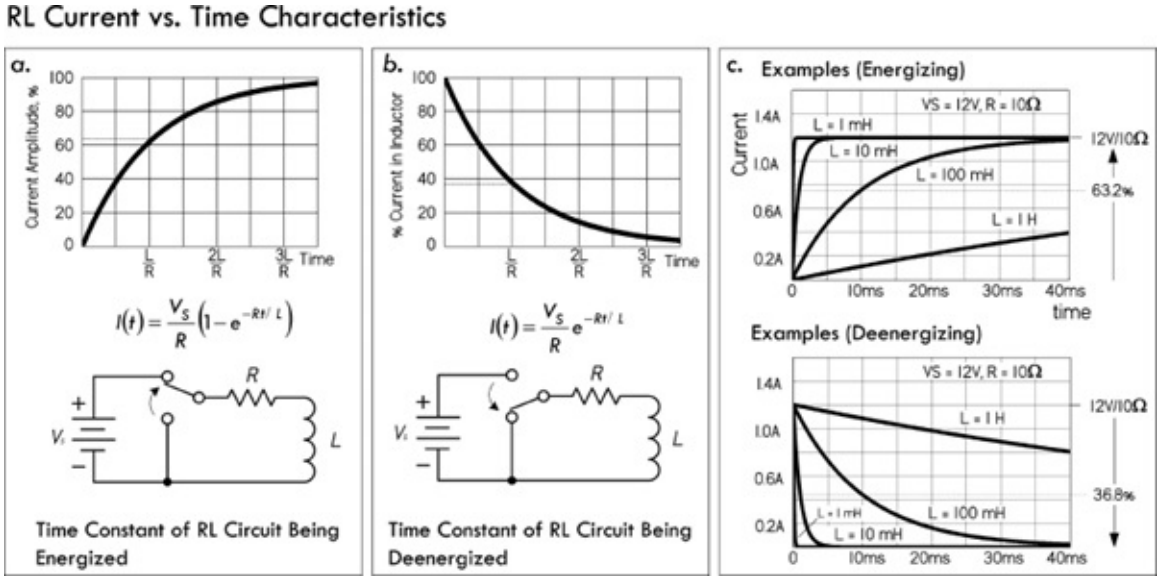


FIGURE 3.87

In the energizing RL circuit, the inductor current and voltage as a function of time after the switch is thrown to the up position is given by the following equation:

$$I(t) = \frac{V_S}{R}\,1 - e^{-Rt/L} \quad \text{(Current through energizing inductor)}$$

$$V_L(t) = V_S e^{-Rt/L} \quad \text{(Voltage across energizing inductor)}$$

$$V_R(t) = V_S(1 - e^{-t(L/R)}) \quad \text{(Voltage across resistor)}$$

The time in seconds required for current to build up to 63.2 percent of the maximum value is called the time constant τ, which is equal to *L/R*. After five time constants, the current is considered to have reached its maximum value.

For a deenergizing RL circuit, things are a bit tricky. However, if we assume when we remove the applied voltage that there is no break in the circuit by moving the discharge to the ground, we can apply the following deenergizing equations:

$$I(t) = \frac{V_S}{R}\,1 - e^{-Rt/L} \quad \text{(Current through deenergizing inductor)}$$

$$V_L(t) = -V_s e^{-Rt/L} \quad \text{(Voltage across energizing inductor)}$$

$$V_R(t) = V_S e^{-Rt/L} \quad \text{(Voltage across resistor)}$$

It takes one time constant τ = *L/R* for the inductor to lose 63.2 percent of its initial current value. After five time constants, the current is considered to have reached its minimum value.

The deenergizing RL circuit requires special attention. In real life, when the current flow through an inductor is suddenly interrupted—say, by flipping a switch—there can be a very large induced voltage generated as the inductor's magnetic field rapidly collapses. The magnitude of the induced voltage can be so great that "electron pressure" at the switch contacts overcomes the work function of the metal, and a spark may jump between the contacts. Section 2.24 describes inductive spikes in greater detail.

3.7.5 Inductive Reactance

An inductor has a frequency-sensitive impedance, or inductive reactance, that increases with applied frequency. Unlike resistance, the inductive reactance doesn't dissipate energy in the form of heat, but temporarily stores energy in a magnetic field and later returns it (or reflects it back) to the source. Inductive reactance has units of ohms, and it increases with frequency according to the following equation:

$$X_L = 2\pi fL \quad \text{(Inductive reactance)}$$

For example, a 1-H coil with an applied frequency of 60 Hz will provide 377 Ω of reactance, while a 10-μH inductor at 20 MHz will provide 1257 Ω of reactance.

Note that the preceding equation is for an ideal inductor. Real inductors have imperfections, such as internal resistance and capacitance that tend to cause the reactance to deviate from the equation, especially at higher frequencies—see Sec. 2.24.

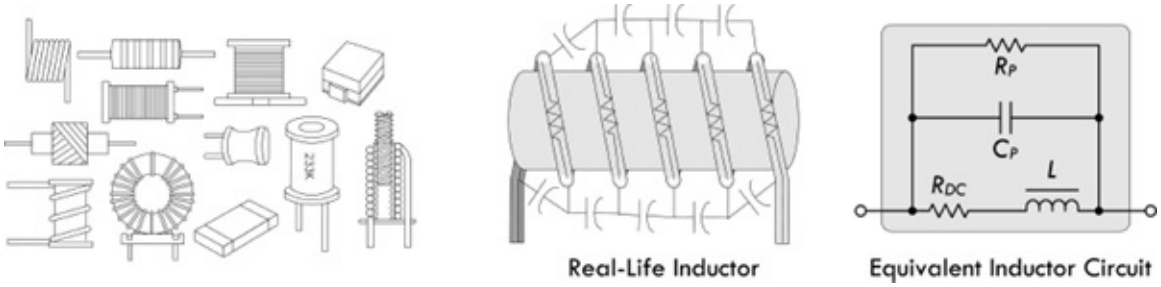


FIGURE 3.88

3.7.6 Real Inductors

There are many different kinds of inductors used for various applications. Selecting the right inductor requires understanding the nonideal characteristics of real inductors. Real inductors have imperfections, such as internal resistance and capacitance, that make them operate slightly differently than the ideal equations predict. Some inductors, due to their construction, may have larger resistive or capacitive components inherent in their design that can cause the inductor to act in a nonlinear fashion when approaching certain frequencies. (See Section 2.24 for details.) Other important differences include current-handling capacity, tolerance, maximum inductance and size, quality factor (*Q*), saturation characteristics, adjustability, radiated electromagnetic interference (EMI), and environmental endurance.

The following provides important inductor specifications that you'll find listed in manufacturer data sheets.

3.7.7 Inductor Specifications

Inductance (*L*): The property of an element that tends to oppose any change in the current flowing through it. The inductance depends on core material, core shape, size, and turn count of the coil. The unit of inductance is the henry (H). Most often, inductances are expressed in microhenries (μH).
1 henry (H) = 10^6 μF
1 millihenry (mH) = 10^3 μH
1 nanohenry (nH) = 10^{-3} μH

Inductive tolerance: Inductive tolerance is the allowed amount of variation from the nominal value specified by the manufacturer. Standard inductance tolerances are typically designated by a tolerance letter: F = ±1%, G = ±2%, H = ±3%, J = ±5%, K = ±10%, L = ±15% (some military products L = ±20%), M = ±20%.

Direct current resistance (*DCR*): The resistance of the inductor winding measured using dc current. The DCR is most often minimized in the design of an inductor and specified as a maximum rating.

Incremental current: The dc bias current flowing through an inductor that causes an inductance drop of 5 percent from the initial zero dc bias inductance value. This current level indicates where the inductance can be expected to drop significantly if the dc bias current is increased further. This applies mostly to ferrite cores rather than to powdered iron. Powdered iron cores exhibit soft saturation characteristics, which means their inductance drop from higher dc levels is much more gradual than that of ferrite cores. The rate at which the inductance will drop is also a function of the core shape.

Maximum dc current (IDC): The level of continuous direct current that can be passed through an inductor with no damage. The dc current level is based on a maximum temperature rise at the maximum rated ambient temperature. For low-frequency currents, the RMS current can be substituted for the dc-rated

current.

Saturation current: The dc bias current flowing through an inductor that causes the inductance to drop by a specified amount from the initial zero dc bias inductance value. Common specified inductance drop percentages are 10 percent and 20 percent. It is useful to use the 10 percent inductance drop value for ferrite cores and 20 percent for powdered iron cores in energy storage applications. The cause of the inductance drop due to the dc bias current is related to the magnetic properties of the core. The core, and some of the space around the core, can store only a given amount of magnetic flux density. Beyond the maximum flux density point, the permeability of the core is reduced; thus, the inductance drops. Core saturation does not apply to air core inductors.

Self-resonant frequency (SRF or f_0*):* The frequency at which the inductor's distributed capacitance resonates with the inductance. At this frequency, the inductance is equal to the capacitance and they cancel each other. As a consequence, at SRF, the inductor acts as a purely resistive high-impedance element. Also at this frequency, the Q value of the inductor is zero. Distributed capacitance is caused by the turns of wire layered on top of each other and around the core. This capacitance is in parallel to the inductance. At frequencies above SRF, the capacitive reactance of the parallel combination will become the dominant component.

*Quality factor (*Q*)*: The measure of the relative losses in the inductor. *Q* is defined as the ratio of inductive reactance to the effective resistance, or X_L/R_E. Since both X_L and R_E are functions of frequency, the test frequency must be given when specifying *Q*. As the self-resonant frequency, Q is zero since the inductance is zero at that point. Ideally, $Q = X_L/R_{DC} = 2\pi f\ \mathrm{L}/R_{\mathrm{DC}}$.

Inductance temperature coefficient: The change in inductance per unit temperature change. Measured under zero bias condition and expressed in parts per million (ppm).

Resistance temperature coefficient: The change in dc wire resistance per unit temperature change. Measured low dc bias (<1 V_{DC}) and expressed in parts per million (ppm).

Curie temperature (TC): The temperature beyond which the core material loses its magnetic properties.

Magnetic saturation flux density (B_{SAT}*):* A core parameter that indicates the maximum flux the material can be induced to hold. At this value of flux density, all magnetic domains within the core are magnetized and aligned.

Electromagnetic interference (EMI): For inductors, this refers to the amount of magnetic field radiated away from the inductor into space. The field may cause interference with other magnetically sensitive components and may require consideration in circuit design and layout.

3.7.8 Types of Inductors

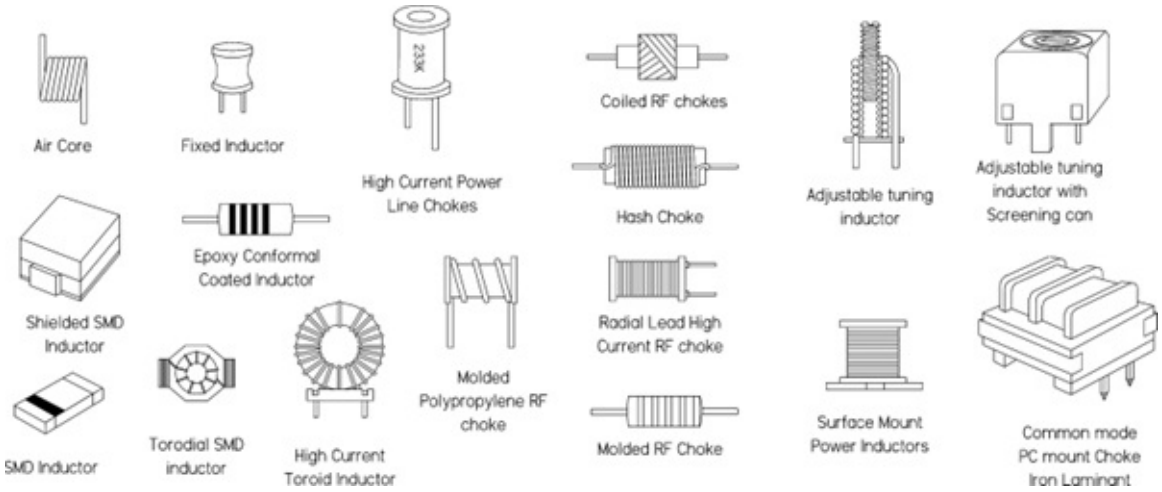


FIGURE 3.89

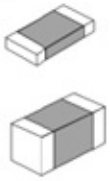
There are many different inductors out there. Some inductors are designed for general-purpose filtering, others for RF/EMI filtering, others for high-current choking, and still others for energy storage (in switching power supplies). Typically, manufacturers break down their inductor stock into the following categories: common-mode, general-purpose, high-current, high-frequency, power, and RF chokes. Fundamental specifications for selection include inductance value, dc-rated current, dc resistance, tolerance, and quality factor (*Q*). Table 3.8 provides a general overview matching inductance characteristics to specified application.

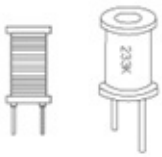
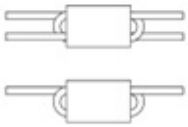
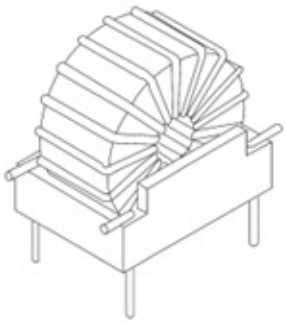
TABLE 3.8 Overview of Typical Applications for Inductors

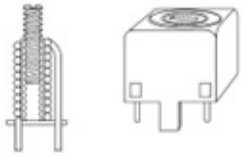
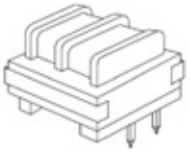
APPLICATION	INDUCTANCE (L)	MAXIMUM DC CURRENT (IDC)	SELF-RESONANT FREQUENCY (SRF)	QUALITY FACTOR (Q)	DC RESISTANCE (RDC)
High-frequency (RF), resonance circuits	Low	Low	Very high	Very high	Low
EM coupling	High	*	High	Low	Very low
Filter circuits	High	High	High	Low	Very low
Switch-mode power supplies, dc/dc converters	*	High	Medium	Low	Low

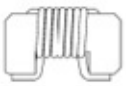
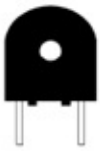
Table 3.9 provides a selection guide to help determine what kind of inductor to use for a particular application. For in-depth details about specific inductors, check out manufacturers' websites and read through their data sheets. (Some noteworthy manufacturers: API Delevan, Bourns, C&D Technologies, Fastron, KOA, JW Miller Magnetics, muRata, Pulse, TRIAD, TDK, and VISHAY.)

TABLE 3.9 Inductor Selection Guide

INDUCTOR TYPE	DESCRIPTION	APPLICATION
Multilayer chip inductors 	Used predominantly in high-density PCB surface-mount circuits where size, interboard magnetic coupling, and vibration are major concerns. They have fewer parasitic characteristics and less resistive loss when compared to most other leaded inductors. Come with excellent Q values, and high-frequency performance, while generating very little noise. Smaller boards and shorter traces also reduce EMI emissions and signal cross-coupling. Come in various current ratings.	Used for EMI/RFI attenuation and suppression, as well as reactive elements in LC resonant oscillators and impedance-matching networks, and other choke-coil type applications. They are found in A/D converters, bandpass filters, pulse generators, RF amplifiers, signal generators, switching power supplies, and telecom.
Molded inductors 	Small with axial leads for PC board mounting. Outer coating protects coil from the environment. Comes in a shielded version as well. Frequency range is typically greater than 50 kHz.	Used in filters, AD converters, AM/FM radio, pulse generators, signal generators, switching power supplies, and telecom.
Shielded inductors 	The magnetic shield is designed to prevent magnetic coupling and RF/EMI interference issues—especially important in densely packed boards where signal corruption is a major concern. Come in surface-mount, axial lead, and other configurations and a variety of current-handling ratings.	Used in high-reliability applications where magnetic coupling is to be avoided. Used for dc/dc converters, computers, telecom equipment, filters, LDC displays, etc.
Conformal coated (dipped) inductors 	An inexpensive inductor that comes with axial or radial leads—similar to molded inductors. Outer coating protects inductor from environments	Typically used in less harsh environments than molded inductors. Usually used in less critical RFI/EMI applications. Higher-Q versions can be used in many of the same applications as molded inductors.

High-current chokes  Hash choke  RF choke 	These devices utilize ferrite or powdered iron cores to achieve large inductance values for low coil count and small volume. Fewer turns translates into lower dc resistance—an important feature for high-current applications.	High-current hash chokes are used for home appliances, communications systems, computer add-ons, dc-fc, switching power supplies, transmitters, and uninterrupted power supplies. Power line chokes are used for filters, power supplies, RFI suppression, power amps, switching regulators, SCR and Triac controls, and speaker crossover networks.
Wideband chokes 	Wideband chokes attenuate unwanted signals in circuits without contributing to power losses at lower frequencies or dc. Impedances range from 20 to 500 Ω and are obtained over frequencies from 1.0 to 400 MHz. Currents are limited only by #22 or #24 AWG tinned copper wire used in construction.	Used mainly in PC boards to filter out EMI and RFI. Also used in RF circuits to suppress parasitic oscillation at VHF and UHF. Used in A/D converters, communications systems, computer add-ons, dc-fc, I/O boards, RF power amplifiers, signal generators, switching power supplies, telecom equipment, and uninterrupted power supplies.
Toroids 	Exhibit high inductance due to powdered iron or ferrite core, and also display good self-shielding due to core shape (good low EMI source). Low coil count reduces dc resistance compared to other solenoid-wound inductors. Come in a variety of types: small surface-mount devices, larger general-purpose devices, and high-power toroids that can handle very large currents. Toroids are also less susceptible to induced noise from other components, as the applied magnetic field induces equal and opposite currents inside the toroid, thus canceling interference	Toroids are used in a variety of different applications. They are used as chokes in ac power lines and to reduce EMI. Also used in home appliances, audio generators, auto electronics, bandpass filters, audiovisual equipment, dc-fc, I/O boards, impedance match transformers, oscillators, pulse generators, switching power supplies, telecom equipment, transmitters, tuned amplifiers, uninterrupted power supplies, VHF/UHF repeaters, ripple filters, etc.
Pot cores 	Provide ultra-high inductance values and high dc current ratings while maintaining a stable inductance due to high saturation currents and self-shielding. Excellent Q values in a small size.	Commonly used in telecom, audio, and automotive applications. Used as dc chokes, differential mode chokes, filters, and in switching circuits.

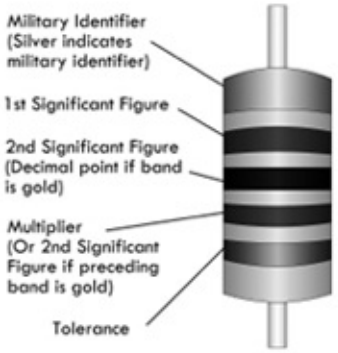
Balun chokes 	Typically used in impedance-matching applications. "Balun" refers to a balanced-unbalanced transformation of impedance levels.	Used in AM/FM radio, television and communications systems, I/O boards, impedance matching, pulse generators, transmitters, and walkie-talkies.
Air coils 	Range from a single loop of wire (for ultra-high frequencies) to larger coils wrapped around nonmagnetic form. Air core inductors are used in high-frequency applications requiring high-Q characteristics, where losses and distortion (common in magnetic core inductors due to hysteresis and eddy currents) cannot be tolerated. Without a magnetic core, however, these inductors have low inductance values. Used primarily for RF applications. Surface-mount air coil are also available.	Used in RF resonant tank circuits for TV tuners, FM stereo receivers, garage door openers, pulse generators, RF power amplifiers, switching power supplies, RC toys, uninterrupted power supplies, and walkie-talkies.
Adjustable inductors (variable or tunable) 	May use a slider contact along coils or, more commonly, a ferrite, powered iron, or brass slug screwed into the center of the coil. The inductance changes as the slug is screwed in and out of the coil. Magnetic cores increase the permeability or inductance, while brass slugs decrease inductance due to eddy current induction in a slug that reduces internal magnetic flux in the core. These devices are used to adjust resonant circuits that require high Q values. Used in circuits requiring very narrow bandwidths.	Typically used at frequencies above 50 kHz. Used in LC resonant tank circuits in AM/FM radio, TV, and other communication systems. Used in garage door openers, I/O boards, oscillators, pulse generators, RF power amplifiers, signal generators, switching power supplies, toys, transmitters, uninterrupted power supplies, VHF/UHF repeaters, and walkie-talkies.
Common-mode chokes  	Common- and differential-mode chokes are used to eliminate noise on a pair of conductors. Common-mode noise is defined as noise that is present in or common to both conductors, and can be the result of induced noise caused by the antenna effect on a conductor or PC trace. Used to prevent EMI and RF interference from power supply lines and for prevention of malfunctioning of electronic equipment. Common-mode chokes usually have ferrite core material well suited to attenuate common-mode current.	Common-mode chokes are very useful devices found in many radio circuits. They may help nearly any interference problem, from cable TV to telephones to audio interference caused by RF picked up on speaker leads. Particularly well-suited for applications such as line filters in switch-mode power supplies, and are also commonly used in desktop computers, industrial electronics, office equipment, and consumer electronics such as TVs and audio equipment.

Ferrite beads (ferrite chokes) 	Unlike a typical core inductor, the bead requires no coiling of wire—though wire/cable from a circuit may be coiled around it or simply pass through it. This effectively increases the inductance of the wire (or wires). The inductance range is rather limited to RF. Used for removing RF energy that exists within a transmission line structure (PCB trace). To remove unwanted RF energy, chip beads are used as high-frequency resistors (attenuators) that allow dc to pass while absorbing RF energy and dissipating it as heat. Come in hollow and leaded core versions.	Ferrite beads are often slipped over cables that radiate RF (computers, dimmers, fluorescent lights, motors, etc.). Ferrite beads are also placed on cables entering receiving equipment to prevent external RF from entering and contaminating signals in the cable runs. The "bumps" on computer and appliance cables (mouse, keyboard, monitor, etc.) are usually a ferrite bead coated by the plastic cord coating.
Ceramic core inductor 	These inductors use a special ceramic core that surpasses many ferrite-core-type inductors in terms of high-frequency operation, low IDC, high SRF, high Q, and tight tolerances.	Used in LC resonant circuits such as oscillators and signal generators. Used in impedance matching, circuit isolation, and RF filtration. Found in mobile phones, Bluetooth devices, wireless instruments, as well as audio, TV, and telecom devices.
Antenna rods ferrite core  Phenolic core 	Commonly used in antenna applications where narrow bandwidths are required. The core material, or "rod," on which the coil is wound is either ferrite, powered iron, or, alternatively, phenolic (essentially an air core). The magnetic core versions are far more popular, but the phenolic cores provide higher operating frequencies.	As antennas, ferrite cores with 800 permeability rods are suitable in 100-kHz to 1-MHz range. The 125 permeability rods are suitable in the 550-kHz to 1.6-MHz range. The 40 permeability rods are suitable in the 30-MHz range. The 20 permeability rods are suitable in the 150-MHz range. As the core's permeability decreases, the operating frequency increases.
Current sense inductor 	Used to sense current passing through conductors. Specified by frequency range. Often come with a center tap.	Used in high-current, low-profile point of load (POL) converters, dc/dc converters in distributed power systems, and dc/dc converters for field-programmable gate arrays. Found in PDAs, notebooks, desktops, servers, and battery-powered devices.

3.7.9 Reading Inductor Labels

If you are lucky, the inductance value and tolerance will be printed on your inductor in an easy-to-read format (e.g., 82µH ± 10%). However, many molded and dipped inductors come with a band color code. For smaller surface-mount inductors, a special shorthand printed code provides the inductance value and tolerance—see Fig. 3.90.

5-Band Inductor Code



Axial-leaded inductors are typically color banded. The military color code consists of 5 bands. A silver band that is double the width of the other four bands is located near one end. When present, the band identifies military radio frequency coils. The next three bands indicate the inductance value in microhenries, while the fourth band indicates tolerance.

For inductances less than 10, when either of the first two inductance bands is gold, the gold band represents a decimal point, while the other two bands represent significant figures. When the inductance is 10 or more, the first two bands represent significant figures, and the third band represents the multiplier.

Example

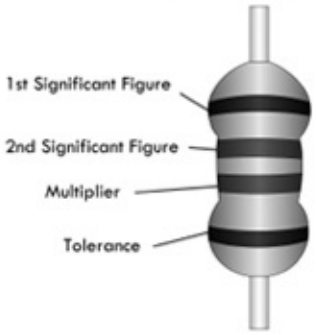


SILVER (Military)
GREEN (5)
GOLD (".")
BLUE (6)
BROWN (±1%)

5.6µH ± 1%

Color	1st S.F.	2nd S.F.	Multiplier	Tolerance
Black	0	0	1	
Brown	1	1	10	±1%
Red	2	2	100	±2%
Orange	3	3	1,000	±3%
Yellow	4	4	10,000	±4%
Green	5	5		
Blue	6	6		
Violet	7	7		
Grey	8	8		
White	9	9		
None				±20%
Silver				±10%
Gold		Dec. Point		±5%

4-Band Inductor Code



The 4-band color code does not contain a military band. The first and second bands represent first and second significant figures. The third band represents the multiplier, where silver = x 0.01, and gold = x 0.1. The fourth band indicates the tolerance, where black is ±20%, silver is ±10%, and gold is ±5%. All values are given in microhenries.

Example



RED (2)
VIOLET (7)
RED (x100)
SILVER (±10%)

2700µH ± 10%

Color	1st S.F.	2nd S.F.	Multiplier	Tolerance
Black	0	0	1	±20%
Brown	1	1	10	
Red	2	2	100	
Orange	3	3	1,000	
Yellow	4	4		
Green	5	5		
Blue	6	6		
Violet	7	7		
Grey	8	8		
White	9	9		
None				
Silver			0.01	±10%
Gold			0.1	±5%

SMD Inductance Code



472K =4700µH ± 10%
221K =220µH ± 10%
22N =22nH ± 20%
5N6F =5.6nH ± 1%
391K = 391µH ± 10%
68M =0.068µH ± 20%

LETTER	TOLERANCE
F	± 1%
G	± 2%
H	± 3%
J	± 5%
K	± 10%
L	± 15%*
M	± 20%

* L = ± 20% for some Military devices

Value Code	Inductance	Value Code	Inductance	Value Code	Inductance	Value Code	Inductance	Value Code	Inductance	Value Code	Inductance
47	0.047µH	1R7	1.7µH	182	1800µH	260	26µH	390	39µH	6R8	6.8µH
68	0.068µH	1N8	1.8nH	R18	180nH	R27	270nH	4N7	4.7nH	601	600µH
82	0.082µH	1R8	1.8µH	181	180µH	271	270µH	4R7	4.7µH	600	60µH
R12	0.12µH	102	1000µH	18N	18nH	27N	27nH	401	400µH	650	65µH
R15	0.15µH	R10	0.1µH	180	18µH	270	27µH	400	40µH	682	6800µH
R18	0.18µH	101	100µH	190	19µH	3N3	3.3nH	421	420µH	681	680µH
R10	0.1µH	10N	10nH	1N0	1nH	3R3	3.3µH	450	45µH	68N	68nH
R22	0.22µH	100	10µH	1R0	1µH	3N9	3.9nH	472	4700µH	680	68µH
R27	0.27µH	110	11µH	2N2	2.2nH	3R9	3.9µH	471	470µH	700	70µH
R33	0.33µH	121	120µH	2R2	2.2µH	301	300µH	47N	47nH	751	750µH
R39	0.39µH	12N	12nH	2N7	2.7nH	300	30µH	470	47µH	750	75µH
R47	0.47µH	120	12µH	2R7	2.7µH	310	31µH	5N6	5.6nH	070	7µH
R56	0.56µH	141	140µH	202	2000µH	332	3300µH	5R6	5.6µH	8N2	8.2nH
R68	0.68µH	152	1500µH	201	200µH	R33	330nH	500	50µH	8R2	8.2µH
R82	0.82µH	R15	150nH	222	2200µH	331	330µH	561	560µH	800	80µH
1N2	1.2nH	151	150µH	R22	220nH	33N	33nH	56N	56nH	821	820µH
1R2	1.2µH	15N	15nH	221	220µH	330	33µH	560	56µH	82N	82nH
1N5	1.5nH	150	15µH	22N	22nH	391	390µH	050	5µH	820	82µH
1R5	1.5µH	170	17µH	220	22µH	39N	39nH	6N8	6.8nH	900	90µH

FIGURE 3.90

3.7.10 Inductor Applications

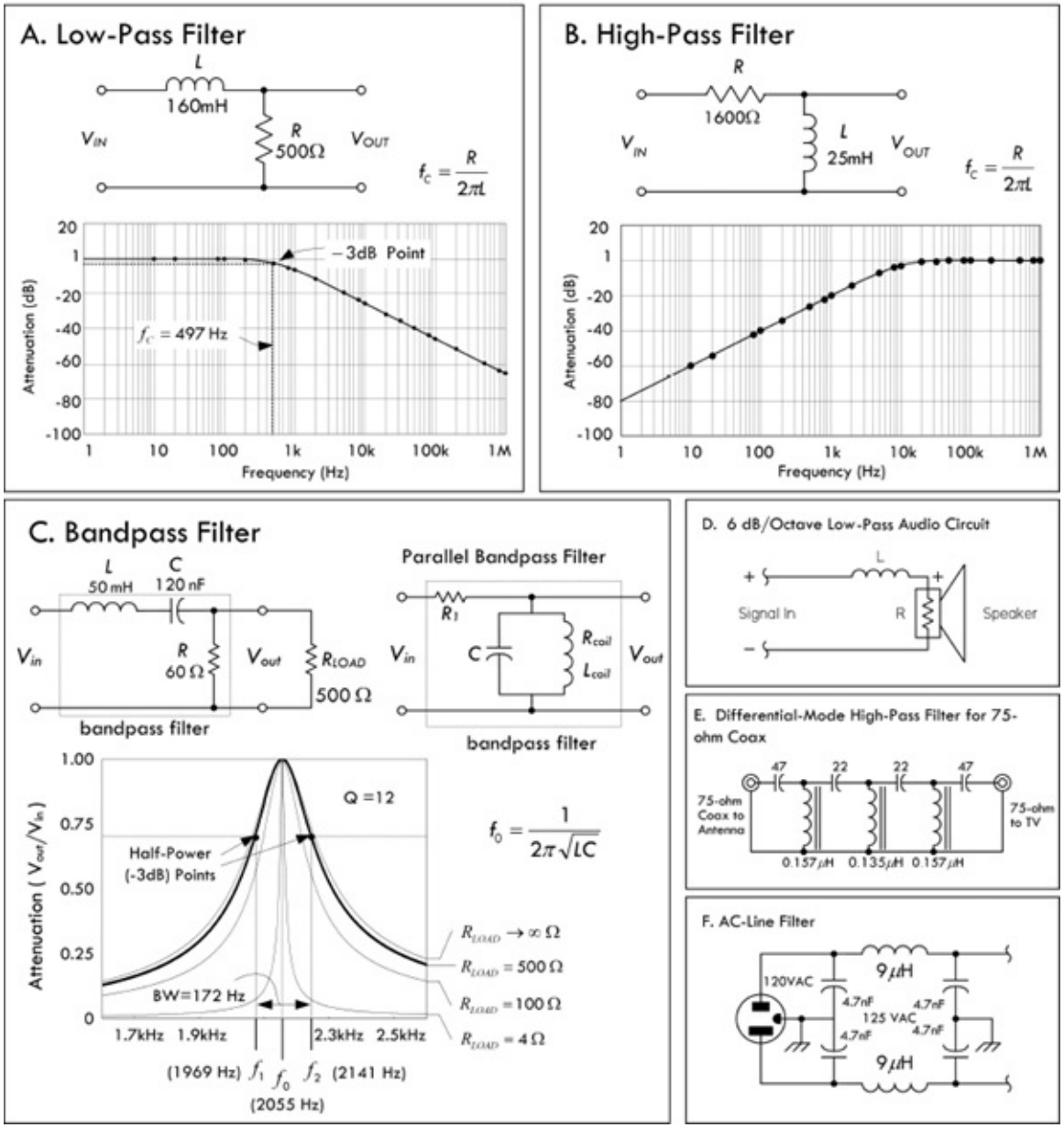


FIGURE 3.91 Filters offer little opposition to certain frequencies while blocking others. In (a), a low-pass filter is constructed using a resistor and inductor. The inductor's impedance increases with frequency, thus preventing high-frequency signals from passing. In (b), a high-pass filter blocks low frequencies—the inductor acts as a low-impedance path to ground for low-impedance signals. (c) shows a bandpass filter, allowing only a very narrow band of frequencies through. See Chap. 2 for the theory. (d) shows a low-pass filter used with a speaker. (e) shows a differential-mode high-pass filter for 75-Ω co-ax. It rejects high-frequency signals picked up by a TV antenna or that leak into a cable TV system. It is ineffective against common-mode signals, however. (f) shows an ac-line filter used to filter RF energy from power lines.

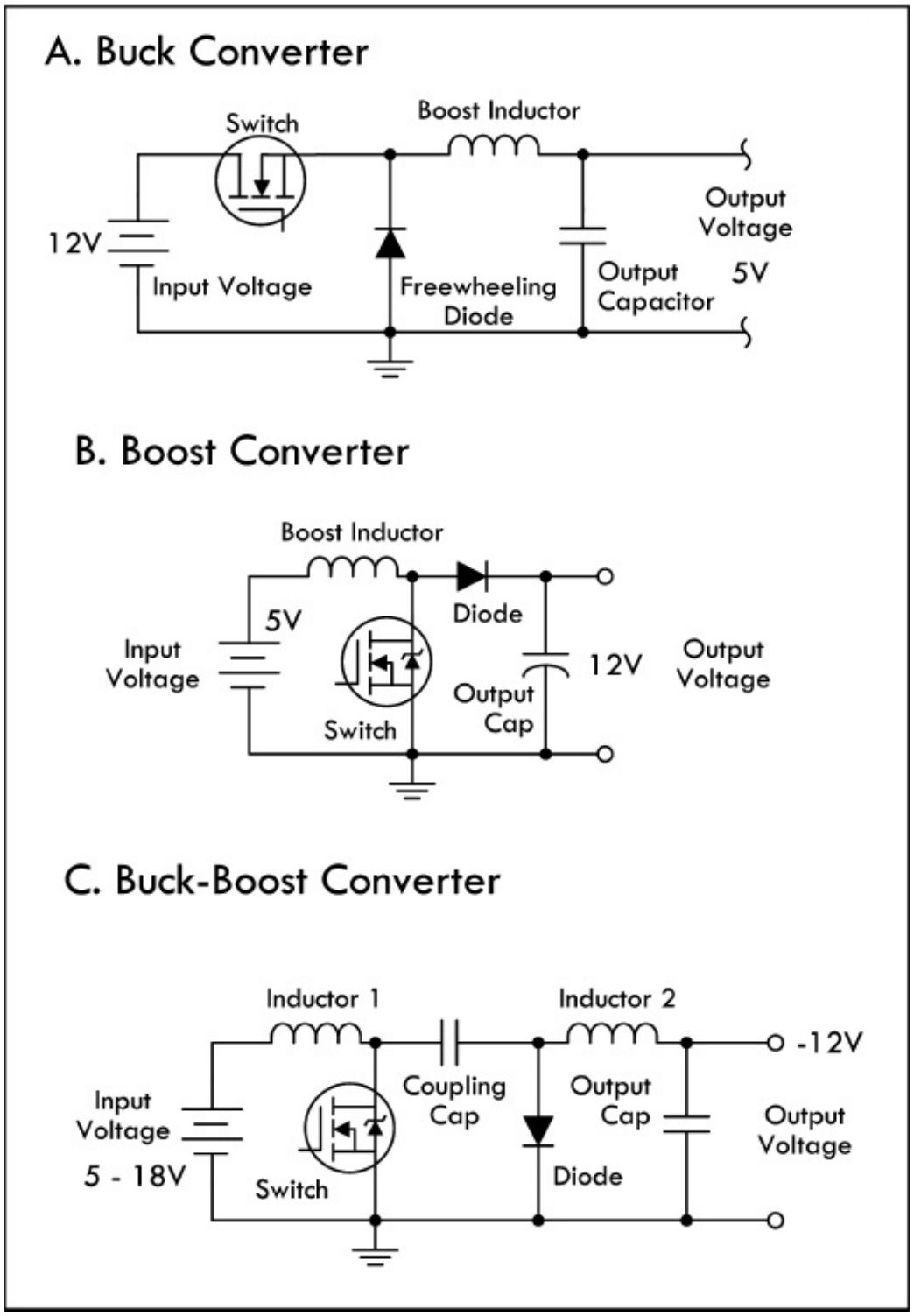


FIGURE 3.92 In switching regulator applications, an inductor is used as an energy storage device, when the semiconductor switch is on, the current in the inductor ramps up and energy is stored. When the switch is turned off, the stored energy is released into the load. Output voltages have a ripple that must be minimized by selecting appropriate inductance and output capacitor values. Figure 3.92a, b, and c shows various switching regulator configurations: *buck* (lower output voltage), *boost* (higher output voltage), and *buck-boost* (opposite polarity). Note that in the boost converter, the boost inductor current does not continuously flow to the load. During the switch "on" period, the inductor current flows to ground and the load current is supplied from the output capacitor.

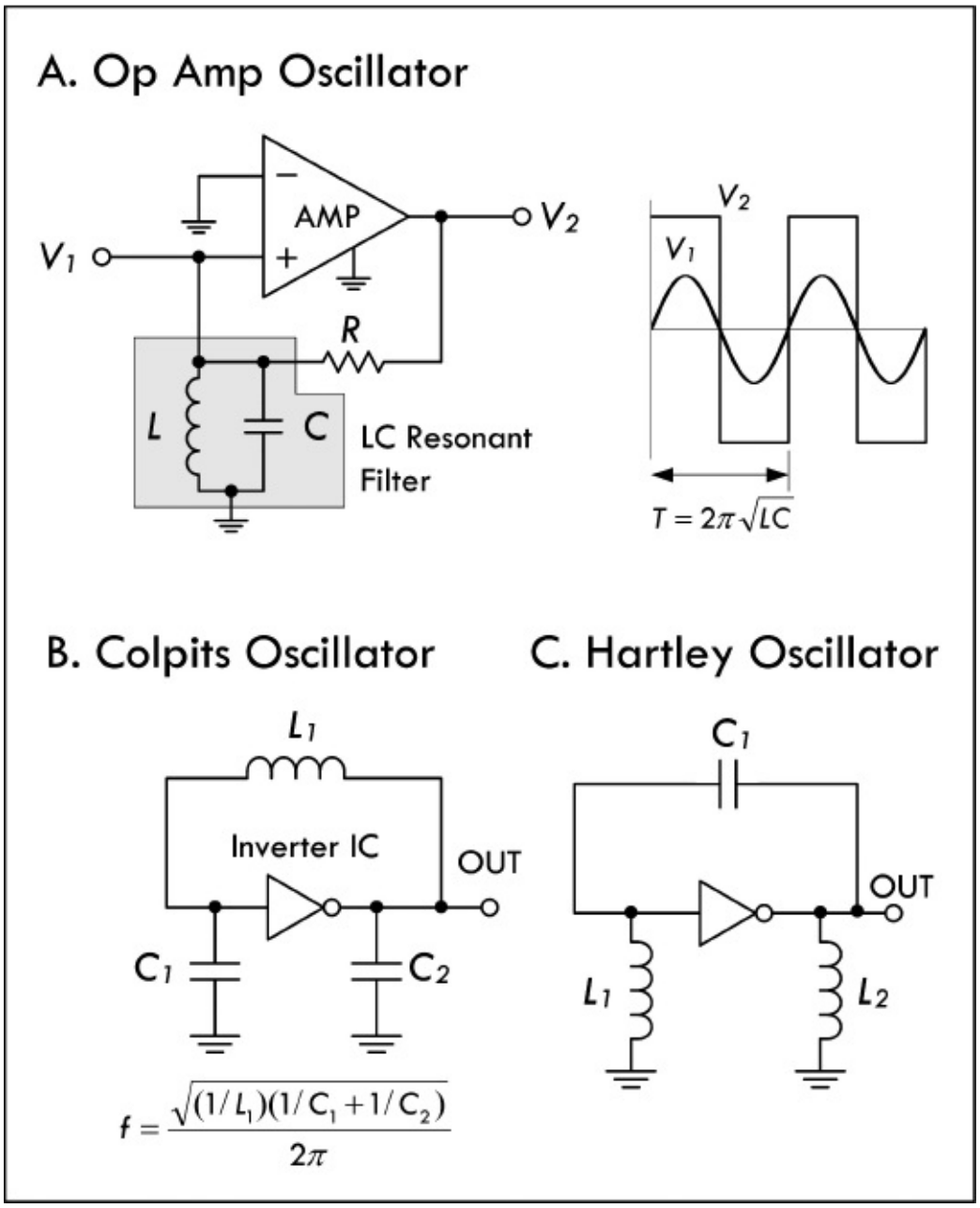


FIGURE 3.93 (*a*) An amplifier with positive feedback can be made to produce an output even in the absence of any input. Such circuits are called oscillators. In (a), an operational amplifier uses an LC resonant filter that eliminates from the amplifier input any frequencies significantly different from the LC resonant filter's resonant frequency: $f_0 = 1/(2\pi\sqrt{LC})$. The amplifier is alternately driven to saturation in the positive and negative direction, so it produces a squarewave at V_2. This squarewave has a Fourier component that is fed back to the noninverting input through resistor *R* in order to keep the oscillation from damping out. A sine wave is generated at V_1.

(b) There are two basic types of LC oscillators, Colpits and Hartley. The Colpits uses two capacitors, as shown here. It is generally favored over the Hartley, shown in (c), because of the simplicity of requiring only one inductor, which is usually more expensive and difficult to obtain than capacitors C_1 and C_2. The frequency of the Colpits oscillator is given by the formula in (b). See Chap. 10 on oscillators for more details.

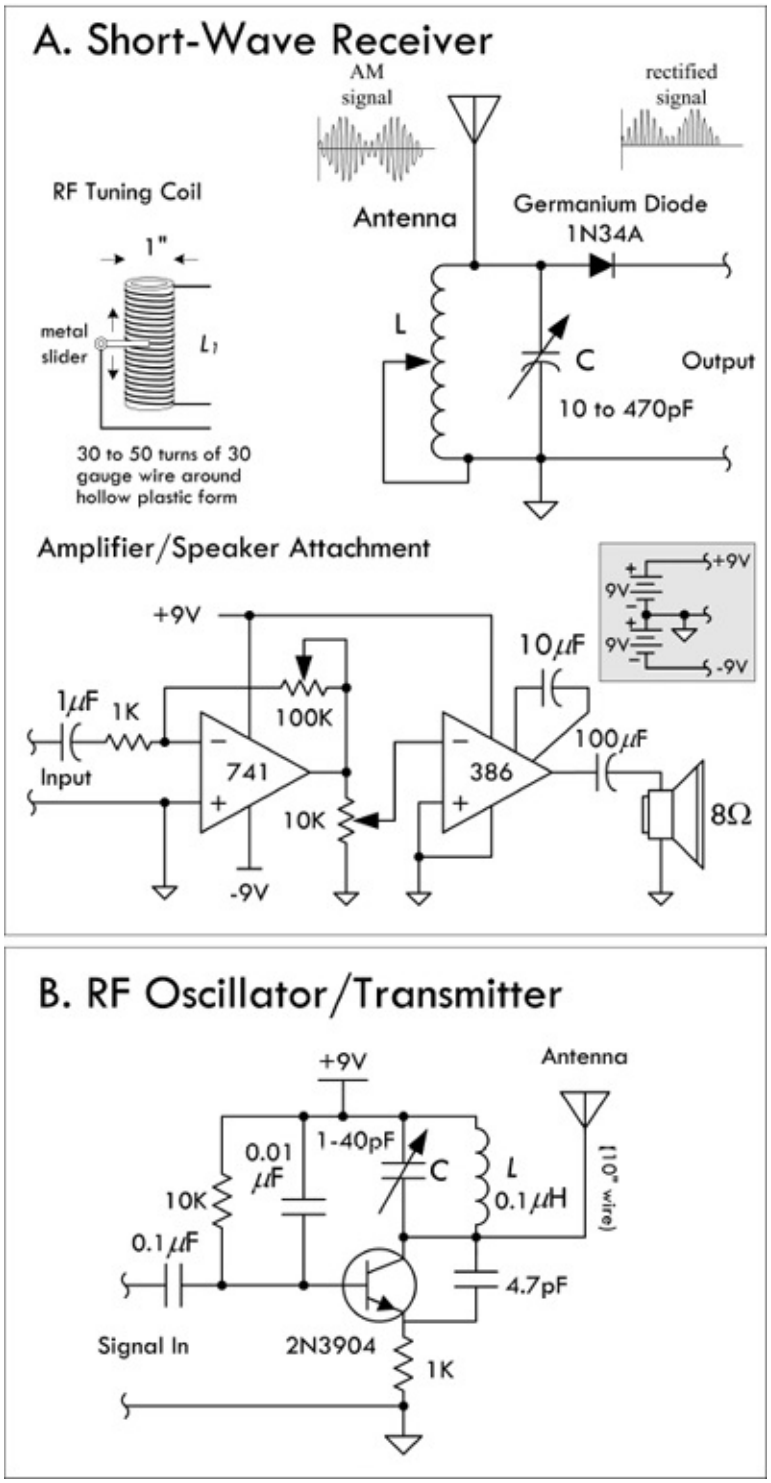


FIGURE 3.94

Here are a few simple radiofrequency circuits that make use of LC resonant filters to tune the circuits.

(a) The simplest radio receiver consists of nothing more than an antenna, a diode (germanium), and a pair of headphones. Such a receiver has no frequency selectivity, however, and will receive several of the strongest AM broadcast stations simultaneously. By adding a variable inductor that forms a resonant LC circuit with the antenna capacitance, a usable radio receiver can be constructed that is capable of tuning in to a number of different stations. (The variable capacitor provides additional tunability.) The encoded audio signal within the AM carrier is demodulated by the diode (the diode generates low-frequency [audio] Fourier components that are absent in the transmitted wave). After passing through the diode, only the positive half of the wave remains. This wave contains low-frequency components in addition to components at the frequency of the carrier wave. With the addition of a low-pass filter, only the low-frequency components remain. The frequency response of headphones and of a human ear will effectively act as a low-pass filter. The demodulated signal can be input into an amplifier to drive a speaker—see amplifier circuit. Real radio AM receivers are much more sophisticated than this, using a superheterodyne design scheme.

(b) A radio transmitter consists of an RF oscillator, one or more amplifier stages, and a modulator. In the simple FM transmitter shown here, the LC resonant filter sets the amplifier oscillatory frequency—the variable capacitor allows for

adjustability. Audio signals that enter will be frequency-modulated into the carrier and radiated as radio waves. An FM radio receiver should be able to pick up the circuit signal. A homemade inductor can be made by tightly winding 10 turns of 22-gauge wire around a ¼-in form.

3.7.11 EMI/EMC Design Tips

The following are some tips to avoid EMI and EMC problems. Tips include proper design techniques for PCB layout, proper power supply considerations, and effective use of filtering components. (Referenced letters that follow refer to Fig. 3.95.)

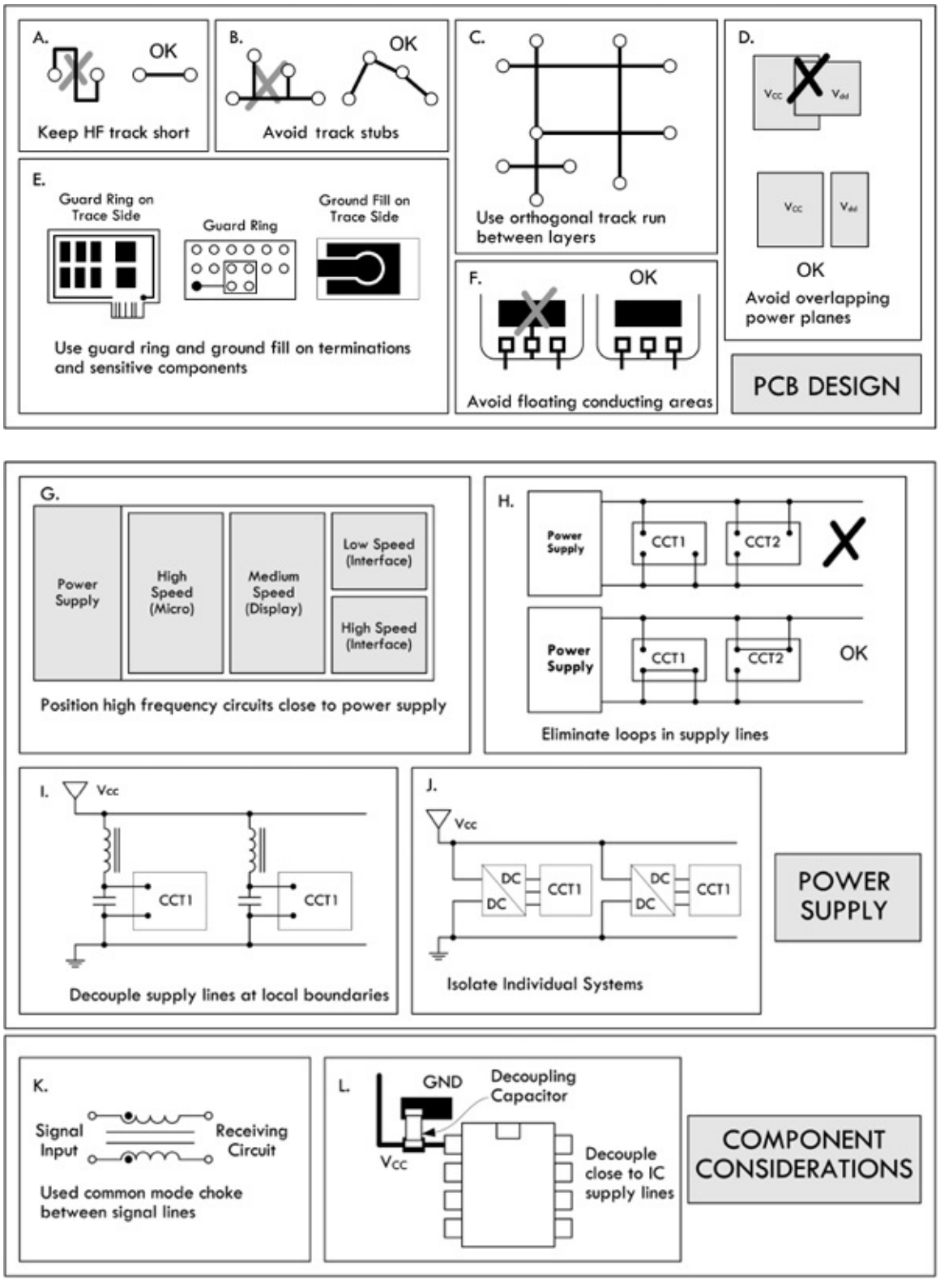


FIGURE 3.95

PCB Design Tips

Avoid slit apertures in PCB layout, such as a ground plane that is divided into two parts. Regions of high impedance are sources of high EMI, so use wide tracks for power lines on the trace sides for increased conduction. Make signal tracks stripline (ground planes above and below) and include ground plane and power plane whenever possible. Keep HF and RF tracks short and lay out the HF tracks first—see (a). In Fig. 3.95. Avoid track stubs, since they cause reflections and harmonics—see (b). Use a surrounding guard ring and ground fill where possible on sensitive components and terminations. A guard ring around trace layers reduces emission out of board. Likewise, connect to ground only at a single point—see (e). Keep separate power planes over a common ground to reduce system noise and power coupling—see (d). Keep track runs as orthogonal as possible between adjacent layers—see (c). Also avoid loop track that acts as a receiving or radiating antenna. Avoid floating conductor areas, since they act as EMI radiators. Connect them to a ground plane otherwise.

Power Supply

Avoid loops in supply lines, as shown in (h). Also decouple supply lines at local boundaries—see (i). Position high-speed circuits closest to the power supply, while placing the slowest section farthest away to reduce power plane transients—see (g). When possible, isolate individual systems on both power supply and signal lines—see (j).

Filtering Components

Position biasing and pull-up/down components close to driver/bias points. Make use of common-mode chokes between current-carrying and signal lines to increase coupling and eliminate stray fields—see (k). Place decoupling capacitors close to chip supply lines to reduce component noise and power line transients—see (l).

These tips were adapted from an Engineering Note, “Electro-Magnetic Interference and Electro-Magnetic Compatibility (EMI/EMC)” written by David B. Fancher, Inductive Products Division, Vishay Dale.

3.8 Transformers

3.8.1 Basic Operations

A basic transformer is a two-port (four-terminal) device capable of transforming an ac input voltage into a higher or lower ac output voltage. Transformers are not designed to raise or lower dc voltages, however, since the conversion mechanism relies on a changing magnetic field generated by a changing current. A typical transformer consists of two or more insulated wire coils that share a common laminated iron core. One of the coils is called the primary (containing N_P turns), while the other coil is called the secondary (containing N_S turns). A simplistic representation of a transformer is shown in Fig. 3.96, along with its schematic symbol.

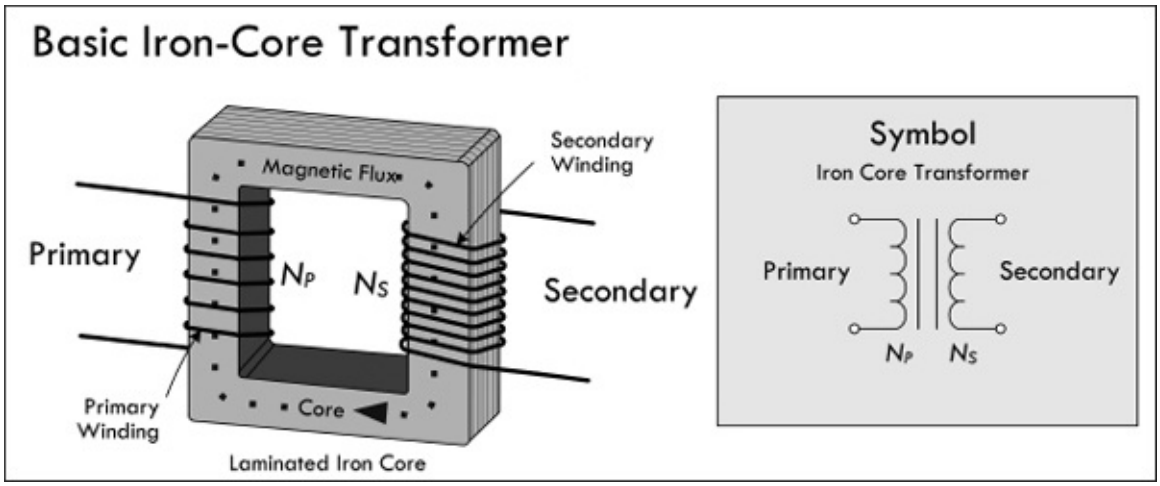


FIGURE 3.96

When an ac voltage is applied across the primary coil of the transformer, an alternating magnetic flux $\Phi_M = \int(V_{IN}/N_p)dt$ emanates from the primary, propagates through the iron-laminated core, and passes through the secondary coil. (The iron core increases the inductance, and the laminations decrease power-consuming eddy currents.) According to Faraday's law of induction, the changing magnetic flux induces a voltage of $V_S = N_S d\Phi_M/dt$, assuming there is perfect magnetic flux coupling (coefficient of coupling $k = 1$). Combining the primary flux equation with the secondary induced voltage equation results in the following useful expression:

$$V_S = V_P\left(\frac{N_S}{N_P}\right)$$

(3.1) Transformer voltage ratio

This equation says that if the number of turns in the primary coil is larger than the number of turns in the secondary coil, the secondary voltage will be smaller than the primary voltage. Conversely, if the number of turns in the primary coil is less than the number of turns in the secondary, the secondary voltage will be larger than the primary.

When a source voltage is applied across a transformer's primary terminals while the secondary terminals are open-circuited (see Fig. 3.97), the source treats the transformer as if it were a simple inductor with an impedance of $Z_P = j\omega L_P = \omega L \angle 90°$, where L_P represents the inductance of the primary coil. This means that the primary current will lag the voltage (source voltage) by 90°, and the primary current will be equal to V_P/Z_P, according to Ohm's law. At the same time, a voltage of $(N_S/N_P)\ V_P$ will be present across the secondary and will be in phase with the primary voltage or 180° out of phase, depending on the secondary coil winding direction or depending on which secondary coil end you choose as a reference (more on this in a moment).

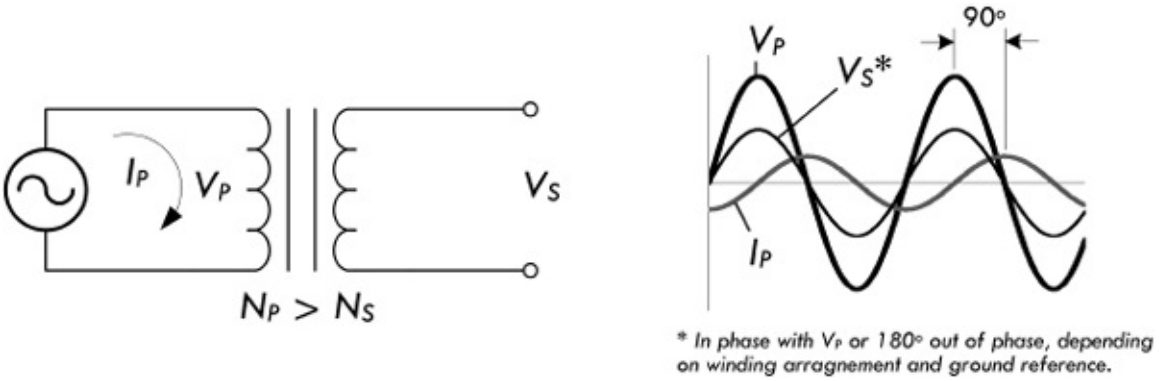


FIGURE 3.97

When there is no load attached to the secondary of a transformer, the current within the primary is called the magnetizing current of the transformer. An ideal transformer, with no internal losses, would consume no power, since the current through the primary inductor would be 90° out of phase with the voltage (in $P = IV$, I is imaginary and the "power" is imaginary or reactive). With no load in the secondary, the only losses in the transformer are associated with those losses in the iron core and losses within the primary coil wire itself.

Example 1: A transformer has a primary of 200 turns and a secondary of 1200 turns. If a 120 VAC is applied to the primary, what voltage appears across the secondary?

Answer: Rearranging Eq. 3.1,

$$V_S = V_P\left(\frac{N_S}{N_P}\right) = 120\ \text{VAC}\left(\frac{1200\ \text{turns}}{200\ \text{turns}}\right) = 720\ \text{VAC}$$

This is an example of a step-up transformer, since the secondary voltage is higher than the primary voltage.

Example 2: Using the same transformer from Example 1, flip it around so the secondary now acts as the primary. What will be the new secondary voltage?

Answer:

$$V_S = V_P\left(\frac{N_S}{N_P}\right) = 120\ \text{VAC}\left(\frac{200\ \text{turns}}{1200\ \text{turns}}\right) = 20\ \text{VAC}$$

This is an example of a step-down transformer, since the secondary voltage is lower than the primary voltage.

As you can see from the previous example, either winding of a transformer can be used as the primary, provided the windings have enough turns (enough inductance) to induce a voltage equal to the applied voltage without requiring an excessive current. The windings must also have insulation with a voltage rating sufficient for the voltage present.

Now let's take a look at what happens when you attach a load to the secondary, as shown in Fig. 3.98.

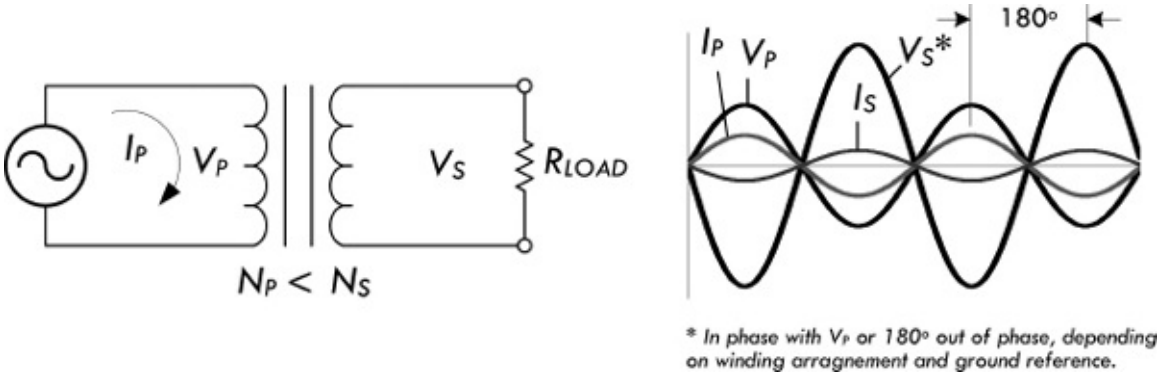


FIGURE 3.98

When a load is attached to the secondary, the secondary current sets up a magnetic field that opposes the field set up by the primary current. For the induced voltage in the primary to equal the applied voltage, the original field must be maintained. The primary must draw enough additional current to set up a field exactly equal and opposite to the field set up by the secondary current. At this point, for practical purposes, we assume that the entire primary current is a result of the secondary load. (This is close to true, since the magnetizing current will be very small in comparison with the primary load current at rated power output.)

Current Ratio

To figure out the relationship between the primary and secondary currents, consider that an ideal transformer is 100 percent efficient (real transformers are around 65 to 99 percent efficient, depending on make), and then infer that all the power dissipated by the load in the secondary will be equal to the power supplied by the primary source. With the help of the generalized power law, we get:

$$P_P = P_S$$
$$I_P V_P = I_S V_S$$

Plugging our transformer voltage equation (3.1) into the V_S term, we get:

$$I_P V_P = I_S \left(V_P \frac{N_S}{N_P} \right)$$

Eliminating the V_P from both sides, we get the following useful current relation:

$$I_P = I_S \left(\frac{N_S}{N_P} \right)$$

(3.2) Ideal transformer current ratio

Example 3: A transformer with a primary of 180 turns and a secondary with 1260 turns is delivering 0.10 A to a load. What is the primary current?

Answer: Rearranging Eq. 3.2 and solving for the primary current:

$$I_P = I_S \left(\frac{N_S}{N_P} \right) = 0.10\text{A} \left(\frac{1260 \text{ turns}}{180 \text{ turns}} \right) = 0.7 \text{ A}$$

Notice from the previous example that even though the secondary voltage is larger than the primary voltage, the secondary current is smaller than the primary current. The secondary current in an ideal transformer is 180° out of phase with the primary current, since the field in the secondary just offsets the field in the primary. The phase relationship between the currents in the windings holds true no matter what the phase difference between the current and the voltage of the secondary. In fact, the phase difference, if any, between voltage and current in the secondary will be reflected back to the primary as an identical phase difference. Note that phase, however, can be selected according to how you pull the secondary out —see the following note.

NOTE ABOUT PHASE

By now you may be a bit annoyed with the notion of phase. For example, to say that the primary voltage is out of phase with the secondary by 180° is smack in the phase to relativity. Couldn't you simply wind the secondary winding in a different direction or, more easily, simply reverse the secondary leads to get an output that is within phase? The answer is yes. It is a relativity game with the transformer pins. Figure 3.99 shows two transformers that are identical in every way except for the winding direction of the secondary. The winding A arrangement, when tested with a common ground and oscilloscope, yields in-phase voltages, while the winding in B yields voltage and currents that are 180° out of phase with the primary. To avoid confusion, a convention is used to keep track of the relative polarity between the leads. This convention makes use of what are called phase dots, which are a pair of dots: one placed on the primary side; the other placed on the secondary side. The similar placement of these dots next to the top ends of the primary and secondary windings tells you that whatever instantaneous voltage polarity is seen across the primary winding will be the same as that across the secondary winding. In other words, the phase shift from primary to secondary will be zero degrees. On the other hand, if dots on each winding of the transformer do not match up, the phase shift is 180° between primary and secondary. Of course, the dot convention only tells you which end of each winding is which, relative to the other winding(s). If you want to reverse the phase relationship, all you have to do is swap the winding connections.

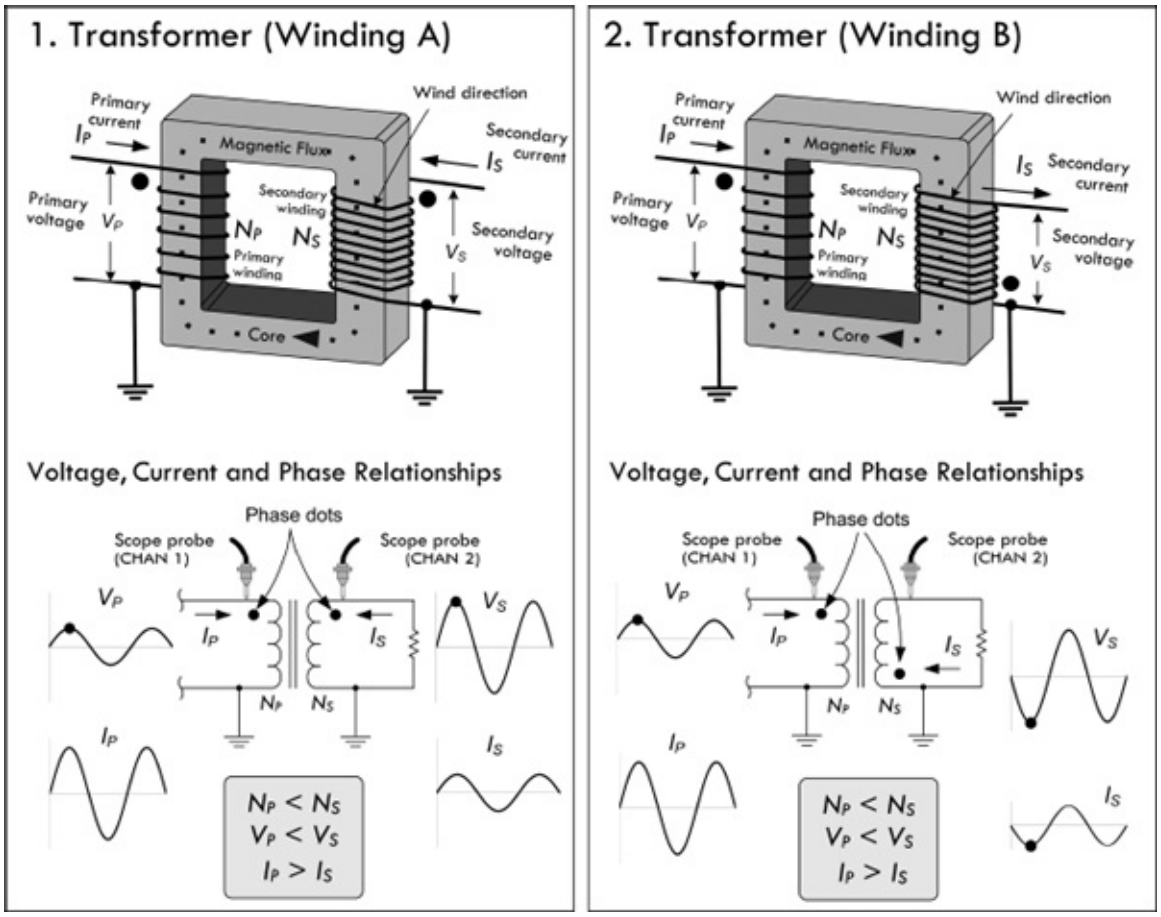


FIGURE 3.99

Power Ratio

A moment ago, when deriving the transformer current equation, we assumed that the transfer of power from primary to secondary was 100 percent efficient. However, it is important to realize that there is always some power loss in the resistance of the coils and in the iron core of the transformer. This means that the power taken from the source is greater than the power used in the secondary. This can be stated by the following expression:

$$P_S = n \times P_P$$

(3.3) Efficiency factor

where P_S is the power output from the secondary, P_P is the power input to primary and n is the efficiency factor. The efficiency n is always less than 1. It is usually expressed as a percentage—for example, 0.75 represents an efficiency of 75 percent.

Example 4: What is the power input to the primary if a transformer has an efficiency of 75 percent and its full load output at the secondary is 100 W?

Answer: Rearranging Eq. 3.3,

$$P_P = \frac{P_S}{n} = \frac{100\ \text{W}}{0.75} = 133\ \text{W}$$

Transformers are typically designed to have highest efficiency at the manufacturer's rated outputs. Above or below the rated output, the efficiency drops. The amount of power a transformer can handle depends on

its own losses (heating of wire and core, etc.). Exceeding the rated power of a transformer can lead to wire meltdown or insulation breakdown. Even when the load is purely reactive, the transformer will still be generating heat loss due to internal resistance of the coils and losses in the core. For this reason, manufacturers also specify a maximum volt-amp rating, or VA-rating, that should not be exceeded.

Impedance Ratio

Using ac Ohm's law, $I_P = V_P/Z_P$, and assuming an ideal transformer, where power from the primary is 100 percent transferred to secondary, we can come up with an equation relating the primary and secondary impedances:

$$P_P = P_S$$

$$I_P V_P = I_S V_S$$

$$\frac{V_P^2}{Z_P} = \frac{V_S^2}{Z_S} \rightarrow (\text{plug in } V_S = V_P(N_S/N_P) \rightarrow \frac{V_P^2}{Z_P} = \frac{V_P^2 (N_S/N_P)^2}{Z_S}$$

Canceling the primary voltage terms, you get the following useful expression:

$$Z_P = Z_S \left(\frac{N_P}{N_S}\right)^2$$

(3.4) Transformer impedance ratio

where Z_P is the impedance looking into the primary terminal from the power source, and Z_S is the impedance of the load connected to the secondary. Figure 3.100 shows an equivalent circuit.

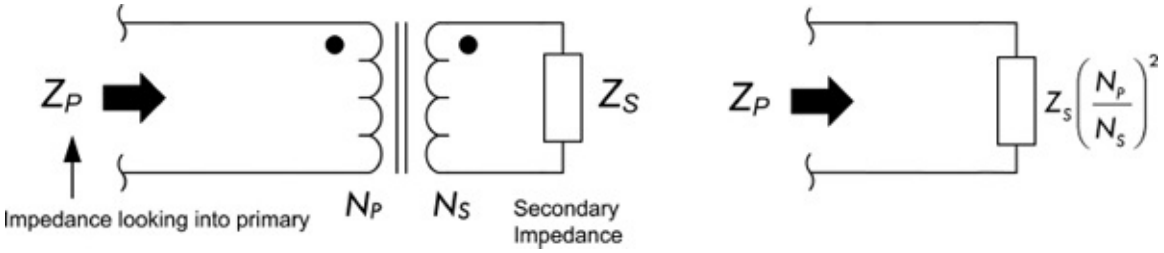


FIGURE 3.100

If the load impedance in the secondary increases, the impedance looking into the primary (from the source's point of view) will also increase in a manner that is proportional to the ratio of the turns squared.

Example 5: A transformer has a primary with 500 turns and a secondary with 1000 turns. What is the primary impedance if a 2000-Ω load impedance is attached to the secondary?

Answer: Using Eq. 3.4:

$$Z_P = 2000\ \Omega \left(\frac{500\ \text{turns}}{1000\ \text{turns}}\right)^2 = 2000\ \Omega(0.5)^2 = 500\ \Omega$$

As you can see, by selecting the proper turns ratio, the impedance of a fixed load can be transformed to any desired value (ideally). If transformer losses can be neglected, the transformed (reflected) impedance

has the same phase angle as the actual load impedance. Hence, if the load is purely resistive, the load presented by the primary to the power source will also be pure resistance. If the load impedance is complex (e.g., inductance and capacitance are thrown in so that load current and voltage are out of phase with each other), then the primary voltage and current will show the same phase angle.

In electronics, there are many instances where circuits require a specific load resistance (or impedance) for optimum performance. The impedance of the actual load dissipating power may differ widely from the impedance of the source. In this case, a transformer can be used to change the actual load into an impedance of desired value. This is referred to as impedance matching. We can rearrange Eq. 3.4 and get:

$$\frac{N_P}{N_S} = \sqrt{\frac{Z_P}{Z_S}} \tag{3.5}$$

where N_P/N_S is the required turns ratio—primary to secondary, Z_P is the primary impedance required, and Z_S is the impedance of the load connected to the secondary.

Example 6: An amplifier circuit requires a 500-Ω load for optimum performance, but is to be connected to an 8.0-Ω speaker. What turns ratio, primary to secondary, is required in the coupling transformer?

Answer:

$$\frac{N_P}{N_S} = \sqrt{\frac{Z_P}{Z_S}} = \sqrt{\frac{500\ \Omega}{8\ \Omega}} = 8$$

Hence, the primary must have eight times as many turns as the secondary.

Knowing what to set the primary count at depends on low internal losses and leakage current and making sure that the primary has enough inductance to operate with low magnetizing current at the voltage applied to the primary.

Example 7: What are the load impedances "seen" by the voltage sources in Fig. 3.101?

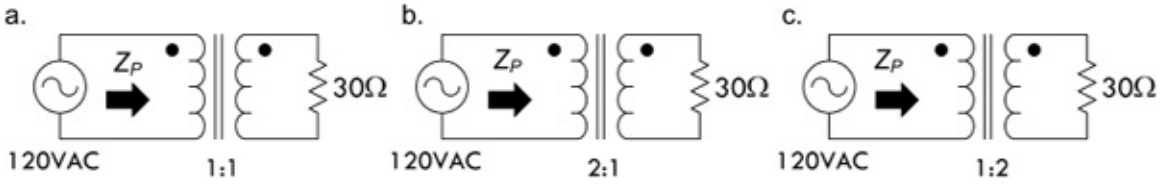


FIGURE 3.101

Answer: (a) 30 Ω, (b) 120 Ω, (c) 8 Ω.

Example 8: If a step-up transformer has a turns ratio of 1:3, what are the voltage ratio, current ratio, and impedance ratio? Assume ratios are given in the form "primary: secondary."

Answer: Voltage ratio is 1:3, current ratio is 3:1, impedance ratio is 1:9.

Transformer Gear Analogy

It is often helpful to think of transformers as gearboxes. For example, in the gearbox analogy in Fig. 3.102, the primary winding is analogous to the input shaft (where the motor is attached) and the secondary winding is analogous to the output shaft. Current is equivalent to shaft speed (rmp) and voltage is equivalent to torque. In a gearbox, mechanical power (speed multiplied by torque) is constant (neglecting losses) and is equivalent to electrical power (voltage times current), which is also constant. The gear ratio is equivalent to the transformer step-up or step-down ratio. A step-up transformer acts like a reduction gear (in which mechanical power is transferred from a small, rapidly rotating gear to a large, slowly rotating gear): it trades current (speed) for voltage (torque), by transferring power from a primary coil to a secondary coil having more turns. A step-down transformer acts similarly to a multiplier gear (in which mechanical power is transferred from a large gear to a small gear): it trades voltage (torque) for current (speed), by transferring power from a primary coil to a secondary coil having fewer turns.

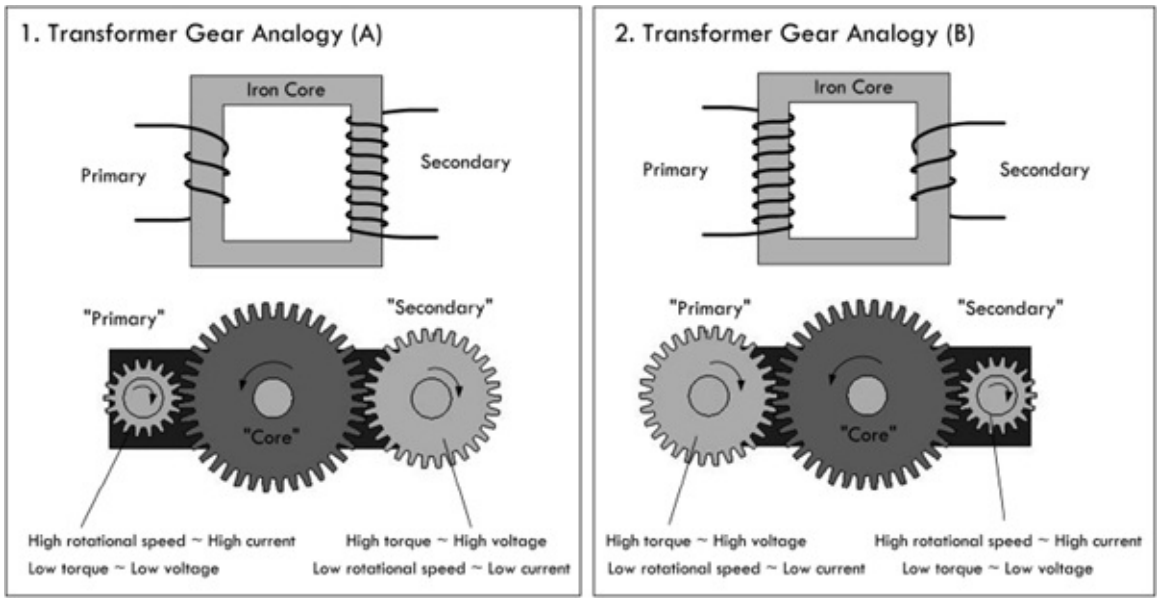


FIGURE 3.102

Center-Tap Transformers

Rarely do you see transformers in the real world with just four leads—two for the primary and two for the secondary. Many commercial transformers employ center taps. A center tap is simply an electrical connection that is made somewhere between the two ends of a transformer winding. By using a center tap, it is possible to utilize only a fraction of the winding voltage. For example, in Fig. 3.103, a transformer's secondary is center-tapped midway between its winding, yielding two output voltages V_{S1} and V_{S2}. If we place a ground reference on the center tap (it is treated now as a common), we see the voltages in terms of phase, as shown in the example circuit in Fig. 3.103. In this case, the two secondary voltages are equal because we assumed that the number of turns on either end of the center tap were the same. In general, the secondary voltages are determined by the turns ratio.

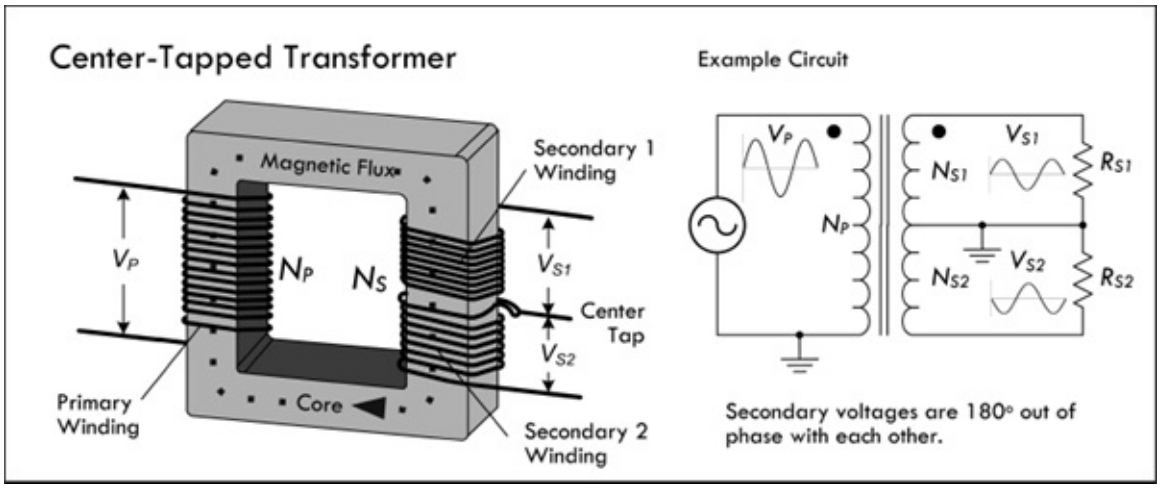


FIGURE 3.103

Center taps can be placed on both the primary side and the secondary side, with multiple taps on either side. For example, a typical power transformer has several secondary windings, each providing a different voltage. Figure 3.104 shows a schematic of a typical power supply transformer. It is possible to join pins with a jumper to get the desired voltage ratios across other pins. Manufacturers will provide you with the voltages between the various tap points, usually specifying CT as the center-tap voltage. Center taps provide flexibility in design and allow varying outputs, which you implement by incorporating switches, for example. We'll see how a center-tap transformer is used to split incoming 240 VAC for the main into two 120-VAC legs within the circuit breaker of your house, and we'll also discover how full-wave center-tap rectifier circuits are used in building dc power supplies.

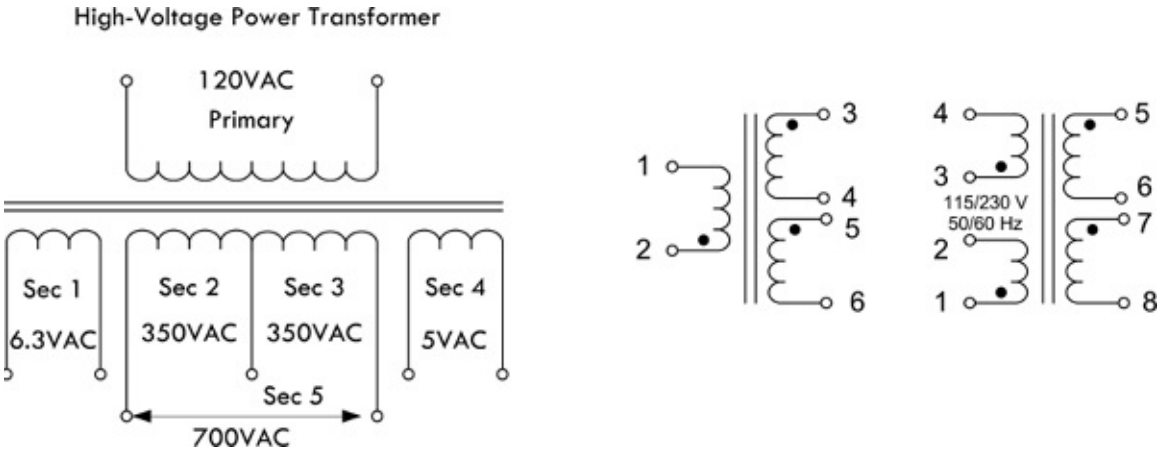


FIGURE 3.104

Real Transformer Characteristics

A perfect or ideal transformer has a primary-to-secondary coupling coefficient of 1. This means that both coils link with all the magnetic flux lines, so that the voltage induced per turn is the same for both coils. This also means that the induced voltage per turn is the same for both primary and secondary coils. Iron core transformers operating at low frequencies come close to being ideal. However, due to various imperfections, such as eddy current, hysteresis losses, internal coil resistance, and skin effects at higher frequencies, this isn't quite true.

In real transformers, not all of the magnetic flux is common to both windings. Flux not associated with linkage is referred to as *leakage flux* and is responsible for a voltage of self-induction. There are small amounts of leakage inductance associated with both windings of a transformer. Leakage inductance acts in exactly the same manner as an equivalent amount of ordinary inductance inserted in series with the circuit. The reactance associated with leakage inductance is referred to as *leakage reactance*, which varies with

transformer build and frequency. Figure 3.105 shows a real-life model of a transformer including leakage reactances for both primary and secondary coils, namely, X_{L1} and X_{L2}. When current flows through a leakage reactance, there is an associated voltage drop. The voltage drop becomes greater with increasing current and increases as more power is taken from the secondary.

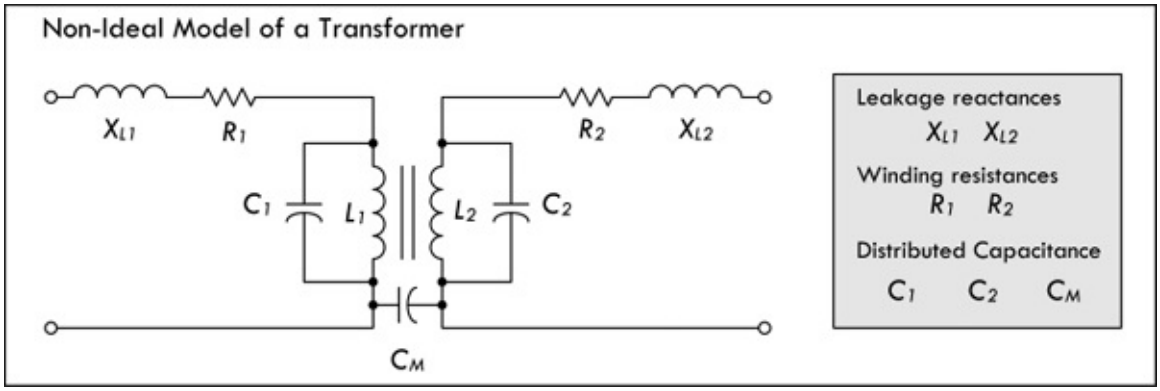


FIGURE 3.105

The internal resistances of a transformer's windings R_1 and R_2 also result in voltage drop when there is current flow. Although these voltage drops are not in phase with those caused by leakage reactance, together they result in a lower secondary voltage under load than is indicated by the transformer turns ratio formula.

Another nonideal characteristic of transformers is stray capacitance. An electric field exists between any two points having a different voltage. When current flows through a coil, each turn has a slightly different voltage than its adjacent turns. This results in capacitance between turns and is modeled by C_1 and C_2 in Fig. 3.105. A mutual capacitance C_M also exists between the primary and secondary windings for the same reason. It is also possible for transformer windings to exhibit capacitance relative to nearby metal, such as a chassis, shield, or even the core itself.

Stray capacitance tends to have little influence in power and audio transformers, but becomes influential as the frequencies increase. In RF applications where transformers are used, the stray capacitance can resonate with either the leakage reactance or, at lower frequencies, the winding reactances, especially under very light or zero-ohm loads. In the frequency region around resonance, transformers do not exhibit behavior as described by the previous transformer equations.

Iron core transformers also experience losses with hysteresis and eddy current, as was discussed in Sec. 2.24. These losses, which add to the required magnetizing current, are equivalent to adding a resistance in parallel to R_1 in Fig. 3.105.

TRANSFORMER PRECAUTIONS

There are three basic rules to observe when using a transformer. First, never apply a voltage that is greatly in excess of the transformer winding ratings. Second, never allow a significant direct current to flow through any winding not designed to handle it. Third, don't operate the transformer at a frequency outside the range specified by the manufacturer. Applying a voltage of, say, 120 VAC to a secondary in hopes of achieving 1200 VAC at the primary is a bad idea—expect smoke and combustion, accompanied by insulation failure. Similar results can be expected with excessive dc current through the primary. In terms of frequency, a 60-Hz transformer driven at 20 Hz will draw too much magnetizing current and will run dangerously hot.

3.8.2 *Transformer Construction*

Cores

Transformers used for power and audio frequencies have cores made of many thin laminations of silicon steel. The laminations reduce eddy currents, as discussed in Sec. 2.24. A typical laminated core is made from E-shaped and I-shaped pieces, sandwiched together, as shown in Fig. 3.106. Transformers made from these cores are therefore often referred to as EI transformers.

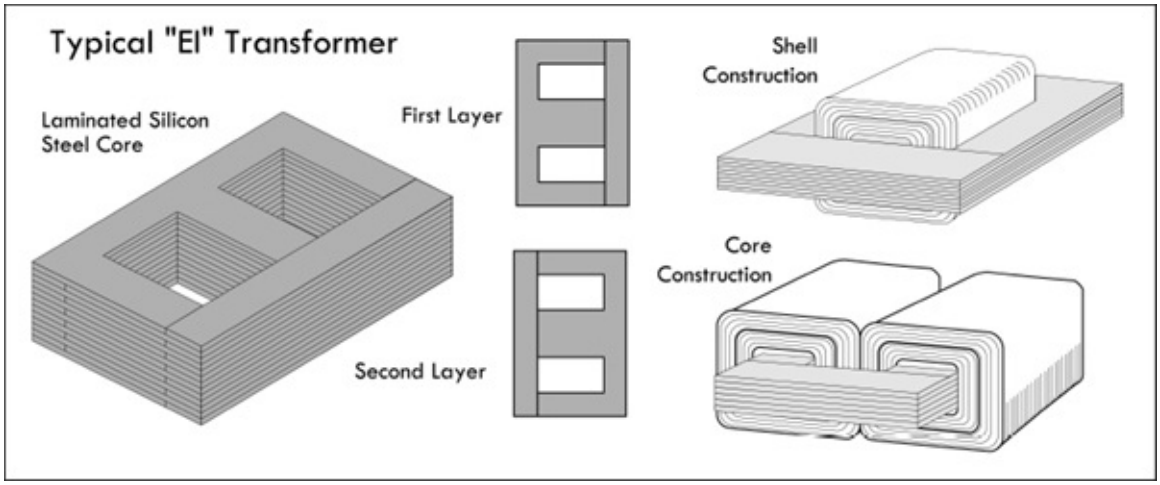


FIGURE 3.106

Two common core shapes in use are shown in Fig. 3.106. In the shell construction, both the primary and the secondary windings are wound around the same inner leg, while in the core construction, primary and secondary windings are wound on separate legs. The core construction is often implemented to minimize capacitive effects between primary and secondary windings, or when one winding is to be operated at very high voltage. The size, shape, and type of core material, as well as the frequency range, influence the required number of turns in each winding. In most transformers, the coils are wound in layers, with a sheet of special paper insulation placed between each layer. A thicker insulation is used between adjacent coils and between the core and the first coil.

Powdered iron cores, with their low eddy current characteristics, are used in transformers that operate above mains frequencies (60 Hz) up to several kilohertz. These cores have a very high permeability and thus provide decent stepping capability for their size. Transformers that are used in even higher-frequency applications, such as RF, often contain cores made from nonconductive magnetic ceramic materials or ferrites.

A common core shape for powdered iron and ferrite core transformers is the toroid, as shown in Fig. 3.107a. The closed ring shape of the toroid eliminates air gaps inherent in the construction of an EI core. The primary and secondary coils are often wound concentrically to cover the entire surface of the core. Ferrite cores are used at higher frequencies, typically between a few tens of kilohertz to a megahertz. In general, toroidal transformers are more efficient (around 95 percent) than cheaper laminated EI transformers; they are more compact (about half the size), weigh less (about half), have less mechanical hum (making them superior in audio applications), and have lower off-load losses (making them more efficient in standby circuits).

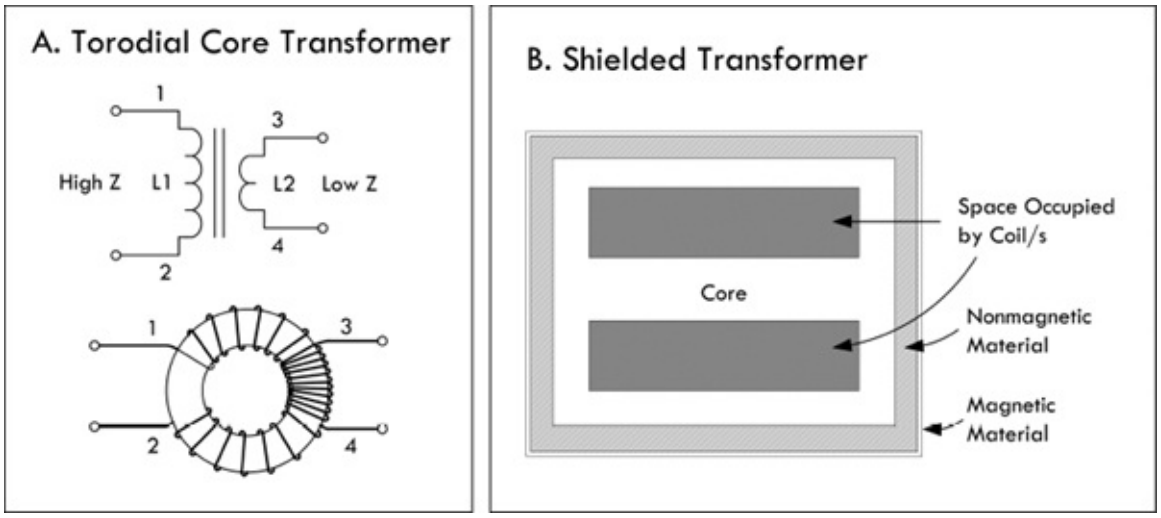


FIGURE 3.107

Shielding

To eliminate mutual capacitance between windings within a transformer, an electrostatic shield is often placed between the windings. Some transformers may incorporate a magnetic shield, as shown in Fig. 3.107b. The magnetic shield helps prevent outside magnetic fields (interference) from inducing currents within the inner windings. The shield also helps prevent the transformer from becoming an interference radiator itself.

Windings

For small-power and signal transformers, the windings are made from solid wire copper, insulated typically with enamel; sometimes additional insulation is used for safety. Larger-power transformers may be wound with copper or aluminum wire, or even strip conductors for very heavy current; in some cases multistrand conductors are used to reduce skin effect losses. High-frequency transformers operating in the kilohertz range often have windings made of Litz wire to minimize skin effects. For signal transformers, the windings may be arranged in a way to minimize leakage inductance and stray capacitance in order to improve high-frequency response.

3.8.3 Autotransformers and Variable Transformers

An autotransformer is similar to a standard transformer; however, it uses only one single coil and a center tap (or taps) to make primary and secondary connections. See Fig. 3.108. As with standard transformers, autotransformers can be used to step up or step down voltages, as well as match impedances; however, they will not provide electrical isolation like a standard transformer, since their primary and secondary are on the same coil—there is no electrical isolation between the two coils.

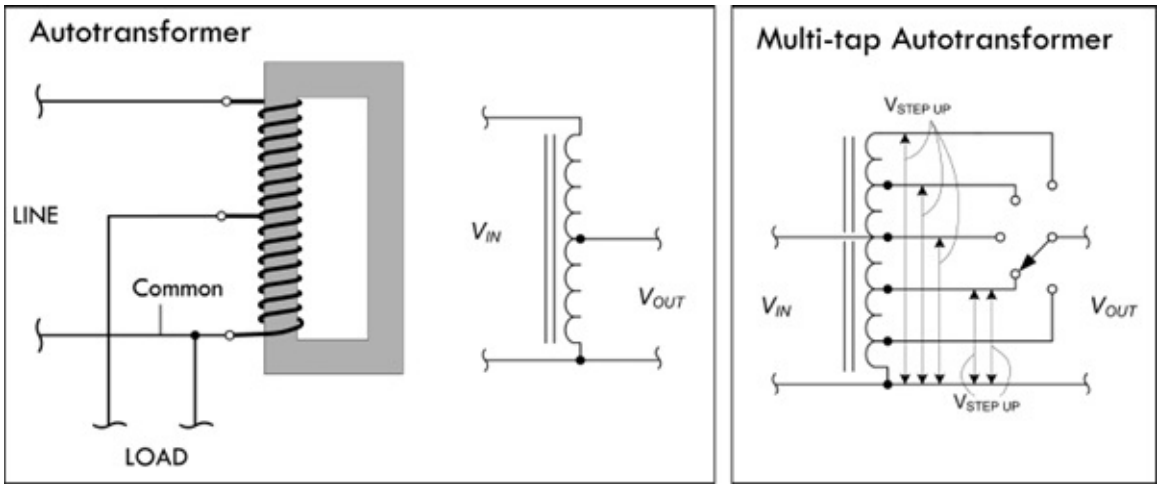


FIGURE 3.108

Although an autotransformer has only one winding, the laws of induction that were used with a standard transformer to step up and step down voltage can be applied just as well. This also applies to principles of current and impedance as a function of the number of winding turns. In Fig. 3.108, the current in the common winding is the difference between the line current (primary current) and the load current (secondary current), since these currents are out of phase. Hence, if the line and load currents are nearly equal, the common section of the winding may be wound with comparatively small wire. The line and load currents will be equal only when the primary (line) and secondary (load) voltages are close in magnitude.

Autotransformers are often used in impedance-matching applications. They are also frequently used for boosting or reducing the power-line voltage by relatively small amounts. Figure 3.108 shows a switch-stepped autotransformer whose output voltage can be set to any number of values determined by the switch contact position.

A variable transformer or Variac (commercial name) is similar to the switch-stepped autotransformer in Fig. 3.108; however, it has a continuous wiper action along a circular coil, as shown in Fig. 3.109. A Variac acts like an adjustable ac voltage source. Its primary is connected to the hot and neutral of the 120-V line voltage, while the secondary leads consist of the neutral and an adjustable wiper that moves along the single core winding.

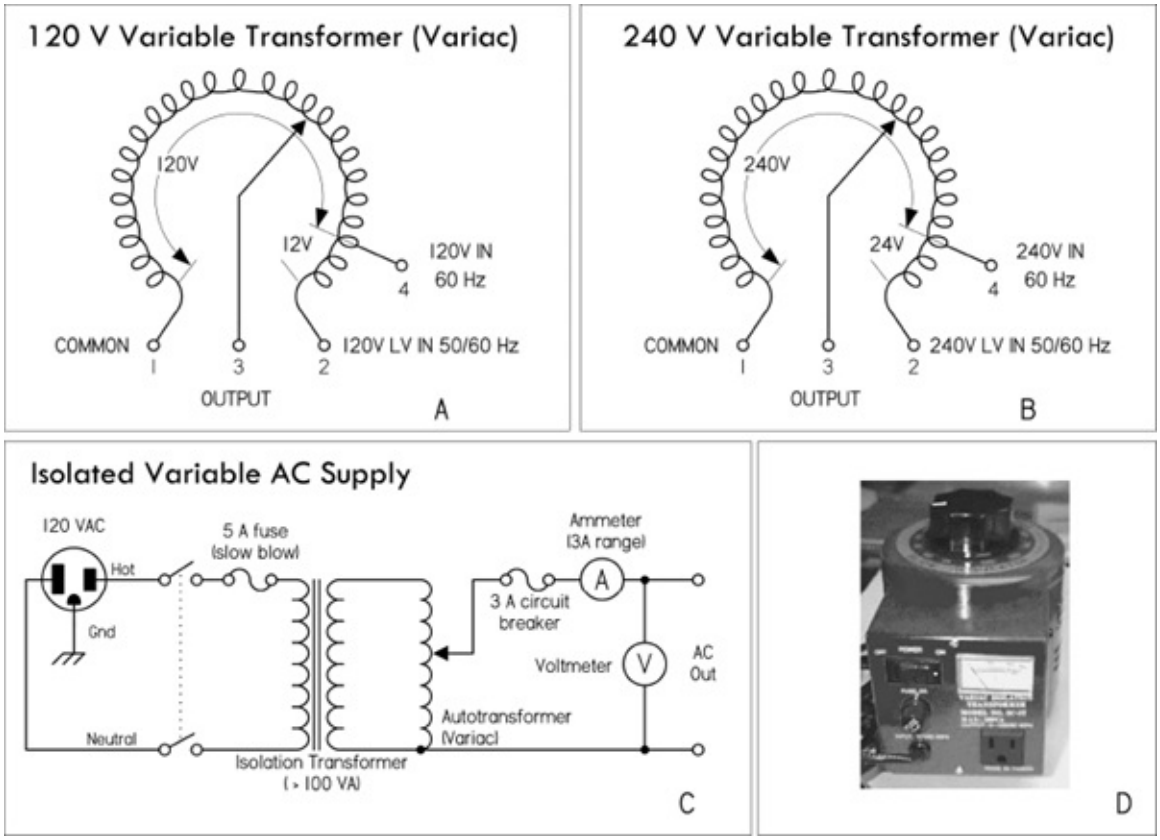


FIGURE 3.109 (a) Nonisolated 120-V Variac whose output voltage is varied by rotating a wiper. (b) Nonisolated 240-V Variac. (c) A homemade variable ac supply with isolation protection provided by means of an isolation transformer. (d) ac power supply that houses an isolation transformer, Variac, switch, fuse, ac outlet, and meter.

Being able to adjust the line voltage is a very useful trick when troubleshooting line-power equipment, where the fuse instantly blows at normal line voltage. Even without a fuse blowing, troubleshooting at around 85 V may reduce the fault current.

It is important to note that a Variac by itself does not provide isolation protection like a standard transformer, since the primary and secondary shared a common winding. It is therefore important, if you plan to work on ungrounded, "hot chassis" equipment, that you place an isolation transformer before the Variac—never after it. If you don't, shock hazards await. Figure 3.109c shows a schematic of such an arrangement. It includes a switch and fuse protection, as well as current and voltage meters, all of which create an adjustable, fully isolated ac power source.

To avoid the hassle of cascading a Variac and isolation transformer together, simply get an ac power supply that houses both elements together in one package—see Fig. 3.109d.

Boosting and Bucking

We just saw how autotransformers are used in applications requiring a slight boost or reduction in voltage to a load. It is possible to accomplish the same effect by using a normal (isolated) transformer with just the right primary/secondary turns ratio. There is still another alternative—use a step-down configuration with secondary winding connected in a series-aiding ("boosting") or series-opposing ("bucking") configuration, as shown in Fig. 3.110.

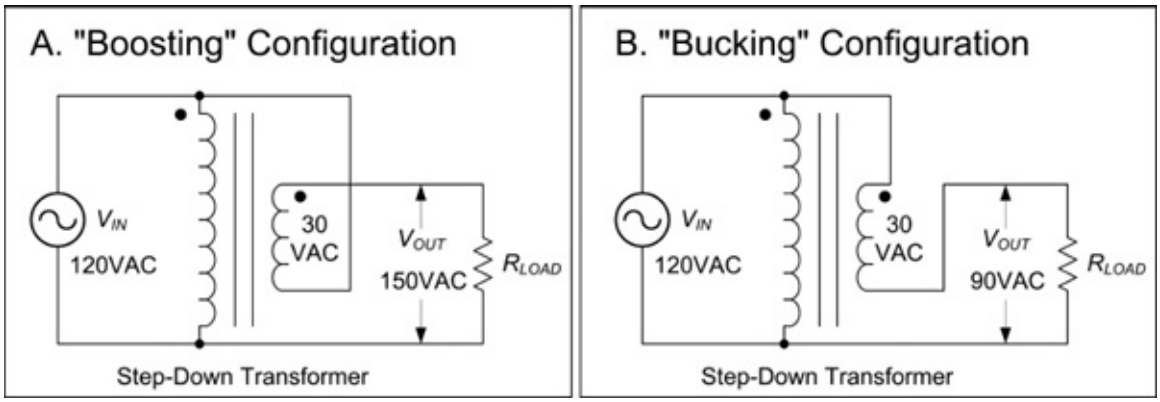


FIGURE 3.110

In the boosting configuration, the secondary coil's polarity is oriented so that its voltage directly adds to the primary voltage. In the bucking configuration, the secondary coil's polarity is oriented so that its voltage directly subtracts from the primary voltage. An autotransformer does the same job as the boosting and bucking functions displayed here, but using only a single winding, making it cheaper and lighter to manufacture.

3.8.4 Circuit Isolation and the Isolation Transformer

Transformers perform an important role in isolating one circuit from another. Figure 3.111 shows an example application that uses a transformer to isolate a load connected to an ac outlet. In this application, there is no need to step up or step down the voltage, so the transformer has a 1:1 winding ratio. Such a transformer is referred to as an *isolation transformer*. In Fig. 3.111, a mains isolation transformer is used to isolate a load from the source, as well as provide ground fault protection. An isolation transformer should be used whenever you work on nongrounded equipment, with no input isolation, such as switch-mode power supplies.

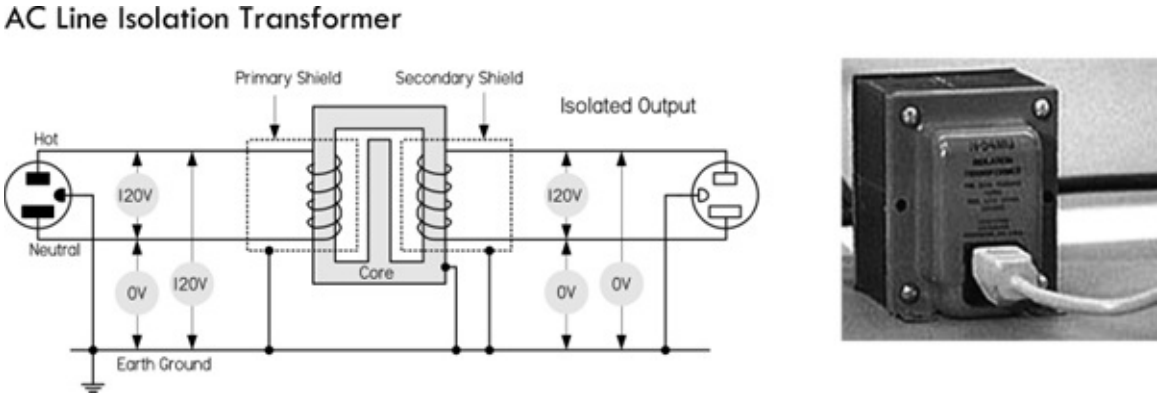


FIGURE 3.111

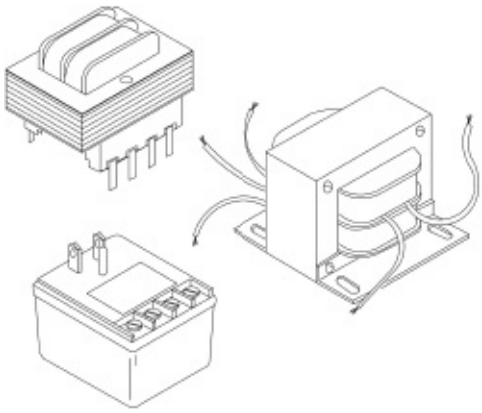
In your home wiring, the neutral (white) and the ground (green) connections are tied together at the main junction box, so they are basically at the same potential—0 V, or earth ground. If you accidentally touch the hot wire while being in contact with a grounded object, current will pass through your body and give you a potentially fatal shock. With an isolation transformer, the secondary winding leads act as a 120-V source and return, similar to the mains' hot and neutral, but with an important difference. Neither the secondary source nor return runs are tied to earth ground! This means that if you touch the secondary source or return while being in contact with a grounded object, no current will flow through your body. Current only wants to pass between the secondary source and return runs. (Note that all transformers provide isolation, not just line isolation transformers. Therefore, equipment with input power transformers already have basic isolation protection built in. Isolation transformers used for laboratory

work are explained in greater detail in Sec. 7.5.12.)

Isolation transformers are also typically constructed with two isolated Faraday shields between the primary and secondary windings. The use of the two shields diverts high-frequency noise, which would normally be coupled across the transformer to ground. Increasing the separation between the two Faraday shields minimizes the capacitance between the two and, hence, the coupling of noise between the two. Therefore, the isolation transformer acts to clean up line power noise before being delivered to a circuit.

3.8.5 Various Standard and Specialized Transformers

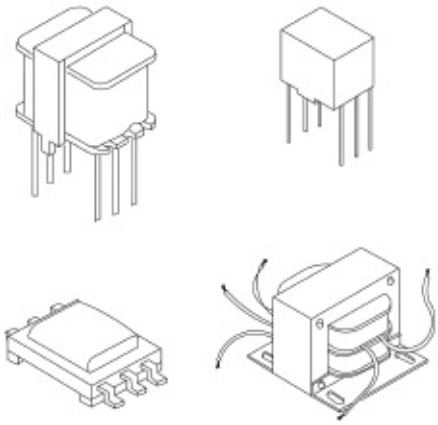
Power Transformers



These transformers are used primarily to reduce line voltage. They come in a variety of different shapes, sizes, and primary and secondary winding ratios. They often come with taps and multiple secondary windings. Color-coded wires are frequently used to indicate primary and secondary leads (e.g., black wires for primary, green for secondary, and yellow for tap lead is one possibility). Other transformers use pins for primary, secondary, and tapped leads, allowing them to be mounted on a PC board. You can also find transformers in wall-mount packages that plug directly into an ac outlet, with screw-in terminals as secondary and tapped leads.

FIGURE 3.112

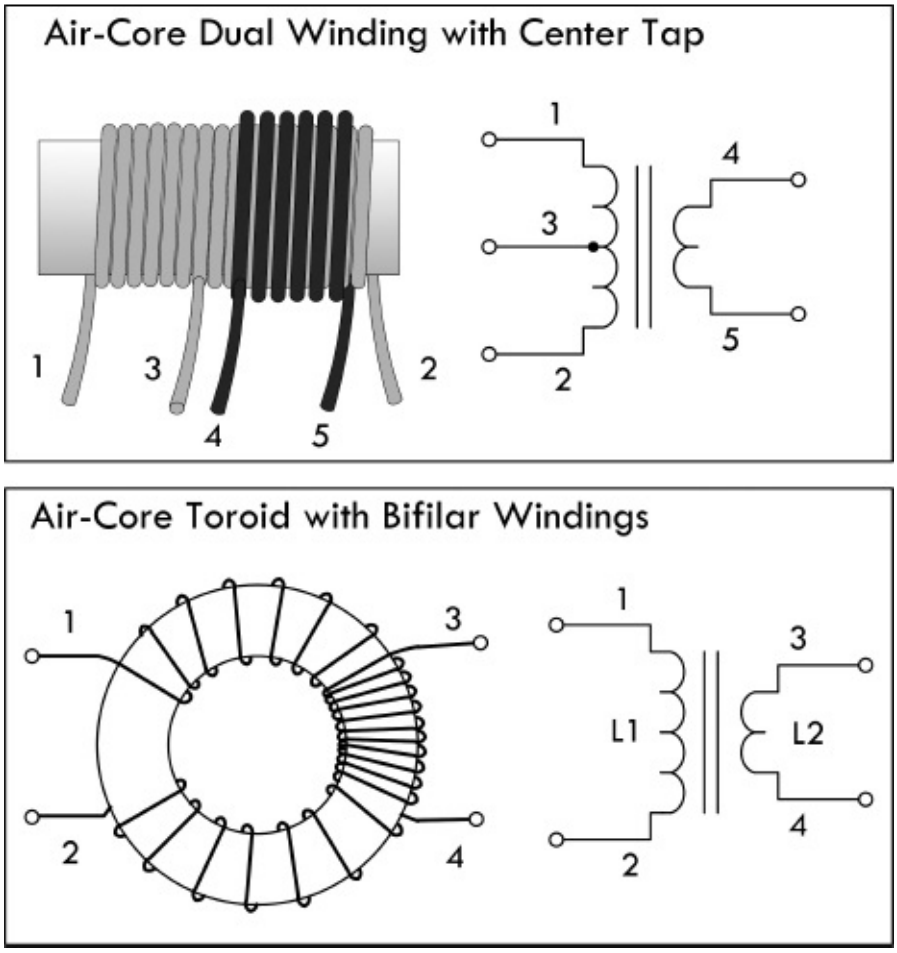
Audio Transformers



Audio transformers are used primarily to match impedances between audio devices (e.g., between microphone and amplifier or amplifier and speaker), though they can be implemented in other ways as well. They work best at audio frequencies from 20 Hz to 20 kHz. Outside this range they will reduce or block signals. They come in a variety of shapes and sizes and typically contain a center tap in both the primary and secondary windings. Some come with color-coded wires to specify leads, while other audio transformers have pinlike terminals that can be mounted on PC boards. Spec tables provide dc resistance values for primary and secondary windings to help you select the appropriate transformer for the particular matching application. Besides performing simple impedance matching, audio transformers can be used to step up or step down a signal voltage, convert a circuit from unbalanced to balanced and vice versa, block dc current in a circuit while allowing ac current to flow, and provide basic isolation between one device and another. Note that audio transformers have a maximum input level that cannot be exceeded without causing distortion. Also, audio transformers cannot step up a signal by more than about 25 dB when used in typical audio circuits. Because of this, an audio transformer cannot be substituted for a microphone preamp. If more than 25 dB of gain is required, an active preamp must be used instead of a transformer.

FIGURE 3.113

Air Core RF Transformers



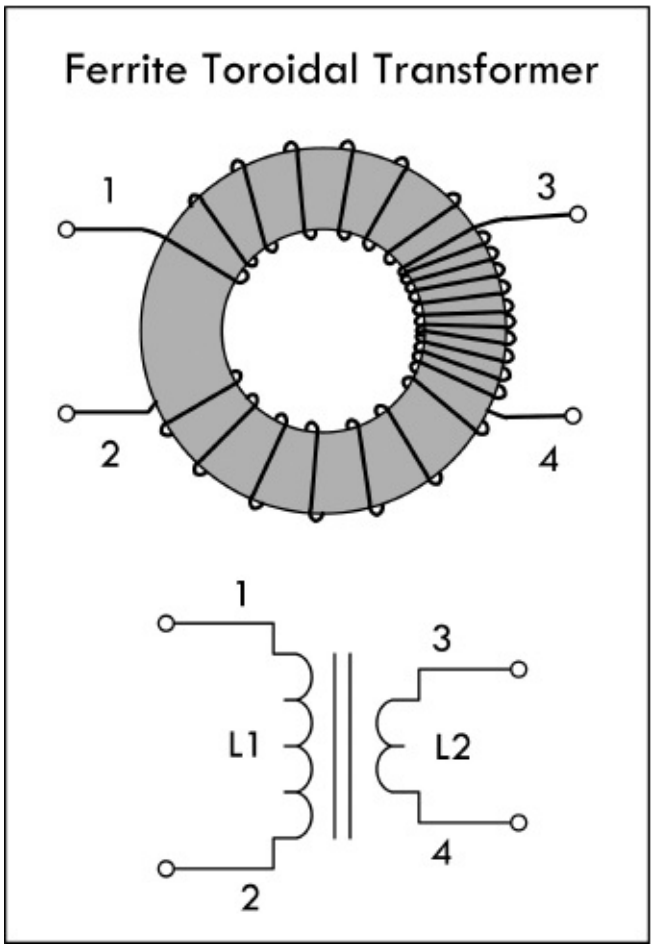
Air core transformers are special devices used in radio-frequency circuits. (They are used for RF coupling, such as antenna tuning and impedance matching.) Unlike steel or ferrite core transformers, the core is made from a nonmagnetic form, usually a hollow tube of plastic. The degree of coupling between windings in an air core transformer is much less than that of a steel core transformer; however, there are no losses associated with eddy currents, hysteresis, saturation, and so on, as is the case with magnetic cores. This becomes critically important in RF applications—at high frequencies, steel core transformers experience significant losses. Toroidal air core transformers aren't common nowadays, except in VHF (very high frequency) work. Today, special coupling networks and RF powdered-iron and ferrite toroids have generally replaced air cores, except in situations where the circuit handles very high power or the coils must be very temperature stable.

FIGURE 3.114

3.8.6 Transformer Applications

There are three principal uses for transformers: to transform voltages and currents from one level to another, to physically isolate the primary circuit from the secondary, and to transform circuit impedances from one level to another. Here are some examples in action.

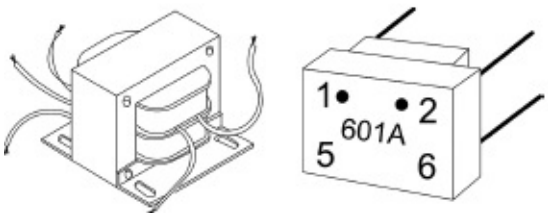
Ferrite and Powdered-Iron Toroidal Transformers



Toroidal ferrite and powdered-iron transformers are used from a few hundred hertz well into the UHF spectrum. The principal advantage of this type of core is self-shielding and low losses due to eddy currents. The permeability/size ratio is also very large, making them compact devices requiring fewer coil turns than traditional transformers. The most common ferrite toroid transformer is the conventional broadband transformer. Broadband transformers provide dc isolation between the primary and secondary circuits. The primary of a step-down impedance transformer is wound to occupy the entire core, with the secondary wound over the primary, as shown in Fig. 3.115. This style of transformer is frequently used in impedance matching. In standard broadcast radio receivers, these transformers operate in a frequency range from 530 to 1550 kHz. In shortwave receivers, RF transformers are subjected to frequencies up to about 20 MHz; in radar, it approaches upward of 200 MHz.

FIGURE 3.115

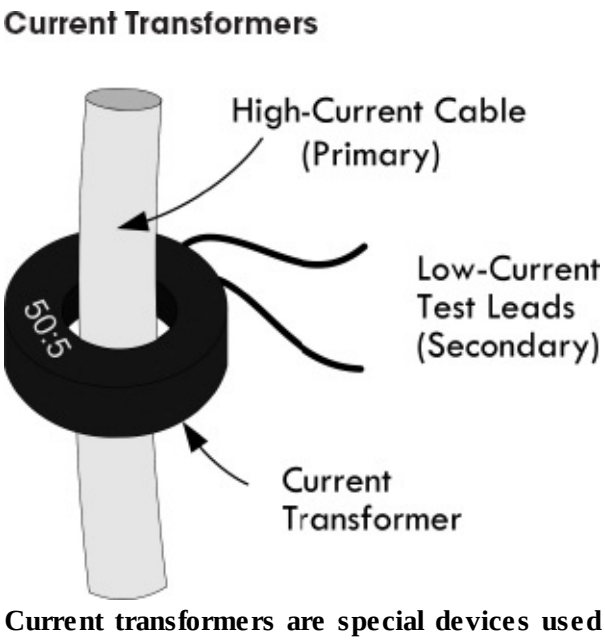
Pulse and Small Signal Transformers



Pulse transformers are special transformers optimized for transmitting rectangular electrical pulses—ones with fast rise and fall times and constant amplitude. A small signal transformer is a small version of a pulse transformer. These devices are used in digital logic and telecom circuits, often for matching logic drivers to transmission lines. Medium-size power versions are used in power-control circuits such as camera flash controllers, while larger-power versions are used in electrical power distribution to interface low-voltage control circuitry with high-voltage power semiconductive gates, such as TRIACs, IGBTs, thyristors, and MOSFETs. Special high-voltage pulse transformers are used to generate high-power pulses for radar, particle accelerators, or other pulsed power applications.

To minimize pulse shape distortion, a pulse transformer requires very low leakage inductance and distributed capacitance, and a high open-circuit inductance. Low coupling capacitance is also important in power-pulse transformer applications to protect circuitry on the primary side from high-power transients created by load.

FIGURE 3.116



Current transformers are special devices used primarily to measure larger currents that would be too dangerous to measure with an ammeter. They are designed to provide a current in their secondary that is proportional to the current flowing in the primary. A typical current transformer resembles a toroidal core inductor with many secondary windings. The primary coil consists of simply passing a single cable-to-be-measured (insulated) through the center of the toroid. The output current through the secondary is many times smaller than the actual current through the cable (primary). These transformers are specified by their input and output current ratio (400:5, 2000:5, etc.). Current transformers designed for electrical supply applications are designed to drive 5-A (full-scale) meters. There are also wideband current transformers used to measure high-frequency waveforms and pulsed currents.

FIGURE 3.117

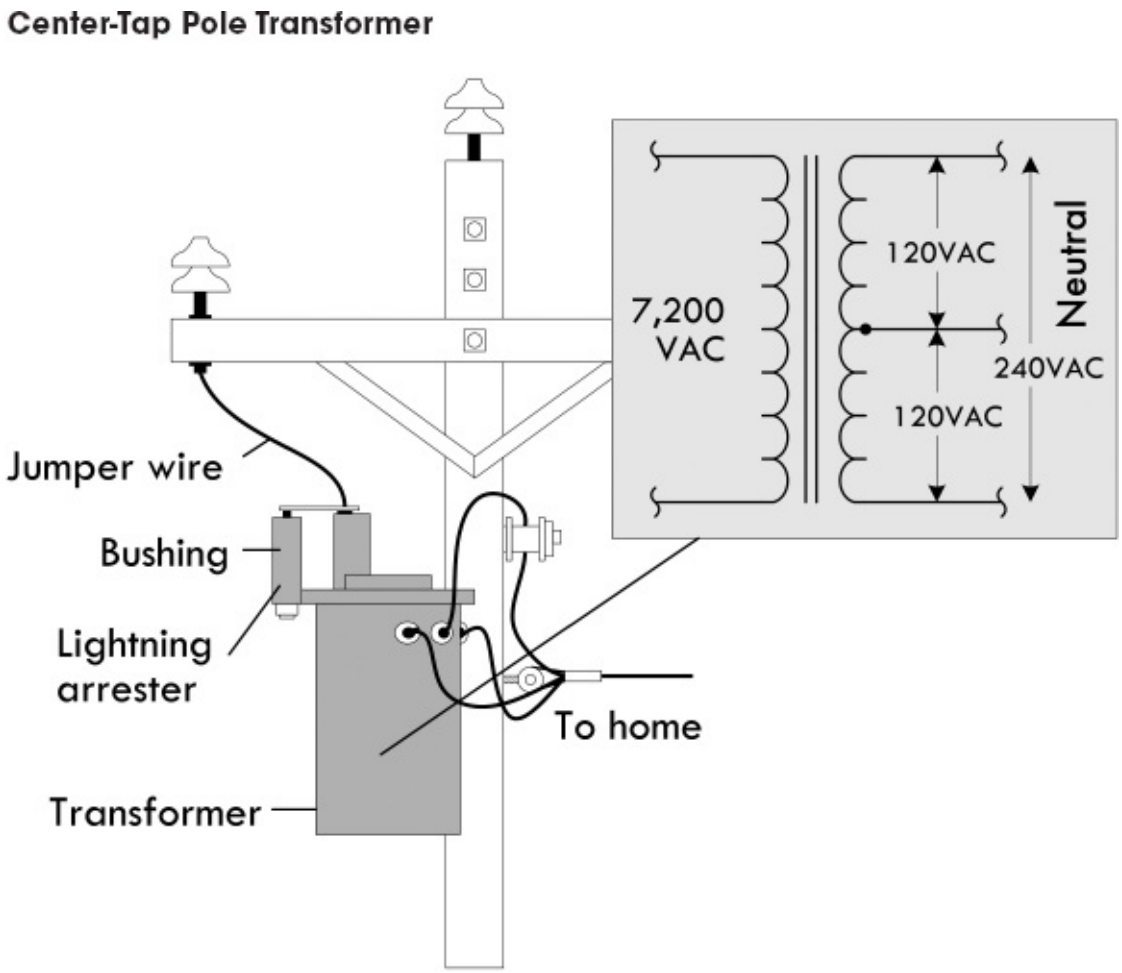


FIGURE 3.118 In the United States, main power lines carry ac voltages upward of several thousand volts. A center-tap pole transformer is used to step down the line voltage to 240 V. The tap then acts to break this voltage up into 120-V portions. Small appliances, such as TVs, lights, and hair dryers can use either the top line and the neutral line or the bottom line and the neutral line. (The neutral is grounded to a ground rod through a link between neutral and ground buses in the breaker box.) Larger

appliances, such as stove terminal as well. See App

Transformer Used for Lands

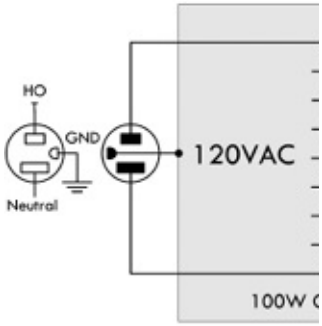


FIGURE 3.119 Here a case don't care if the vo output. Most commercial with multiple outputs. T lamps if there is an anticip the total load should n transformer should not b lamp dimming.

Step-Down Transformer fo

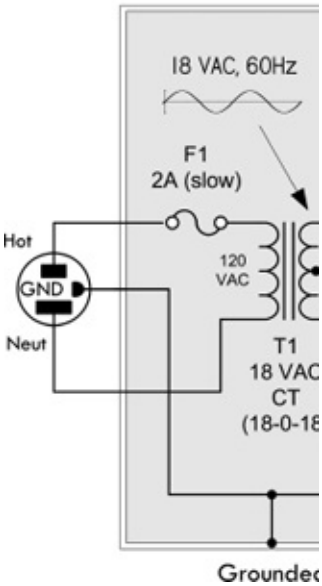


FIGURE 3.120 Transf transformer is used to cr coil end and the center section and eliminate pos voltages appear dc. The for more details.

CHAPTER 4

Semiconductors

4.1 Semiconductor Technology

The most important and perhaps most exciting electrical devices used today are built from semiconductive materials. Electronic devices, such as diodes, transistors, thyristors, thermistors, photovoltaic cells, phototransistors, photoresistors, lasers, and integrated circuits, are all made from semiconductive materials, or semiconductors.

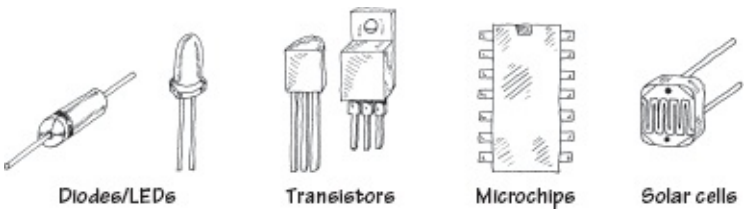


FIGURE 4.1

4.1.1 What Is a Semiconductor?

Materials are classified by their ability to conduct electricity. Substances that easily pass an electric current, such as silver and copper, are referred to as *conductors*. Materials that have a difficult time passing an electric current, such as rubber, glass, and Teflon, are called *insulators*. There is a third category of material whose conductivity lies between those of conductors and insulators. This third category of material is referred to as a *semiconductor*. A semiconductor has a kind of neutral conductivity when taken as a group. Technically speaking, semiconductors are defined as those materials that have a conductivity σ in the range between 10^{-7} and 10^{3} mho/cm (see Fig. 4.2). Some semiconductors are pure-elemental structures (e.g., silicon, germanium), others are alloys (e.g., nichrome, brass), and still others are liquids.

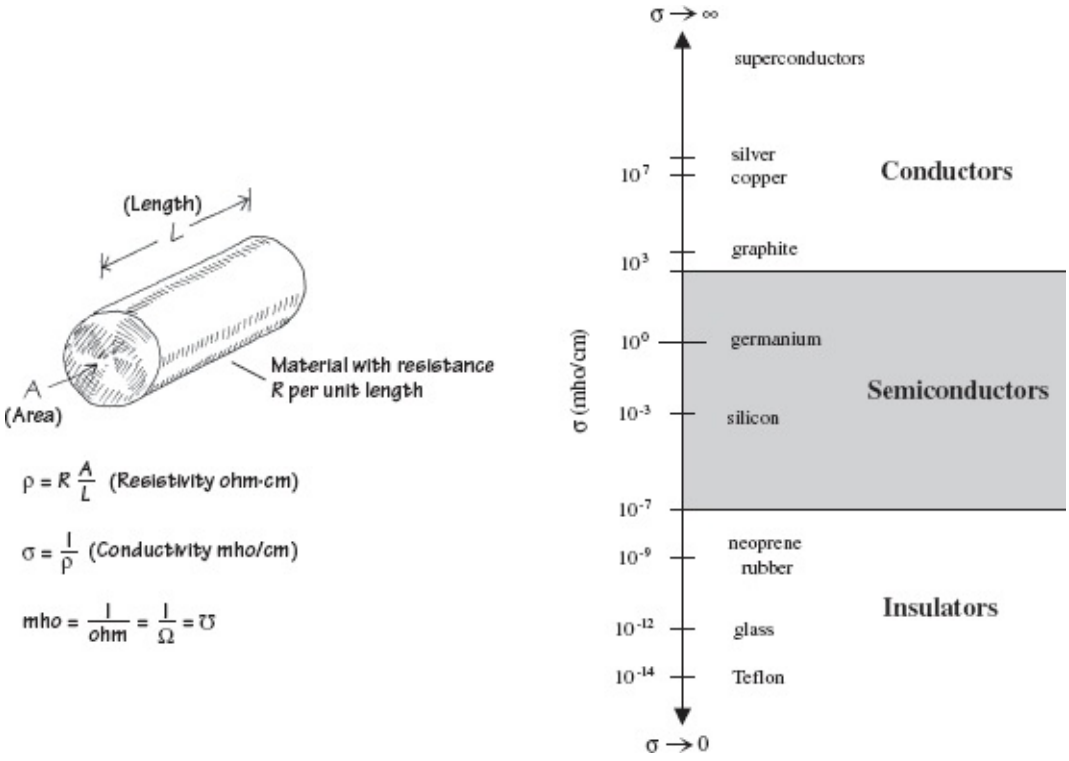


FIGURE 4.2

Silicon

Silicon is the most important semiconductor used in building electrical devices. Other materials such as germanium and selenium are sometimes used, too, but they are less popular. In pure form, silicon has a unique atomic structure with very important properties useful in making electrical devices.

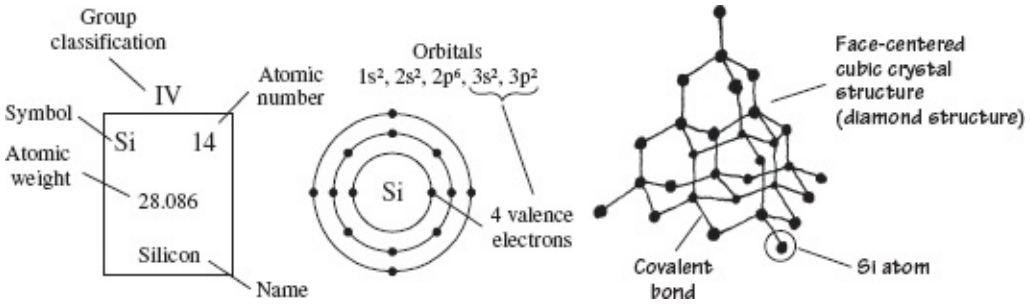


FIGURE 4.3

Silicon is ranked second in the order of elements appearing in the earth's crust, an average of 27 percent occurring in igneous rocks. It is estimated that a cubic mile of seawater contains about 15,000 tons of silicon. It is extremely rare to find silicon in its pure crystalline form in nature, and before it can be used in making electronic devices, it must be separated from its binding elements. After individuals—chemists, material scientists, etc.—perform the purification process, the silicon is melted and spun into a large "seed" crystal. This long crystal can then be cut up into slices or wafers that semiconductor-device designers use in making electrical contraptions.

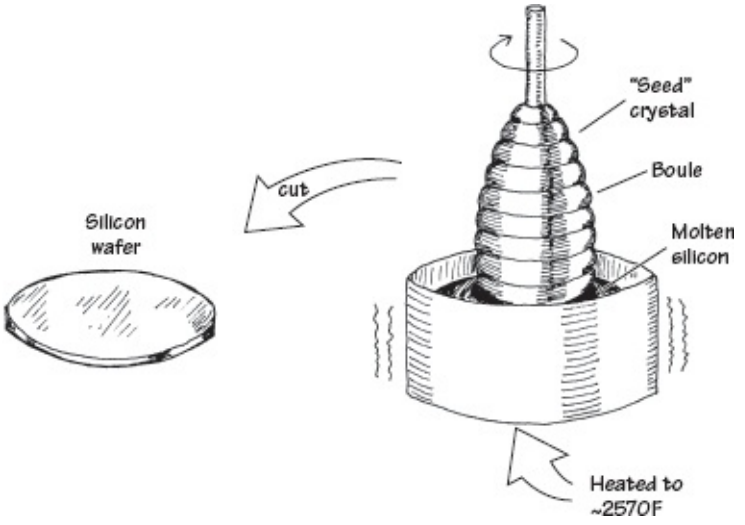


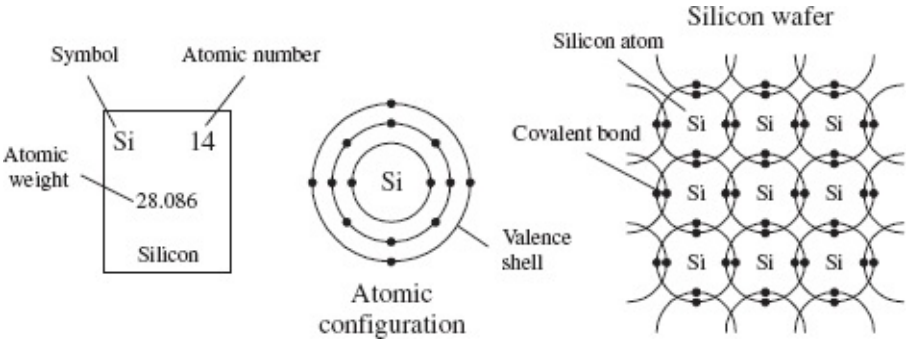
FIGURE 4.4

For the semiconductor-device designer, a silicon wafer alone does not prove very useful. A designer would not use the silicon wafer in its pure form to build a device; it just does not have quite the right properties needed to be useful. A semiconductor-device designer is looking for a material that can be made to alter its conductive state, acting as a conductor at one moment and as an insulator at another moment. For the material to change states, it must be able to respond to some external force applied at will, such as an externally applied voltage. A silicon wafer alone is not going to do the trick. In fact, a pure silicon wafer acts more as an insulator than a conductor, and it does not have the ability to change conductive states when an external force is applied. Every designer today knows that a silicon wafer can be transformed and combined with other transformed silicon wafers to make devices that have the ability to alter their conductive states when an external force is applied. The transforming process is referred to as *doping*.

Doping

Doping refers to the process of “spicing up” or adding ingredients to a silicon wafer in such a way that it becomes useful to the semiconductor-device designer. Many ingredients can be added in the doping process, such as antimony, arsenic, aluminum, and gallium. These ingredients provide specialized characteristics such as frequency response to applied voltages, strength, and thermal integrity, to name a few. By far, however, the two most important ingredients that are of fundamental importance to the semiconductor-device designer are boron and phosphorus.

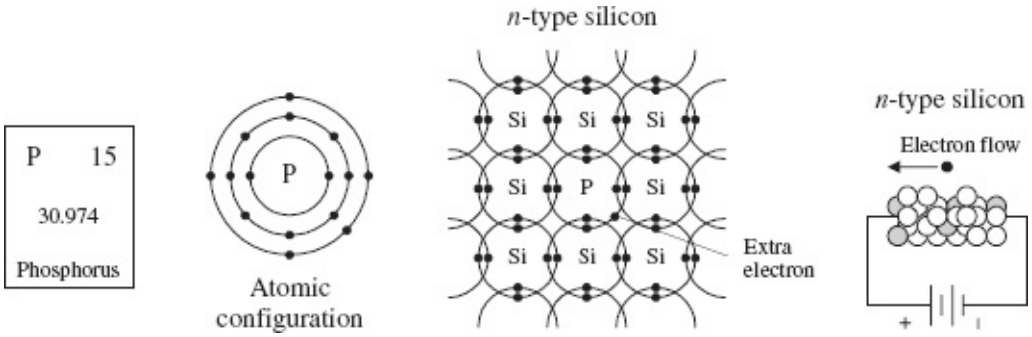
When a silicon wafer is doped with either boron or phosphorus, its electrical conductivity is altered dramatically. Normally, a pure silicon wafer contains no free electrons; all four of its valence electrons are locked up in covalent bonds with neighboring silicon atoms (see Fig. 4.5). Without any free electrons, an applied voltage will have little effect on producing an electron flow through the wafer.



A silicon wafer in pure form doesn’t contain any free charge carriers; all the electrons are locked up into covalent bonds between neighboring atoms.

FIGURE 4.5

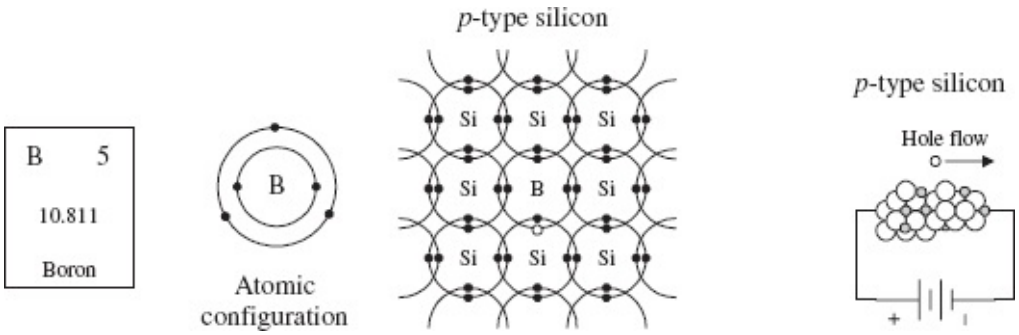
However, if phosphorus is added to the silicon wafer, something very interesting occurs. Unlike silicon, phosphorus has five valence electrons instead of four. Four of its valence electrons will form covalent bonds with the valence electrons of four neighboring silicon atoms (see Fig. 4.6). However, the fifth valence electron will not have a "home" (binding site) and will be loosely floating about the atoms. If a voltage is applied across the silicon-phosphorus mixture, the unbound electron will migrate through the doped silicon toward the positive voltage end. By supplying more phosphorus to the mixture, a larger flow of electrons will result. Silicon that is doped with phosphorus is referred to as *n*-type silicon, or negative-charge-carrier-type silicon.



A phosphorus atom added to a silicon wafer provides an extra unbound electron that aids in conduction. Silicon doped with phosphorus is called re-type silicon.

FIGURE 4.6

Now, if you take pure silicon and add some boron, you will see a different kind of conduction effect. Boron, unlike silicon or phosphorus, contains only three valence electrons. When it is mixed with silicon, all three of its valence electrons will bind with neighboring silicon atoms (see Fig. 4.7). However, there will be a vacant spot—called a *hole*—within the covalent bond between one boron and one silicon atom. If a voltage is applied across the doped wafer, the hole will move toward the negative voltage end, while a neighboring electron will fill in its place. Holes are considered positive charge carriers even though they do not contain a physical charge per se. Instead, it only appears as if a hole has a positive charge because of the charge imbalance between the protons within the nucleus of the silicon atom that receives the hole and the electrons in the outer orbital. The net charge on a particular silicon atom with a hole will appear to be positive by an amount of charge equivalent to one proton (or a "negative electron"). Silicon that is doped with boron is referred to as *p*-type silicon, or positive-charge-carrier-type silicon.



When boron is added to silicon, a hole is formed. This hole acts like a positive charge (see text) that aids in conduction. Silicon doped with boron is called *p*-type silicon.

FIGURE 4.7

As you can see, both *n*-type and *p*-type silicon have the ability to conduct electricity; one does it with extra unbound electrons (*n*-type silicon), and the other does it with holes (*p*-type silicon).

A Note to Avoid Confusion

Boron atoms have three valence electrons, not four like silicon. This means that the combined lattice structure has fewer free valence electrons as a whole. However, this does not mean that a *p*-type silicon semiconductor has an overall positive charge; the missing electrons in the structure are counterbalanced by the missing protons in the nuclei of the boron atoms. The same idea goes for *n*-type silicon, but now the extra electrons within the semiconductor are counterbalanced by the extra protons within the phosphorus nuclei.

Another Note to Avoid Confusion (Charge Carriers)

What does it mean for a hole to flow? I mean, a hole is nothing, right? How can nothing flow? Well, it is perhaps misleading, but when you hear the phrase "hole flow" or "flow of positive charge carriers in *p*-type silicon," electrons are in fact flowing. You may say, doesn't that make this just like the electron flow in *n*-type silicon? No. Think about tipping a sealed bottle of water upside down and then right side up (see Fig. 4.8). The bubble trapped in the bottle will move in the opposite direction of the water. For the bubble to proceed, water has to move out of its way. In this analogy, the water represents the electrons in the *p*-type silicon, and the holes represent the bubble. When a voltage is applied across a *p*-type silicon semiconductor, the electrons around the boron atom will be forced toward the direction of the positive terminal. Now here is where it gets tricky. A hole next to a boron atom will be pointing toward the negative terminal. This hole is just waiting for an electron from a neighboring atom to fall into it, due in part to the lower energy configuration there. Once an electron, say, from a neighboring silicon atom, falls into the hole in the boron atom's valence shell, a hole is briefly created in that silicon atom's valence shell. The electrons in the silicon atom lean toward the positive terminal, and the newly created hole leans toward the negative terminal. The next silicon atom over will let go of one of its electrons, and the electron will fall into the hole, and the hole will move over again—the process continues, and it appears as if the hole flows in a continuous motion through the *p*-type semiconductor.

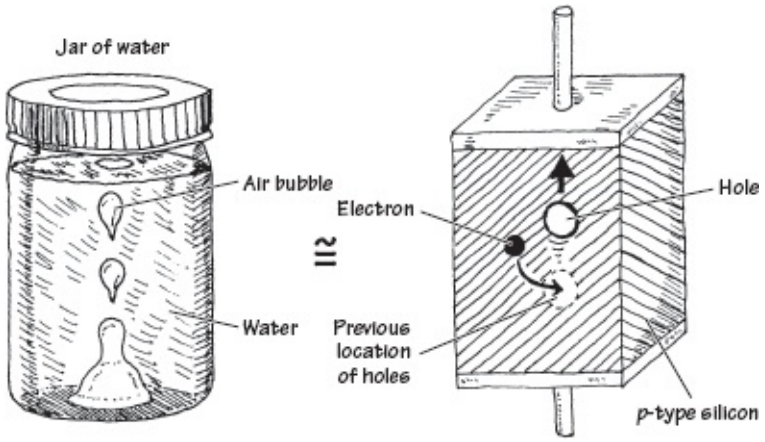


FIGURE 4.8

A Final Note to Avoid Confusion

And finally, why is a hole called a positive charge carrier? How can "nothing" carry a positive charge? Well, what's going on here is this: A hole, as it moves through the mostly silicon-based crystal, causes a brief alteration in the electrical field strength around the silicon atom in the crystal where it happens to be situated. When an electron moves out of the way, thus creating a new hole, this silicon atom as a whole will be missing an electron, and hence the positive charge from the nucleus of the silicon atom will be felt (one of the protons is not counterbalanced). The "positive charge carrier" attributed to holes comes from this effective positive nuclear charge of the protons fixed within the nucleus.

4.1.2 Applications of Silicon

You may be asking yourself, Why are these two new types of silicon (*n*-type and *p*-type) so useful and interesting? What good are they for semiconductor-device designers? Why is there such a fuss over them? These doped silicon crystals are now conductors. Big deal, right? Yes, we now have two new conductors, but the two new conductors have two unique ways of passing an electric current—one does it with holes, the other with electrons. This is very important.

The manners in which *n*-type and *p*-type silicon conduct electricity (electron flow and hole flow) are very important in designing electronic devices such as diodes, transistors, and solar cells. Some clever people figured out ways to arrange slabs, chucks, strings, etc. made of *n*-type and *p*-type silicon in such a way that when an external voltage or current is applied to these structures, unique and very useful features result. These unique features are made possible by the interplay between hole flow and electron flow between the *n*-type and *p*-type semiconductors. With these new *n*-type/*p*-type contraptions, designers began building one-way gates for current flow, opening and closing channels for current flow controlled by an external electrical voltage and/or current. Folks figured out that when an *n*-type and a *p*-type semiconductor were placed together and a particular voltage was applied across the slabs, light, or photons, could be produced as the electrons jumped across the junction between the interface. It was noticed that this process could work backward as well. That is, when light was exposed at the *np* junction, electrons were made to flow, thus resulting in an electric current. A number of clever contraptions have been built using *n*-type and *p*-type semiconductor combinations. The following chapters describe some of the major devices people came up with.

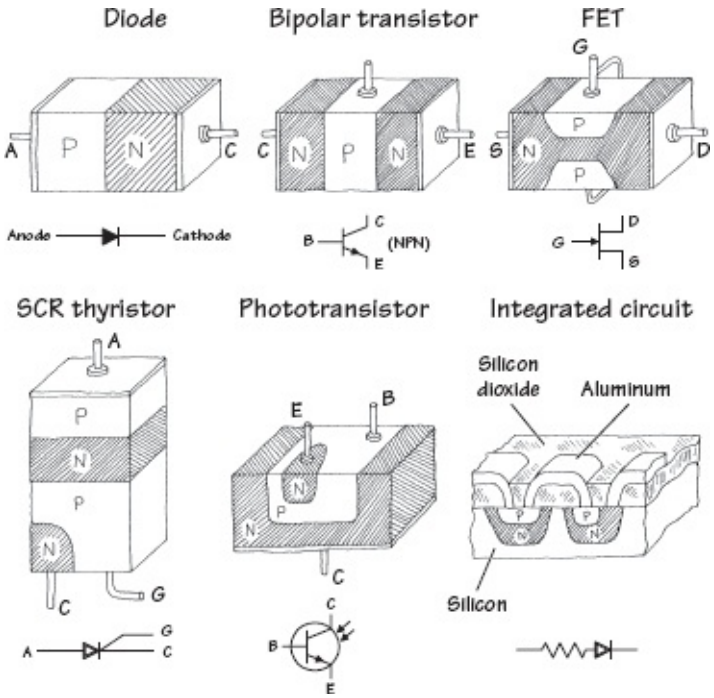


FIGURE 4.9

4.2 Diodes

A *diode* is a two-lead semiconductor device that acts as a one-way gate to electric current flow. When a diode's *anode* lead is made more positive in voltage than its *cathode* lead—a condition referred to as *forward biasing*—current is permitted to flow through the device. However, if the polarities are reversed (the anode is made more negative in voltage than the cathode)—a condition referred to as *reversed biasing*—the diode acts to block current flow.



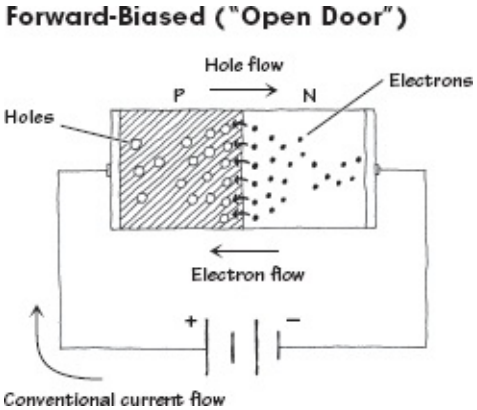
FIGURE 4.10

Diodes are used most commonly in circuits that convert ac voltages and current into dc voltages and currents (e.g., ac/dc power supply). Diodes are also used in voltage-multiplier circuits, voltage-shifting circuits, voltage-limiting circuits, and voltage-regulator circuits.

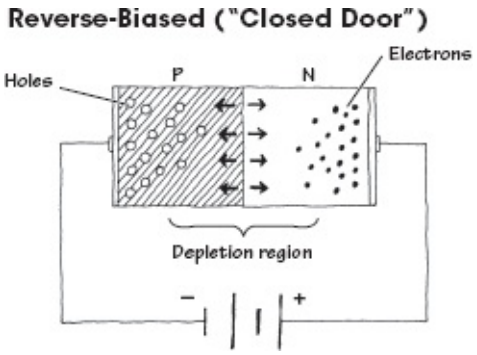
4.2.1 How p-n Junction Diodes Work

A *p-n junction diode* (*rectifier diode*) is formed by sandwiching together *n*-type and *p*-type silicon. In practice, manufacturers grow an *n*-type silicon crystal and then abruptly change it to a *p*-type crystal. Then either a glass or plastic coating is placed around the joined crystals. The *n* side is the cathode end, and the p side is the anode end.

The trick to making a one-way gate from these combined pieces of silicon is getting the charge carriers in both the *n*-type and *p*-type silicon to interact in such a way that when a voltage is applied across the device, current will flow in only one direction. Both *n*-type and *p*-type silicon conducts electric current; one does it with electrons (*n*-type), and the other does it with holes (*p*-type). Now the important feature to note here, which makes a diode work (act as a one-way gate), is the manner in which the two types of charge carriers interact with each other and how they interact with an applied electrical field supplied by an external voltage across its leads. Below is an explanation describing how the charge carriers interact with each other and with the electrical field to create an electrically controlled one-way gate.



When a diode is connected to a battery, as shown here, electrons from the *n* side and holes from the *p* side are forced toward the center (*pn* interface) by the electrical field supplied by the battery. The electrons and holes combine, and current passes through the diode. When a diode is arranged in this way, it is said to be *forward-biased*.



When a diode is connected to a battery, as shown here, holes in the *n* side are forced to the left, while electrons in the *p* side are forced to the right. This results in an empty zone around the *p-n* junction that is free of charge carriers, better known as the *depletion region*. This depletion region has an insulative quality that prevents current from flowing through the diode. When a diode is arranged in this way, it is said to be *reverse-biased*.

FIGURE 4.11

A diode's one-way gate feature does not work all the time. That is, it takes a minimal voltage to turn it on when it is placed in forward-biased direction. Typically for silicon diodes, an applied voltage of 0.6 V or greater is needed; otherwise, the diode will not conduct. This feature of requiring a specific voltage to turn the diode on may seem like a drawback, but in fact, this feature becomes very useful in terms of acting as a voltage-sensitive switch. Germanium diodes, unlike silicon diodes, often require a forward-biasing voltage of only 0.2 V or greater for conduction to occur. Figure 4.12 shows how the current and voltage are related for silicon and germanium diodes.

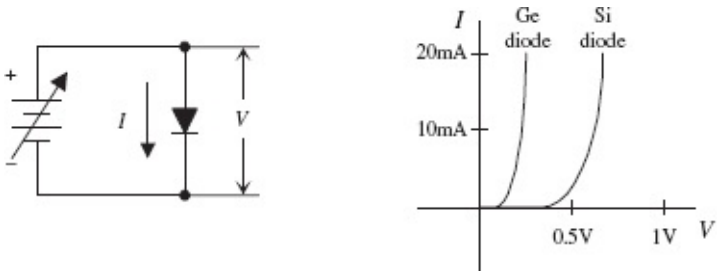


FIGURE 4.12

Another fundamental difference between silicon diodes and germanium diodes, besides the forward-biasing voltages, is their ability to dissipate heat. Silicon diodes do a better job of dissipating heat than germanium diodes. When germanium diodes get hot—temperatures exceeding 85°C—the thermal vibrations affect the physics inside the crystalline structure to a point where normal diode operation becomes unreliable. Above 85°C, germanium diodes become worthless.

4.2.2 Diode Water Analogy

A *diode* (or *rectifier*) acts as a one-way gate to current flow—see the water analogy in Fig. 4.13. Current flows in the direction of the arrow, from anode (+) to cathode (−), provided *the forward voltage* V_F across it exceeds what's called the *junction threshold voltage*. As a general rule of thumb, silicon p-n junction diodes have about a 0.6-V threshold, germanium diodes a 0.2-V threshold, and Schottky diodes a 0.4-V threshold. However, don't take this rule too seriously, because with real-life components, you'll find these thresholds may be a few tenths of a volt off. For example, it's entirely possible for a p-n junction diode's threshold to be anywhere between 0.6 and 1.7 V; for germanium, 0.2 to 0.4 V; and for Schottky diodes, 0.15 to 0.9 V.

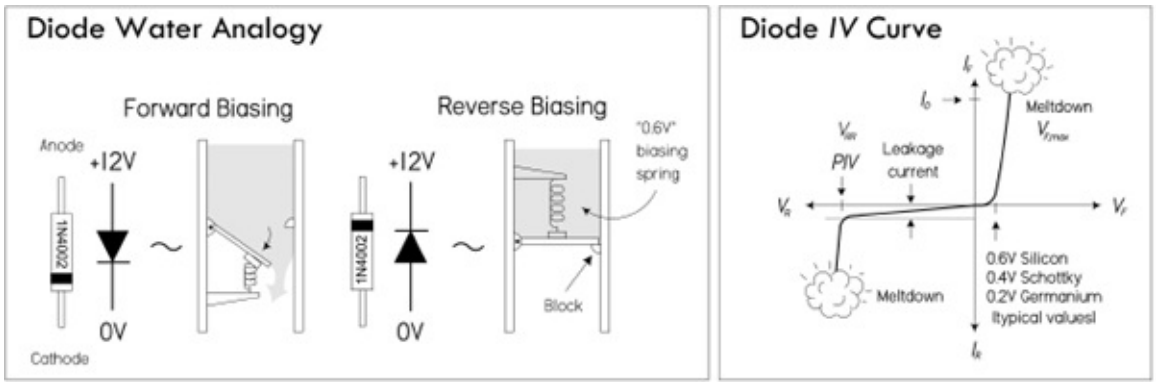


FIGURE 4.13

Note that if you actually put 12 V across a forward-biased diode as shown in Fig. 4.13, a very large current would flow, probably destroying the diode. Also, the axes of Fig. 4.13 are not to scale.

In terms of limits, avoid supplying a diode with a forward current I_F beyond its *peak current rating*

($I_{O\max}$). If you do, you'll get internal junction meltdown. Likewise, avoid applying a reverse voltage V_R any bigger than the diode's *peak inverse voltage* (PIV) *rating*. This, too, can render a diode worthless. See the graph in Fig. 4.13.

4.2.3 Kinds of Rectifiers/Diodes

There are numerous types of diodes, each specifically designed to work better in one application or another. Diodes for high-power applications (switching, power supplies, etc.) which draw lots of current or rectify high voltages typically go by the name *rectifier diodes*. On the other hand, diodes that go by names such as *signal*, *switching*, *fast recovery*, or *high speed* are designed to provide a low internal capacitance (they store less charge but usually have weaker junctions for large currents). At high speeds, these diodes will reduce RC switching time constants, which means fewer time delays and lower signal losses.

Schottky diodes have a particularly low junction capacitance and faster switching (~10 ns) when compared to silicon p-n junction diodes due to their special metal-semiconductor-junction interface. They also have a lower junction threshold voltage—as low as 0.15 V, but usually a bit more (0.4 V average). Both these characteristics enable them to detect low-voltage, high-frequency signals that ordinary p-n junction diodes would not see. (A Schottky with a 0.3-V threshold can pass signals greater than 0.3 V, but a silicon p-n junction diode with a 0.7-V threshold can only pass signals greater than 0.7 V.) For this reason, Schottky diodes are very popular in low-voltage signal rectifiers in RF circuits, signal switching in telecommunication, small dc/dc converters, small low-voltage power supplies, protection circuits, and voltage clamping arrangements. Their high-current density and low voltage drop also make them great in power supplies, since they generate less heat, requiring smaller heat sinks to be used in design. Therefore, you'll find both rectifier and fast-switching Schottky diodes listed in the catalogs.

In terms of *germanium diodes*, they are used mostly in RF signal detection and low-level logic design due to their small threshold voltage of about 0.2 V. You do not see them in high-current rectifying applications, since they are weaker and leak more than silicon diodes when temperatures rise. In many applications, a good Schottky signal diode can replace a germanium diode.

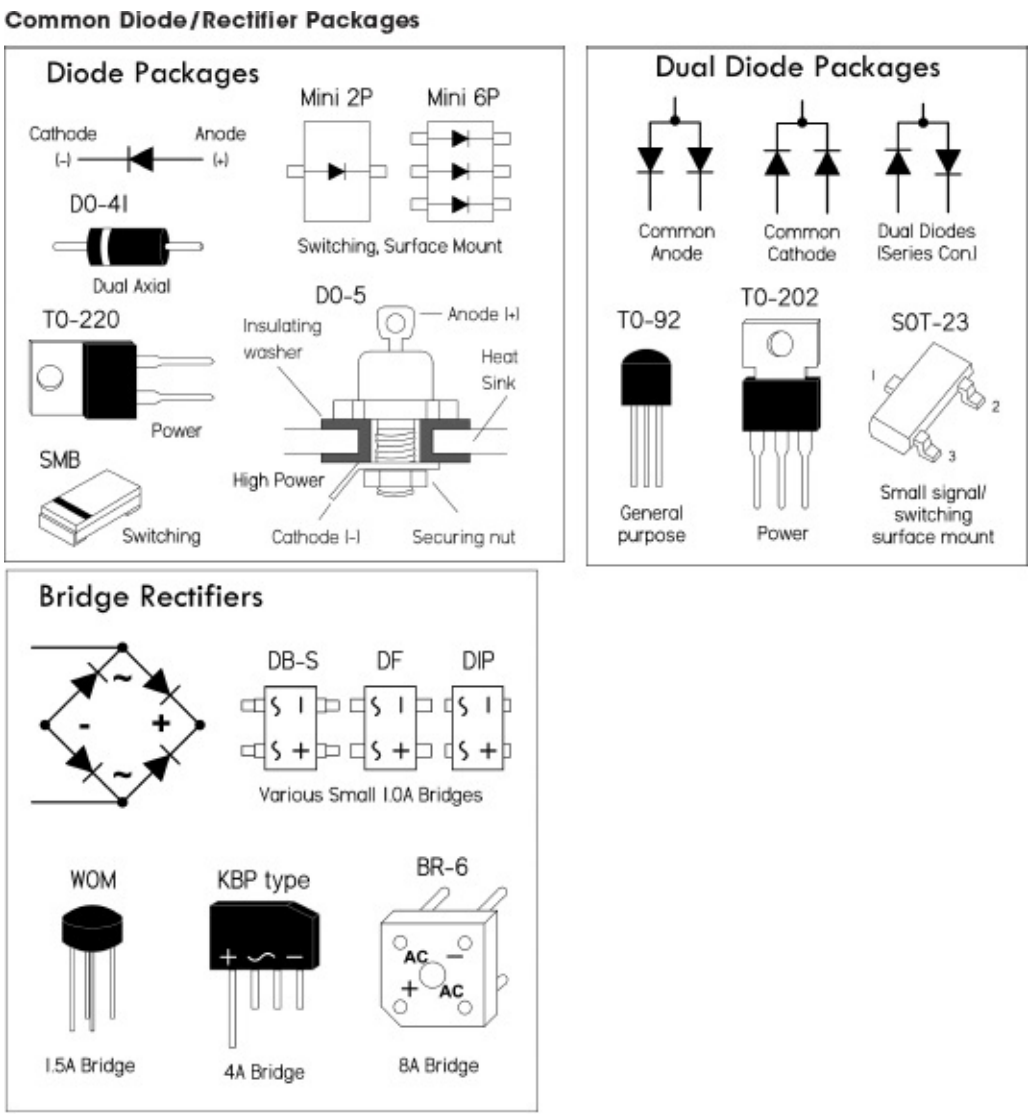


FIGURE 4.14

4.2.4 Practical Considerations

Five major specs to consider when choosing a diode are peak inverse voltage, PIV; forward current-handling capacity, $I_{O(\max)}$; response speed t_R (time for diode to switch on and off); reverse-leakage current, $I_{R(\max)}$; and maximum forward-voltage drop, $V_{F(\max)}$. Each of these characteristics can be manipulated during the manufacturing process to produce the various special-purpose diodes. In rectification applications (e.g., power supplies, transient protection), the most important specs are PIV and current rating. The peak negative voltages that are stopped by the diode must be smaller in magnitude than the PIV, and the peak current through the diode must be less than $I_{O(\max)}$. In fast and low-voltage applications, t_R and V_F become important characteristics to consider. In the applications section that follows, you'll get a better sense of what all these specs mean.

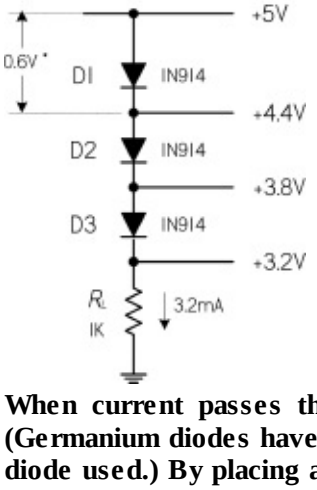
TABLE 4.1 Selection of Popular Diodes

DEVICE	TYPE	PEAK INVERSE VOLTAGE PIV (V)	MAX. FORWARD CURRENT $I_{O(MAX)}$	MAX. REVERSE CURRENT $I_{R(MAX)}$	PEAK SURGE CURRENT I_{FSM}	MAX. VOLTAGE DROP V_F (V)
1N34A	Signal (Ge)	60	8.5 mA	15 μA		1.0
1N67A	Signal (Ge)	100	4.0 mA	5 μA		1.0
1N191	Signal (Ge)	90	5.0 mA			1.0
1N914	Fast Switch	90	75 mA	25 nA		0.8
1N4148	Signal	75	10 mA	25 nA	450 mA	1.0
1N4445	Signal	100	100 mA	50 nA		1.0
1N4001	Rectifier	50	1 A	0.03 mA	30 A	1.1
1N4002	Rectifier	100	1 A	0.03 mA	30 A	1.1
1N4003	Rectifier	200	1 A	0.03 mA	30 A	1.1
1N4004	Rectifier	400	1 A	0.03 mA	30 A	1.1
1N4007	Rectifier	1000	1 A	0.03 mA	30 A	1.1
1N5002	Rectifier	200	3 A	500 μA	200 A	
1N5006	Rectifier	600	3 A	500 μA	200 A	
1N5008	Rectifier	1000	3 A	500 μA	200 A	
1N5817	Schottky	20	1 A	1 mA	25 A	0.75
1N5818	Schottky	30	1 A		25 A	
1N5819	Schottky	40	1 A		25 A	0.90
1N5822	Schottky	40	3 A			
1N6263	Schottky	70	15 mA		50 mA	0.41
5052-2823	Schottky	8	1 mA	100 nA	10 mA	0.34

Diodes come in a variety of different packages. Some are standard two-lead deals; others are high-power packages with heat-sink attachments (e.g., TO-220, DO-5). Some come in surface-mount packages, and others contain diode arrays in IC form, used for switching applications. Dual-diode and diode-bridge rectifiers also come in a variety of package sizes and shapes for varying power levels.

4.2.5 Diode/Rectifier Applications

Voltage Dropper

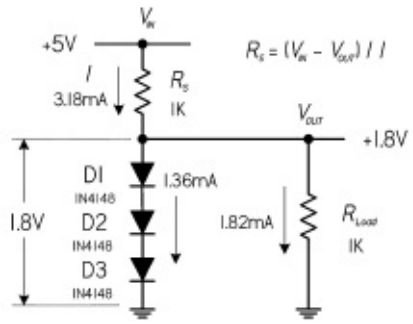


When current passes through a diode, there is a voltage drop across it of about 0.6 V, for a silicon p-n junction diode. (Germanium diodes have around a 0.2-V drop; Schottky, around 0.4 V—all these values vary slightly, depending on the specific diode used.) By placing a number of diodes in series, the total voltage drop across the combination is the sum of the individual voltage drops across each diode. Voltage droppers are often used in circuits where a fixed small voltage difference between two sections of a circuit is needed. Unlike resistors that can be used to lower the voltage, the diode arrangement typically doesn't waste as much power to resistive heating and will supply a stiffer regulated voltage that is less dependent on current variations. Later in this chapter, you'll see that a single zener diode can often replace a multiple series diode arrangement like the one

shown here.

FIGURE 4.15

Voltage Regulator



Here's a spin-off of the last circuit, making use of the three diodes to create a simple regulated (steady) low-voltage output equal to the sum of the threshold voltages of the diodes: 0.6 V + 0.6 V + 0.6 V = 1.8 V. The series resistor is used to set the desired output current (I) and should be less than the value calculated using the following formula, but not so low that it exceeds the power ratings of itself and the diodes:

$$R_S = \frac{V_{in} - V_{out}}{I}$$

Diodes and the series resistor must have proper power ratings for the amount of current drawn. Use *P = IV*. Note that for higher-power critical voltage sources, you'll typically use a zener diode regulator or, more commonly, a special regulator IC instead.

FIGURE 4.16

Reverse-Polarity Protection

Battery reversal or power polarity reversal can be fatal to portable equipment. The best design is to use a mechanical block to safeguard against reverse installation. However, even momentarily fumbling around and making contacts can lead to problems. This is especially true for one or more single-cell battery applications that use AA-alkaline, NiCad, and NiMH batteries. For these systems you must ensure that any flow of reverse current is low enough to avoid damaging the circuit or the battery.

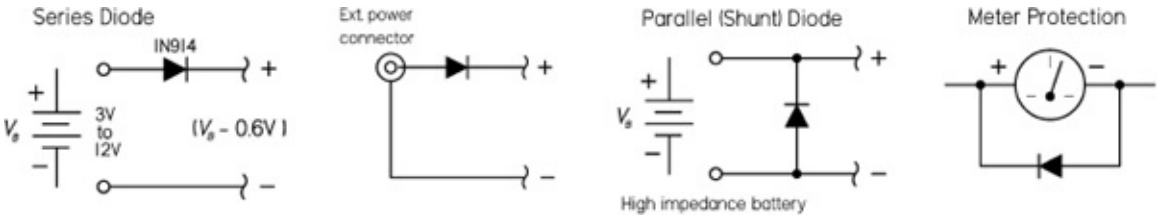


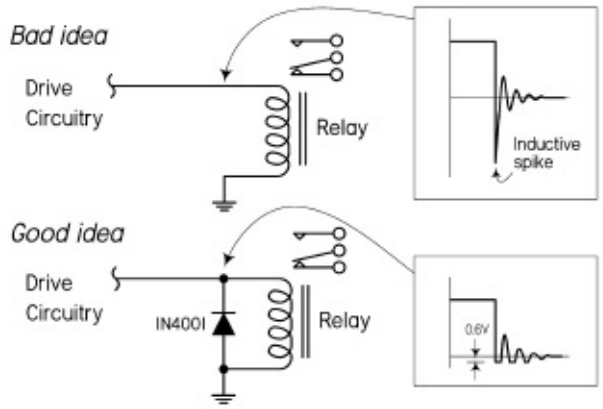
FIGURE 4.17 ***Series diode:*** **This is the simplest battery-reversal protection. It can be used with external power connections, too—plug-and-jack. The diode allows current from a correctly installed battery to flow to load, but blocks current flow from a backward-installed battery. The drawback with a series diode is that the diode must handle the full load current. Also, the forward voltage drop of the diode shortens the equipment's operating time—cuts off about 0.6 V right away. Schottky diodes with low thresholds can do better. See Problem 1 at the end of Section 4.2 for another reverse-polarity protection circuit.**

Parallel diode: **In applications that call for alkaline or other batteries that have high output impedances, you can guard against reverse installations by using a parallel (shunt) diode, while eliminating the diode's voltage drop. This approach protects the load but draws high current from the backward-installed battery. The diode must be properly rated for current and power. Another application of the parallel diode is in meter protection, where it acts to divert large currents entering the negative terminal of the meter.**

Note: **In more sophisticated battery-powered designs, special ICs or transistor arrangements are used to provide essentially zero voltage drop protection, while providing a number of other special features, such as reverse polarity protection, thermal shutdown, and voltage level monitor.**

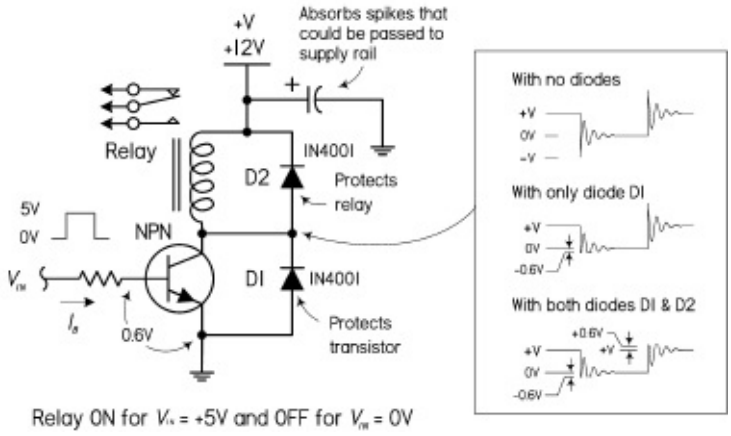
Transient Suppression with Fly-Back Diodes

Transient Protection



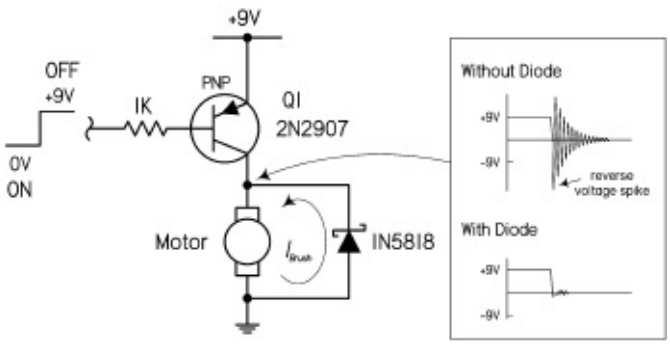
When current flowing through an inductor is suddenly switched off, the collapsing magnetic field will generate a high-voltage spike in the inductor's coils. This voltage spike or transient may have an amplitude of hundreds or even thousands of volts. This is particularly common within relay coils. A diode—referred to as a fly-back diode for this type of application—placed across the relay's coil can protect neighboring circuitry by providing a short circuit for the high-voltage spike. It also protects the relay's mechanical contacts, which often get viciously slapped shut during an inductive spike. Notice, however, that the diode is ineffective during turn-on time. Select a rectifier diode with sufficient power ratings (1N4001, 1N4002, etc.). Schottky rectifiers (e.g., 1N5818) work well, too.

Transistor Relay Driver with Protection Diodes



Here's a more practical example for driving a relay that contains an extra diode placed across a transistor driver in order to protect the transistor from damage due to inductive spikes generated from the relay's coil when the transistor is turned off. This arrangement also deadens spikes during turn-on time. This dual diode arrangement is sometimes used in voltage regulator circuits, where one diode is wired from the output to the input and another is wired from ground to the output. This prevents any attached loads from sending damaging spikes back into the IC's output.

Motor Inductive Kickback Protection



Here's another example of how inductive kickback from a motor that is running and then suddenly turned off can generate a voltage transient that can potentially damage connected electronics—in this case, a 2N2907 transistor. The diode reroutes or shorts the induced voltage to the opposite terminal of the motor. Here a 1N5818 Schottky diode is used—though you could use other p-n junction types, too. The Schottky diode happens to be a bit faster and will clip the transient voltage a bit lower down—at around 0.4 V.

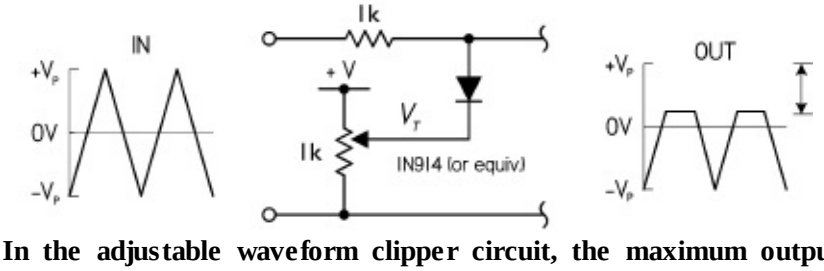
Note: **Devices such as TVs and varistors are specially designed for transient suppression. See the section on transient suppressors later in this chapter.**

FIGURE 4.18

Diode Clamps

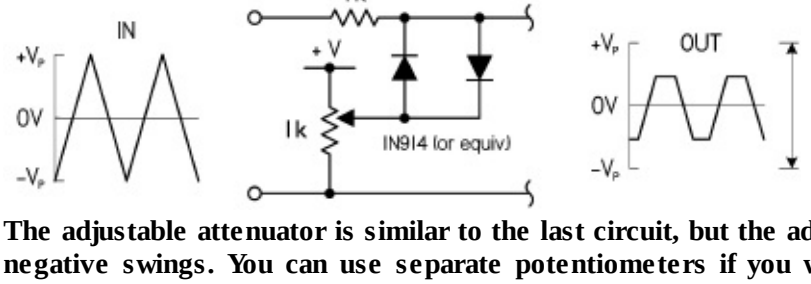
Diode clamps are used to clip signal levels, or they can shift an ac waveform up or down to create what's called a pulsing dc waveform—one that doesn't cross the 0-V reference.

Adjustable Waveform Clipper



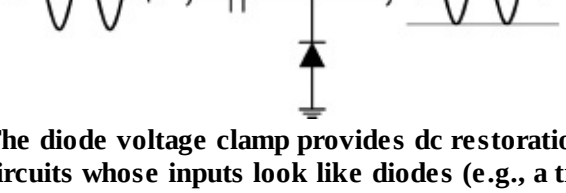
In the adjustable waveform clipper circuit, the maximum output is clipped to a level determined by the resistance of the potentiometer. The idea is to set the negative end of the diode to about 0.6 V lower than the maximum desired output level, to account for the forward voltage drop of the diode. That's what the potentiometer is intended to do. +V should be a volt or so higher than the peak input voltage.

Adjustable Attenuator



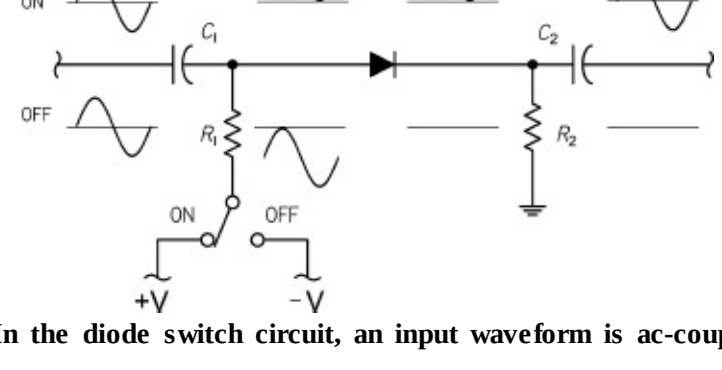
The adjustable attenuator is similar to the last circuit, but the additional opposing diode allows for clipping on both positive and negative swings. You can use separate potentiometers if you want separate positive and negative clipping level control. +V should be a volt or so higher than the peak input voltage.

Diode Voltage Clamp (DC Restoration)



The diode voltage clamp provides dc restoration of a signal that has been ac-coupled (capacitively coupled). This is important for circuits whose inputs look like diodes (e.g., a transistor with grounded emitter); otherwise, an ac-couple signal would fade away.

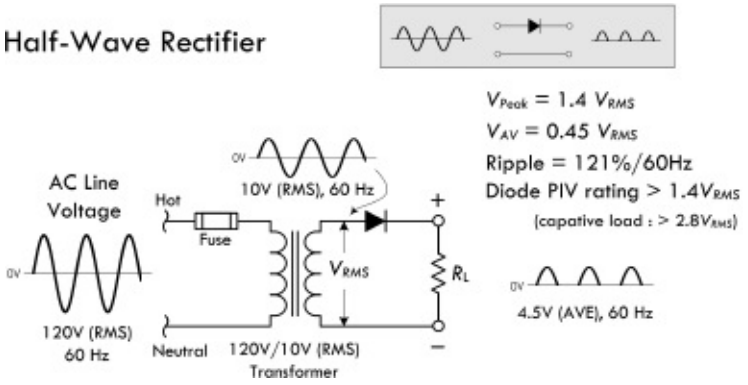
Diode Switch



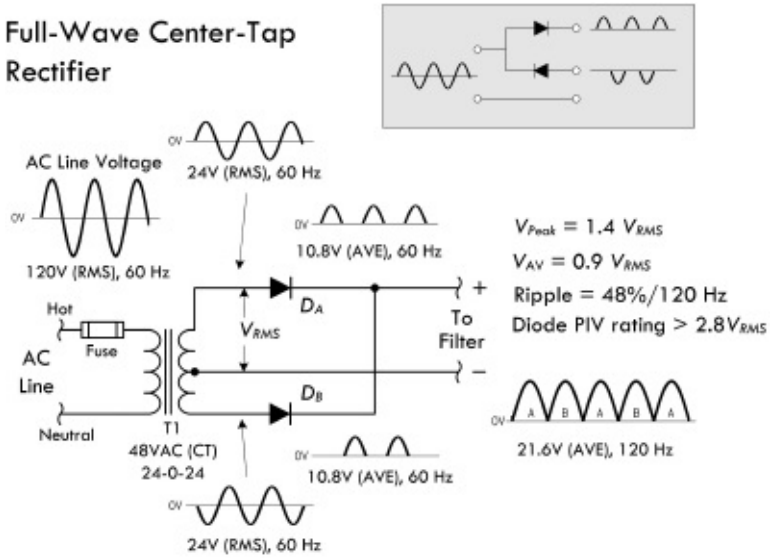
In the diode switch circuit, an input waveform is ac-coupled to the diode through C1 at the input and C_2 at the output. R_2

provides a reference for the bias voltage. When the switch is thrown to the ON position, a positive dc voltage is added to the signal, forward-biasing the diode and allowing the signal to pass. When the switch is thrown to the OFF position, the negative dc voltage added to the signal reverse-biases the diode and the signal does not get through.

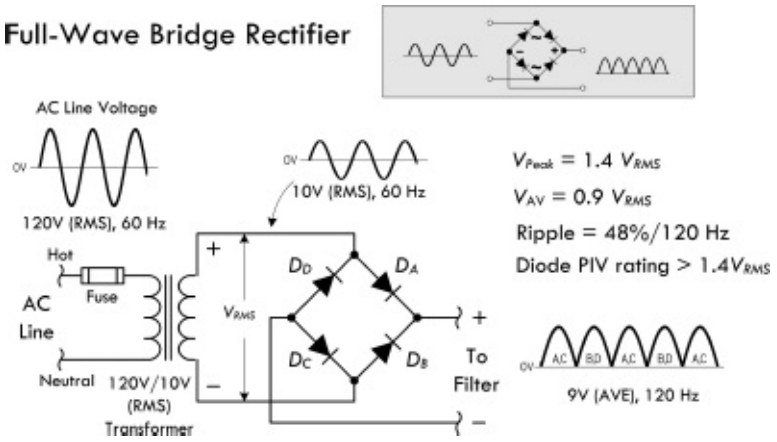
FIGURE 4.19



Half-wave rectifier: **Used to transform an ac signal into pulsing dc by blocking the negative swings. A filter is usually added (especially in low-frequency applications) to the output to smooth out the pulses and provide a higher average dc voltage. The peak inverse voltage (PIV)—the voltage that the rectifier must withstand when it isn't conducting—varies with load, and must be greater than the peak ac voltage (1.4 × V_{rms}). With a capacitor filter and a load drawing little or no current, it can rise to 2.8 × V_{rms} (capacitor voltage minus the peak negative swing of voltage from transformer secondary).**



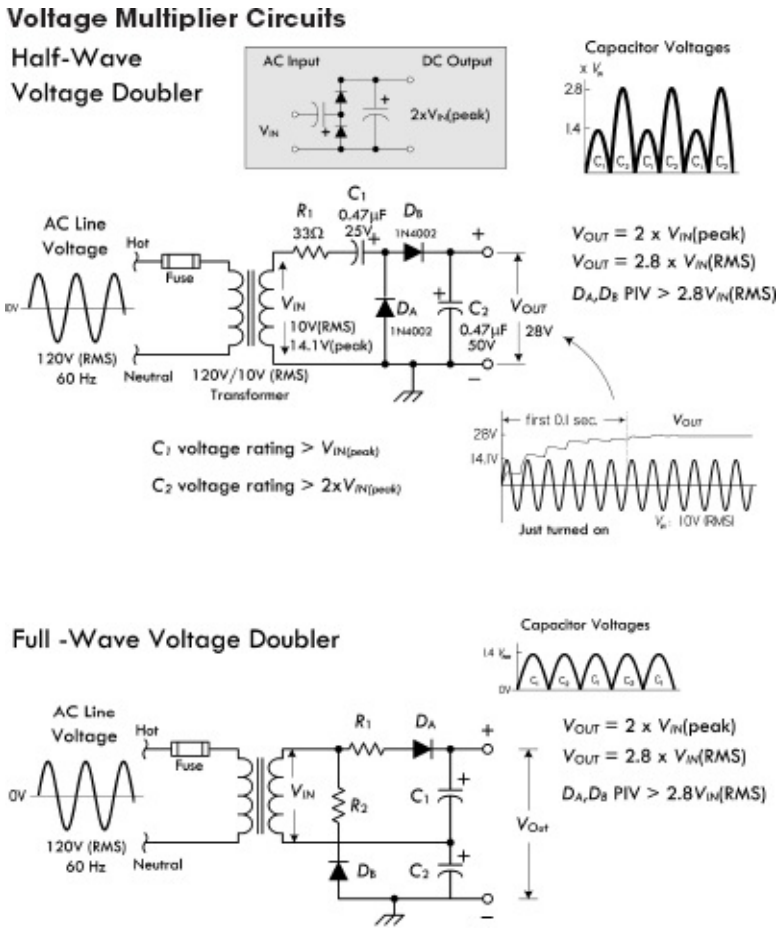
Full-wave center-tap rectifier: **This commonly used circuit is basically two combined half-wave rectifiers that transform both halves of an ac wave into a pulsing dc signal. When designing power supplies, you need only two diodes, provided you use a center-tap transformer. The average output voltage is 0.9 V_{rms} of half the transformer secondary; this is the maximum that can be obtained with a suitable choke-input filter. The peak output voltage is 1.4 × V_{rms} of half the transformer secondary; this is the maximum voltage that can be obtained from a capacitor-input filter. The PIV impressed on each diode is independent of the type of load at the output. This is because the peak inverse voltage condition occurs when diode A conducts and diode B does not conduct. As the cathodes of diodes A and B reach a positive peak (1.4 V_{rms}), the anode of diode B is at the negative peak, also 1.4 V_{rms}, but in the opposite direction. The total peak inverse voltage is therefore 2.8 V_{rms}. The frequency of the output pulses is twice that of the half-wave rectifier, and thus comparatively less filtering is required. Since the diodes work alternately, each handles half of the load current. The current rating of each rectifier need be only half the total current drawn from the supply.**



Full-wave bridge rectifier: **This rectifier produces a similar output as the last full-wave rectifier, but doesn't require a center-tap transformer. To understand how the device works, follow the current through the diode one-way gates. Note that there will be at least a 1.2-V drop from zero-to-peak input voltage to zero-to-peak output voltage (there are two 0.6-V drops across a pair of diodes during a half cycle). The average dc output voltage into a resistive load or choke-input filter is 0.9 × V_{rms} of the transformer's secondary; with a capacitor filter and a light load, the maximum output voltage is 1.4 × V_{rms}. The inverse voltage across each diode is 1.4 V_{rms}; there the PIV of each diode is more than 1.4 V_{rms}.**

See the following text for the pros and cons of the various rectifier configurations.

FIGURE 4.20



Half-wave voltage doubler: **Takes an ac input voltage and outputs a dc voltage that is approximately equal to twice the input's peak voltage (or 2.8 times the input's RMS voltage). (The actual multiplication factor may differ slightly, depending on the capacitor, resistor, and load values.) In this circuit, we take V_{IN} to mean the secondary voltage from the transformer. During the first negative half cycle, D_A conducts, charging C_1 to the peak rectified voltage V_{IN} (peak), or 1.4 V_{IN} (RMS). During the positive half cycle of the secondary voltage, D_A is cut off and D_B conducts, charging capacitor C_2. The voltage delivered to C_2 is the sum of the transformer peak secondary voltage, V_{IN} (peak) plus the voltage stored in C_1, which is the same, so the sum gives 2 V_{IN} (peak), or 2.8 V_{IN} (RMS). On the next negative cycle, D_B is nonconducting and C_2 will discharge into an attached load. If no load is present, the capacitors will remain charged—C_1 to 1.4 V_{IN} (RMS), C_2 to 2.8 V_{IN} (RMS). When a load is connected to the output, the voltage across C_2 drops during the negative half cycle and is recharged up to 2.8 V_{IN} (RMS) during the positive half cycle. The output waveform across C_2 resembles that of a half-wave rectifier circuit because C_2 is pulsed once**

every cycle. Figure 4.21 shows levels to which the two capacitors are charged through-out the cycle. In actual operation, the capacitor will not discharge all the way to zero, as shown.

Full-wave doubler: **During the positive half cycle of the transformer secondary voltage, D_A conducts discharging C_1 to V_{IN} (peak) or 1.4 V_{IN} (RMS). During the negative half cycle, D_B conducts, charging C_2 to the same value. The output voltage is the sum of the two capacitor voltages, which will be 2 V_{IN} (peak) or 2.8 V_{IN} (RMS) under no-load conditions. The graph shows each capacitor alternately receiving a charge once per cycle. The effective filter capacitance is that of C_1 and C_2 in series, which is less than the capacitance of either C_1 or C_2 alone. R_1 and R_2 are used to limit the surge current through the rectifiers. Their values are based on the transformer voltage and the rectifier surge current rating, since at the instant the power supply is turned on, the filter capacitors look like a short-circuited load. Provided the limiting resistors can withstand the surge current, their current-handling capacity is based on the maximum load current from the supply. The peak inverse voltage across each diode is 2.8 V_{IN} (RMS).**

FIGURE 4.21

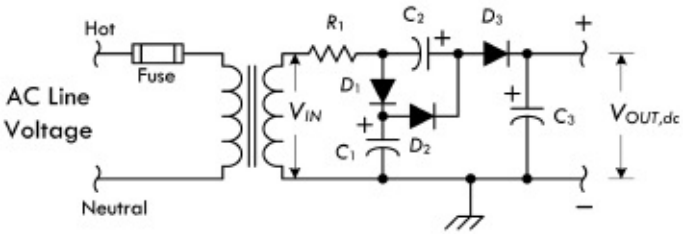
Pros and Cons of the Rectifier Circuits

Comparing the full-wave center-tap rectifier and the full-wave bridge rectifier, you'll notice both circuits have almost the same rectifier requirements. However, the center-tap version has half the number of diodes as the bridge. These diodes will require twice the maximum inverse voltage ratings of the bridge diodes (PIV > 2.8 V_{rms}, as opposed to >1.4 V_{rms}). The diode current ratings are identical for the two circuits. The bridge makes better use of the transformer's secondary than the center-tap rectifier, since the transformer's full winding supplies power during both half cycles, while each half of the center-tap circuit's secondary provides power only during its positive half-cycle. This is often referred to as the *transformer utilization factor*, which is unity for the bridge configuration and 0.5 for the full-wave center-tap configuration.

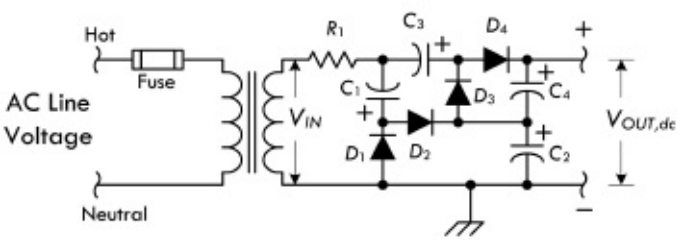
The bridge rectifier is often not as popular as the full-wave center-tap rectifier in high-current, low-voltage applications. This is because the two forward-conducting series-diode voltage drops in the bridge introduce a volt or more of additional loss, and thus consume more power (heat loss) than a single diode would within a full-wave rectifier.

In regard to the half-wave configuration, it's rarely used in 60-Hz rectification for other than bias supplies. It does see considerable use, however, in high-frequency switching power supplies in what are called forward converter and fly-back converter topologies.

Voltage Tripler



Voltage Quadrupler



Voltage tripler: **On one half of the ac cycle, C_1 and C_3 are charged to V_{IN} (peak) through D_1, D_2, and D_3. On the opposite half of the cycle, D_2 conducts and C_2 is charged to twice V_{IN} (peak), because it sees the transformer plus the charge in C_1 as its source. (D_1 is cut off during this half cycle.) At the same time, D_3 conducts, and with the transformer and the charge in C_2 as the source, C_3 is charged to three times the transformer voltage.**

Voltage quadrupler: **Works in a similar manner as the previous one. In both these circuits, the output voltage will approach an**

exact multiple of the peak ac voltage when the output current drain is low and the capacitance values high.

Capacitance values are usually 20 to 50 µF, depending on the output current drain. Capacitor dc ratings are related to VIN (peak) by:

C_1—Greater than V_{IN} (peak) or 0.7 V_{IN} (RMS)

C_2—Greater than 2 V_{IN} (peak) or 1.4 V_{IN} (RMS)

C_3—Greater than 3 V_{IN} (peak) or 2.1 V_{IN} (RMS)

C_4—Greater than 4 V_{IN} (peak) or 2.8 V_{IN} (RMS)

FIGURE 4.22

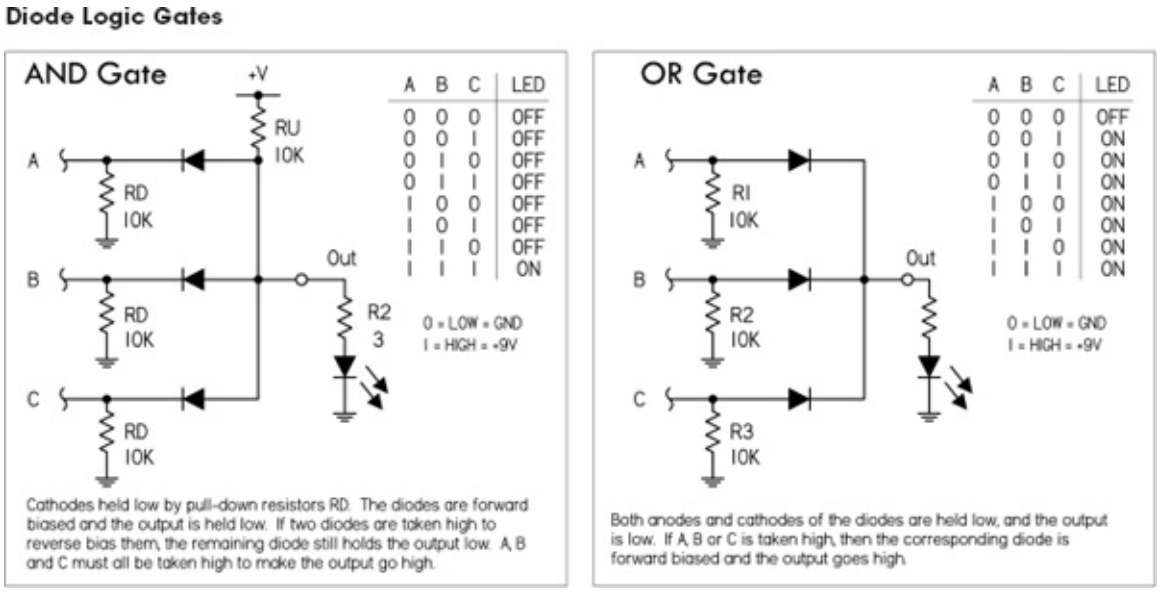
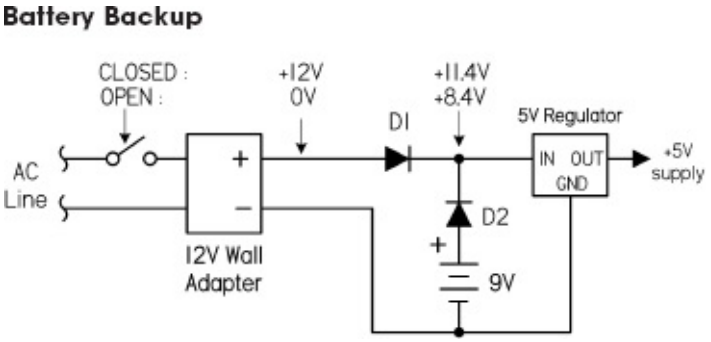
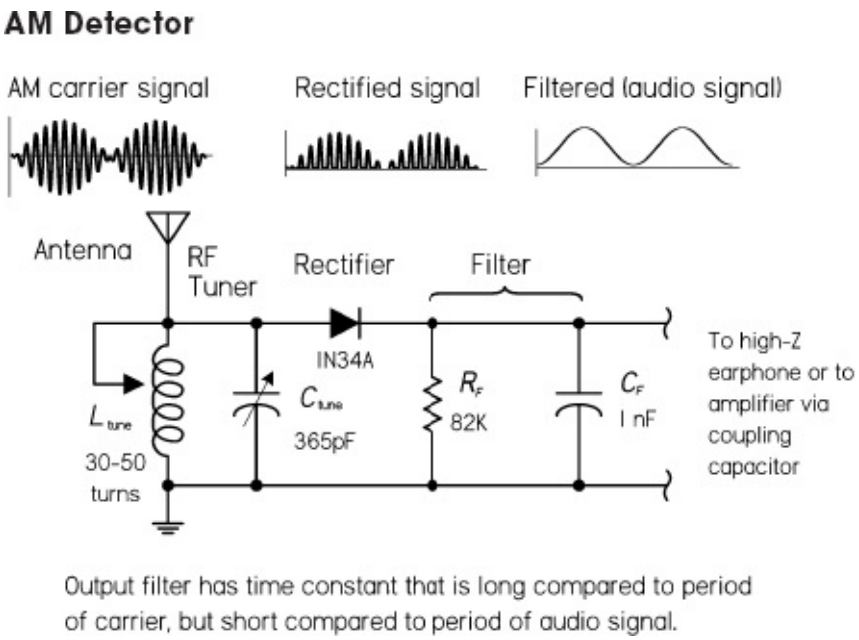


FIGURE 4.23 These simple diode logic gates are useful for learning the basics of digital logic, and can also be adapted for non-logic-level electronics (e.g., higher-voltage and power analog-like circuits)—see the following battery-backup example (Fig. 4.24). When designing high-power circuits, make sure your diodes have the proper PIV and current ratings for the job. It's also important to note that the recovery time of power diodes won't be as fast as digital logic ICs or fast-switching diodes.



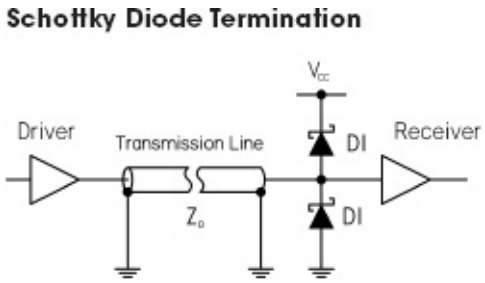
Devices are powered by a wall adapter with battery backup, typically diode-OR for the battery and wall-adapter connections, as shown in Fig. 4.24. Normally if the switch is closed, power is delivered to the load from the 12-V wall adapter through D_1; D_2 is reverse-biased (off), since its negative end is 2.4V more positive than its positive end. If power is interrupted (switch opened), D_1 stops conducting, and the battery kicks in, sending current through D_2 into the load; D_1 blocks current from flowing back into the wall adapter. There is a penalty for using diodes for battery backup, however, since the diode in series with the battery limits the minimum voltage at which the battery can supply power (around a 0.6-V drop for silicon p-n junction, 0.4 V for Schottky). Better battery-backup designs implement transistors or special ICs that contain an internal comparator which switches over battery power through a low-resistance transistor without the 0.6-V penalty. Check out MAXIM's website for some example ICs.

FIGURE 4.24



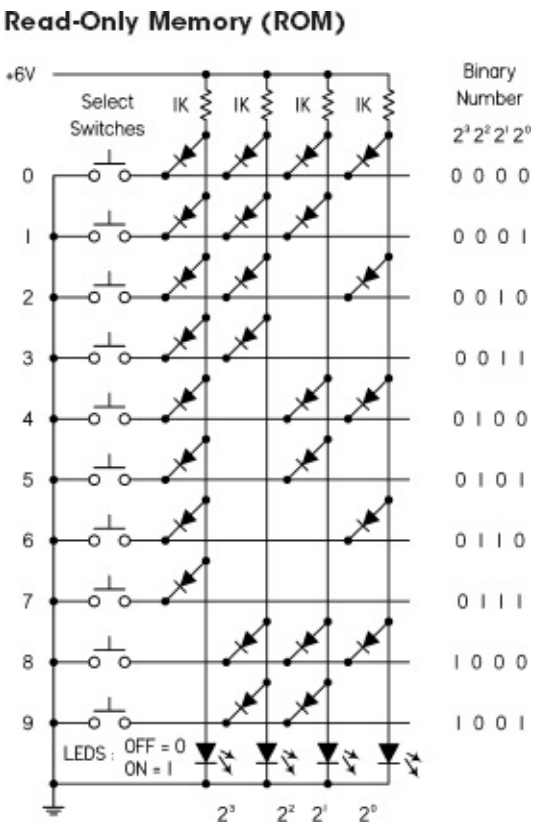
Diodes are often used in the detection of amplitude modulated (AM) signals, as demonstrated in the simple AM radio in Fig. 4.25. Within an AM radio signal, an RF carrier signal of constant high frequency (550 to 1700 kHz) has been amplitude modulated with an audio signal (10 to 20,000 Hz). The audio information is located in duplicate in both upper and lower sidebands, or the envelope of the AM signal. Here, an antenna and LC-tuning circuit act to "resonate" in on the specific carrier frequency of interest (transform radio signal into corresponding electrical signal). A signal diode (e.g., 1N34) is then used to rectify out the negative portion of the incoming signal so it can be manipulated by the next dc stages. The rectified signal is then stripped of its high-frequency carrier by passing through a low-pass filter. The output signal is the audio signal. This signal can be used to drive a simple crystal earpiece, a modern sensitive headphone, or a telephone receiver earpiece. (Low-impedance earphones or speakers will require additional amplification via a coupling capacitor of 1 µF or so.)

FIGURE 4.25



Schottky diode termination can be used to counteract the high-speed transmission line effects, which cause over/undershoots from signal reflections, reduce noise margins, and destroy timing. These types of distortions can cause false triggering in clock lines and erroneous data on address, data, and control lines, as well as contribute significantly to clock and signal jitter. For applications where transmission line impedance is variable or unknown, it's not possible to specify a termination resistance value—an alternative is needed. The Schottky diode termination has the ability to maintain signal integrity, save significant power, and permit flexible system design. A Schottky diode termination consists of a diode series combination, where one diode clamps to V_{CC}, or supply voltage, and the other to ground. The diodes at the end of the transmission line minimize the effect of reflection via a clamping operation. The top diode clamps voltages that exceed V_{CC} by the forward-bias threshold limit. This clamping will minimize overshoots caused by reflections. For falling edge signals, a clamp diode to ground affects a similar termination. This clamping function does not depend on matching the transmission line characteristic impedance, making it useful in situations where the line impedance is unknown or variable.

FIGURE 4.26



This circuit is a simple read-only memory (ROM) made with diodes. Here, the ROM acts as a decimal-to-binary encoder. With no buttons pressed, all LEDs are lit. If 1 is pressed, current from the supply is diverted away from the 2^3, 2^2, and 2^1 lines via the diodes to ground, but is allowed to pass on the 20 line, thus presenting 0001 on the LED readout. In reality, using a PROM such as this for encoding—or anything else, for that matter—isn't practical. Usually there is a special encoder IC you buy or you simply take care of the encoding—say, with a multiplexed keypad that's interfaced with a microcontroller—the actual encoding is taken care of at the programming level. At any rate, it's a fun circuit, and this gives you a basic idea of how read-only memory works.

FIGURE 4.27

4.2.6 Zener Diodes

A zener diode acts like a two-way gate to current flow. In the forward direction, it's easy to push open; only about 0.6 V—just like a standard diode. In the reverse direction, it's harder to push open; it requires a voltage equal to the *zener's breakdown voltage* V_Z. This breakdown voltage can be anywhere between 1.8 and 200 V, depending on the model (1N5225B = 3.0 V, 1N4733A = 5.1 V, 1N4739A = 9.1 V, etc.). Power ratings vary from around 0.25 to 50 W.

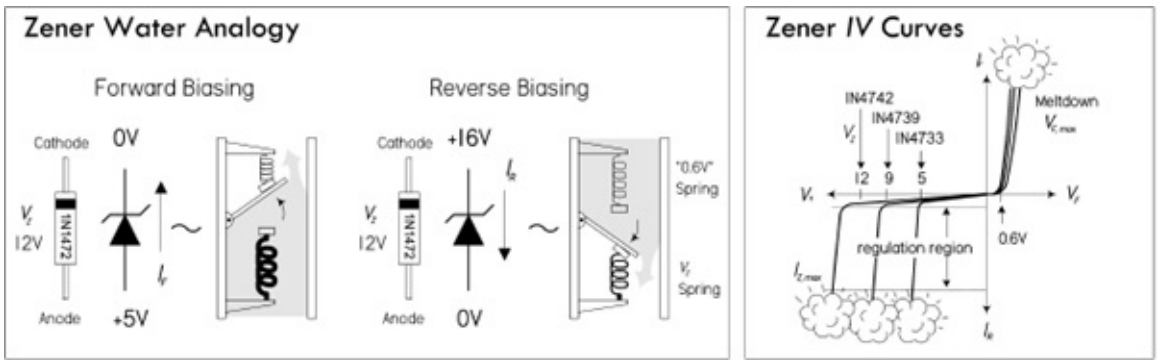


FIGURE 4.28 The reverse-bias direction is the standard configuration used in most applications, along with a series resistor. In this configuration, the zener diode acts like a pressure release value, passing as much current as necessary to keep the voltage across it constant, equal to V_Z. In other words, it can act as a voltage regulator. See application in Fig. 4.29.

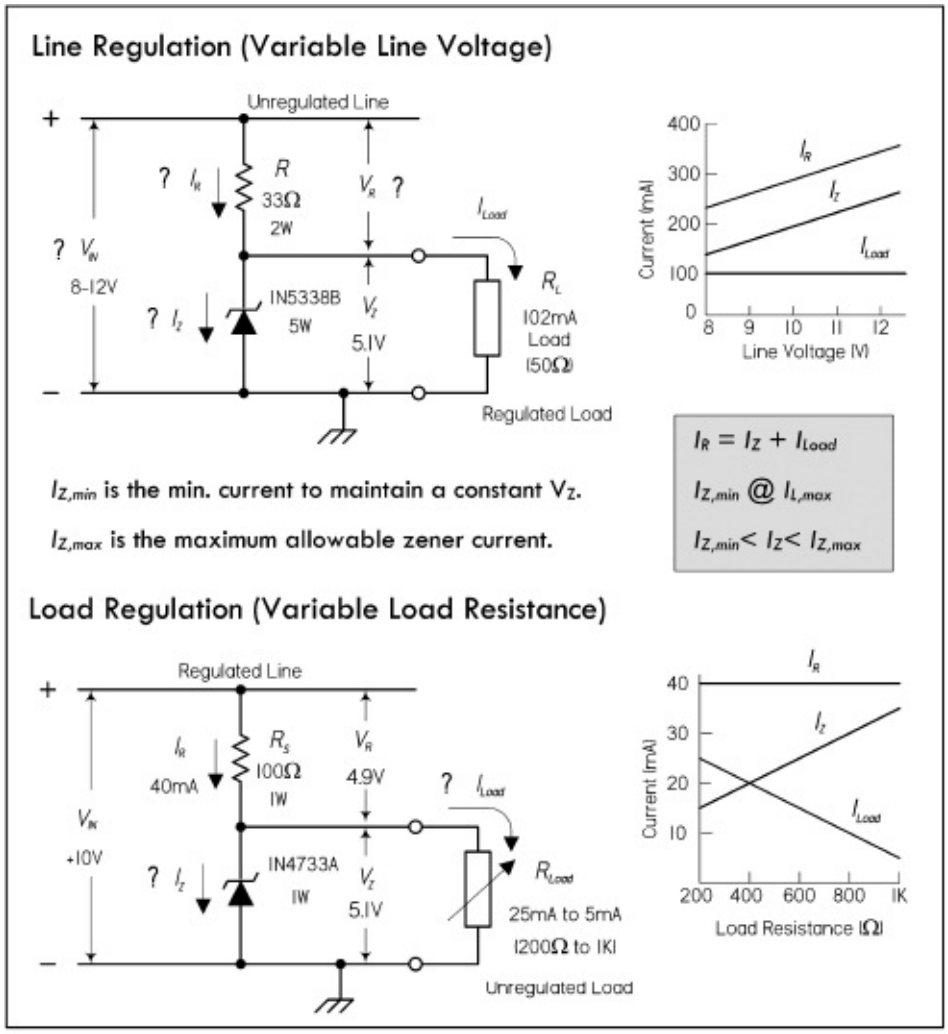


FIGURE 4.29 ***Line regulation example:*** **If the line voltage increases, it will cause an increase in line current. Since the load voltage is constant (maintained by the zener), the increase in line current will result in an increase in zener current, thus maintaining a constant load current. If the line voltage decreases, less line current results, and less current is passed by the zener. See graph in Fig. 4.29, top right.**

Load regulation example: **If the load voltage attempts to decrease as a result of decreased load resistance (increased load current), the increase in load current is offset by the decrease in zener current. The voltage across the load will remain fairly constant. If the load voltage attempts to increase due to an increase in load resistance (decrease in load current), the decrease in load current is offset by an increase in zener current. See graph in Fig. 4.29, bottom right.**

The following formulas can be used when selecting the component values:

$$R_S = \frac{V_{in,min} - V_Z}{I_{Z,min} + I_{L,max}}; \quad P_R = \frac{(V_{in,max} - V_Z)^2}{R_S}$$

$$P_{Z,max} = V_Z \frac{(V_{in,max} - V_Z)}{R_S}$$

See Problem 3 at end of this section for a design example.

Note that zener regulators are somewhat temperature dependent and aren't the best choice for critical applications. A linear regulator IC, though more expensive, is less dependent on temperature variations due to an internal error amplifier. They do typically use an internal zener to supply the reference, nonetheless.

Zener Voltage Regulator

These circuits act as voltage regulators, preventing any supply voltage or load current variations from pulling down the voltage supplied to the load. The following explains how the zener diode compensates for both line and load variations.

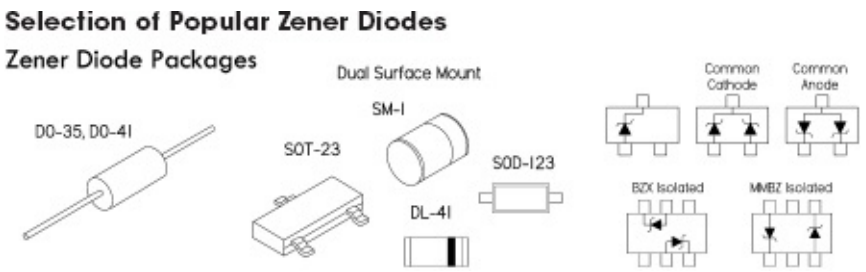


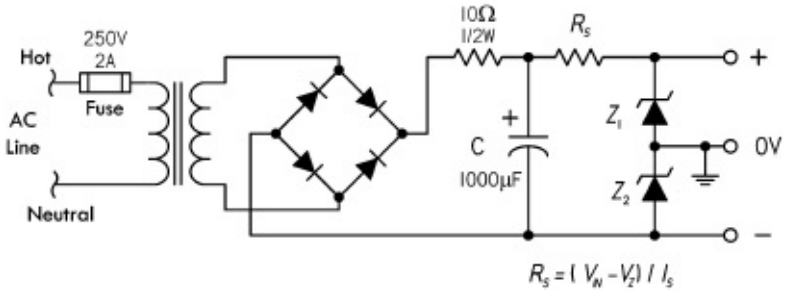
FIGURE 4.30

TABLE 4.2

ZENER VOLTS VZ VOLTS	CASE AND POWER RATING					
	AXIAL LEAD			SURFACE MOUNT		
	500 MW	1 W	5 W	200 MW	500 MW	1 W
2.4	1N5221B			BZX84C2V4, MMBZ5221B	BZT52C2V4	
2.7	1N5222B			BZX84C2V7	BZT52C2V7	
3.0	1N5225B			BZX84C3V0, MMBZ52251B	BZT52C3V0, ZMM5225B	
3.3	1N5226B	1N4728A	1N5333B	BZX84C3V3, MMBZ5226B	BZT52C3V3, ZMM5226B	ZM4728A
3.6	1N5227B	1N4729A	1N5334B	BZX84C3V6, MMBZ5227B	BZT52C3V6, ZMM5227B	
3.9	1N5228B	1N4730A	1N5335B	BZX84C3V9, MMBZ5228B	BZT52C3V9, ZMM5228B	ZM4730A
4.3	1N5229B	1N4731A	1N5336B		BZT52C4V3, ZMM5229B	ZM4731A
4.7	1N5230B	1N4732A	1N5337B	BZX84C4V7, MMBZ5230B	BZT52C4V7, ZMM5230B	ZM4732A
5.1	1N5231B	1N4733A	1N5338B	BZX84C5V1, MMBZ5231B	BZT52C5V1, ZMM5231B	SMAZ5V1, ZM4733A
5.6	1N5232B	1N4734A	1N5339B	BZX84C5V6, MMBZ5232B	BZT52C5V6, ZMM5232B	SMAZ5V6, ZM4734A
6.0	1N5233B		1N5340B		BZT52C6V0, ZMM52330B	
6.2	1N5234B	1N4735A	1N5341B	BZX84C6V2, MMBZ5234B	BZT52C6V2, ZMM5234B	SMAZ6V2, ZM4735A
6.8	1N5235B	1N4736A	1N5342B	BZX84C6V8, MMBZ5235B	BZT52C6V8, ZMM5235B	SMAZ6V8, ZM4736A
7.5	1N5236B	1N4737A	1N5343B	BZX84C7V5, MMBZ5236B	BZT52C7V5, ZMM5236B	SMAZ7V5, ZM4737A
8.2	1N5237B	1N4738A	1N5344B	BZX84C8V2, MMBZ5237B	BZT52C8V2, ZMM5237B	SMAZ8V2, ZM4738A
8.7	1N5238B		1N5345B		BZT52C8V7, ZMM5238B	
9.1	1N5239B	1N4739A	1N5346B	BZX84C9V1, MMBZ5239B	BZT52C9V1, ZMM5239B	SMAZ9V1, ZM4739A
10.0	1N5240B	1N4740A	1N5347B	BZX84C10	BZT52C10, ZMM5240B	SMAZ10, ZM4740A
11	1N5241B	1N4741A	1N5348B	BZX84C11, MMBZ5241B	BZT52C11, ZMM5241B	ZM4741A
12	1N5242B	1N4742A	1N5349B	BZX84C12, MMBZ5242B	BZT52C12, ZMM5242B	SMAZ12, ZM4742A
13	1N5243B	1N4743A	1N5350B	MMBZ5243B	BZT52C13, ZMM5243B	ZM4743A
14	1N5244B		1N5351B		BZT52C14, ZMM5244B	
15	1N5245B	1N4744A	1N5352B	BZX84C15, MMBZ5245B	BZT52C15, ZMM5245B	SMAZ15, ZM4744A
16	1N5246B	1N4745A	1N5353B	BZX84C16, MMBZ5246B	BZT52C16, ZMM5246B	SMAZ16, ZM4745A
17	1N5247B		1N5354B		ZMM5247B	
18	1N5248B	1N4746A	1N5355B	BZX84C18, MMBZ5248B	BZT52C18, ZMM5248B	SMAZ18, ZM4746A
19	1N5249B		1N5356B		ZMM5249B	
20	1N5250B	1N4747A	1N5357B	BZX84C20, MMBZ5250B	BZT52C20, ZMM5250B	SMAZ20, ZM4747A
22	1N5251B	1N4748A	1N5358B	BZX84C22, MMBZ5251B	BZT52C22, ZMM5251B	SMAZ22, ZM4748A
24	1N5252B	1N4749A	1N5359B	BZX84C24, MMBZ5252B	BZT52C24, ZMM5252B	SMAZ24, ZM4749A
25	1N5253B		1N5360B		ZMM5253B	
27	1N5254B	1N4750A	1N5361B	BZX84C27, MMBZ5254B	BZT52C27, ZMM5254B	SMAZ27, ZM4750A
28	1N5255B		1N5362B	MMBZ5255B	ZMM5255B	
30	1N5256B	1N4751A	1N5363B	BZX84C30	BZT52C30, ZMM5256B	SMAZ30, ZM4751A
33	1N5257B	1N4752A	1N5364B	BZX84C33	BZT52C33, ZMM5257B	SMAZ33, ZM4752A
36	1N5258B	1N4753A	1N5365B	BZX84C36, MMBZ5258B	BZT52C36, ZMM5258B	SMAZ36, ZM4753A
39	1N5259B	1N4754A	1N5366B	BZX84C39, MMBZ5259B	BZT52C39, ZMM5259B	SMAZ39, ZM4754A
43	1N5260B	1N4755A	1N5367B		BZT52C43, ZMM5260B	ZM4755A
47	1N5261B	1N4756A	1N5368B		BZT52C47, ZMM5261B	ZM4756A
51	1N5262B	1N4757A	1N5369B		BZT52C51, ZMM5262B	ZM4757A
56	1N5263B	1N4758A	1N5370B			ZM4758A
60	1N5264B		1N5371B			
62	1N5265B	1N4759A	1N5372B		ZMM5265B	ZM4759A
68	1N5266B	1N4760A	1N5373B		ZMM5266B	ZM4760A
75	1N5267B	1N4761A	1N5374B			ZM4761A
82	1N5268B	1N4762A	1N5375B			ZM4762A
87	1N5269B					
91	1N5270B	1N4763A	1N5377B			ZM4763A
100	1N5271B	1N4764A	1N5378B			ZM4764A

4.2.7 Zener Diode Applications

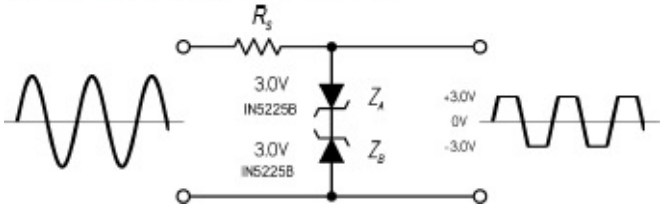
Split Supply from Single Transformer Winding



Here's a method for obtaining a split supply from a non-center-tapped transformer using two zener diodes. Z_1 and Z_2 are selected of equal voltage and power rating for desired split voltage and load. As with the previous example, the temperature dependency of the zener diodes makes this arrangement less accurate than a supply that uses two separate regulator ICs. However, it's a simple alternative for noncritical applications. See Chap. 11 on power supplies.

FIGURE 4.31

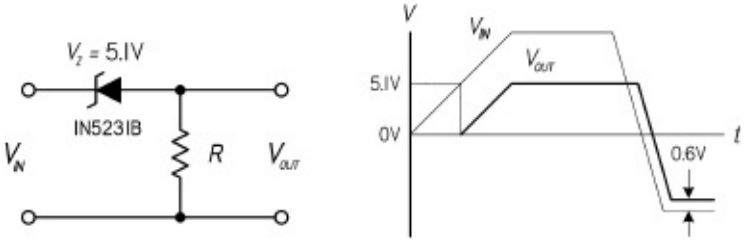
Waveform Modifier and Limiter



Two opposing zener diodes act to clip both halves of an input signal. Here a sine wave is converted to a near squarewave. Besides acting to reshape a waveform, this arrangement can also be placed across the output terminal of a dc power supply to prevent unwanted voltage transients from reaching an attached load. The breakdown voltages in that case must be greater than the supply voltage, but smaller than the maximum allowable transient voltage. A single bidirectional TVS does the same thing—see the section on transient suppressors.

FIGURE 4.32

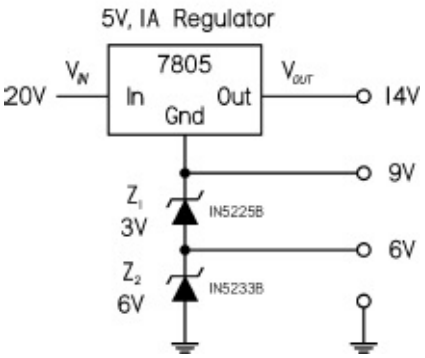
Voltage Shifter



This circuit shifts the input voltage down by an amount equal to the breakdown voltage of the zener diode. As the input goes positive, the zener doesn't go into breakdown until it reaches 5.1 V (for the 1N5281B). After that, the output follows the input, but shifted 5.1 V below it. When the input goes negative, the output will follow the input, but shifted by 0.6 V—the forward threshold voltage drop of the zener.

FIGURE 4.33

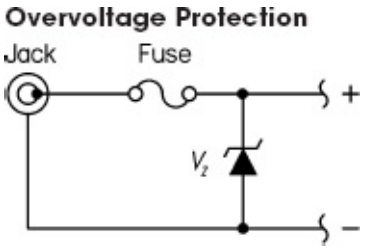
Voltage Regulator Booster



Zener diodes can be used to raise the level of a voltage regulator and obtain different regulated voltage outputs. Here 3-V and 6-V zener diodes are placed in series to push the reference ground of a 5 V regulator IC up 9 V to a total of 14 V. Note that in

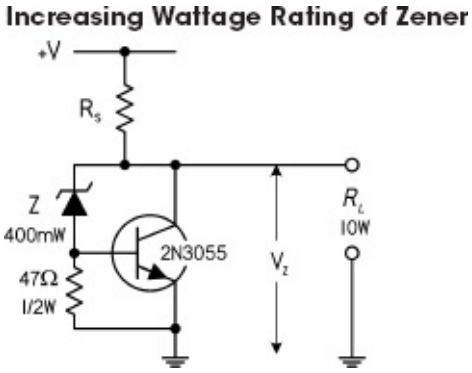
real designs, capacitors may be required at the input and output. See the section on voltage regulator ICs.

FIGURE 4.34



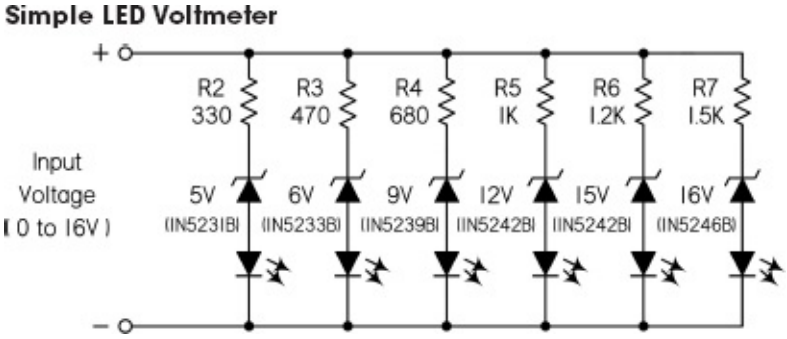
If excessive voltage is applied to the jack (say, via an incorrectly rated wall plug-in supply), the zener diode will conduct until the fuse is blown. The breakdown voltage of the zener should be slightly above the maximum tolerable voltage that the load can handle. Either a fast- or a slow-blow fuse can be used, depending on the sensitivity of the load. The current and voltage ratings of the fuse must be selected according to the expected limits of the application. Note that there are other, similar overvoltage protection designs that use special devices, such as TVSs and varistors. These devices are cheap and are very popular in design today.

FIGURE 4.35



Here's a simple circuit that effectively increases the wattage rating (current-handling capacity) of a zener diode by letting a power transistor take care of the majority of the regulating current. The zener itself takes a small portion of the total current and creates a base voltage/current (with the help of the base-to-ground resistor) that changes the collector-to-emitter current flow according to any variations in line or load current.

FIGURE 4.36



Here's a simple circuit voltmeter that uses the sequence of zener diodes with increasing breakdown voltages. LEDs glow in sequence as the input voltage rises. It's okay to use different zener diodes so long as the series resistors limit current through LED to a safe level. Most LEDs are happiest around 20 mA or so. You can calculate the worst-case scenario to be at the 5 V LED leg when V_{in} = 16 V. If you're looking for more sophistication, you can always use an analog-to-digital converter, along with a microcontroller and LCD or LED display.

FIGURE 4.37

4.2.8 Varactor Diodes (Variable Capacitance Diodes)

A *varactor* or variable capacitance diode (also called a *varicap*) is a diode whose junction capacitance can be altered with an applied reverse voltage. In this way, it acts as a variable capacitor. As the applied

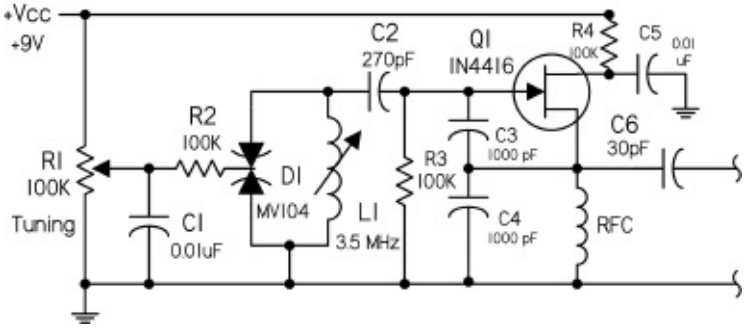
reverse voltage increases, the width of its junction increases, which decreases its capacitance. The typical capacitance range for varactors ranges from a few picofarads to over 100 pF, with a maximum reverse voltage range from a few volts to close to a hundred volts, depending on device. (Many standard diodes and zener diodes can be used as inexpensive varactor diodes, though the relationship between reverse voltage and capacitance isn't always as reliable.)

The low capacitance levels provided by a varactor usually limit its use to high-frequency RF circuits, where the applied voltage is used to change the capacitance of an oscillator circuit. The reverse voltage may be applied via a tuning potentiometer, which acts to change the overall frequency of an oscillator, or it may be applied by a modulating signal (e.g., audio signal), which acts to FM-modulate the oscillator's high-frequency carrier. See the examples that follow.

When designing with varactor diodes, the reverse-bias voltage must be absolutely free of noise, since any variation in the bias voltage will cause changes in capacitance. Unwanted frequency shifts or instability will result if the reverse-bias voltage is noisy. Filter capacitors are used to limit such noise.

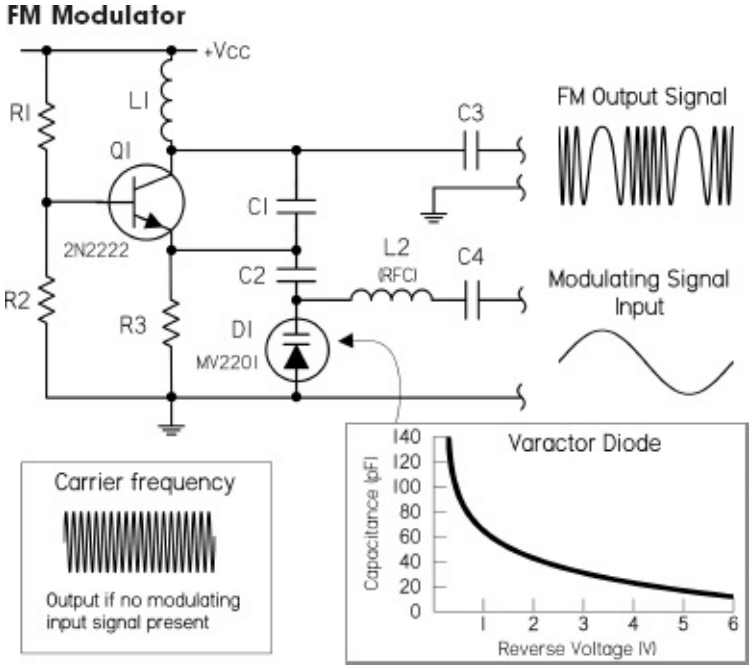
Varactors come in both single and dual forms. The dual varactor configuration contains two varactors in series-opposing, with common anodes and separate cathodes. In this configuration, the varactors acts as series capacitors that change capacitance levels together when a voltage is applied to the common anode lead. See Fig. 4.39.

Oscillator with Pot-Controlled Varactor Tuning



Unlike the preceding circuit, this circuit acts simply as a variable high-frequency oscillator, whose frequency is varied via a potentiometer (R_1). The voltage from the pot is applied to a dual varactor diode D_1 through a low-frequency filter (C_1, R_2) to ensure that the varactor bias is clean dc. This alters the effective capacitance of the D_1-L_1 tuned circuit, which changes the frequency of the entire oscillator. C_2 and C_6 are dc-blocking (ac-coupling) capacitors. Q_1 is an N-channel JFET in common drain configuration with feedback to the gate through C_3. R_3 is the gate bias resistor. R_4 is the drain voltage resistor with filter capacitor CF.

FIGURE 4.39



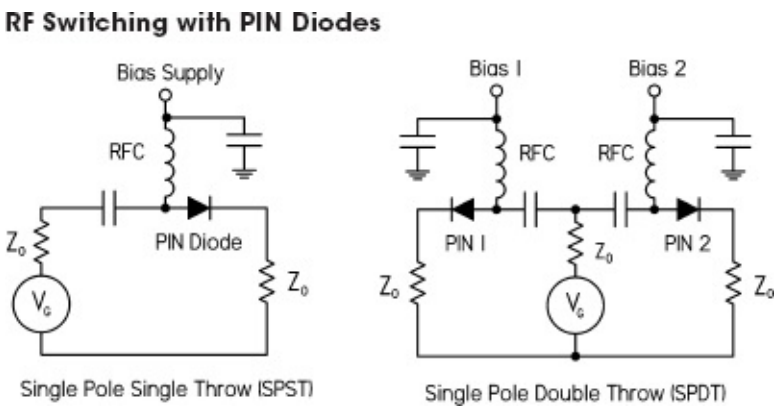
FM modulation: **FM (frequency modulation) is produced when the frequency of a carrier is changed instantaneously according to the magnitude of an applied modulating signal. (The frequency of the carrier is usually in the megahertz, while the modulating signal is typically in the hertz to kilohertz range, e.g., audio modulating radio signal.) One way to produce FM is to use a voltage-controlled oscillator. The oscillator will have an output frequency proportional to the modulating signal's amplitude. As the amplitude of the modulating signal increases, the frequency of the carrier increases. Here a Colpitts LC oscillator uses a varactor diode in place of one of its regulator capacitors that form the tuned circuit. The modulating voltage is applied across the diode and changes the diode's capacitance in proportion. This causes the oscillator frequency to change, thus generating FM in the process. L_2 (RFC) is a radiofrequency choke that prevents high-frequency signals from feeding backing into the modulating source. C_3 and C_4 are ac-coupling capacitors. The rest of the components go into making the Colpitts oscillator.**

FIGURE 4.38

4.2.9 PIN Diodes

PIN diodes are used as RF and microwave switches. To high-frequency signals, the PIN diode acts like a variable resistor whose value is controlled by an applied dc forward-bias current. With a high dc forward bias, the resistance is often less than an ohm. But with a small forward bias, the resistance appears very large (kiloohms) to high-frequency signals. PIN diodes are constructed with a layer of intrinsic (undoped) semiconductor placed between very highly doped *p*-type and *n*-type material, creating a PIN junction.

In terms of application, PIN diodes are used primarily as RF and microwave switches—even at high power levels. A common application is their use as transmit/receive switches in transceivers operating from 100 MHz and up. They are also used as photodetectors in fiber-optic systems. For the most part, you'll never need to use them, unless you are a graduate student in electrical engineering or physics, or are working for a high-tech firm.



At RF frequencies, switching is very finicky, requiring special design techniques to minimize signal contamination and degradation. Here are two switching circuits that make use of PIN diodes. In the SPST switch circuit, a signal from a RF generator (VG), can be allowed to pass, or can be prevented from passing to the load by applying a bias voltage to the PIN diode. The RFC is a high-frequency choke to prevent RF from entering bias supply, while the capacitor to ground is used to supply clean dc at the bias input. The SPDT switch circuit is similar to the first, but with, of course, two bias inputs.

FIGURE 4.40

4.2.10 Microwave Diodes (IMPATT, Gunn, Tunnel, etc.)

There are a number of diodes that you'll probably never have to use, but they are around nevertheless. These diodes are used for very special purposes at the high-frequency end—microwave and millimeter wave (>20 GHz) range, often in microwave amplifiers and oscillators. Most standard diodes and bipolar transistors usually won't cut it at such high speeds, due to the relatively slow diffusion or migration of charge carriers across semiconductor junctions. With the tunnel, Gunn, IMPATT, and other diodes, the variable effects that lead to useful alterations in, say, an amplifier's gain or an oscillator's resonant frequency involve entirely different physics—physics that allows for alterations at essentially the speed of light. The physics may be electron tunneling (through electrostatic barrier separating *p*-type and *n*-type regions, rather than being thermionically emitted over the barrier, as generally occurs in a diode)—tunnel diode. Or it may be due to a negative resistance at forward biasing because of an increase in effective mass (slowing down) of electrons due to complex conduction band symmetry—Gunn diodes. It may also be a negative resistance resulting in electrons moving to higher, less mobile bands, reducing current flow with applied forward bias—IMPATT diodes. Anyway, you get the idea—it's hairy high-frequency stuff that should probably be left to the experts. (Note: TRAPATT and Baritt diodes are also used in microwave applications.)

4.2.11 Problems

Problem 1: What does this circuit do? What's the final output voltage? What are the individual voltage drops across each diode with plug tip-positive and plug tip-negative? (Assume each diode has a 0.6 V forward voltage drop.) To prevent diode meltdown, what would be the minimum load resistance, assuming 1N4002 diodes?

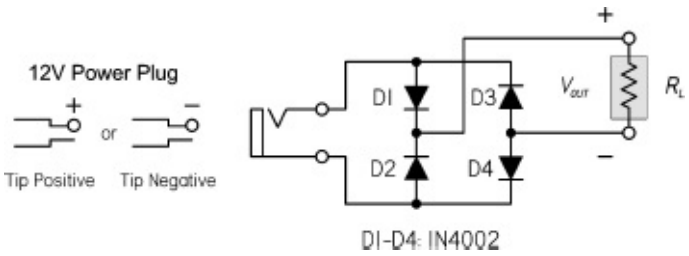


FIGURE 4.41

Answer: Polarity protection circuit that will output the same polarity regardless of the polarity applied to input. The final output voltage is 11.4 V. Tip-positive: VD_1 = 0.6 V, VD_2 = 11.4 V, VD_3 = 11.4 V, VD_4 = 0.6 V; Tip-negative: VD_1 = 11.4 V, VD_2 = 0.6 V, VD_3 = 0.6 V, VD_4 = 11.4 V. Load resistance should not drop below 11.4 Ω, assuming 1N4002 diodes, since they have a maximum current rating of 1 A. It's a good idea to keep the current to around 75 percent of the maximum value for safety, so 15 Ω would be a better limit.

Problem 2: What does the output look like for the circuit to the left in Fig. 4.42? What happens if a load of 2.2K is attached to the output?



FIGURE 4.42

Answer: Clamp circuit, where the output is shifted so that it's practically pure alternating dc, for the exception of a 0.6 V negative dip due to the diode drop. This gives a maximum peak of 27.6 V and a minimum of -0.6 V. (Recall $V_{peak} = 1.41 \times V_{rms}$.) With the load attached, the output level decreases slightly—the capacitor/load resistor acts like a high-pass filter, with a cutoff frequency of $1/(2\pi RC)$. In simulation, the output goes to 8.90 V(RMS) or 24.5 peak, −0.6 V minimum.

Problem 3: A 10- to 50-mA load requires a regulated 8.2 V. With a 12 V ± 10 percent power supply and 8.2 V zener diode. What series resistance is required? Assume from the data sheets (or experimentation) that the zener diode's minimum regulation current is 10 mA. Determine the power ratings for the resistor and zener diode.

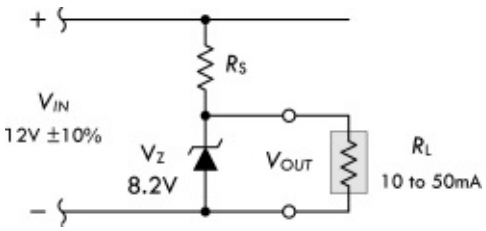
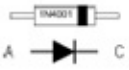
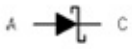
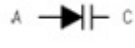


FIGURE 4.43

Answer: $V_{in,max} = 13.2$ V, $V_{in,min} = 10.8$ V: $R_S = (10.8\text{ V} - 8.2\text{ V})/(10\text{ mA} + 50\text{ mA}) = 43\ \Omega$; $P_R = (13.2\text{ V} - 8.2\text{ V})^2/(43\ \Omega) = 0.58$ W; $P_Z = 8.2\text{ V}(13.2\text{ V} - 8.2\text{ V})/(43\ \Omega) = 0.95$ W. See Fig. 4.29 for details.

TABLE 4.3

DIODE TYPE	SYMBOL	MODE OF OPERATION
p-n Junction		Acts as one-way gate to current-flow, from anode (A) to cathode (C). Comes in silicon and germanium types. Both require a forward-bias voltage to conduct; typically 0.6 to 1.7 V for silicon, and 0.2 to 0.4 V for germanium. Used in rectification, transient suppression, voltage multiplication, RF demodulation, analog logic, clamps, fast switches, and voltage regulation.
Schottky		Similar in operation to p-n junction diode, but designed with special metal semiconductor junction instead of a p-n junction. This provides for extremely low junction capacitance that stores less charge. Results of this junction yield quicker switching times, useful in fast clamping and high-frequency applications approaching the gigahertz range. Also, generally has a lower forward-bias voltage of around 0.4 V (average)—but can be from 0.15 to 0.9 V or more. Used in similar applications as p-n junction diode, but offers better low-signal level detection, speed, and low-power loss in rectification due to low forward threshold.
Zener		Conducts from A to C like p-n junction diode, but will also conduct from C to A if the applied reverse voltage is greater than the zener's breakdown voltage rating V_Z. Acts like a voltage-sensitive control valve. Comes with various breakdown voltages—1.2 V, 3.0 V, 5.1 V, 6.3 V, 9 V, 12 V, etc., and power ratings. Applications include voltage regulation, waveform clipping, voltage shifting, and transient suppression.
LED & Laser		Light-emitting diode (LED) emits a near constant wavelength of light when forward-biased (A > C) by a voltage of about 1.7 V. Comes in various wavelengths (IR through visible), sizes, power ratings, etc. Used as indicator and emitting source in IR and light-wave communications. Laser diode is similar to LED, but provides a much narrower wavelength spectrum (about 1 nm compared to around 40 nm for LED), usually in the IR region. They have very fast response times (1ns). These features provide clean signal characteristics useful in fiber-optic systems, where minimized dispersion effects, efficient coupling, and limited degradation over long distances are important. They are also used in laser pointers, CD/DVD players, bar-code readers, and in various surgical applications.
Photo		Generates a current when exposed to light, or can be used to alter current flow passing through it when the light intensity changes. Operates in reverse-bias direction (current flows from C to A) when exposed to light. Current increases with light intensity. Very fast response times (ns). Not as sensitive as phototransistors, but their linearity can make them useful in simple light meters.
Varactor (Varicap)		Acts like a voltage-sensitive variable capacitor, whose capacitance decreases as the reverse-bias voltage on the diode increases. Designed with a junction specifically formulated to have a relatively large range of capacitance values for a modest range of reverse-bias voltages. Capacitance range in the picofarad range, so they are usually limited to RF applications, such as tuning receivers and generating FM.
PIN, IMPATT Gunn, Tunnel, etc.		Most of these are resistance devices used in RF, microwave, and millimeter wave applications (e.g., amplifiers and oscillators). Unique conduction physics yields much faster response times when compared to standard diodes that use charge carrier dispersion across a p-n junction.

4.3 Transistors

Transistors are semiconductor devices that act as either electrically controlled switches or amplifier controls. The beauty of transistors is the way they can control electric current flow in a manner similar to the way a faucet controls the flow of water. With a faucet, the flow of water is controlled by a control knob. With a transistor, a small voltage and/or current applied to a control lead acts to control a larger electric flow through its other two leads.

Transistors are used in almost every electric circuit you can imagine. For example, you find transistors in switching circuits, amplifier circuits, oscillator circuits, current-source circuits, voltage-regulator circuits, power-supply circuits, digital logic ICs, and almost any circuit that uses small control signals to control larger currents.

4.3.1 Introduction to Transistors

Transistors come in a variety of designs and come with unique control and current-flow features. Most transistors have a variable current-control feature, but a few do not. Some transistors are normally off until a voltage is applied to the base or gate, whereas others are normally on until a voltage is applied. (Here, *normally* refers to the condition when the control lead is open circuited. Also, *on* can represent a variable amount of current flow.) Some transistors require both a small current and a small voltage applied to their control lead to function, whereas others only require a voltage. Some transistors require a negative voltage and/or output current at their base lead (relative to one of their other two leads) to function, whereas others require a positive voltage and/or input current at their base.

The two major families of transistors include *bipolar transistors* and *field-effect transistors* (FETs). The major difference between these two families is that bipolar transistors require a biasing input (or output) current at their control leads, whereas FETs require only a voltage—practically no current. [Physically speaking, bipolar transistors require both positive (holes) and negative (electrons) carriers to operate, whereas FETs only require one charge carrier.] Because FETs draw little or no current, they have high input impedances ($\sim 10^{14}\ \Omega$). This high input impedance means that the FET's control lead will not have much influence on the current dynamics within the circuit that controls the FET. With a bipolar transistor, the control lead may draw a small amount of current from the control circuit, which then combines with the main current flowing through its other two leads, thus altering the dynamics of the control circuit.

In reality, FETs are definitely more popular in circuit design today than bipolar transistors. Besides drawing essentially zero input-output current at their control leads, they are easier to manufacture, cheaper to make (require less silicon), and can be made extremely small—making them useful elements in integrated circuits. One drawback with FETs is in amplifier circuits, where their transconductance is much lower than that of bipolar transistors at the same current. This means that the voltage gain will not be as large. For simple amplifier circuits, FETs are seldom used unless extremely high input impedances and low input currents are required.

Table 4.4 provides an overview of some of the most popular transistors. Note that the term *normally* used in this chart refers to conditions where the control lead (e.g., base, gate) is shorted (is at the same potential) with one of its channel leads (e.g., emitter, source). Also, the terms *on* and *off* used in this chart are not to be taken too literally; the amount of current flow through a device is usually a variable quantity, set by the magnitude of the control voltage. The transistors described in this chart will be discussed in greater detail later on in this chapter.

TABLE 4.4 Overview of Transistors

TRANSISTOR TYPE	SYMBOL	MODE OF OPERATION
Bipolar		Normally off, but a small input current and a small positive voltage at its base (B)—relative to its emitter (E)—turns it on (permits a large collector-emitter current). Operates with $V_C > V_E$. Used in switching and amplifying applications.
		Normally off, but a small output current and negative voltage at its base (B)—relative to its emitter (E)—turns it on (permits a large emitter-collector current). Operates with $V_E > V_C$. Used in switching and amplifying applications.
Junction FET		Normally on, but a small negative voltage at its gate (G)—relative to its source (S)—turns it off (stops a large drain-source current). Operates with $V_D > V_S$. Does not require a gate current. Used in switching and amplifying applications.
		Normally on, but a small positive voltage at its gate (G)—relative to its source (S)—turns it off (stops a large source-drain current). Operates with $V_S > V_D$. Does not require a gate current. Used in switching and amplifying applications.
Metal oxide semiconductor FET (MOSFET) (depletion)		Normally on, but a small negative voltage at its gate (G)—relative to its source (S)—turns it off (stops a large drain-source current). Operates with $V_D > V_S$. Does not require a gate current. Used in switching and amplifying applications.
		Normally on, but a small positive voltage at its gate (G)—relative to its source (S)—turns it off (stops a large source-drain current). Operates with $V_S > V_D$. Does not require a gate current. Used in switching and amplifying applications.
Metal oxide semiconductor FET (MOSFET) (enhancement)		Normally off, but a small positive voltage at its gate (G)—relative to its source (S)—turns it on (permits a large drain-source current). Operates with $V_D > V_S$. Does not require a gate current. Used in switching and amplifying applications.
		Normally off, but a small negative voltage at its gate (G)—relative to its source (S)—turns it on (permits a large source-drain current). Operates with $V_S > V_D$. Does not require a gate current. Used in switching and amplifying applications.
Unijunction FET (UJT)		Normally a very small current flows from base 2 (B_2) to base 1 (B_1), but a positive voltage at its emitter (E)—relative to B_1 or B_2—increases current flow. Operates with $V_{B2} > V_{B1}$. Does not require a gate current. Only acts as a switch.

4.3.2 Bipolar Transistors

Bipolar transistors are three-terminal devices that act as electrically controlled switches or as amplifier controls. These devices come in either *npn* or *pnp* configurations, as shown in Fig. 4.44. An *npn* bipolar transistor uses a small input current and positive voltage at its base (relative to its emitter) to control a much larger collector-to-emitter current. Conversely, a pnp transistor uses a small output base current and negative base voltage (relative its emitter) to control a larger emitter-to-collector current.

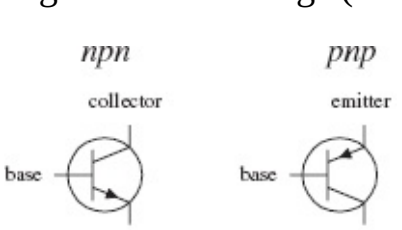


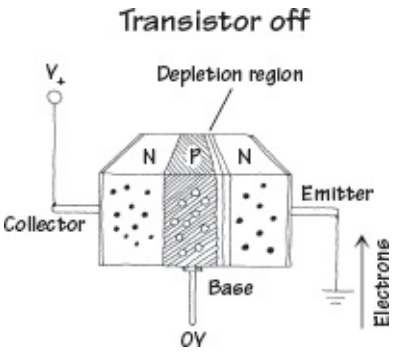
FIGURE 4.44

Bipolar transistors are incredibly useful devices. Their ability to control current flow by means of applied control signals makes them essential elements in electrically controlled switching circuits, current-regulator circuits, voltage-regulator circuits, amplifier circuits, oscillator circuits, and memory circuits.

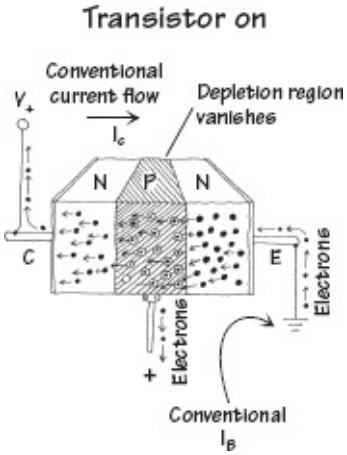
How Bipolar Transistors Work

Here is a simple model of how an *npn* bipolar transistor works. (For a *pnp* bipolar transistor, all

ingredients, polarities, and currents are reversed.)



An *npn* bipolar transistor is made by sandwiching a thin slice of *p* semiconductor between two *n*-type semiconductors. When no voltage is applied at the transistor's base, electrons in the emitter are prevented from passing to the collector side because of the *p*-*n* junction. (Remember that for electrons to flow across a *p*-*n* junction, a biasing voltage is needed to give the electrons enough energy to "escape" the atomic forces holding them to the *n* side.) Notice that if a negative voltage is applied to the base, things get even worse—the *p*-*n* junction between the base and emitter becomes reverse-biased. As a result, a depletion region forms and prevents current flow.



If a positive voltage (of at least 0.6 V) is applied to the base of an *npn* transistor, the *pn* junction between the base and emitter is forward-biased. During forward bias, escaping electrons are drawn to the positive base. Some electrons exit through the base, but—this is the trick—because the *p*-type base is so thin, the onslaught of electrons that leave the emitter get close enough to the collector side that they begin jumping into the collector. Increasing the base voltage increases this jumping effect and hence increases the emitter-to-collector electron flow. Remember that conventional currents are moving in the opposite direction to the electron flow. Thus, in terms of conventional currents, a positive voltage and input current applied at the base cause a "positive" current *I* to flow from the collector to the emitter.

FIGURE 4.45

Theory

Figure 4.46 shows a typical characteristic curve for a bipolar transistor. This characteristic curve describes the effects the base current I_B and the emitter-to-collector voltage V_{EC} have on the emitter/collector currents I_E and I_C. (As you will see in a second, I_C is practically equal to I_E.)

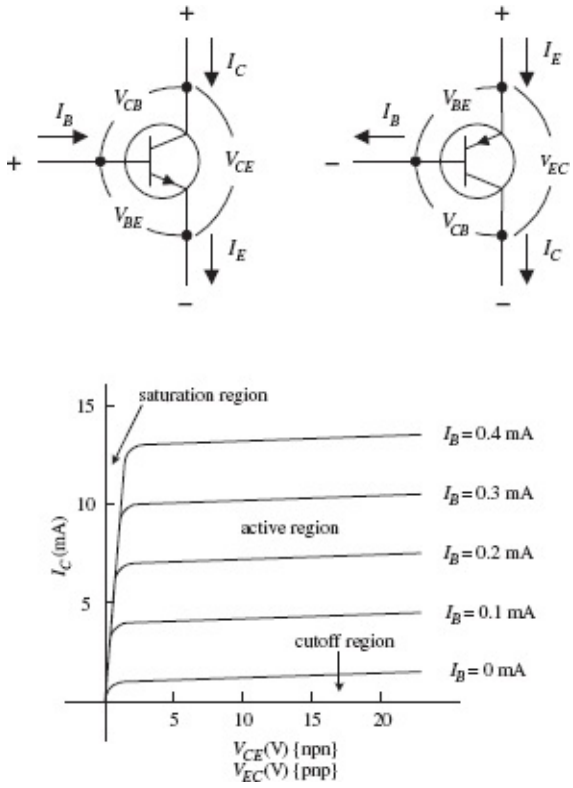


FIGURE 4.46

Some important terms used to describe a transistor's operation include saturation region, cutoff region, active mode/region, bias, and quiescent point (Q-point). *Saturation region* refers to a region of operation where maximum collector current flows and the transistor acts much like a closed switch from collector to emitter. *Cutoff region* refers to the region of operation near the voltage axis of the collector characteristics graph, where the transistor acts like an open switch—only a very small leakage current flows in this mode of operation. *Active mode/region* describes transistor operation in the region to the right of saturation and above cutoff, where a near-linear relationship exists between terminal currents (I_B, I_C, I_E). Bias refers to the specific dc terminal voltages and current of the transistor to set a desired point of active-mode operation, called the *quiescent point*, or *Q-point*.

The Formulas

The fundamental formula used to describe the behavior of a bipolar transistor (at least within the active region) is

$$I_C = h_{FE} I_B = \beta I_B$$

SOME IMPORTANT RULES

Rule 1 For an *npn* transistor, the voltage at the collector V_C must be greater than the voltage at the emitter V_E by at least a few tenths of a volt; otherwise, current will not flow through the collector-emitter junction, no matter what the applied voltage is at the base. For *pnp* transistors, the emitter voltage must be greater than the collector voltage by a similar amount.

Rule 2 For an *npn* transistor, there is a voltage drop from the base to the emitter of 0.6 V. For a *pnp* transistor, there is a 0.6-V rise from base to emitter. In terms of operation, this means that the base voltage V_B of an *npn* transistor must be at least 0.6 V greater than the emitter voltage V_E; otherwise, the transistor will pass an collector-to-emitter current. For a *pnp* transistor, V_B must be

at least 0.6 V less than V_E; otherwise, it will not pass a collector-to-emitter current.

where I_B is the base current, I_C is the collector current, and h_{FE} (also referred to as β) is the *current gain*. Every transistor has its own unique h_{FE}. The h_{FE} of a transistor is often taken to be a constant, typically around 10 to 500, but it may change slightly with temperature and with changes in collector-to-emitter voltage. (A transistor's h_{FE} is given in transistor spec tables.) A simple explanation of what the current-gain formula tells you is this: If you take a bipolar transistor with, say, an h_{FE} of 100 and then feed (*npn*) or sink (*pnp*) a 1-mA current into (*npn*) or out of (*pnp*) its base, a collector current of 100 mA will result. Now, it is important to note that the current-gain formula applies only if rules 1 and 2 are met, i.e., assuming the transistor is within the active region. Also, there is a limit to how much current can flow through a transistor's terminals and a limit to the size of voltage that can be applied across them. I will discuss these limits later in the chapter (Fig. 4.47).

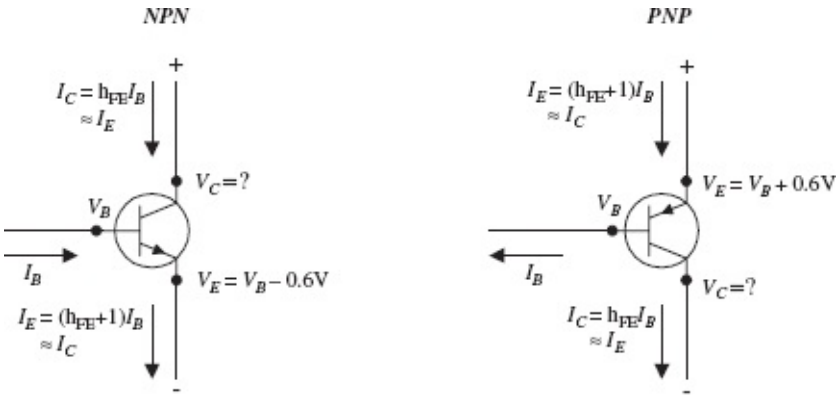


FIGURE 4.47

Now, if you apply the law of conservation of current (follow the arrows in Fig. 4.47), you get the following useful expression relating the emitter, collector, and base currents:

$$I_E = I_C + I_B$$

If you combine this equation with the current-gain equation, you can come up with an equation relating the emitter and base currents:

$$I_E = (h_{FE} + 1)I_B$$

As you can see, this equation is almost identical to the current-gain equation ($I_C = h_{FE}I_B$), with the exception of the +1 term. In practice, the +1 is insignificant as long as h_{FE} is large (which is almost always the case). This means that you can make the following approximation:

$$IE \approx IC$$

Finally, the second equation below is simply rule 2 expressed in mathematical form:

$$V_{BE} = V_B - V_E = +0.6 \text{ V } (npn)$$
$$V_{BE} = V_B - V_E = -0.6 \text{ V } (pnp)$$

Figure 4.48 shows how all the terminal currents and voltages are related. In the figure, notice that the collector voltage has a question mark next to it. As it turns out, the value of V_C cannot be determined directly by applying the formulas. Instead, V_C's value depends on the network that is connected to it. For

example, if you consider the setup shown in Fig. 4.48, you must find the voltage drop across the resistor in order to find the collector voltage. By applying Ohm's law and using the current-gain relation, you can calculate V_C. The results are shown in the figure.

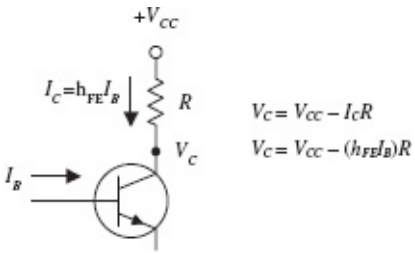


FIGURE 4.48

It is important to note that the equations here are idealistic in form. In reality, these equations may result in "unreal" answers. For instance, they tend to "screw up" when the currents and voltages are not within the bounds provided by the characteristic curves. If you apply the equations blindly, without considering the operating characteristics, you could end up with some wild results that are physically impossible.

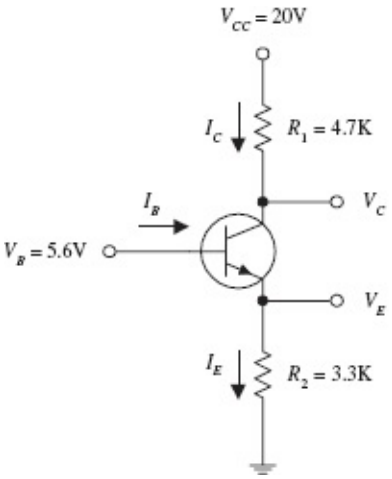
One final note with regard to bipolar transistor theory involves what is called *transresiststance* r_{tr}. Transresistance represents a small resistance that is inherently present within the emitter junction region of a transistor. Two things that determine the transresistance of a transistor are temperature and emitter current flow. The following equation provides a rough approximation of the r_{tr}:

$$r_{tr} = \frac{0.026\,\text{V}}{I_E}$$

In many cases, r_{tr} is insignificantly small (usually well below 1000 Ω) and does not pose a major threat to the overall operation of a circuit. However, in certain types of circuits, treating r_{tr} as being insignificant will not do. In fact, its presence may be the major factor determining the overall behavior of a circuit. We will take a closer look at transresistance later on in this chapter.

Here are a couple of problems that should help explain how the equations work. The first example deals with an *npn* transistor; the second deals with a *pnp* transistor.

EXAMPLE 1 Given V_{CC} = +20 V, V_B = 5.6 V, R_1 = 4.7 kΩ, R_2 = 3.3 kΩ, and h_{FE} = 100, find V_E, I_E, I_B, I_C, and V_C.



$V_E = V_B - 0.6\,\text{V}$

$V_E = 5.6\,\text{V} - 0.6\,\text{V} = 5.0\,\text{V}$

$$I_E = \frac{V_E - 0\,\text{V}}{R_2} = \frac{5.0\,\text{V}}{3300\,\Omega} = 1.5\,\text{mA}$$

$$I_B = \frac{I_E}{(1 + h_{FE})} = \frac{1.5\,\text{mA}}{(1 + 100)} = 0.015\,\text{mA}$$

$I_C = I_E - I_B \approx I_E = 1.5\,\text{mA}$

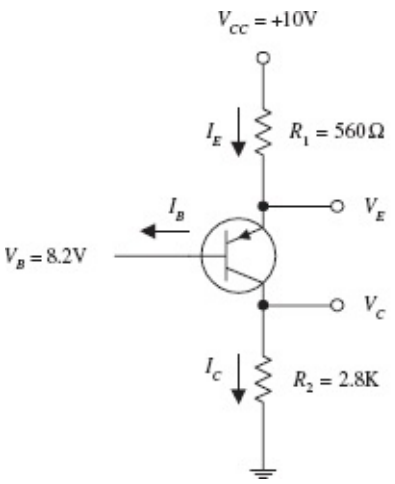
$V_C = V_{CC} - I_C R_1$

$V_C = 20\,\text{V} - (1.5\,\text{mA})(4700\,\Omega)$

$V_C = 13\,\text{V}$

FIGURE 4.49

EXAMPLE 2 Given V_{CC} = + 10 V, V_B = 8.2 V, R_1 = 560 Ω, R_2 = 2.8 kΩ, and h_{FE} = 100, find V_E, I_E, I_B, I_C, and V_C.



$$V_E = V_B + 0.6\text{ V}$$

$$V_E = 8.2\text{ V} + 0.6\text{ V} = 8.8\text{ V}$$

$$I_E = \frac{V_{CC} - V_E}{R_1} = \frac{10\text{ V} - 8.8\text{ V}}{560\,\Omega} = 2.1\text{ mA}$$

$$I_B = \frac{I_E}{(1 + h_{FE})} = \frac{2.1\text{ mA}}{(1 + 100)} = 0.02\text{ mA}$$

$$I_C = I_E - I_B \approx I_E = 2.1\text{ mA}$$

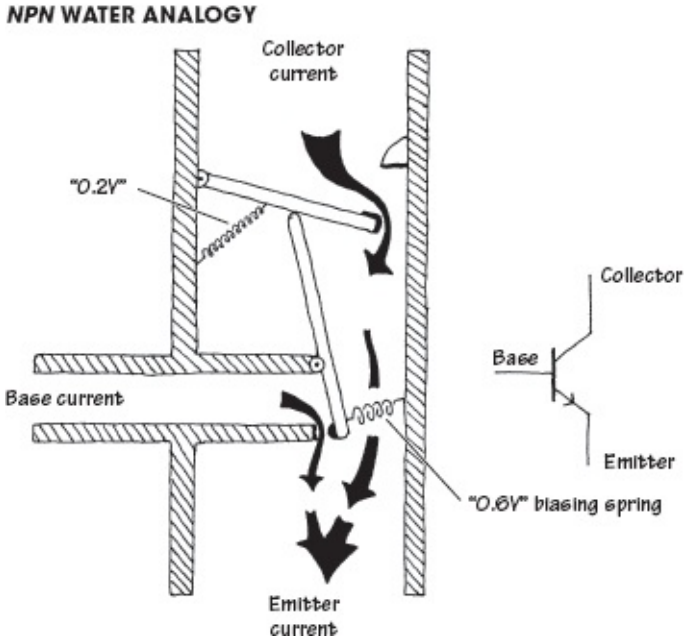
$$V_C = 0\text{ V} + I_C R_2$$

$$V_C = 0\text{ V} + (2.1\text{ mA})(2800\,\Omega)$$

$$V_C = 5.9\text{ V}$$

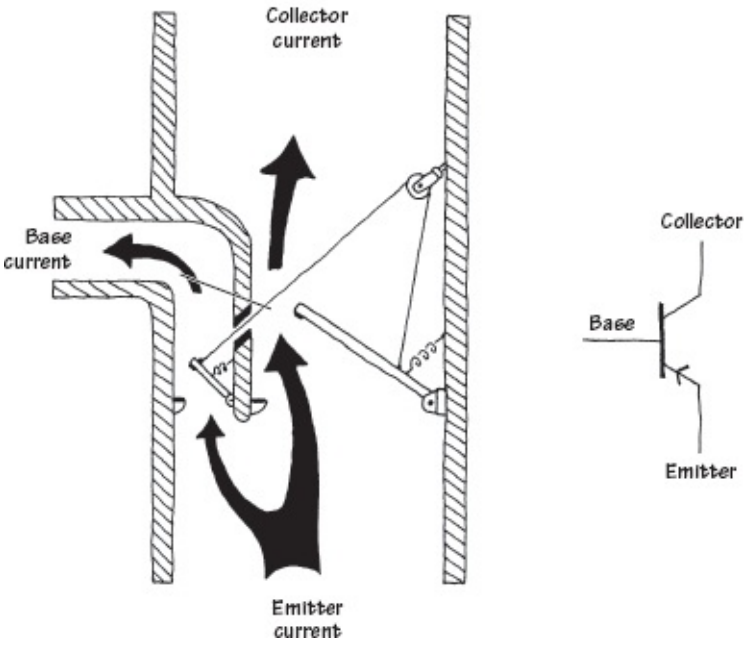
FIGURE 4.50

Bipolar Transistor Water Analogy



The base of the *npn* water transistor represents the smaller tube entering the main device from the left side. The collector is represented by the upper portion of the vertical tube, while the emitter is represented by the lower portion of the vertical tube. When no pressure or current is applied through the "base" tube (analogous to an *npn* transistor's base being open circuited), the lower lever arm remains vertical while the top of this arm holds the upper main door shut. This state is analogous to a real bipolar *npn* transistor off state. In the water analogy, when a small current and pressure are applied to the base tube, the vertical lever is pushed by the entering current and swings counterclockwise. When this lever arm swings, the upper main door is permitted to swing open a certain amount that is dependent on the amount of swing of the lever arm. In this state, water can make its way from the collector tube to the emitter tube, provided there is enough pressure to overcome the force of the spring holding the door shut. This spring force is analogous to the 0.6 V biasing voltage needed to allow current through the collector-emitter channel. Notice that in this analogy, the small base water current combines with the collector current.

PNP WATER ANALOGY

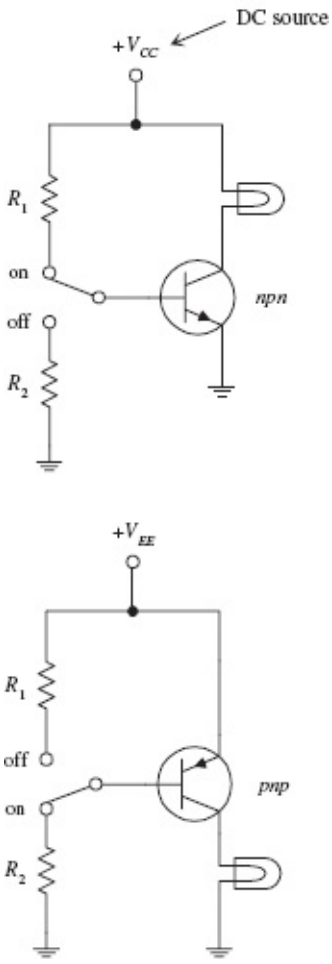


The main feature to note here is the need for a lower pressure at the base for the *pnp* water transistor to turn on. By allowing current to flow out the base tube, the lever moves, allowing the emitter-collector door to open. The degree of openness varies with the amount of swing in the lever arm, which corresponds to the amount of current escaping through the base tube. Again, note the 0.6 V biasing spring.

FIGURE 4.51

Basic Operation

TRANSISTOR SWITCH



Here, an *npn* transistor is used to control current flow through a light bulb. When the switch is thrown to the on position, the transistor is properly biased, and the collector-to-emitter channel opens, allowing current to flow from V_{CC} through the light bulb and into ground. The amount of current entering the base is determined by

$$I_B = \frac{V_E + 0.6\text{V}}{R_1} = \frac{0\text{V} + 0.6\text{V}}{R_1}$$

To find the collector current, you can use the current-gain relation ($I_C = h_{FE}I_B$), provided that there is not too large a voltage drop across the light bulb (it shouldn't cause V_C to drop below 0.6 V + V_E). When the switch is thrown to the off position, the base is set to ground, and the transistor turns off, cutting current flow to the light bulb. R_2 should be large (e.g., 10 kΩ) so that very little current flows to ground.

In the *pnp* circuit, everything is reversed; current must leave the base in order for a collector current to flow.

FIGURE 4.52

CURRENT SOURCE

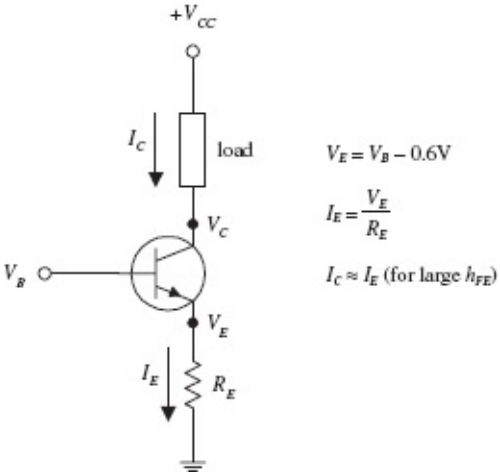


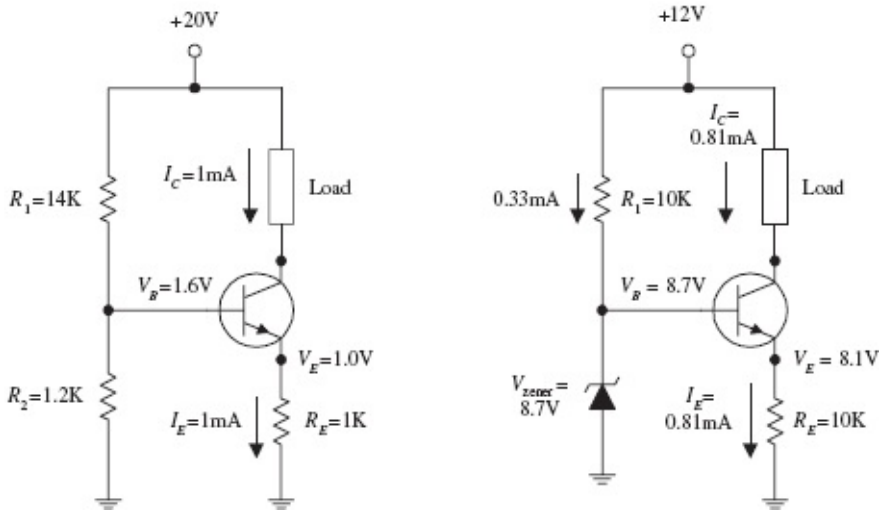
FIGURE 4.53

The circuit here shows how an *npn* transistor can be used to make a simple current source. By applying a small input voltage and current at the transistor's base, a larger collector/load current can be controlled. The collector/load current is related to the base voltage by

$$I_C = I_{\text{load}} = \frac{V_B - 0.6\text{V}}{R_E}$$

The derivation of this equation is shown with the figure.

CURRENT BIASING METHODS



Two common methods for biasing a current source are to use either a voltage-divider circuit (shown in the leftmost circuit) or a zener diode regulator (shown in the right-most circuit). In the voltage-divider circuit, the base voltage is set by R_1 and R_2 and is equal to

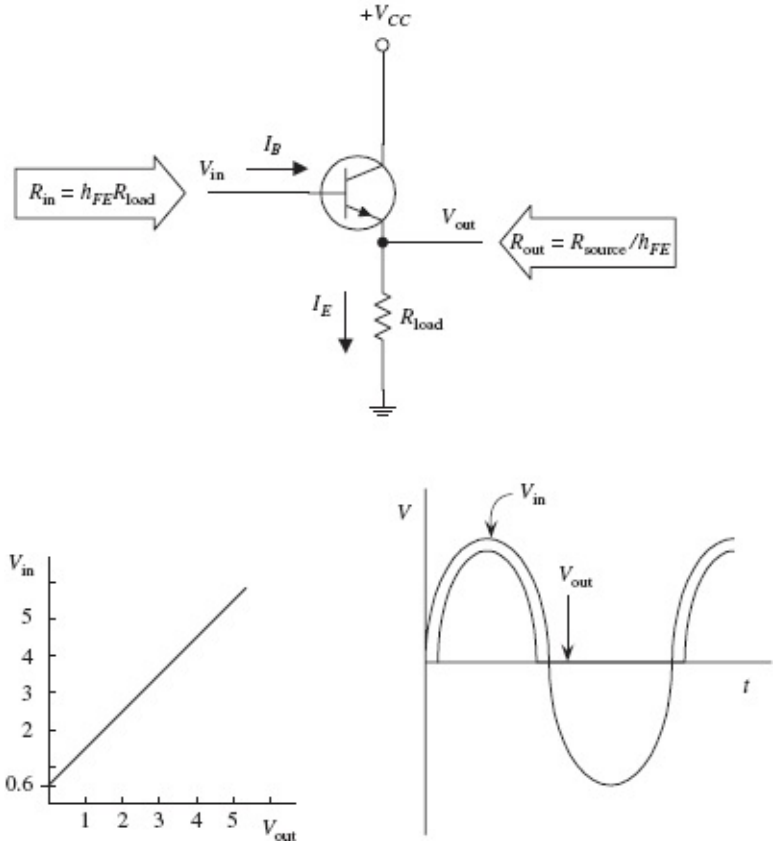
$$V_B = \frac{R_2}{R_1 + R_2} V_{cc}$$

In the zener diode circuit, the base voltage is set by the zener diode's breakdown voltage such that

$V_B = V_{zener}$

FIGURE 4.54

EMITTER FOLLOWER



The network shown here is called an *emitter follower*. In this circuit, the output voltage (tapped at the emitter) is almost a mirror image of the input (output "follows" input), with the exception of a 0.6 V drop in the output relative to the input (caused by base-emitter *pn* junction). Also, whenever $V_B \leq 0.6$ V (during negative swings in input), the transistor will turn off (the *pn* junction is reversed-biased). This effect results in clipping of the output (see graph). At first glance, it may appear that the emitter follower is useless—it has no voltage gain. However, if you look at the circuit more closely, you will see that it has a much larger input impedance than an output impedance, or more precisely, it has a much larger output current (I_E) relative to an input current (I_B). In other words, the emitter follower has current gain, a feature that is just as important in applications as voltage gain. This means that this circuit requires less power from the signal source (applied to V_{in}) to drive a load than would otherwise be required if the load were to be powered directly by the source. By manipulating the transistor gain equation and using Ohm's law, the input resistance and output resistance are:

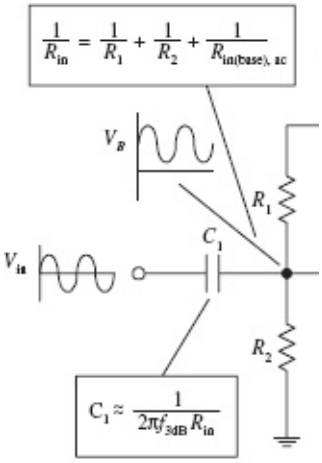
$$R_{in} = h_{FE}R_{load}$$

$$R_{out} = \frac{R_{source}}{h_{FE}}$$

FIGURE 4.55

EMITTER-FOLLOWER (COMMON-COLLECTOR) AMPLIFIER

The circuit shown here is called a *common-collector amplifier,* which has current gain but no voltage gain. It makes use of the emitter-follower arrangement but is modified to avoid clipping during negative input swings. The voltage divider (R_1 and R_2) is used to give the input signal (after passing through the capacitor) a positive dc level or operating point (known as the *quiescent point*). Both the input and output capacitors are included so that an ac input-output signal can be added without disturbing the dc operating point. The capacitors, as you will see, also act as filtering elements.



EXAMPLE

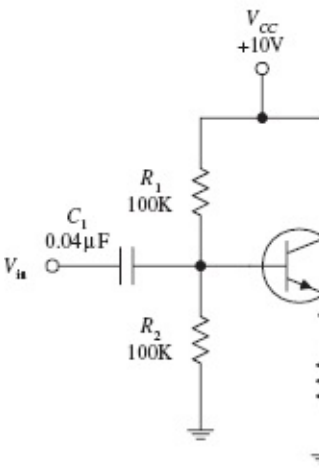


FIGURE 4.56

To design a common-coll
100, and a desired f_{3dB} pc

1. **Choose a quiescent cu**
2. **Next, select $V_E = \frac{1}{2}V_{CC}$**
   set V_E = 5 V and still g

$$R_E = \frac{\frac{1}{2}V_{CC}}{I_Q} = \frac{5\text{ V}}{1\text{ mA}} = 5\text{ k}$$

3. **Next, set the $V_B = V_{EE}$**
   the voltage divider (R
   substituting into it V_B =

$$\frac{R_2}{R_1} = \frac{V_B}{V_{CC} - V_B} = \frac{V_E + 0.}{V_{CC} - (V_E -}$$

Fortunately, you can r
isn't too dramatic. The
the dc (quiescent) inp
loading conditions):

$$\frac{R_1 R_2}{R_1 + R_2} \le \frac{1}{10} R_{in(base),dc}$$

$$\frac{R}{2} \le \frac{1}{10} R_{in(base),dc} \qquad \text{(using}$$

Here, $R_{in(base),dc} = h_{FE}R_E$,

CHAPTER 5

Optoelectronics

Optoelectrics is a branch of electronics that deals with light-emitting and light-detecting devices. Light-emitting devices, such as lamps and light-emitting diodes (LEDs), create electromagnetic energy (e.g., light) by using an electric current to excite electrons into higher energy levels (when an electron changes energy levels, a photon is emitted). Light-detecting devices such as phototransistors and photoresistors, on the other hand, are designed to take incoming electromagnetic energy and convert it into electric currents and voltages. This is usually accomplished using photons to liberate bound electrons within semiconductor materials. Light-emitting devices typically are used for illumination purposes or as indicator lights. Light-detecting devices are used primarily in light-sensing and communication devices, such as dark-activated switches and remote controls. This chapter examines the following optoelectronic devices: lamps, LEDs, photoresistors, photodiodes, solar cells, phototransistors, photothyristors, and optoisolators.

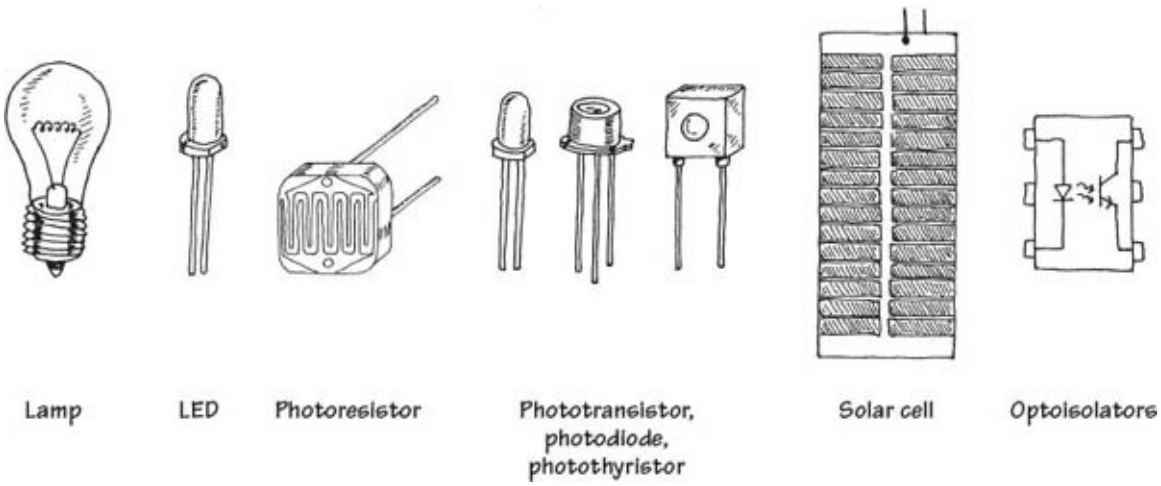


FIGURE 5.1

5.1 A Little Lecture on Photons

Photons are the elemental units of electromagnetic radiation. White light, for example, is composed of a number of different kinds of photons; some are blue photons, some are red photons, etc. It is important to note that there is no such thing as a white photon. Instead, when the combination of the various colored photons interacts with our eye, our brain perceives what we call *white light*.

Photons are not limited to visible light alone. There are also radiofrequency photons, infrared photons, microwave photons, and other kinds of photons that our eyes cannot detect.

In terms of the physics, photons are very interesting creatures. They have no rest mass, but they do carry momentum (energy). A photon also has a distant wavelike character within its electromagnetic bundle. The wavelength of a photon (horizontal distance between consecutive electrical or magnetic field peaks) depends on the medium in which it travels and on the source that produced it. It is this wavelength

that determines the color of a photon. A photon's frequency is related to its wavelength by $\lambda = v/f$, where v is the speed of the photon. In free space, v is equal to the speed of light ($c = 3.0 \times 10^8$ m/s), but in other media, such as glass, v becomes smaller than the speed of light. A photon with a large wavelength (or small frequency) is less energetic than a photon with a shorter wavelength (or a higher frequency). The energy of a photon is equal to $E = hf$, where h is Planck's constant (6.63×10^{-34} J·s).

The trick to "making" a photon is to accelerate/decelerate a charged particle. For example, an electron that is made to vibrate back and forth within an antenna will produce radiofrequency photons that have very long wavelengths (low energies) when compared with light photons. Visible light, on the other hand, is produced when outer-shell electrons within atoms are forced to make transitions between energy levels, accelerating in the process. Other frequency photons may be created by vibrating or rotating molecules very quickly, while still others, specifically those with very high energy (e.g., gamma rays), can be created by the charge accelerations within the atomic nuclei.

Figure 5.2 shows the breakdown of the electromagnetic spectrum. *Radiofrequency* photons extend from a few hertz to about 10^9 Hz (wavelengths from kilometers to about 0.3 m). They are often generated by alternating currents within power lines and electric circuits such as radio and television transmitters.

Microwave photons extend from about 10^9 up to 3×10^{11} Hz (wavelengths from 30 cm to 1 mm). These photons can penetrate the earth's atmosphere and hence are used in space vehicle communications, radio astronomy, and transmitting telephone conversations to satellites. They are also used for cooking food. Microwaves are often produced by atomic transitions and by electron and nuclear spins.

Infrared photons extend from about 3×10^{11} to 4×10^{14} Hz. Infrared radiation is created by molecular oscillations and is commonly emitted from incandescent sources such as electric heaters, glowing coals, the sun, human bodies (which radiate photons in the range of 3000 to 10,000 nm), and special types of semiconductor devices.

Light photons comprise a narrow frequency band from about 3.84×10^{14} to about 7.69×10^{14} Hz and are generally produced by a rearrangement of outer electrons in atoms and molecules. For example, in the filament of an incandescent light bulb, electrons are randomly accelerated by applied voltages and undergo frequent collision. These collisions result in a wide range of electron acceleration, and as a result, a broad frequency spectrum (within the light band) results, giving rise to white light.

Ultraviolet photons extend from approximately 8×10^{14} to 3.4×10^{16} Hz and are produced when an electron in an atom makes a long jump down from a highly excited state. The frequency of ultraviolet photons—unfortunately for us—tend to react badly with human cell DNA, which in turn can lead to skin cancer. The sun produces a large output of ultraviolet radiation. Fortunately for us, protective ozone molecules in the upper atmosphere can absorb most of this ultraviolet radiation by converting the photon's energy into a vibrating motion within the ozone molecule.

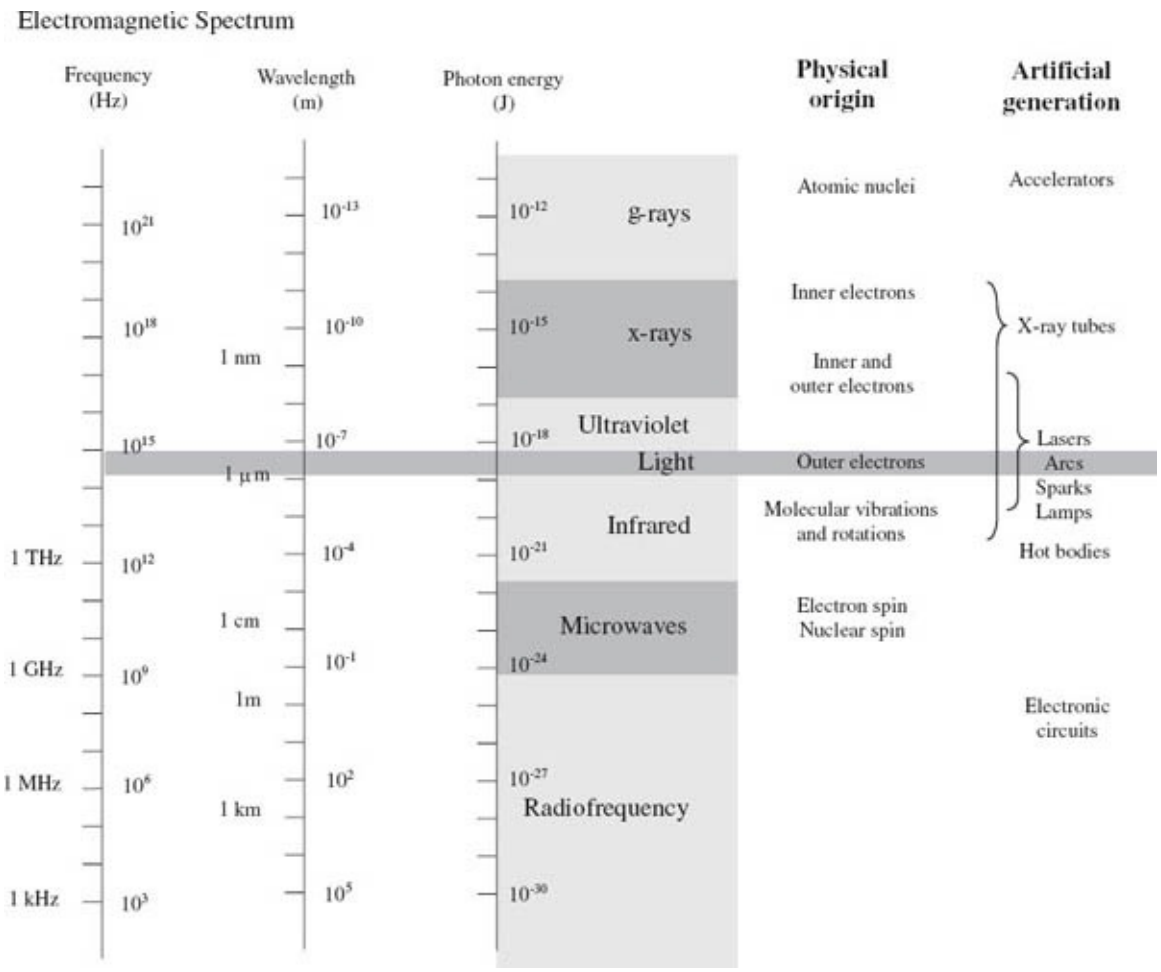


FIGURE 5.2

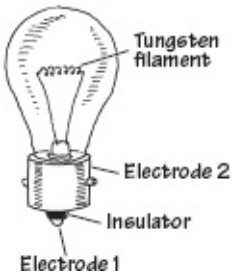
X-rays are highly energetic photons that extend from about 2.4×10^{16} to 5×10^{19} Hz, making their wavelengths often smaller than the diameter of an atom. One way of producing x-rays is to rapidly decelerate a high-speed charged particle. X-rays tend to act like bullets and can be used in x-ray imagery.

Gamma rays are the most energetic of the photons, whose frequency begins around 5×10^{19} Hz. These photons are produced by particles undergoing transitions within the atomic nuclei. The wavelike properties of gamma rays are extremely difficult to observe.

5.2 Lamps

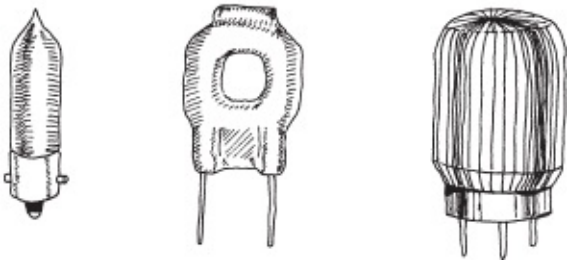
Lamps are devices that convert electric current into light energy. One approach used in the conversion process is to pass a current through a special kind of wire filament. As current collides with the filament's atoms, the filament heats up, and photons are emitted. (As it turns out, this process produces a variety of different wavelength photons, so it appears that the emitted light is white in color.) Another approach used to produce light involves placing a pair of electrodes a small distance apart within a glass gas-enclosed bulb. When a voltage is set across the electrodes, the gas ionizes (electrons are stripped from the gas atoms) emitting photons in the process. Here's an overview of some of the major kinds of lamps.

Incandescent



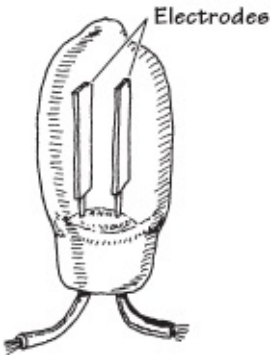
These lamps use a tungsten wire filament to produce a glowing white light when current passes through it. The filament is enclosed in an evacuated glass bulb filled with a gas such as argon, krypton, or nitrogen that helps increase the brilliance of the lamp and also helps prevent the filament from burning out (as would be the case in an oxygen-rich environment). Incandescent lamps are used in flashlights, home lighting, and as indicator lights. They come in a variety of different sizes and shapes, as well as various current, voltage, and candlelight power ratings.

Halogen



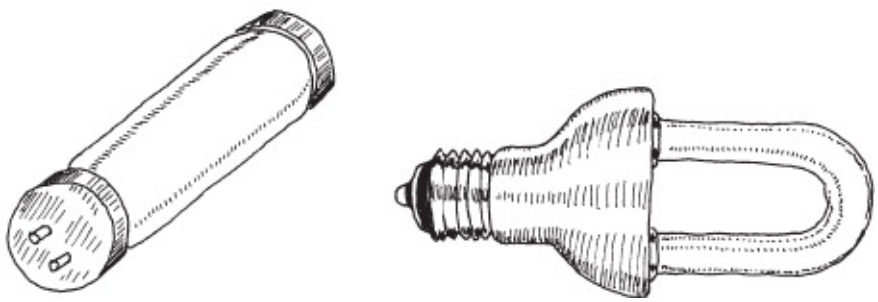
Similar to a typical incandescent lamp, these lamps provide ultrabright output light. Unlike a typical incandescent lamp, the filament is coated on the inside of a quartz bulb. Within this bulb, a halogen gas, such as bromine or iodine, is placed. These lamps are used in projector lamps, automotive headlights, strobe lights, etc.

Gas-Discharge



This lamp produces a dim, pale light that results from the ionization of neon gas molecules between two electrodes within the bulb. Types of gas-discharge lamps include neon, xenon flash, and mercury vapor lamps. Gas-discharge lamps have a tendency to suddenly switch on when a particular minimum operating voltage is met. For this reason, they are sometimes used in triggering and voltage-regulation applications. They are also used as indicator lights and for testing home ac power outlets.

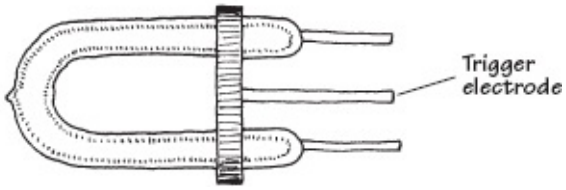
Fluorescent



This is a lamp consisting of a mercury vapor–filled glass tube whose inner wall is coated with a material that fluoresces. At either end of the tube are cathode and anode incandescent filaments. When electrons emitted from an incandescent cathode electrode collide with the mercury atoms, ultraviolet (UV) radiation is emitted. The UV radiation then collides with the lamp's florescent coating, emitting visible light in the process. Such lamps require an auxiliary glow lamp with bimetallic contacts and a choke

placed in parallel with the cathode and anode to initiate discharge within the lamp. These are highly efficient lamps that are often used in home lighting applications.

Xenon Flash Lamp



This is a gas-discharge lamp that is filled with a xenon gas that ionizes when a particular voltage is applied across its electrodes. These lamps come with three leads: an anode, a cathode, and a trigger-voltage lead. Normally, a particular voltage is applied across the anode and cathode leads, and the lamp is off. However, when a particular voltage is applied to the trigger lead, the gas suddenly ionizes and releases an extremely bright flash. These lamps are used in photographic applications and in special-effects lighting projects.

FIGURE 5.3

Technical Stuff About Light Bulbs

A lamp's brightness is measured in what is called the *mean spherical candle power* (MSCP). Bulb manufacturers place a lamp at the center of an integrating sphere that averages the lamp's light output over its surface. The actual value of the MSCP for a lamp is a function of color temperature of the emitting surface of the lamp's filament. For a given temperature, doubling the filament's surface area doubles the MSCP. Other technical things to consider about lamps include voltage and current ratings, life expectancy, physical geometry of the bulb, and filament type. Figure 5.4 shows a number of different bulb types.

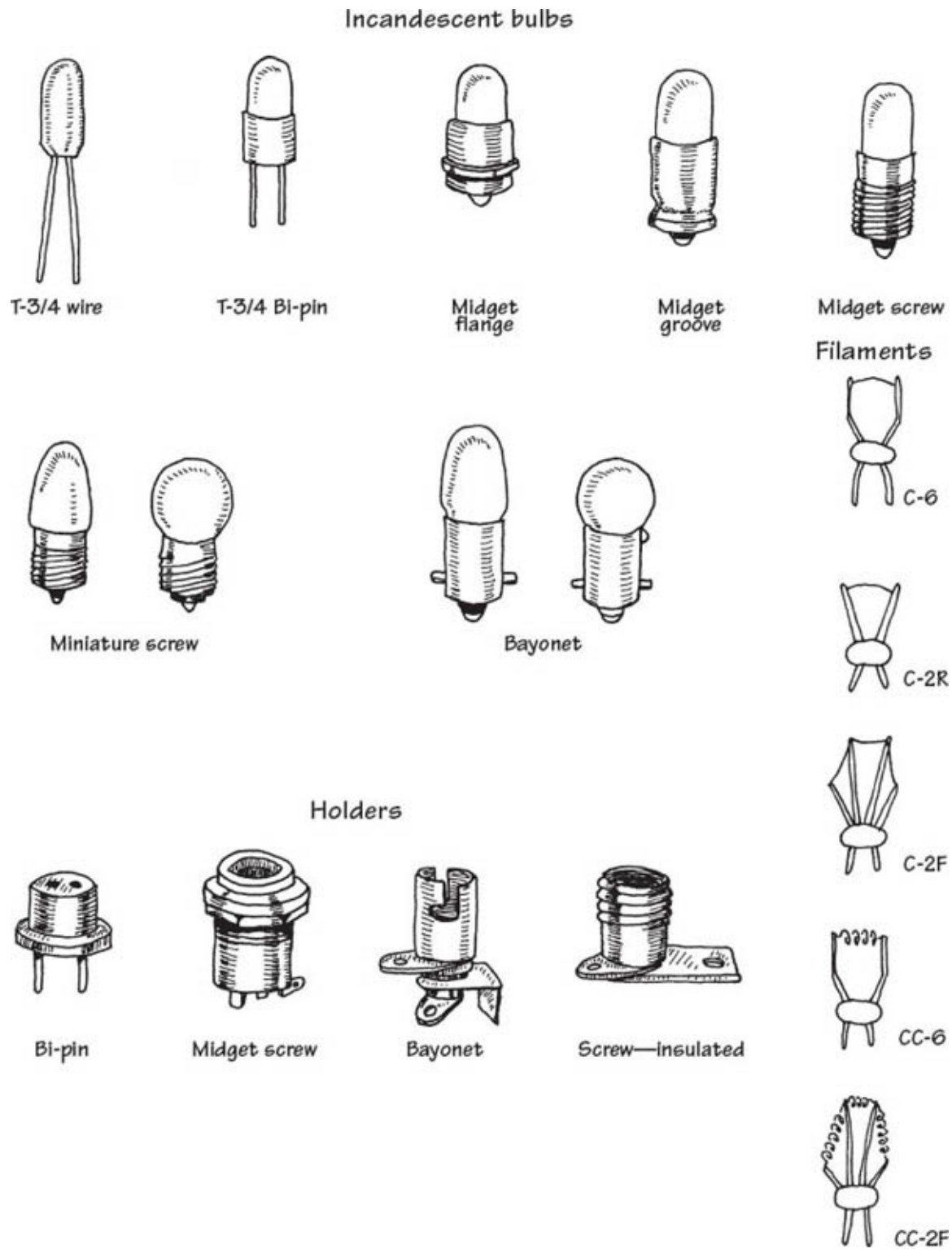


FIGURE 5.4

In recent years, incandescent light bulbs have become something of a rarity. In domestic lighting situations, they have almost completely been replaced by compact fluorescent designs, and indeed are not even for sale in some European countries. There is also a trend toward using arrays of LEDs for illumination in place of incandescent bulbs (see the next section). In applications that formerly would have used incandescent indicator lights, LEDs now rule supreme. LEDs have a much longer life, lower power consumption, and much greater tolerance of physical and thermal shock.

5.3 Light-Emitting Diodes

Light-emitting diodes (LEDs) are two-lead devices that are similar to pn-junction diodes, except that they

are designed to emit visible or infrared light. When a LED's anode lead is made more positive in voltage than its cathode lead (by at least 0.6 to 2.2 V), current flows through the device and light is emitted. However, if the polarities are reversed (anode is made more negative than the cathode), the LED will not conduct, and hence it will not emit light. The symbol for an LED is shown in Fig. 5.5.

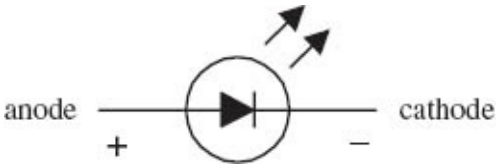


FIGURE 5.5

LEDs are available in a wide range of colors. Historically, red was the first LED color. Yellow and green and infrared LEDs followed. It was not until the 1990s that blue LEDs became available. These days, LEDs are available in pretty much any color, including white.

There are also high-powered LEDs that are used for illumination and organic LEDs (OLEDs) made from polymers that can be built into displays.

Often, LEDs (especially infrared LEDs) are used as transmitting elements in remote-control circuits (e.g., TV remote control). The receiving element in this case may be a phototransistor that responds to changes in LED light output by altering the current flow within the receiving circuit.

5.3.1 How an LED Works

The light-emitting section of an LED is made by joining *n*-type and *p*-type semiconductors together to form a *pn* junction. When this *pn* junction is forward-biased, electrons in the *n* side are excited across the *pn* junction and into the *p* side, where they combine with holes. As the electrons combine with the holes, photons are emitted. Typically, the *pn*-junction section of an LED is encased in an epoxy shell that is doped with light-scattering particles to diffuse the light and make the LED appear brighter. Often a reflector placed beneath the semiconductor is used to direct light upward. The cathode and anode leads are made from a heavy-gauge conductor to help wick heat away from the semiconductor.

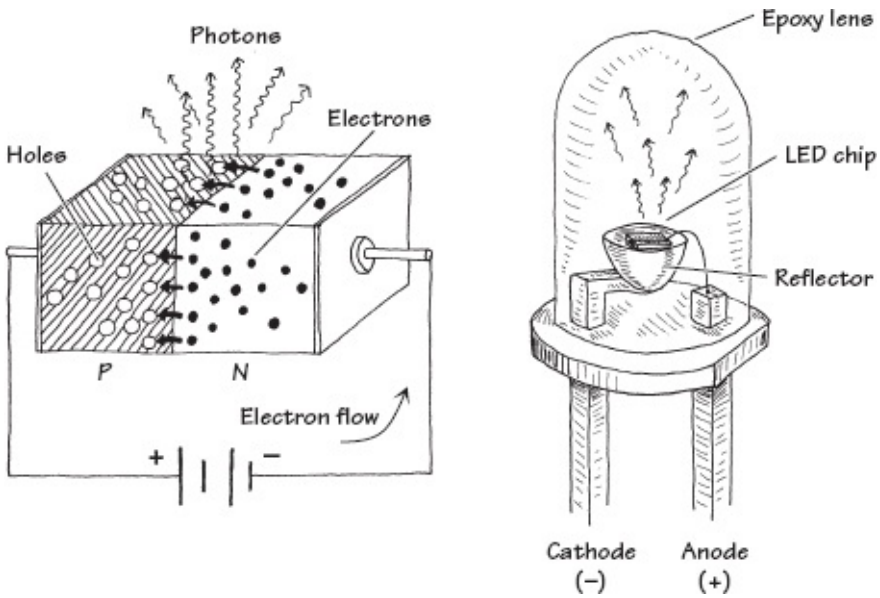
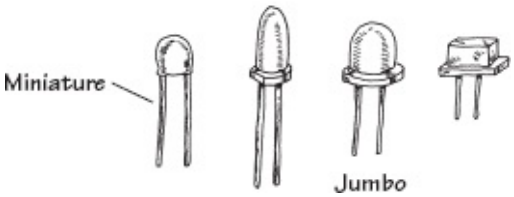


FIGURE 5.6

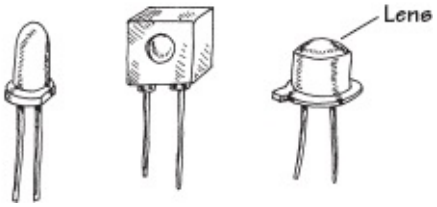
5.3.2 Kinds of LEDs

Visible-Light LEDs



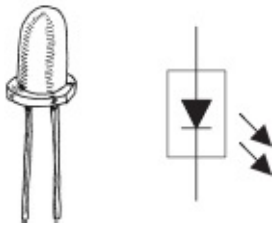
These LEDs are inexpensive and durable devices that typically are used as indicator lights. Common colors include green (~565 nm), yellow (~585 nm), orange (~615 nm), and red (~650 nm). Maximum forward voltages are about 1.8 V, with typical operating currents from 10 to 30 mA.

Infrared LEDs



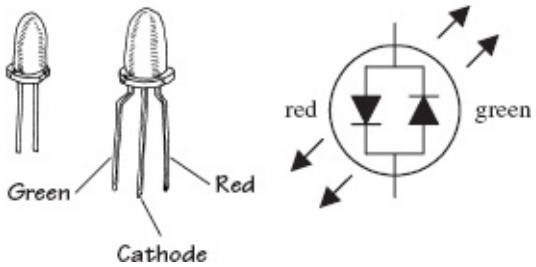
These LEDs are designed to emit infrared photons that have wavelength between approximately 880 and 940 nm. They are used in conjunction with a photosensor (e.g., photodiode, photoresistor, phototransistor) in remote-control circuits (e.g., TV remotes, intrusion alarms). They tend to have a narrower viewing angle when compared with a visible-light LED so that transmitted information can be directed efficiently. Photon output is characterized in terms of output power per specific forward current. Typical outputs range from around 0.50 mW/20 mA to 8.0 mW/50 mA. Maximum forward voltages at specific forward currents range from about 1.60 V at 20 mA to 2.0 V at 100 mA.

Blinking LEDs



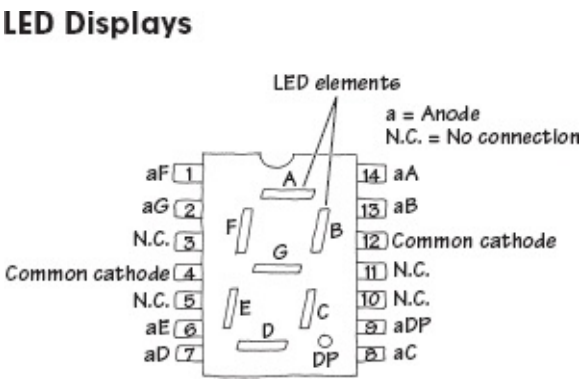
These LEDs contain a miniature integrated circuit within their package that causes the LED to flash from 1 to 6 times each second. They are used primarily as indicator flashers, but may be used as simple oscillators.

Multicolor LEDs



Multicolor LEDs mount a number of LEDs in a single package. The color they produce is a mixture of the colors produced by each LED. The ultimate multicolor LED is an RGB LED that has red, green, and blue LEDs within a single package. This allows any color to be mixed by controlling the brightness of each of the LEDs, with white light being produced if all the LEDs are on with equal intensity. The packages usually have a common anode or common cathode connection.

A variation on this has red and green LEDs back to back to indicate polarity (the polarity of the supply voltage will determine which LED is forward-biased and on, and which is reverse-biased and off).



These are used for displaying numbers and other characters. In the LED display shown here, seven individual LEDs are used to make up the display. When a voltage is applied across one of the LEDs, a portion of the 8 lights up. Unlike liquid-crystal displays, LED displays tend to be more rugged, but they also consume more power.

As well as such seven-segment LEDs, LED packages containing LEDs in a matrix are also available.

FIGURE 5.7

5.3.3 More on LEDs

LEDs emit visible light, infrared radiation, or even ultraviolet radiation when forward-biased. Visible single-tone LEDs emit relatively narrow bands of green, yellow, orange, red, and blue light (with spectrum spread usually less than 40 nm at 90 percent peak intensity). Infrared diodes emit one of several bands beyond red. White LEDs provide a variety of wavelengths to mimic white light and are used in low-level lighting applications, such as backlighting, headlamps, and nightlights.

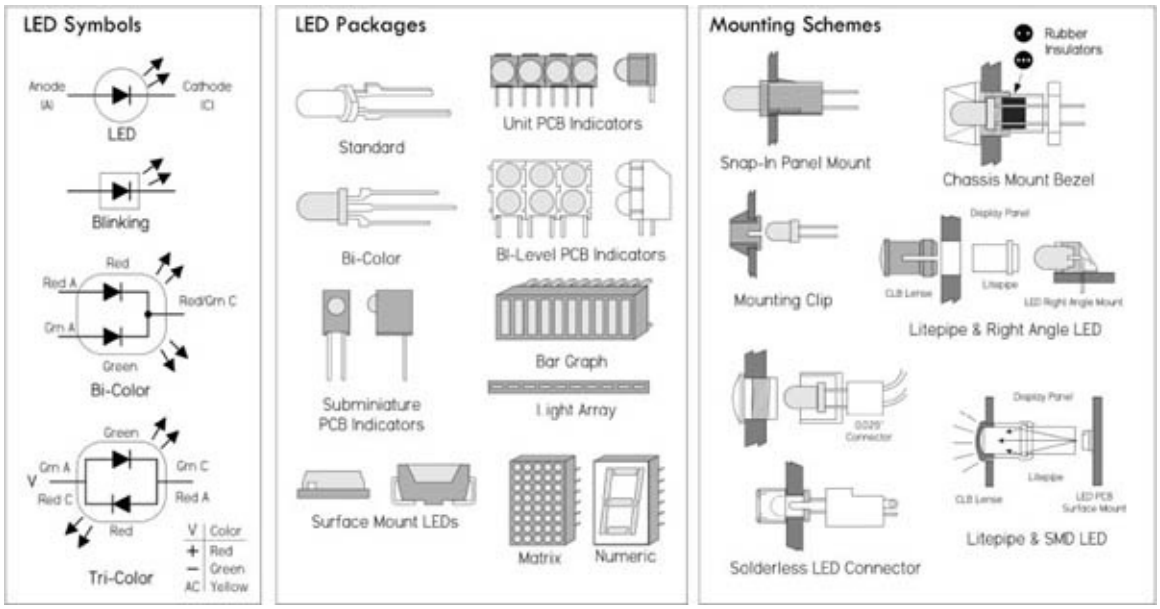


FIGURE 5.8

High-power LEDs (HPLEDs) are now available. These have forward currents of hundreds of mA to more than 1 A. Such LEDs are very bright but also generate a lot of heat. They must be mounted on a heat sink to prevent thermal destruction.

LEDs have very fast response times, excellent efficiency, and long lifetimes. They are current-dependent devices whose light output is directly proportional to the forward current.

To light an LED, apply a voltage greater than the LED's forward voltage V_{LED}, and limit current flow via a series resistor to a level below the LED's maximum rating, usually to I_{LED}—the manufacturer's recommended value. The following equation is used to select the series resistor:

$$R_S = \frac{V_{IN} - V_{LED}}{I_{LED}}$$

If you want brightness control, throw in a 1-K potentiometer in series, as shown in Fig. 5.10.

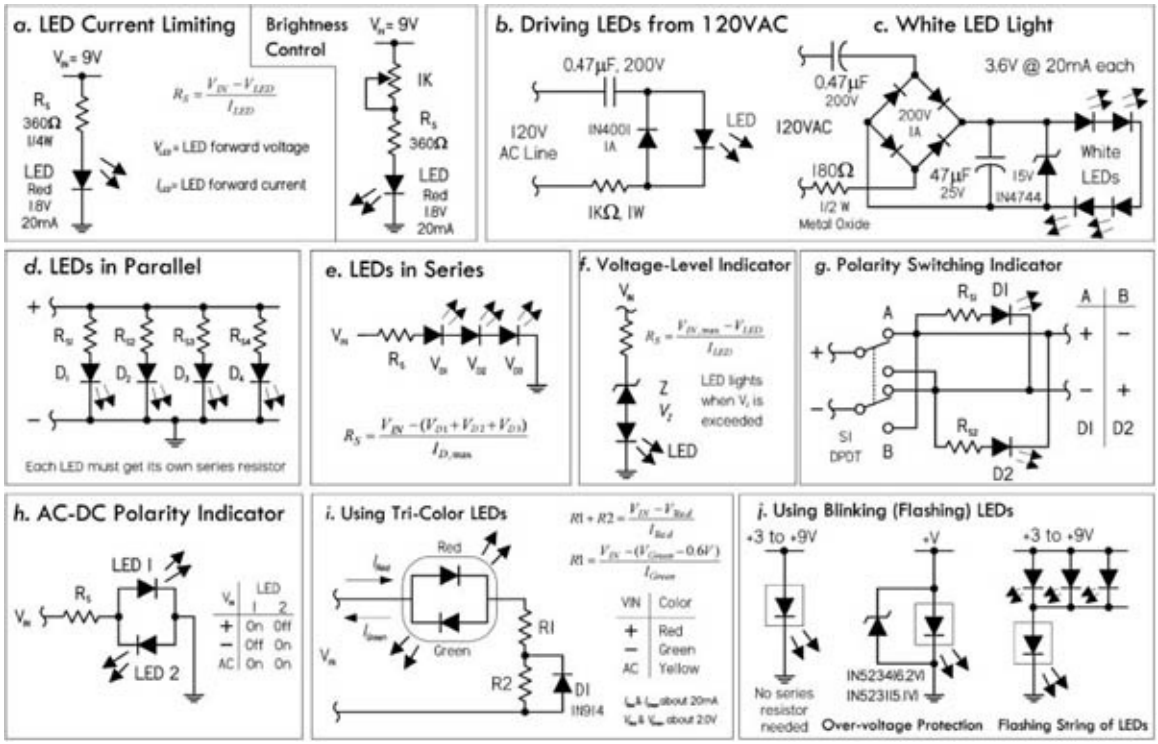


FIGURE 5.10 **(*a*) It's important to keep the LED current flow below the maximum rating. Select a series current-limiting resistor according to the provided formula. A 1K pot can be added to provide variable light-intensity control. (*b*) Here's a circuit that can be used to power LEDs from an ac line. The key is the capacitor, which acts to attenuate the ac signal, and the resistor, which acts to limit the current level. Since the capacitor must pass current in both directions, a small diode is connected in parallel with the LED to provide a path for the negative half cycle and also to limit the reverse voltage across the LED. A second LED with polarity reverse may be substituted for the diode, or a tricolor LED could be used, which would appear orange with alternating current. The resistor is chosen to limit the worst-case inrush current to about 150 mA, which will drop to less than 30 mA in a millisecond, as the capacitor charges. The 0.47-µF capacitor has a reactance of 5640 Ω at 60 Hz, so the LED current is about 20 mA half wave, or 10 mA average. A larger capacitor will increase the current and a smaller one will reduce it. The capacitor must be a nonpolarized type with a voltage rating of 200 V or more. (*c*) This is also an ac-line powered arrangement used to create a nightlight using white LEDs. Like the previous circuit, the 0.47-µF input capacitor attenuates the ac voltage level, and the 180-Ω resistor acts to limit current. The bridge rectifier with attached filter capacitor creates a near constant dc voltage, while the zener diode acts to regulate the voltage level. In this case, there are four white LEDs in series, each with a 3.4-V drop, for a total voltage drop of 13.6 V; therefore, a 15-V zener is used. (*d*) Do not put LEDs in parallel with each other without using individual dropping resistors for each LED. Although this usually works, it is not reliable. LEDs become more conductive as they warm up, which may lead to unstable current distribution through paralleled LEDs. LEDs in parallel need their own individual dropping resistors, as shown in the figure. Series strings can be paralleled if each string has its own dropping resistors. (*e*) It's okay to put LEDs in series with a single common series resistor. Simply add up the voltage drops of all the diodes in series, and use that value in place of V_{LED} in the series-resistor formula. It's a good idea not to exceed 80 percent of the supply voltage in order to maintain good stability and predicable current consumption. (*f*) Here, a zener diode is used to create a simple voltage-level indicator: when the voltage exceeds the breakdown voltage of the zener, current flows, lighting the LED. (*g*) This circuit reverses the polarity at the output using a DPDT switch. D_1 goes on when the polarity is one way, and D_2 goes on when the polarity is another way. (*h*) Two LEDs in reverse-parallel make up a simple polarity indicator. Note that both LEDs appear on with ac. (*i*) Basic circuit showing how to use tricolor LEDs. (*j*) Blinking LEDs don't require a series resistor, but don't apply more than the recommended supply voltage—3 to 9 V is a safe range. A zener placed in reverse parallel with the blinking LED can be used for overvoltage protection. A number of standard LEDs can be made to blink using a blinking LED. You can also use the blinking LED to drive transistors that can be made to switch on and off as the blinking LED flashes.**

V_{LED} varies with LED color. Typical V_{LED} values are 1.7 V for non-high-brightness red, 1.9 V for high-brightness high-efficiency low-current red, 2 V for orange and yellow, 2.1 V for green, 3.4 to 3.6 V for bright white and most blue types, and 6 V for 430-nm blue. Given the preceding voltage drops, it's a good idea to use at least a 3-V supply voltage for lower-voltage LEDs, 4.5 V for 3.4-V types, and 6 V for 430-

nm blue. If you don't know the recommended I_{LED} value of a given LED, it's usually safe to assume it will be around 20 mA. Table 5.1 shows the range of LED types and corresponding characteristic values.

TABLE 5.1 Diode Characteristics

WAVELENGTH	COLOR NAME	FWD VOLTAGE (V_F AT 20 MA)	INTENSITY (5-MM LEDS)	LED DYE MATERIAL
940	Infrared	1.5	16 mW @ 50 mA	GaAlAs/GaAs
880	Infrared	1.7	18 mW @ 50 mA	GaAlAs/GaAs
850	Infrared	1.7	26 mW @ 50 mA	GaAlAs/GaAs
660	Ultra Red	1.5–1.8	200 mcd @ 50 mA	GaAlAs/GaAs
635	High Eff. Red	2.0	200 mcd @ 20 mA	GaAsP/GaP
633	Super Red	2.2	3500 mcd @ 20 mA	InGaAlP
620	Super Orange	2.2	4500 mcd @ 20 mA	InGaAlP
612	Super Orange	2.2	6500 mcd @ 20 mA	InGaAlP
605	Orange	2.1	160 mcd @ 20 mA	GaAsP/GaP
595	Super Yellow	2.2	5500 mcd @ 20 mA	InGaAlP
592	Super Pure Yellow	2.1	7000 mcd @ 20 mA	InGaAlP
585	Yellow	2.1	100 mcd @ 20 mA	GaAsP/GaP
574	Super Lime Yellow	2.4	1000 mcd @ 20 mA	InGaAlP
570	Super Lime Green	2.0	1000 mcd @ 20 mA	InGaAlP
565	High Efficiency Green	2.1	200 mcd @ 20 mA	GaP/GaP
560	Super Pure Green	2.1	350 mcd @ 20 mA	InGaAlP
555	Pure Green	2.1	80 mcd @ 20 mA	GaP/GaP
525	Aqua Green	3.5	10,000 mcd @ 20 mA	SiC/GaN
505	Blue Green	3.5	2000 mcd @ 20 mA	SiC/GaN
470	Super Blue	3.6	3000 mcd @ 20 mA	SiC/GaN
430	Ultra Blue	3.8	100 mcd @ 20 mA	SiC/GaN
370–400	UV LED	3.9	NA	GaN
4500K	"Incandescent White"	3.6	2000 mcd @ 20 mA	SiC/GaN
6500K	Pale White	3.6	4000 mcd @ 20 mA	SiC/GaN
8000K	Cool White	3.6	6000 mcd @ 20 mA	SiC/GaN

Some other specifications to consider include power dissipation (100 mW typical), reverse voltage rating, operating temperature (-40 to +85°C typical), pulse current (100 mA typical), luminous intensity (given in millicandles, mcd), viewing angle (given in degrees), peak emission wavelength, and spectral width (20 to 40 nm typical).

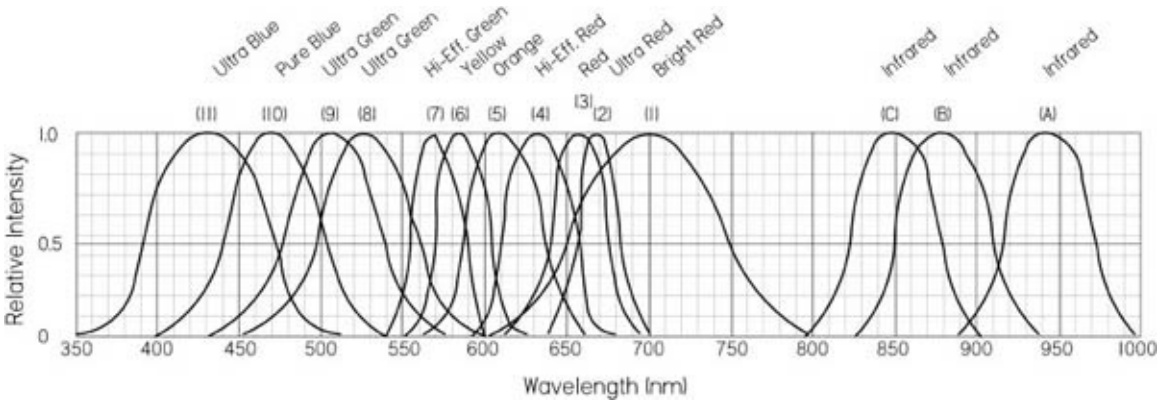


FIGURE 5.9

5.3.4 LED Applications

In Chap. 13, we will look at how you drive LED displays from a microcontroller, including techniques such as multiplexing, Charlieplexing, and driving RGB LEDs with PWM signals to mix colors.

5.3.5 Laser Diodes

Laser diodes are light-emitting diodes with two "mirrors" on the surface of the diode to create a laser cavity. When the diode is forward-biased, charges are injected into the active area of the junction, while electrons and holes recombine in the junction, creating spontaneous emission of photons. These photons can cause other electron-hole pairs to recombine by stimulated emission. When the current is high enough, the device lases. Laser diodes are driven by low-voltage power and usually incorporate optical feedback from a monitor photodiode (commonly built into the same laser diode package) to regulate the laser diode current.

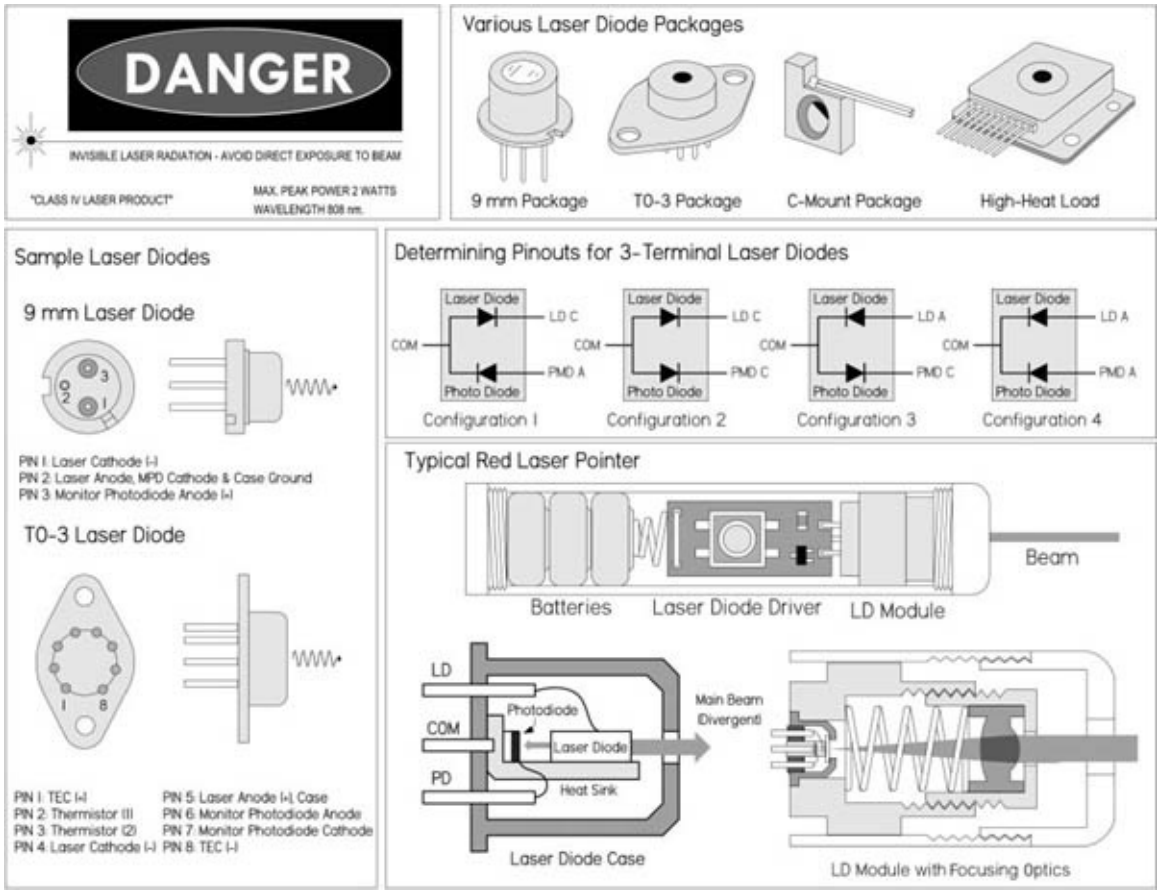


FIGURE 5.11 Figure shows a rough diagram of a laser diode of the type found in a laser pointer or CD player. The size of the package is typically 5 to 10 mm overall, but the laser diode chip is less than 1 mm in length. The maxim beam as it emerges from the laser diode is wedge-shaped and highly divergent (unlike a helium-neon laser) with a typical spread of 10 by 30°. External optics are required to produce anything approaching a parallel (collimated) beam. A simple (spherical) short-focal-length convex lens will work reasonably well for this purpose, but diode laser modules and laser pointers might use a lens that at least on the surface is aspheric (not ground to a spherical shape, as are most common lenses). The beam from the back end of the laser diode chip hits a built-in photodiode that is normally used in an optoelectronic feedback loop to regulate current and thus beam power.

Compared to LEDs, laser diodes have quicker response times and very narrow spectrum spread (around 1 nm), and can focus radiation to a spot as small as 1 μm in diameter—even for a cheap laser diode using simple optics found in a CD player. Unlike gas lasers, however, a laser diode's output beam

is divergent, typically elliptical or wedge-shaped, and astigmatic, requiring refocusing.

The output wavelength of a laser diode is usually fixed to a single mode: for example red (635 nm, 670 nm), infrared (780 nm, 800 nm, 900 nm, 1,550 nm, etc.), and green, blue, violet. However, multimode laser diodes also exist where the emission spectrum consists of several individual spectral lines with a dominant line (line with greatest intensity) occurring at the nominal wavelength of the device. Multimode laser diodes are often desirable because problems with mode hops are suppressed—consequently, they generally have a better signal-to-noise ratio. Mode hops are slight changes in the wavelength caused by thermal expansion of the laser cavity.

Typical optical output for low-powered laser diodes range from around 1 mW to 5 mW, while high-power laser diodes can reach 100 W or more. The highest-power units consist of arrays of laser diodes—not a single device alone.

Laser diodes are used in CD players, CD-ROM drives, and DVD and Blu-ray players. They are also used in laser printers, laser fax machines, laser pointers, sighting and alignment scopes, measurement equipment, high-speed fiber-optic and free-space communication systems; as pump source for other lasers; in bar-code and UPC scanners; and in high-performance imagers. In those applications requiring high-speed modulation or pulse rates (into the gigahertz range), special integrated driver chips are needed to control the laser diode drive current (more on that in a bit).

A variety of small laser diodes are found in CD players, CD-ROM drives, laser printers, and bar-code scanners. The most common laser diodes around are those used in CD players and CD-ROM drives. These produce a mostly invisible beam in the near infrared spectrum at 780 nm. The optical output from the actual laser diode itself may be up to 5 mW, but once it passes through the optics leading to the CD, the power drops to 0.3 to 1 mW. Higher-power IR laser diodes found in read-write drives have more power output—up to 30 mW or so. Even higher-power blue laser diodes can be found in Blu-ray players.

Visible laser diodes have replaced the helium-neon lasers used in bar-code scanners, laser pointers, positioning devices in medicine (e.g., CT and MRI scanners), and many other applications. The first visible-light laser diode emitted with a wavelength around 670 nm in the deep red spectrum. More recently, 650-and 635-nm red laser diodes have dropped in price. 635 to 650-nm laser diodes are used in DVD technology. The shorter wavelength compared to 780 nm is one of the several improvements that enable DVDs to store about eight times the amount of information as CDs (4 to 5 GB per layer and up to two layers on each side of the disc, as compared to a typical CD that stores only around 650 MB). Like the IR laser diodes, the visible laser diodes have a typical maximum power around 3 to 5 mW, and cost about $10 to $50 for the basic laser diode device—more with optics and drive electronics. Higher-power types are also available but can cost upward of several hundred dollars for something like a 20-mW module. Very high-power diode lasers using arrays of laser diodes or laser diode bars with power output of watts or greater may cost thousands of dollars.

It’s important to note that you should never look into a laser beam or any secular reflection of the laser beam—“you might poke your eye out.” Also, laser diodes are extremely sensitive to electrostatic discharge (ESD), so it’s important to use grounding straps and grounded equipment when working with them, as well as following manufacturer’s suggested handling precautions.

Laser Diode Drive Circuits

You should never drive a laser diode without the proper drive circuitry. Without the proper drive circuitry, you can run into all sorts of problems stemming from swings in operating temperature with unstable injection current. The results can lead to a fried laser diode, or one whose life is short. For this reason, it’s important to have a driver circuit that can provide stable current without the possibility of supply transients screwing things up. Two basic techniques used to achieve stable optical output from a laser diode are described here.

Automatic current control (ACC) or constant current circuit: This involves driving the laser diode without a photodiode feedback loop; the laser diode is simply driven at constant current. This is a simple method to use, but the optical output will fluctuate as the laser diode temperature changes. However, there are circuits and corresponding laser diodes that are designed to control the diode's operating temperature (without using the photodiode) that are very popular. The constant current circuit with temperature control provides faster control loop and precision current reference for accurately monitoring laser current. In addition, in many cases, the laser diode's internal photodiode may exhibit drift and have a poor noise characteristic. If performance of the internal photodiode is inferior, the laser diode's optical output is liable to be noisy and unstable. Constant current operation without temperature control, however, is generally not a good idea, except in cheap, low-power situations (cheap laser pointers, etc.). If the operating temperature of the laser diode decreases significantly, the optical power output will increase, and this could easily exceed its maximum rating.

Automatic power control (APC) circuit: This drive circuit is based on a photodiode feedback loop that monitors the optical output and provides a control signal for the laser diode to maintain a constant optical output level. Constant power control prevents the possibility of the optical power output increasing as the laser diode's temperature decreases. However, when operating in the constant power mode and without temperature control, mode hops and changes in wavelength will still occur. Also, if the diode's heat sink is inadequate and the temperature is allowed to rise, the optical power will decrease. In turn, the drive circuit will increase the injection current, attempting to maintain the optical power at a constant level. Without an absolute current limit, thermal runaway is possible and the laser may be damaged or destroyed in the process.

Regardless of the type of drive circuit used, the key is to prevent the drive current from overshooting the maximum operating level. Exceeding the maximum optical output, even for a nanosecond, will damage the mirror coatings on the laser diode end facets. Your typical laboratory power supply should not be used to directly drive a laser diode—it simply doesn't provide enough protection. Typical drive circuits incorporate slow-start circuitry, capacitive filtering, and other provisions to eliminate supply spikes, surges, and other switching transients.

Figure 5.12 shows a few do-it-yourself laser diode supplies. Though these drive circuits will work with many low-powered laser diodes that don't require modulation, it's worth your while to check out the laser data manufacturer's data sheets for recommended drive circuits. Creating your own drive circuits can be very tricky, and you'll probably end up frying a few expensive laser diodes in the process. Of course, you can also buy a laser diode drive chip, which may be the best bet if you're doing something a bit more complex than creating a laser pointer. These drive chips may support high-bit-rate modulation in addition to providing the constant current needed for optically stable power. Other types of chips can be adapted to linear and switching regulators. Some companies worth checking out include MAXIM, Linear Technology, Sharp, Toshiba, Mitsubishi, Analog Devices, and Burr-Brown. Often, these manufacturers provide free samples.

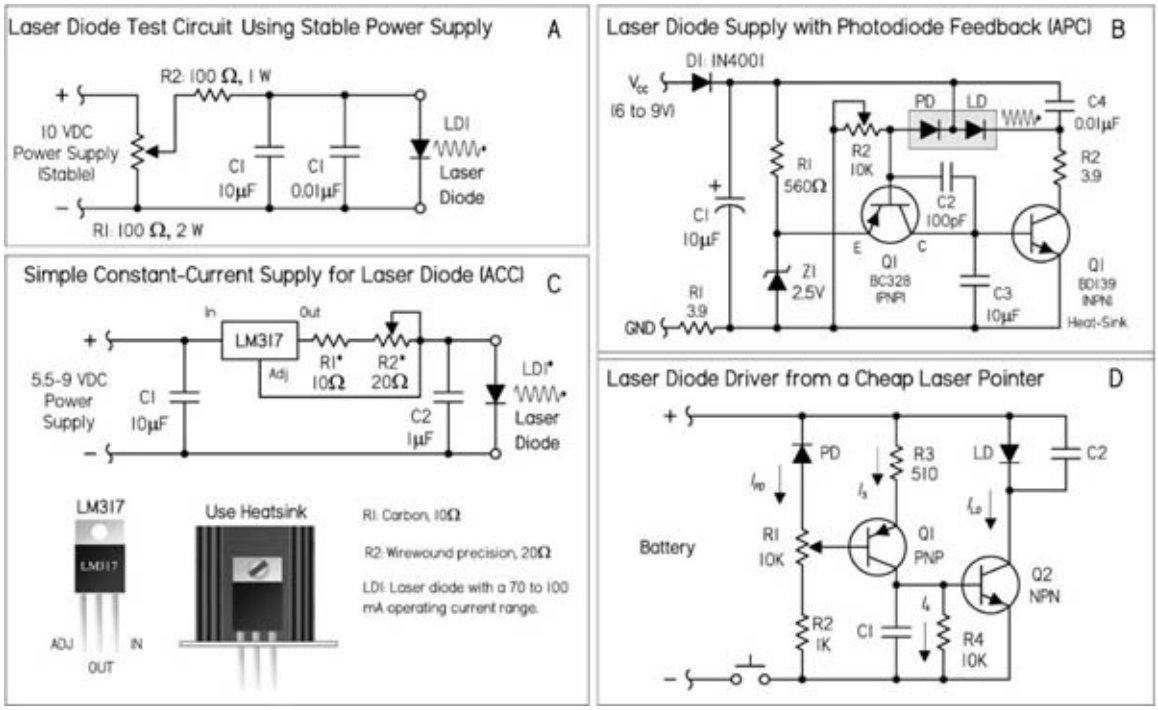


FIGURE 5.12 (*a*) This circuit can be used to identify the proper connections and polarity and then to drive the laser diode for testing purposes. Here a 0- to 10-VDC supply with current-limiting resistor is in series with the diode. If your power supply has a current limiter, set it at 20 to 25 mA to start. You can always increase it later. R_2 limits the maximum current. If you know the specs for your diode, this is a good idea (and to protect your power supply, too). You can always reduce its value if your laser diode requires more than about 85 mA (with R_2 = 100 Ω). (*b*) This runs on a (wall adapter) power supply from 6 to 9 V. There is heavy capacitive filtering in this circuit. Changes would be needed to enable this circuit to be modulated at any reasonable rate. Note that C_4 is estimated; also, an LM341 shunt regulator set up as a 2.5-V reference could replace the 2.5-V zener. (*c*) The resistor value depends on your specific laser diode current requirements. Power is provided by a 5.5- to 9-VDC battery. For the resistor, a small 10-Ω carbon resistor in series with a precision 20-Ω adjustable resistor can be used. It's a good idea to place three ordinary diodes in series instead of a laser diode, then measure the current through them and adjust the resistor until the necessary current level is reached, in this case 50 to 60 mA. You can turn up the current, never exceeding your diode's maximum limit. The dim glow will increase in intensity, but at some point a distinctive step in intensity is reached. (*d*) This circuit is from an inexpensive laser pointer. It includes some capacitive filtering, as well as a power-adjust pot R_1. Unlike the previous circuit, this one doesn't have any absolute reference, so power output will be dependent on the battery voltage to some extent. It's possible to modulate this module at a reasonable frequency by removing or greatly reducing the value of the filter capacitor C_1.

DRIVE CIRCUIT PRECAUTIONS

Even with a suitable drive circuit, watch out for intermittent or unreliable connections between the laser diode and the drive circuit. An intermittent contact in the photodiode feedback circuit will usually destroy a laser diode. Even if a power-control potentiometer's wiper breaks contact with the resistive element, there can be problems. Also, never use a switch or relay to make or break the connection between the drive circuit and the laser diode. The following are some other laser diode precautions:

Power measurement: It's not safe to assume that the optical power output of a laser diode will match what's stated in the manufacturer's minimum-maximum data—each diode will have a unique operating characteristic as a result of manufacturing tolerances. For accuracy, you must measure the output with an optical power meter or a calibrated photodiode. Remember, once the laser diode is past the threshold point, stimulated emission is achieved and the optical output increases significantly for a small increase in forward current. Therefore, a very slight increase in drive current may cause the optical output to exceed the absolute maximum. Also, make sure to include optical losses through any lenses or other components when making measurements and calculations.

Operating temperature and heat sinks: In most applications, laser diodes require heat sinks, especially when operating continuously. Without a heat sink, the laser diode junction temperature will quickly

increase, causing optical output to degrade. If the laser diode temperature continues to rise, exceeding the maximum operating temperature, the diode can be catastrophically damaged or the long-term performance may degrade significantly. Generally, a lower operating temperature will help extend the diode's lifetime. Visible laser diodes with lower wavelengths (e.g., ~635 nm) are typically more sensitive to temperature than the IR laser diodes. Often, thermoelectric cooling is required to keep temperatures down. When using heat sinks, a small amount of non-silicone-type heat sink compound improves thermal conductivity between laser diode and heat sink.

Windows: Keep laser diode windows and other opticals in the path clean. Dust or fingerprints will cause diffraction or interference in the laser output that can result in lower output or anomalies in the far-field pattern. The window should be cleaned using a cotton swab and ethanol.

The Easy Way Out: Laser Diode Modules and Laser Pointers

If all you really want is a visible laser to shoot around, a commercial diode laser module or some brands of laser pointers (those that include optical feedback based on laser power regulation) may be the ticket. Both the modules and laser pointers include a driver circuit capable of operating reliably on unregulated low-voltage dc input and a collimating lens matched to the laser diode. Many of the modules will permit fine adjustment of the lens position to optimize the collimation or permit focusing to a point at a particular distance. However, neither the module nor the pointer is designed to be modulated at any more than a few hertz, due to heavy internal filtering designed to protect the laser diode from power spikes. Therefore, they are generally not suited for laser communication applications. In general, it's a whole lot easier starting out with the module or pointer rather than a laser diode and homemade power supply, or even a commercial driver, if it isn't explicitly designed for your particular laser diode. There is no way to know how reliable or robust an inexpensive laser pointer will be, or if the beam quality will be acceptable. Diode laser modules are generally more expensive and of higher quality than the pointers, so they may be better for serious applications. Also, consider a helium-neon laser, since even the cheapest type is likely to generate a beam with better beam quality than the typical diode laser module or laser pointer.

Laser Diode Specifications

Lasing wavelength, λp: The wavelength of light emitted by the laser diode. For a single-mode device, this is the wavelength of the single spectral line of the laser output. For a multimode device, this is the wavelength of the spectral line with the greatest intensity.

Threshold current, I_{th}: The boundary between spontaneous emission and the stimulated emission shown on the optical power output versus forward current curve. Below the threshold current point, the output resembles the incoherent output from an LED; at or above the specified threshold current, the device begins to produce laser output. Once past the threshold point, stimulated emission is achieved and the optical output increases significantly for small increases in forward current.

Operating current, I_{op}: The amount of forward current through the laser diode necessary to produce the specified typical optical output at a specified operating temperature.

Operating voltage, V_{op}: The forward voltage across the laser diode when the device produces its specified typical optical output at a specified operating temperature.

Optical power output, P_O: Maximum allowable instantaneous optical power output in either continuous or pulse operation.

Operating temperature range: Range of case temperatures within which the device may be safely operated.

Monitor current, I_m: The current through the photodiode, at a specified reverse-bias voltage, when the laser diode is producing its typical optical power output.

Photodiode dark current, $I_{D(PD)}$: The current through the reverse-biased internal monitor photodiode when the laser is not emitting.

Reverse voltage, V_R: Maximum allowable voltage when reverse bias is applied to the laser diode or photodiode. For laser diodes with an internal monitor photodiode, the reverse voltage is specified for the laser diode as $V_{L(\mathrm{LD})}$ and for the photodiode as $V_{R(\mathrm{PD})}$.

Aspect ratio, AR: The ratio of the laser diode's divergence angles, $\theta\perp$ (perpendicular) and $\theta \parallel$ (parallel). A diode with a 30° perpendicular divergence and a 10° parallel divergence has an elliptical beam with an aspect ratio of 3:1.

Astigmatism, A_s or *D_{as}:* The laser beam appears to have different source points for the directions perpendicular and parallel to the junction plane. The astigmatic distance is defined as the distance between the two apparent sources. A laser diode with a large amount of astigmatism must have the astigmatism corrected (or reduced) if the laser diode output is to be accurately focused—otherwise, the resulting focused beam will be astigmatic.

Beam divergence, $\theta\perp$ and $\theta\parallel$: Also referred to as *radiation angles*. The beam divergence is measured as the full angle and at the half-maximum intensity point, known as full-width half-maximum, or FWHM. Angular specifications are provided for both the perpendicular axis and the parallel axis.

Polarization ratio: The output from a single-cavity laser diode is linearly polarized parallel to the laser junction. Spontaneous emission with a random polarization and/or with a polarization perpendicular to the laser junction is also present. The polarization ratio is defined as the parallel component divided by the perpendicular component. For a diode operating near its maximum power, the ratio is typically greater than 100:1. When operating near the threshold point, the ratio would be considerably lower as the spontaneous emission becomes more significant.

Slope efficiency, SE: Also referred to as *differential efficiency*. This is the mean value of the incremental change in optical power for an incremental change in forward current when the device is operating in the lasing region of the optical power output versus forward current curve.

Rise time: Time required for the optical output to rise from 10 percent to 90 percent of its maximum value.

Positional accuracy (D_x, D_y, D_x): Also referred to as *emission point accuracy*. These specifications define the positional accuracy of the laser diode emitter with respect to the device package. D_x and D_y are measured as the planar displacement of the chip from the physical axis of the package. D_z is measured perpendicular to the reference surface. Specifications may list both angular error expressed in degrees and the linear error in microns.

5.4 Photoresistors

Photoresistors are light-controlled variable resistors. These are also known as light-dependent resistors (LDRs). In terms of operation, a photoresistor is usually very resistive (in the megaohms) when placed in the dark. However, when it is illuminated, its resistance decreases significantly; it may drop as low as a few hundred ohms, depending on the light intensity. In terms of applications, photoresistors are used in light- and dark-activated switching circuits and in light-sensitive detector circuits. Figure 5.13 shows the symbol for a photoresistor.

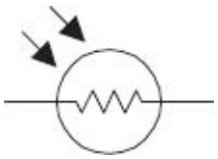
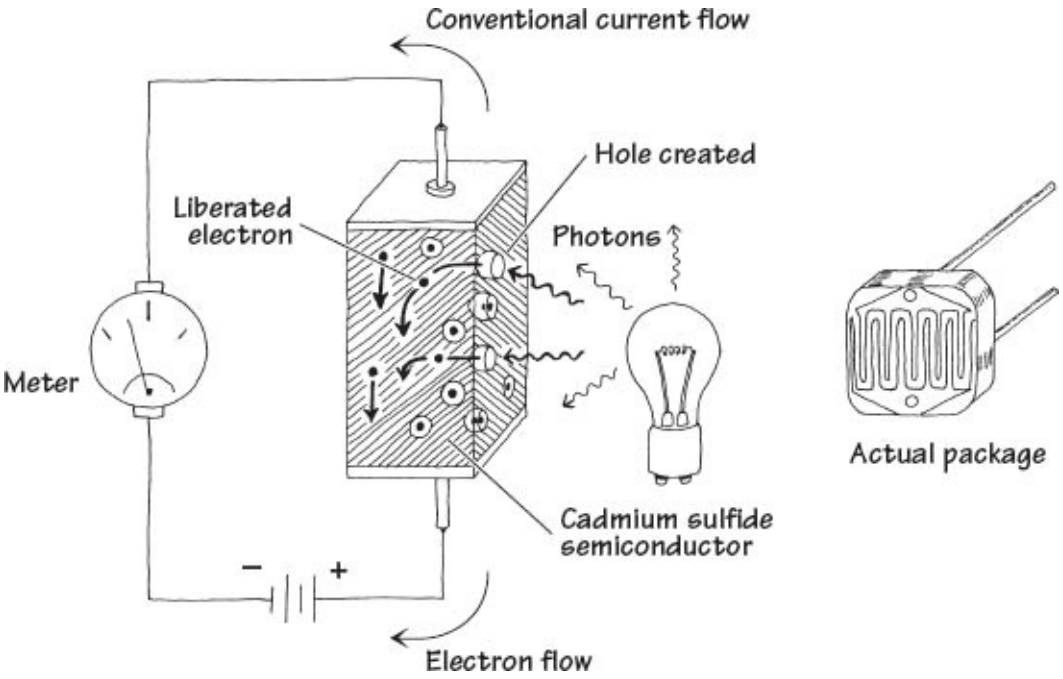


FIGURE 5.13

5.4.1 How a Photoresistor Works



Photoresistors are made from a special kind of semiconductor crystal, such as cadmium sulfide (for light) or lead sulfide (for infrared). When this semiconductor is placed in the dark, electrons within its structure do not want to flow through the resistor because they are too strongly bound to the crystal's atoms. However, when illuminated, incoming photons of light collide with the bound electrons, stripping them from the binding atom, thus creating a hole in the process. These liberated electrons can now contribute to the current flowing through the device (the resistance goes down).

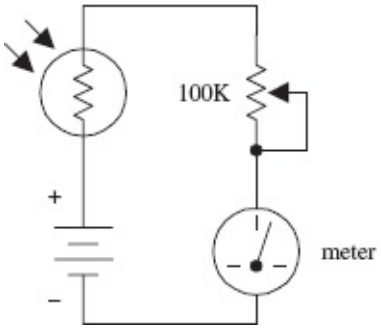
FIGURE 5.14

5.4.2 Technical Stuff

Photoresistors may require a few milliseconds or more to fully respond to changes in light intensity and may require a few seconds to return to their normal *dark resistance* once light is removed. In general, photoresistors pretty much all function in a similar manner. However, the sensitivity and resistance range of a photoresistor may vary greatly from one device to the next. Also, certain photoresistor may respond better to light that contains photons within a particular wavelength of the spectrum. For example, cadmium sulfide photoresistors respond best to light within the 400- to 800-nm range, whereas lead sulfide photoresistors respond best to infrared photons.

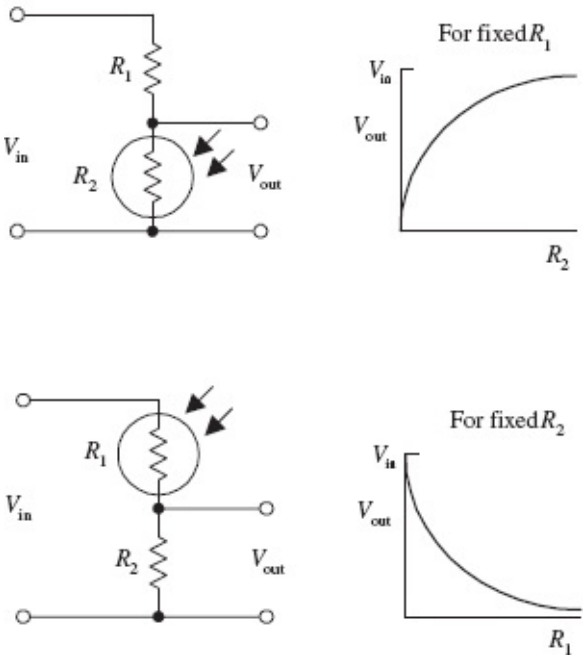
5.4.3 Applications

Simple Light Meter



Here, a photoresistor acts as the light-sensing element in a simple light meter. When dark, the photoresistor is very resistive, and little current flows through the series loop; the meter is at its lowest deflection level. When an increasingly bright light source is shone on the photoresistor, the photoresistor's resistance begins to decrease, and more current begins to flow through the series loop; the meter starts to deflect. The potentiometer is used to adjust the sensitivity of the meter.

Light-Sensitive Voltage Divider

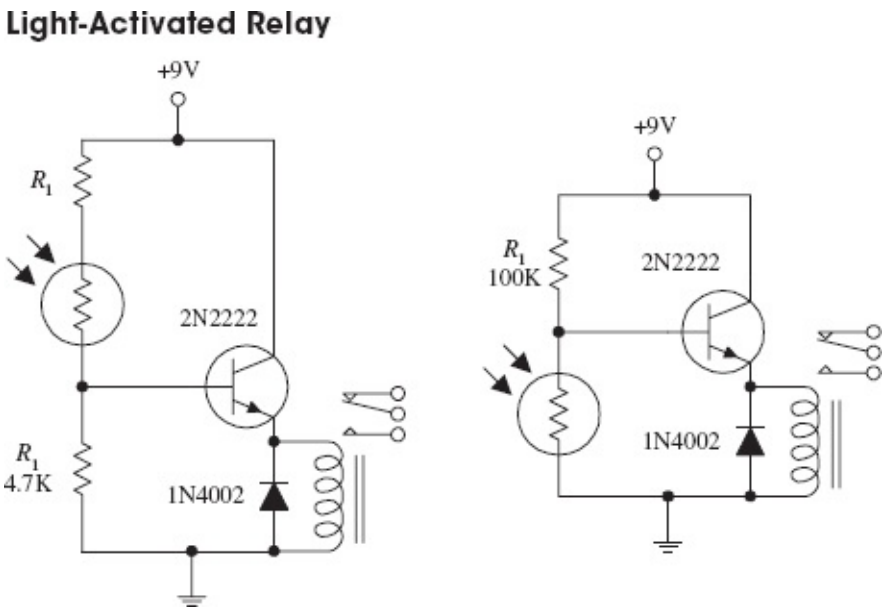


These circuits resemble the voltage-divider circuit described in Chap. 3. As before, the output voltage is given by

$$V_{out} = \frac{R_2}{R_1 + R_2} V_{in}$$

As the intensity of light increases, the resistance of the photoresistor decreases, so V_{out} in the top circuit gets smaller as more light hits it, whereas V_{out} in the lower circuit gets larger.

A voltage divider like this is usually used when using the photoresistor with a microcontroller.



The two circuits shown here make use of light-sensitive voltage dividers to trip a relay whenever the light intensity changes. In the light-activated circuit, when light comes in contact with the photoresistor, the photoresistor's resistance decreases, causing an increase in the transistor's base current and voltage. If the base current and voltage are large enough, the transistor will allow enough current to pass from collector to emitter, triggering the relay. The dark-activated relay works in a similar but opposite manner. The value of R_1 in the light-activated circuit should be around 1 kΩ but may need some adjusting. R_1 in the dark-activated circuit (100 kΩ) also may need adjusting. A 6- to 9-V relay with a 500-Ω coil can be used in either circuit.

FIGURE 5.15

5.5 Photodiodes

Photodiodes are two-lead devices that convert light energy (photon energy) directly into electric current. If the anode and cathode leads of a photodiode are joined together by a wire and then the photodiode is placed in the dark, no current will flow through the wire. However, when the photodiode is illuminated, it suddenly becomes a small current source that pumps current from the cathode through the wire and into the anode. Figure 5.16 depicts the symbol for a photo diode.

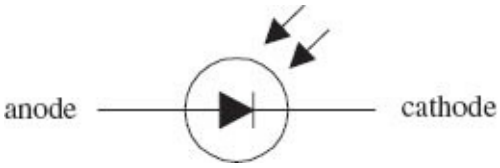


FIGURE 5.16

Photodiodes are used most commonly to detect fast pulses of near-infrared light used in wireless communications. They are often found in light-meter circuits (e.g., camera light meters, intrusion alarms, etc.) because they have very linear light/current responses.

5.5.1 How a Photodiode Works

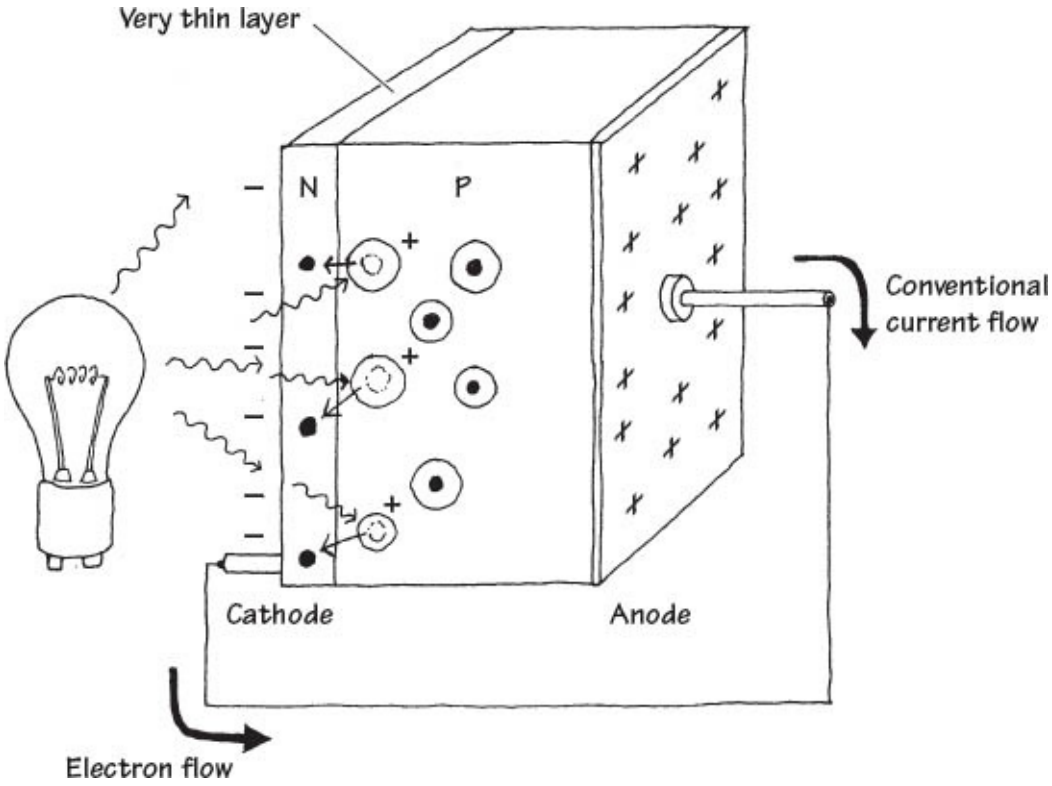
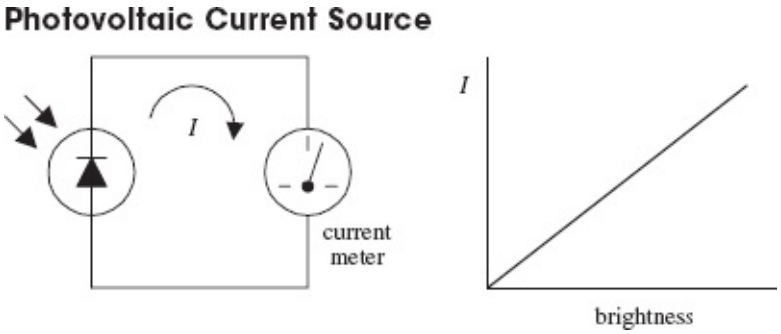


FIGURE 5.17

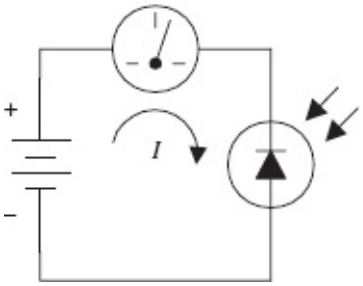
A photodiode is built by sandwiching a very thin *n*-type semiconductor together with a thicker *p*-type semiconductor. (The n side has an abundance of electrons; the *p* side has an abundance of holes.) The *n* side of the combination is considered the cathode, while the *p* side is considered the anode. Now, if you shine light on this device, a number of photons will pass through the *n*-semiconductor and into the *p*-semiconductor. Some of the photons that make it into the *p* side will then collide with bound electrons within the *p*-semiconductor, ejecting them and creating holes in the process. If these collisions are close enough to the *pn* interface, the ejected electrons will then cross the junction. What you get in the end is extra electrons on the *n* side and extra holes on the *p* side. This segregation of positive and negative charges leads to a potential difference being formed across the *p*-*n* junction. Now if you connect a wire from the cathode (*n* side) to the anode (*p* side), electrons will flow through the wire, from the electron-abundant cathode end to the hole-abundant anode end (or if you like, a conventional positive current will flow from the anode to cathode). A commercial photodiode typically places the *p*-*n* semiconductor in a plastic or metal case that contains a window. The window may contain a magnifying lens and a filter.

5.5.2 Basic Operations



Here, a photodiode acts to convert light energy directly into electric current that can be measured with a meter. The input intensity of light (brightness) and the output current are nearly linear.

Photoconductive Operation



Individual photodiodes may not produce enough current needed to drive a particular light-sensitive circuit. Usually they are incorporated along with a voltage source. Here, a photodiode is connected in reversed-biased direction with a battery. When dark, a small current called the *dark current* (within the nA range) flows through the photodiode. When the photodiode is illuminated, a larger current flows. This circuit, unlike the preceding circuit, uses the battery for increased output current. A resistor placed in series with the diode and battery can be used to calibrate the meter. Note that if you treat the photodiode as if it were an ordinary diode, conduction will not occur; it must be pointed in the opposite direction.

FIGURE 5.18

5.5.3 Kinds of Photodiodes

Photodiodes come in all shapes and sizes. Some come with built-in lenses, some come with optical filters, some are designed for high-speed responses, some have large surface areas for high sensitivity, and some have small surface areas. When the surface area of a photodiode increases, the response time tends to slow down. Table 5.2 presents a sample portion of a specifications table for a photodiode.

TABLE 5.2 Part of a Specifications Table for a Photodiode

MNFR #	DESCRIPTION	REVERSE VOLTAGE (V) V_R	MAX. DARK CURRENT (NA) I_D	MIN. LIGHT CURRENT (μA) I_L	POWER DISSIPATION (MW) P_D	RISE TIME (ns) t_r	TYPICAL DETECTION ANGLE (°)	TYPICAL PEAK EMISSION WAVELENGTH (NM) λ_p
NTE3033	Infrared	30	50	35	100	50	65	900

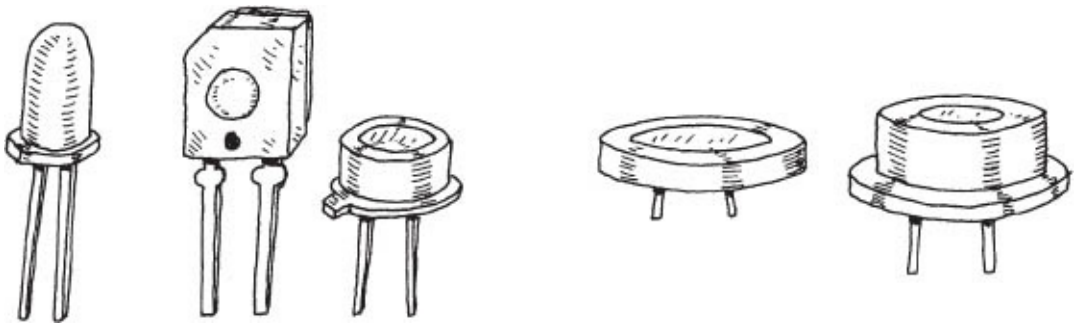


FIGURE 5.19

5.6 Solar Cells

Solar cells are photodiodes with very large surface areas. The large surface area of a solar cell makes the device more sensitive to incoming light, as well as more powerful (larger currents and voltages) than photodiodes. For example, a single silicon solar cell may be capable of producing a 0.5-V potential that can supply up to 0.1 A when exposed to bright light.

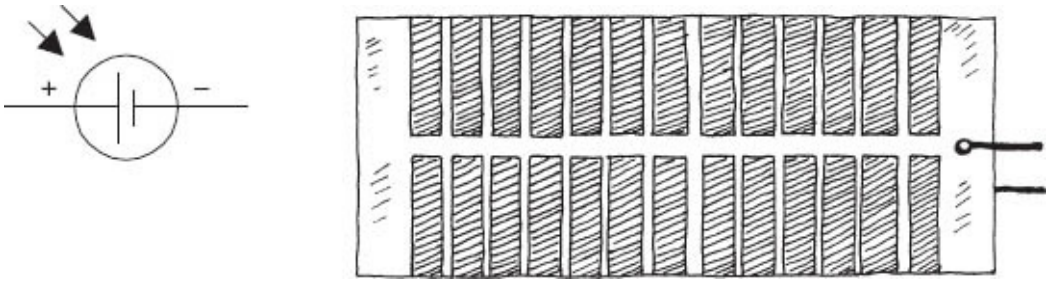
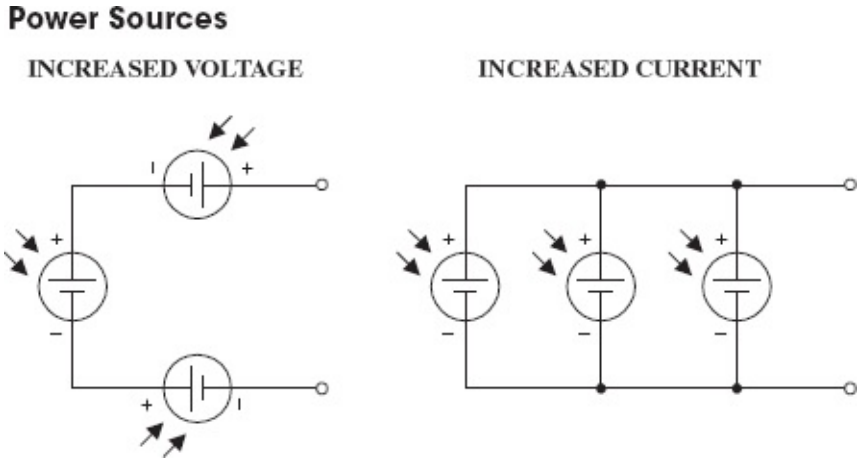


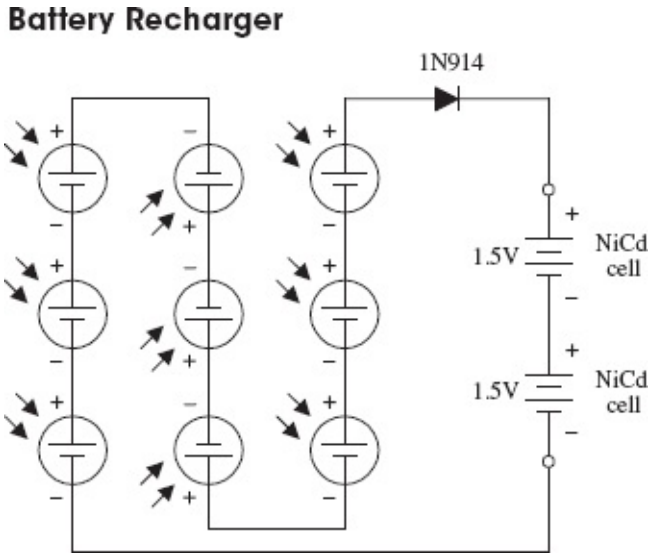
FIGURE 5.20

Solar cells can be used to power small devices such as solar-powered calculators or can be added in series to recharge nickel cadmium batteries. Often solar cells are used as light-sensitive elements in detectors of visible and near-infrared light (e.g., light meters, light-sensitive triggering mechanism for relays). Like photodiodes, solar cells have a positive and negative lead that must be connected to the more positive and more negative voltage regions within a circuit. The typical response time for a solar cell is around 20 ms.

5.6.1 Basic Operations



Like batteries, solar cells can be combined in series or parallel configurations. Each solar cell produces an open-circuit voltage from around 0.45 to 0.5 V and may generate as much as 0.1 A in bright light. By adding cells in series, the output voltage becomes the sum of the individual cell voltages. When cells are placed in parallel, the output current increases.



The circuit here shows how nine solar cells placed in series can be used to recharge two 1.5-V NiCd cells. (Each cell provides 0.5 V, so the total voltage is 4.5 V minus a 0.6-V drop due to the diode.) The diode is added to the circuit to prevent the NiCd cells from discharging through the solar cell during times of darkness. It is important not to exceed the safe charging rate of NiCd cells. To slow the charge rate, a resistor placed in series with the batteries can be added.

FIGURE 5.21

5.7 Phototransistors

Phototransistors are light-sensitive transistors. A common type of phototransistor resembles a bipolar transistor with its base lead removed and replaced with a light-sensitive surface area. When this surface area is kept dark, the device is off (practically no current flows through the collector-to-emitter region). However, when the light-sensitive region is exposed to light, a small base current is generated that controls a much larger collector-to-emitter current. Field-effect phototransistors (photoFETs) are light-sensitive field-effect transistors. Unlike photobipolar transistors, photoFETs use light to generate a gate voltage that is used to control a drain-source current. PhotoFETs are extremely sensitive to variations in light but are more fragile (electrically speaking) than bipolar phototransistors.

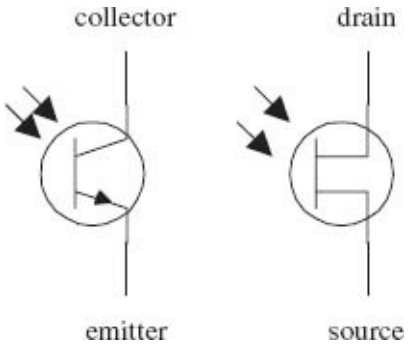


FIGURE 5.22

5.7.1 How a Phototransistor Works

Figure 5.23 shows a simple model of a two-lead bipolar phototransistor. The details of how this device works are given below.

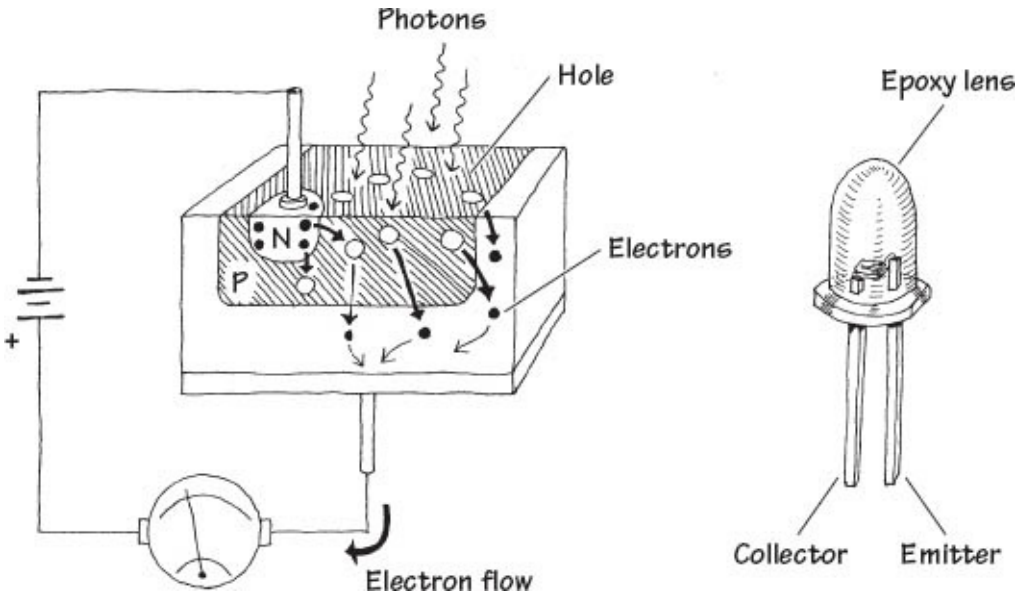
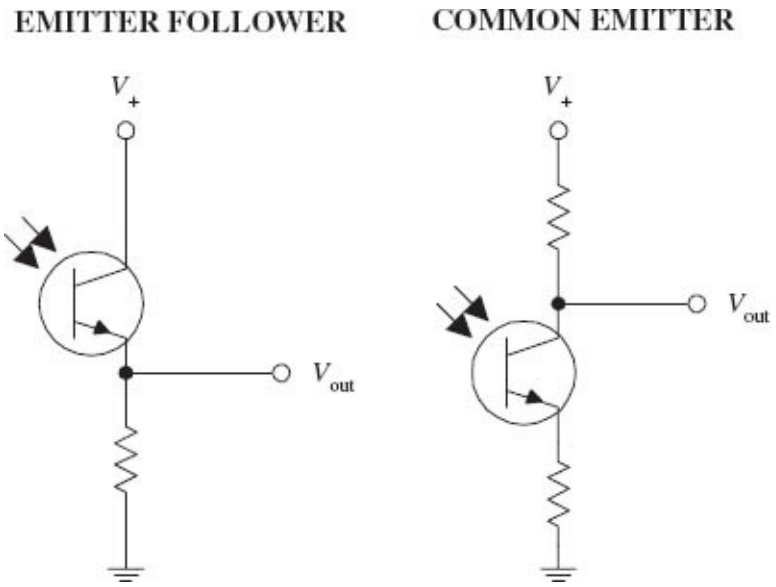


FIGURE 5.23

The bipolar phototransistor resembles a bipolar transistor (with no base lead) that has an extra large *p*-type semiconductor region that is open for light exposure. When photons from a light source collide with electrons within the *p*-type semiconductor, they gain enough energy to jump across the *pn*-junction energy barrier—provided the photons are of the right frequency/energy. As electrons jump from the *p* region into

the lower *n* region, holes are created in the *p*-type semiconductor. The extra electrons injected into the lower *n*-type slab are drawn toward the positive terminal of the battery, while electrons from the negative terminal of the battery are drawn into the upper *n*-type semiconductor and across the *n*-*p* junction, where they combine with the holes. The net result is an electron current that flows from the emitter to the collector. In terms of conventional currents, everything is backward. That is, you would say that when the base region is exposed to light, a positive current *I* flows from the collector to the emitter. Commercial phototransistors often place the *pnp* semiconductor in an epoxy case that also acts as a magnifying lens. Other phototransistors use a metal container and a plastic window to encase the chip.

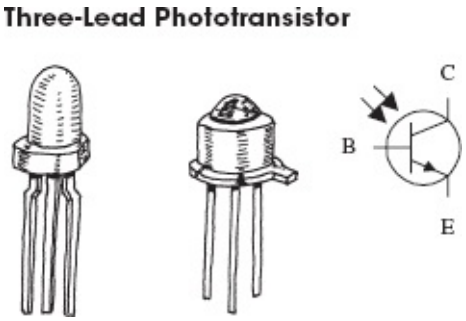
5.7.2 Basic Configurations



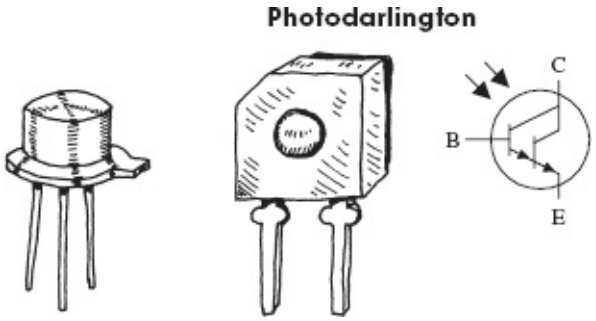
In many ways, a phototransistor is much like an ordinary bipolar transistor. Here, you can see the emitter-follower (current gain, no voltage gain) and the common-emitter amplifier (voltage gain) configurations. The emitter-follower and common-emitter circuits are discussed in Chap. 4.

FIGURE 5.24

5.7.3 Kinds of Phototransistors



Two-lead phototransistors may not be able to inject enough electrons into the base region to promote a desired collector-to-emitter current. For this reason, a three-lead phototransistor with a base lead may be used. The extra base lead can be fed external current to help boost the number of electrons injected into the base region. In effect, the base current depends on both the light intensity and the supplied base current. With optoelectronic circuits, three-lead phototransistors are often used in place of two-lead devices, provided the base is left untouched.



This is similar to a conventional bipolar Darlington transistor but is light-sensitive. Photodarlingtons are much more sensitive to light than ordinary phototransistors, but they tend to have slower response times. These devices may or may not come with a base lead.

FIGURE 5.25

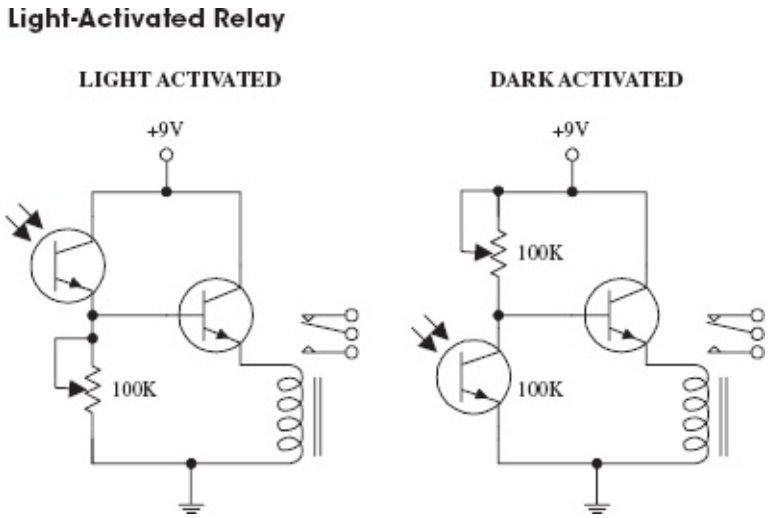
5.7.4 Technical Stuff

Like ordinary transistors, phototransistors have maximum breakdown voltages and current and power dissipation ratings. The collector current I_C through a phototransistor depends directly on the input radiation density, the dc current gain of the device, and the external base current (for three-lead phototransistors). When a phototransistor is used to control a collector-to-emitter current, a small amount of leakage current, called the *dark current* I_D, will flow through the device even when the device is kept in the dark. This current is usually insignificant (within the nA range). Table 5.3 presents a portion of a typical data sheet for phototransistors.

TABLE 5.3 Part of a Specifications Table for Phototransistors

MNFR #	DESCRIPTION	COLLECTOR TO BASE VOLTAGE (V) BV_{CBO}	MAX. COLLECTOR CURRENT (mA) I_C	MAX. COLLECTOR DARK CURRENT (nA) I_D	MIN. LIGHT CURRENT (mA) I_L	MAX. POWER DISSIPATION (mW) P_D	TYPICAL RESPONSE TIME (μS)
NTE3031	*npn*, Si, visible and IR	30 (V_{CEO})	40	100 at 10 V V_{CE}	1	150	6
NTE3036	*npn*, Si, Darlington, visible and near IR	50	250	100	12	250	151

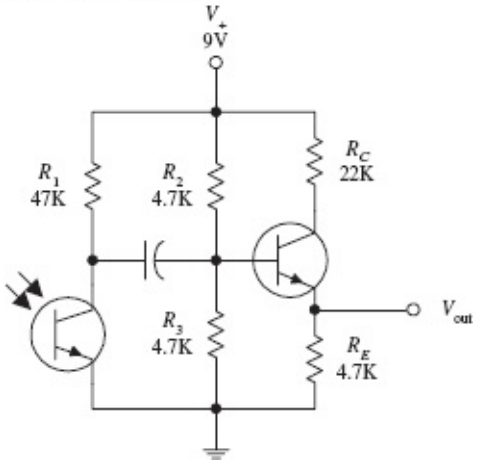
5.7.5 Applications



Here, a phototransistor is used to control the base current supplied to a power-switching transistor that is used to supply

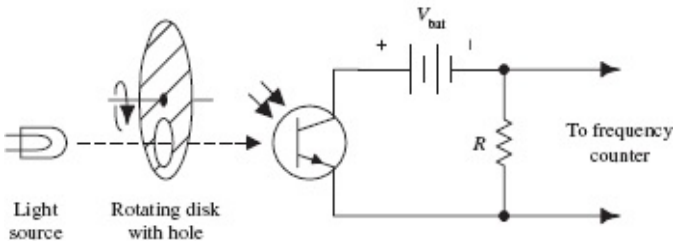
current to a relay. In the light-activated circuit, when light comes in contact with the phototransistor, the phototransistor turns on, allowing current to pass from the supply into the base of the power-switching transistor. The power-switching transistor then turns on, and current flows through the relay, triggering it to switch states. In the dark-activated circuit, an opposite effect occurs. When light is removed from the phototransistor, the phototransistor turns off, allowing more current to enter the base of the power-switching transistor. The 100-k potentiometers are used to adjust the sensitivity of both devices by controlling current flow through the phototransistor.

Receiver Circuit



The circuit shown here illustrates how a phototransistor can be used as a modulated lightwave detector with an amplifier section (current gain amplifier). R_2 and R_3 are used to set the dc operating point of the power-switching transistor, and R_1 is used to set the sensitivity of the phototransistor. The capacitor acts to block unwanted dc signals from entering the amplifier section.

Tachometer



Here is a simple example of how a phototransistor can be used as a simple frequency counter, or tachometer. A rotating disk with a hole (connected to a rotating shaft) will pass light through its hole once every revolution. The light that passes will then trigger the phototransistor into conduction. A frequency counter is used to count the number of electrical pulses generated.

FIGURE 5.26

5.8 Photothyristors

Photothyristors are light-activated thyristors. Two common photothyristors include the light-activated SCR (LASCR) and the light-activated triac. A LASCR acts like a switch that changes states whenever it is exposed to a pulse of light. Even when the light is removed, the LASCR remains on until the anode and cathode polarities are reversed or the power is removed. A light-active triac is similar to a LASCR but is designed to handle ac currents. The symbol for a LASCR is shown below.

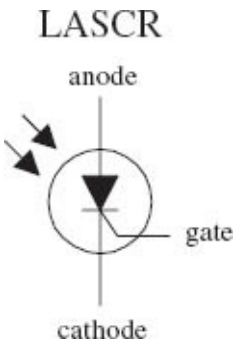
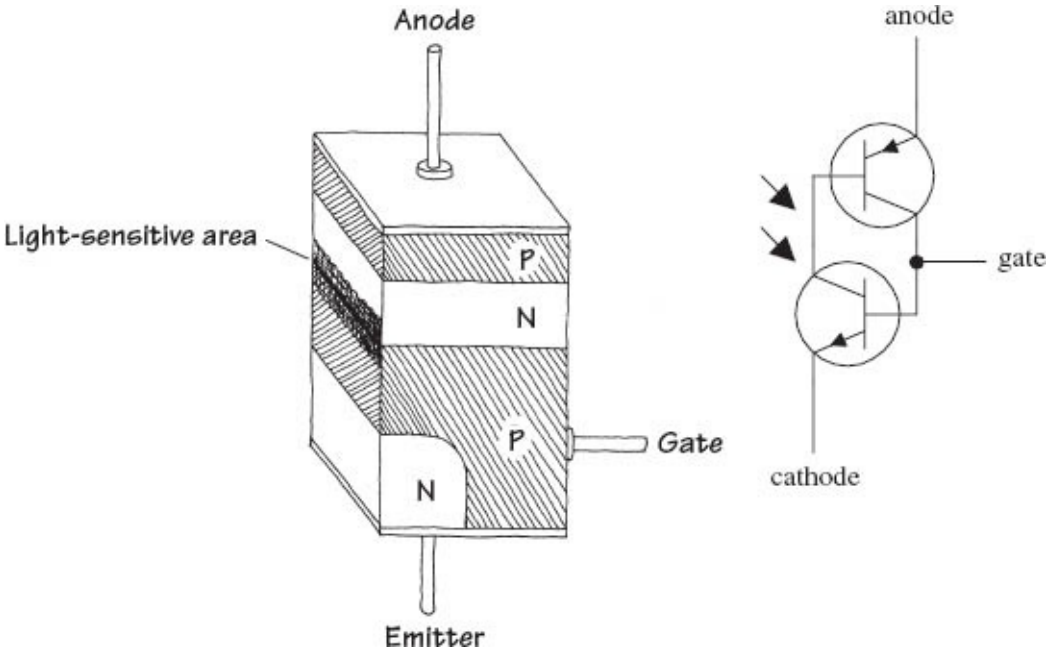


FIGURE 5.27

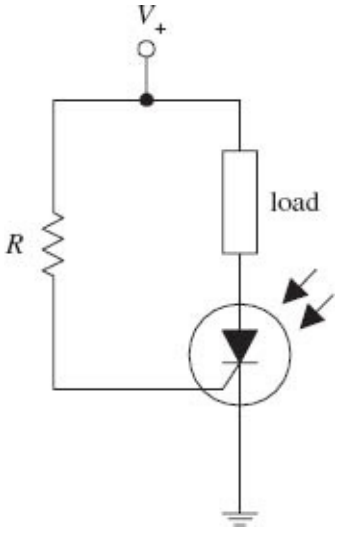
5.8.1 How LASCRs Work



The equivalent circuit shown here helps explain how a LASCR works. Again, like other *p-n* junction optoelectronic device, a photon will collide with an electron in the *p*-semiconductor side, and an electron will be ejected across the *p-n* junction into the *n* side. When a number of photons liberate a number of electrons across the junction, a large enough current at the base is generated to turn the transistors on. Even when the photons are eliminated, the LASCR will remain on until the polarities of the anode and cathode are reversed or the power is cut. (This results from the fact that the transistors' bases are continuously simulated by the main current flowing through the anode and cathode leads.)

FIGURE 5.28

5.8.2 Basic Operation



When no light is present, the LASCR is off; no current will flow through the load. However, when the LASCR is illuminated, it turns on, allowing current to flow through the load. The resistor in this circuit is used to set the triggering level of the LASCR.

FIGURE 5.29

5.9 Optoisolators

Optoisolators/optocouplers are devices that interconnect two circuits by means of an optical interface.

For example, a typical optoisolator may consist of an LED and a phototransistor enclosed in a light-tight container. The LED portion of the optoisolator is connected to the source circuit, whereas the phototransistor portion is connected to the detector circuit. Whenever the LED is supplied current, it emits photons that are detected by the phototransistor. There are many other kinds of source-sensor combinations, such as LED-photodiode, LED-LASCR, and lamp-photoresistor pairs.

In terms of applications, optoisolators are used frequently to provide electrical isolation between two separate circuits. This means that one circuit can be used to control another circuit without undesirable changes in voltage and current that might occur if the two circuits were connected electrically. Isolation couplers typically are enclosed in a dark container, with both source and sensor facing each other. In such an arrangement, the optoisolator is referred to as a *closed pair* (see Fig. 5.30*a*). Besides being used for electrical isolation applications, closed pairs are also used for level conversions and solid-state relaying. A *slotted coupler/interrupter* is a device that contains an open slot between the source and sensor through which a blocker can be placed to interrupt light signals (see Fig. 5.30*b*). These devices are frequently used for object detection, bounce-free switching, and vibration detection. A *reflective pair* is another kind of optoisolator configuration that uses a source to emit light and a sensor to detect that light once it has reflected off an object. Reflective pairs are used as object detectors, reflectance monitors, tachometers, and movement detectors (see Fig. 5.30*c*).

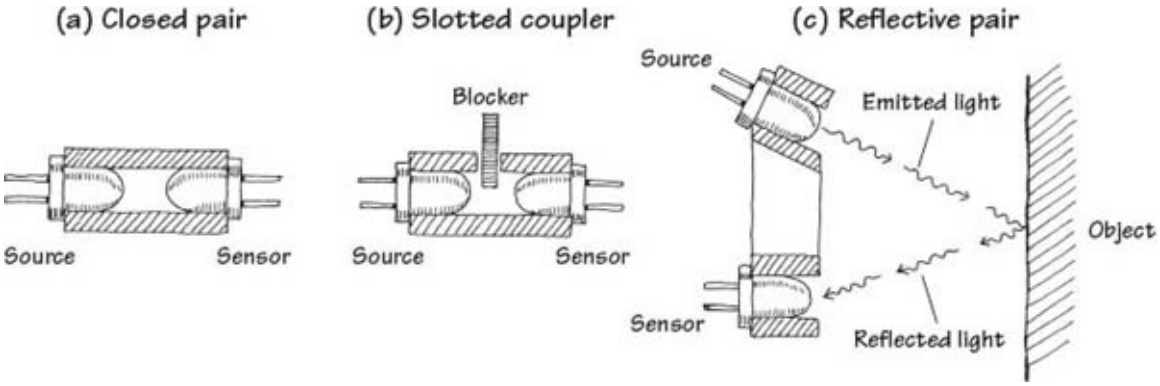


FIGURE 5.30

5.9.1 Integrated Optoisolators

Closed-pair optoisolators usually come in integrated packages. Figure 5.31 shows two sample optoisolator ICs.

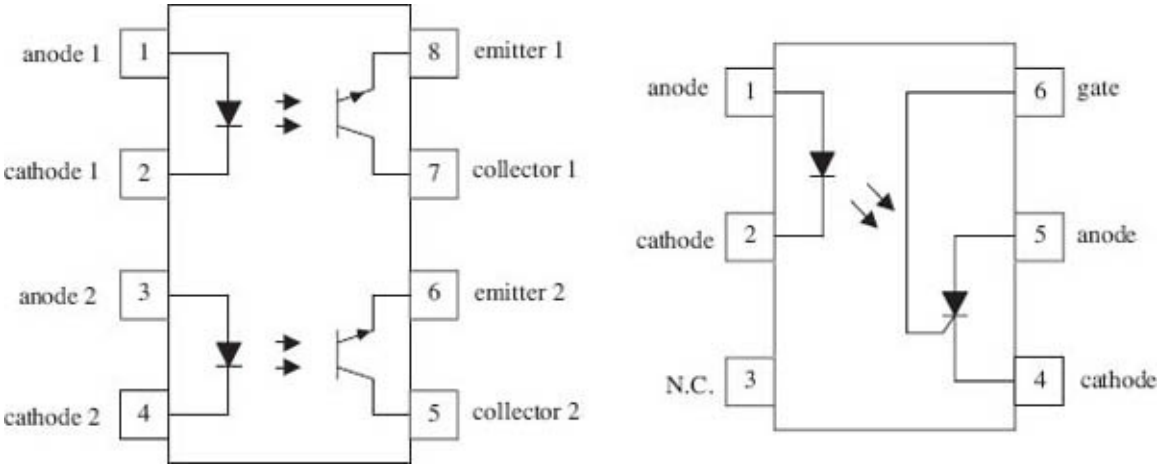
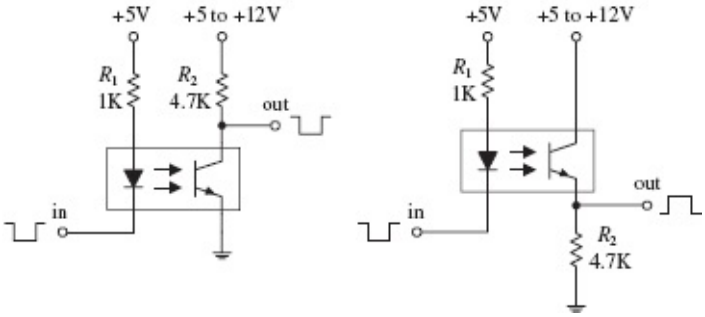


FIGURE 5.31

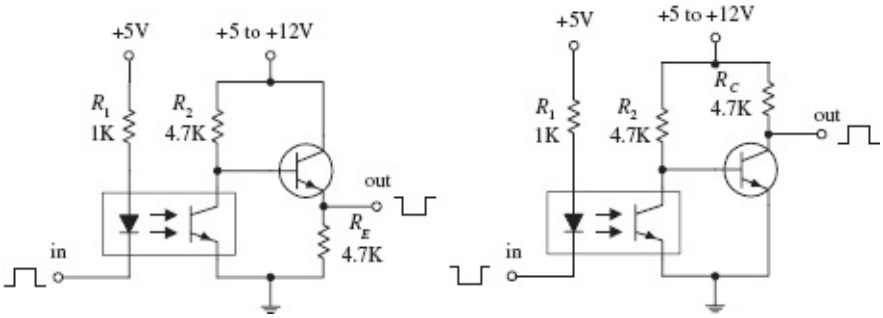
5.9.2 *Applications*

Basic Isolators/Level Shifters



Here, a diode/phototransistor optoisolator is used to provide electrical isolation between the source circuit and the sensor circuit, as well as providing a dc level shift in the output. In the leftmost circuit, the output is noninverted, while the output in the rightmost circuit is inverted.

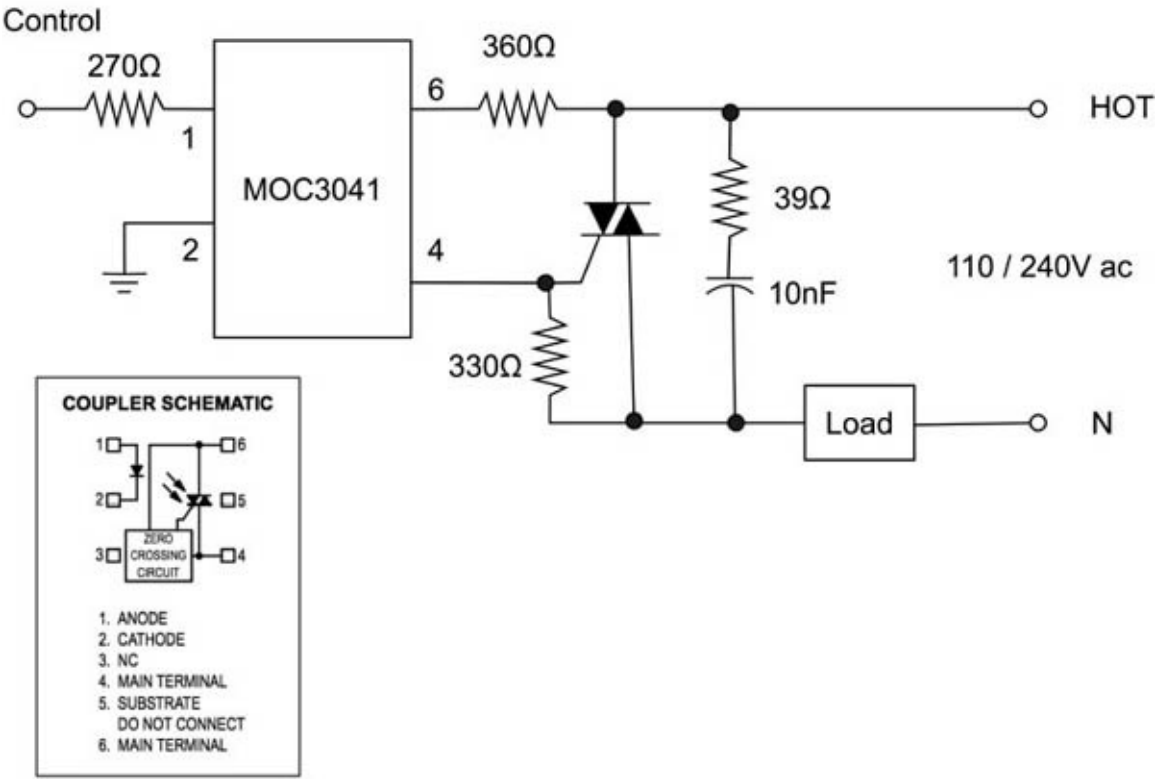
Optocoupler with Amplifier



In optoelectronic applications, the phototransistor section of an optoisolator may not be able to provide enough power-handling capacity to switch large currents. The circuits to the right incorporate a power-switching transistor to solve this problem.

FIGURE 5.32

Solid-State Relay



Optoisolated triacs like the MOC3041 have zero-crossing detection built in so that they turn on only at the zero-crossing point, minimizing current surges. This device can be used to control a regular triac to switch 110 V ac.

FIGURE 5.33

5.10 Optical Fiber

While strictly speaking, optical fibers are “optical” rather than “optoelectrical,” they are often used with photodiodes and LEDs/laser diodes as a medium through which to carry data encoded on beams of light. Figure 5.34 shows how an optical fiber works.

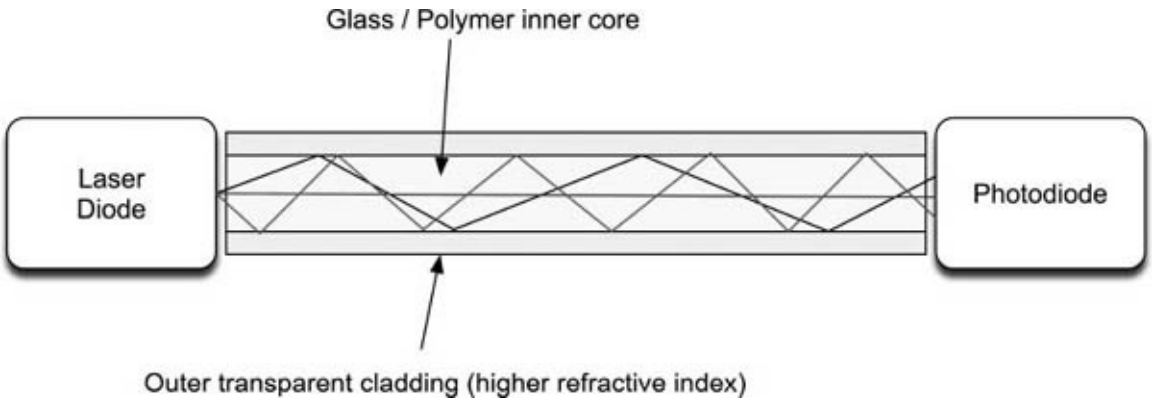


FIGURE 5.34 Optical fiber

The boundary between the two layers, with an outer layer of higher reflective index, ensures that the fiber as a whole acts as a wave guide, with minimal losses of light as it bounces down the fiber. Therefore, optical fibers are much better than wires for transmitting data over long distances. They can also achieve staggering bandwidths, being immune to the induction and capacitive problems associated with wire. As such, they are heavily employed in the telecom industry, especially for high-bandwidth “pipes” between cities and also as submarine cables. It is becoming more and more common for fiber to replace copper in the final mile of telecom connections, actually terminating in homes to provide very high-speed Internet and cable TV.

As Fig. 5.34 implies, light will follow various paths through the fiber, and these paths will be of different lengths, limiting the effective bandwidth. This effect will reduce as the diameter of the internal core reduces. Taken to its limit, a single-mode fiber has a core of between 8 μm and 10.5 μm, and a cladding material diameter of 125 μm. This can be thought of as allowing only the straight-through path of light shown in Fig. 5.34. These fibers can achieve bandwidths in the order of 50 Gb/s over hundreds of miles.

Data are sent using a light source at the sending end (LED or laser diode) and a photodiode or transistor at the other end. Laser diodes are used in telecom systems because the coherent light that they produce travels better, but LEDs are also used in low-cost systems, such as fiber-based audio links in consumer digital audio systems.

CHAPTER 6

Sensors

Sensors are devices that measure a physical property, such as temperature, humidity, stress, and so on. As this is an electronics book, we are concerned with converting such measurements into electrical signals. We have also stretched the definition of a sensor a little to include things such as global positioning systems (GPSs) that locate your position in space.

Many sensors will provide an output signal that is simply a voltage proportional to the property being measured. Others are digital devices that provide digital data about the property. In both cases, these measurements will often be the input to a microcontroller.

6.1 General Principals

Before we look at some of the vast array of sensors that are available, we will cover some of the general principles that underpin all sensing.

6.1.1 Precision, Accuracy, and Resolution

When looking at a sensor, it is important to understand the difference between precision and accuracy, particularly as precision often gives a misleading impression of the accuracy of the sensor. A classic example of this is with digital weighting scales. These may confidently tell you that you weigh 85.7 kilograms when you actually weigh 92.1 kilograms. This is only an error of 10 percent or so, and may not matter if you always use the same scales, but as an absolute measurement, it might be well off the mark.

So precision is a matter of the number of digits reported by the sensor. For an analog sensor, the readings are continuous.

When it comes to digital sensors, there will be a number of bits of resolution. That might be 8 bits (1 in 256), 12 bits (1 in 4096), or higher; that is, the reading is no longer continuous, but has been quantized.

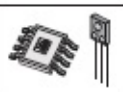
TYPES OF SENSORS			Sensors in bold are covered in this chapter.
Category of Sensor		**What It Does**	**Example Devices**
Position Measuring Devices		Designed to detect and respond to changes in angular position or in linear position of the device.	**Potentiometer** **Linear Position Sensor** **Hall Effect Position Sensor** **Magnetoresistive Angular** **Encoders:** **-Quadrature** **-Incremental Rotary** **-Absolute Rotary** **-Optical**
Proximity, Motion Sensors		Designed to detect and respond to movement outside of the component but within the range of the sensor.	**Ultrasonic Proximity** **Optical Reflective** **Optical Slotted** **PIR (Passive Infrared)** **Inductive Proximity** **Capacitive Proximity** **Reed Switch** **Tactile Switch**
Inertial Devices		Inertia Devices designed to changes in the physical movement of the sensor.	**Accelerometer** **Potentiometer** **Inclinometer** **Gyroscope** **Vibration Sensor/Switch** **Tilt Sensor** **Piezo Shock Sensor** **LVDT/RVDT**
Pressure/Force		Pressure Devices designed to detect a force being exerted against it.	**IC Barometer** **Strain Gauge** **Pressure potentiometer** **LVDT** **Silicon transducer** **Piezoresistive sensor** **Capacitive transducer**
Optical Devices		Optical Devices designed to detect the presence of light or a change in the amount of light on the sensor.	**LDR** **Photodiodes** **Phototransistors** **Photo interrupters** **Reflective Sensors** **IrDA Transceiver** **Solar Cells** **LTV (Light to Voltage) Sensors**
Image, Camera Devices		Image, Camera Devices designed to detect and change a viewable image into a digital signal.	**CMOS Image Sensor**
Magnetic Devices		Magnetic Devices designed to detect and respond to the presence of a magnetic field.	**Hall Effect sensor** **Magnetic Switch** **Linear Compass IC** **Reed Sensor**
Media Devices		Media Devices designed to detect and respond to the presence or the amount of a physical substance on the sensor.	**Gas** **Smoke** **Humidity, Moisture** **Dust** **Float Level** **Fluid Flow**
Current and Voltage Devices		Current Devices designed to detect and respond to changes in the flow of electricity in a wire or circuit.	**Hall Effect current sensor** **DC current sensor** **AC current sensor** **Voltage Transducer**
Temperature		Temperature Devices designed to detect the amount of heat using different techniques and in different mediums.	**Thermistor NTC** **Thermistor PTC** **Resistance Temp Detectors (RTD)** **Thermocouple** **Thermopile** **Digital IC** **Analog IC** **Infrared Thermometer/Pyrometer**
Specialized		Specialized Devices designed to provide detection, measurement, or response in specialized situations, which also may include multiple functions.	**Audio Microphone** **Geiger-Müller Tube** **Chemical**

FIGURE 6.1 Sensor types

6.1.2 The Observer Effect

The observer effect states that the act of observing a property changes it. This is the case in a car tire, for example. When you measure the pressure with a conventional tire gauge, you will let a little of the air out, thus changing the pressure that you are measuring. For most sensors, this change will have a negligible effect, but it is something worth keeping at the back of your mind when deciding whether you are getting a true reading.

6.1.3 Calibration

If your invention is a low-cost consumer item that will be mass-produced, then it will be much cheaper and easier to produce if there is no calibration to do. In fact, individual calibration to take differences in individual sensors into account is likely to be prohibitively expensive.

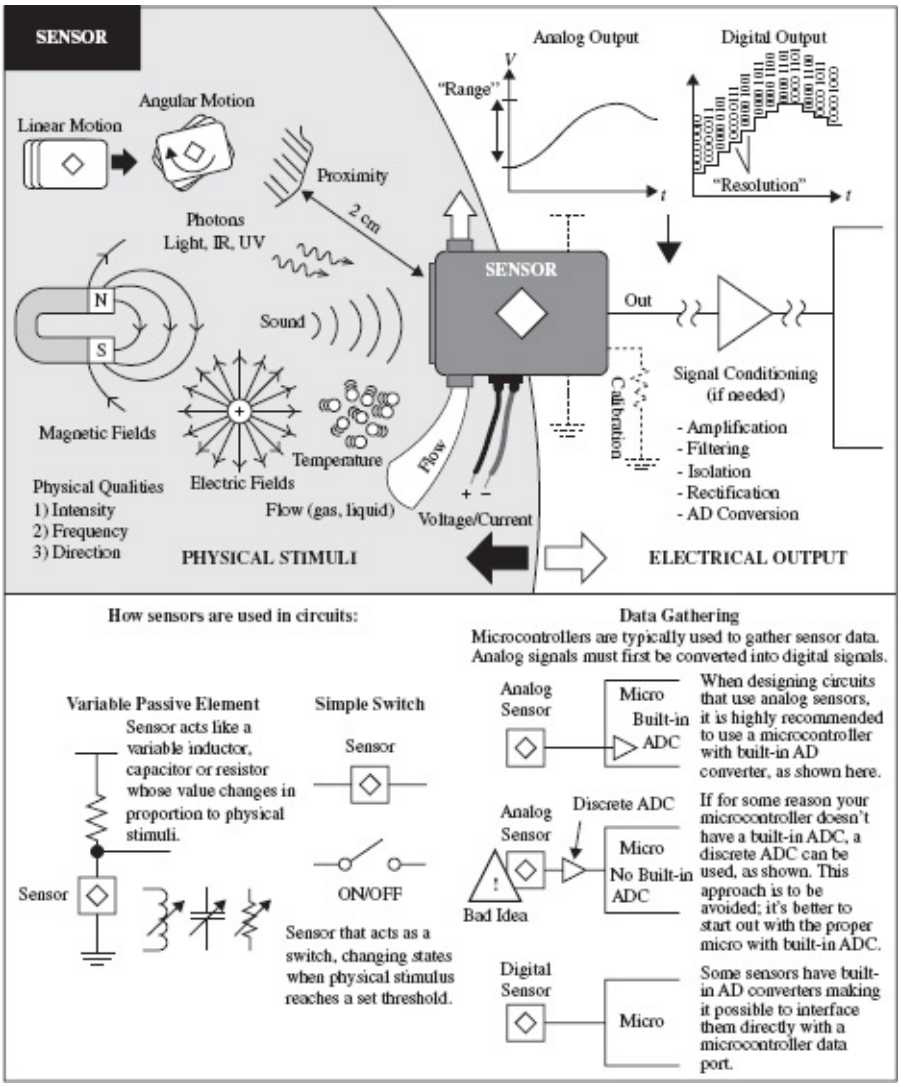


FIGURE 6.2 Sensors

If, on the other hand, your invention is specialized, high-value equipment, then individual calibration of sensors is possible.

The principal of calibration is the same whatever property the sensor measures. The idea is that you take a number of readings from the sensor of unknown accuracy while the sensor is measuring a standard value of known accuracy. So, to calibrate a temperature sensor, your accurate standard might be boiling distilled water at 100°C, or more likely, the temperature in a highly accurate oven constructed for the purpose of calibrating the sensor. This oven will itself need to have been calibrated against an even more accurate standard.

When you have discovered the degree to which your sensor deviates from the standard, then you can compensate for it in some way. Since the sensor will almost certainly be supplying information to a microcontroller, a common way to make the calibration is to change the values in a lookup table. This table contains accurate values against raw values. For example, the raw value straight from a 12-bit analog-to-digital converter is a number between 0 and 1023. These numbers, perhaps in increments of 5, might be the left-hand side of the table, with the right-hand side of the table containing the equivalent temperature as a decimal in degrees Celsius. Interpolation can be used in the gaps between the raw readings by assuming that the small segment of the curve is linear (see Fig. 6.3).

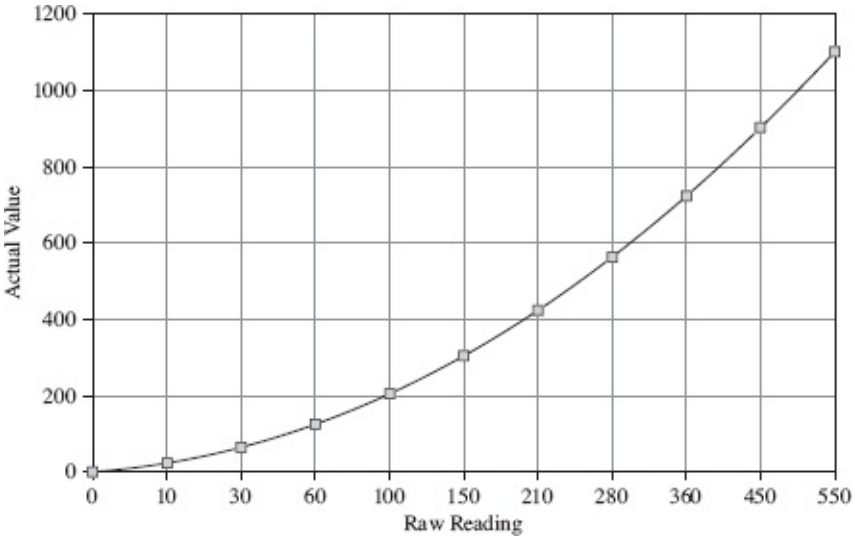


FIGURE 6.3 Actual reading against raw value for a fictional sensor

Some IC sensors are actually calibrated individually during the manufacturing process, with the values for the lookup table being written into the read-only memory (ROM) of the sensor. This allows accurate sensing at low cost.

6.2 Temperature

Temperature sensors are one of the most common types of sensors. Computers and many types of equipment will sense their own temperature to prevent overheating. In addition, as well as electronic thermometers, there are thermostats that keep temperature constant by controlling the power to a load, usually just turning it on and off. Figure 6.4 shows a variety of temperature sensors.



FIGURE 6.4 Temperature sensors

6.2.1 Thermistors

The term “thermistor” is a combination of the words “thermal” and “resistor.” A thermistor is a resistor whose resistance changes markedly with temperature. Thermistors come in two flavors: negative temperature coefficient (NTC) and positive temperature coefficient (PTC). An NTC thermistor’s resistance will decrease with increasing temperature, and a PTC thermistor will behave the opposite way, with the resistance increasing as temperature increases. The NTC is the more common type.

The relationship between temperature and resistance in a thermocouple is not a linear one. Even over relatively short temperature ranges—say 0 to 100°C, a linear approximation will introduce considerable errors (see Fig. 6.5).

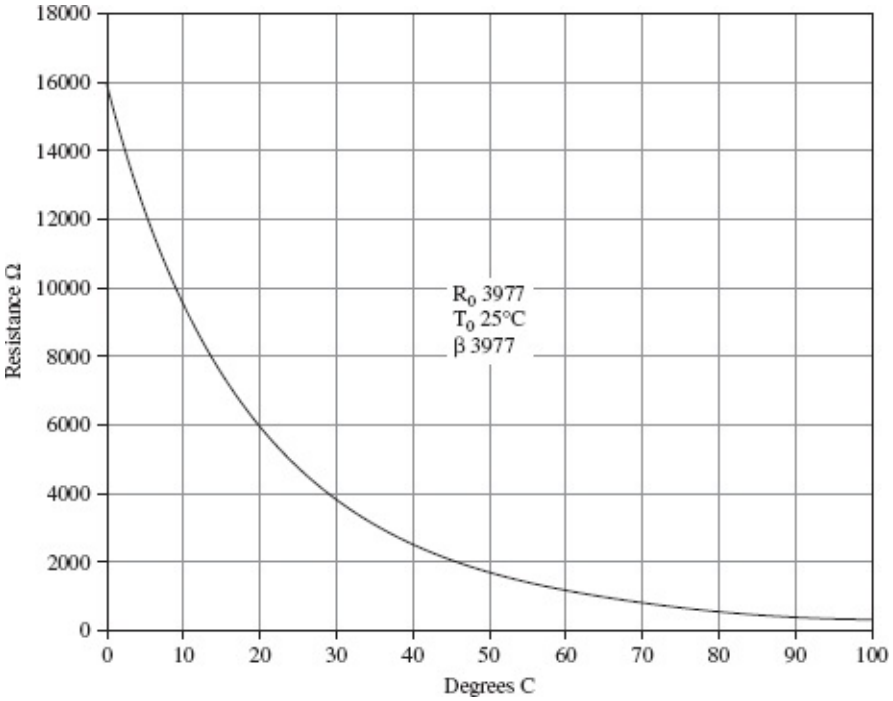


FIGURE 6.5 Resistance against temperature for an NTC thermistor

The Steinhart-Hart equation is a third-order approximation used to determine the temperature as a function of the thermistor's resistance. It works for both NTC and PTC devices. It is normally stated as:

$$\frac{1}{T} = A + B\ln(R) + C\ln^3(R) \tag{6.1}$$

T is the temperature in Kelvin; R is the resistance of the thermistor; and A, B, and C are constants specific to that thermistor. The manufacturer of a thermistor will give all three constant values.

An alternative and more common model for this relationship uses a single parameter (*Beta*) and assumes constants of T_0 and R_0, where T_0 is usually 25°C and R_0 is the resistance of the thermistor at that temperature.

Using this equation, the temperature can be approximated to:

$$\frac{1}{T} = \frac{1}{Beta}\ln\left(\frac{R}{R_0}\right) + \frac{1}{T_0} \tag{6.2}$$

Rearranging this, we can also derive an expression for *R:*

$$R = R_0 e^{Beta\left(\frac{1}{T} - \frac{1}{T_0}\right)} \tag{6.3}$$

As well as providing *Beta*, T_0, and R_0, the data sheet will also specify a temperature range and accuracy.

To use a thermistor as a thermometer for input to a microcontroller, a voltage is required that can be measured by the analog-to-digital converter of the microcontroller. Figure 6.6 shows a typical arrangement using a potential divider with a fixed-value resistor of the same value as R_0.

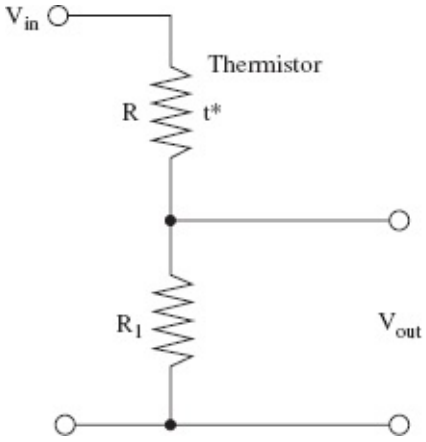


FIGURE 6.6 Using a thermistor as a thermometer

If we place a NTC thermistor at the top of the potential divider, then as the temperature of the thermistor increases, its resistance falls and V_{out} rises.

Looking at Fig. 6.6:

$$V_{out} = \frac{R_1}{R_1 + R} V_{in} \tag{6.4}$$

Combining Formula 6.3 with Formula 6.4, we have:

$$V_{out} = \frac{R_1}{R_1 + \left(R_0 e^{Beta\left(\frac{1}{T} - \frac{1}{T_0}\right)} \right)} V_{in} \tag{6.5}$$

Example 1: Using a potential divider as shown in Fig. 6.3 with a fixed resistor of 4.7 kΩ and an NTC thermistor with a T_0 of 25°C, an R_0 of 4.7 kΩ, and a *Beta* of 3977, what is the formula for calculating V_{out}?

Answer 1: Just substituting the values into the equation, we get:

$$V_{out} = \frac{(5 \times 4700)}{4700 + \left(4700 e^{3977\left(\frac{1}{T} - \frac{1}{(25+273)}\right)} \right)}$$

$$V_{out} = \frac{5}{1 + \left(e^{3977\left(\frac{1}{T} - \frac{1}{(25+273)}\right)} \right)}$$

Note that 273 is added to the temperature in °C to give a temperature in °K.

Example 2: Using the same setup as Example 1, what is V_{out} at 25°C?

Answer 2: In one way, this is a trick question, since by definition the thermistor's resistance at 25°C will be 4.7 kΩ, so the voltage will be 2.5 V. However, we can apply the formula as a sanity check:

$$V_{out} = \frac{5}{1+\left(e^{3977 \times 0}\right)}$$

$$V_{out} = \frac{5}{1+1} = 2.5V$$

Example 3: Using the same setup as Examples 1 and 2, what is V_{out} at 0°C?

Answer 3:

$$V_{out} = \frac{5}{1+\left(e^{3977\left(\frac{1}{273} - \frac{1}{(25+273)}\right)}\right)}$$

$$V_{out} = \frac{5}{1+\left(e^{3977\left(\frac{1}{273} - \frac{1}{(25+273)}\right)}\right)}$$

$$V_{out} = \frac{5}{4.34} = 1.15V$$

6.2.2 Thermocouples

While thermistors are good for measuring relatively small ranges in temperature—typically –40 to +125°C, for much higher temperatures and temperature ranges, thermocouples are used (see Fig. 6.7).

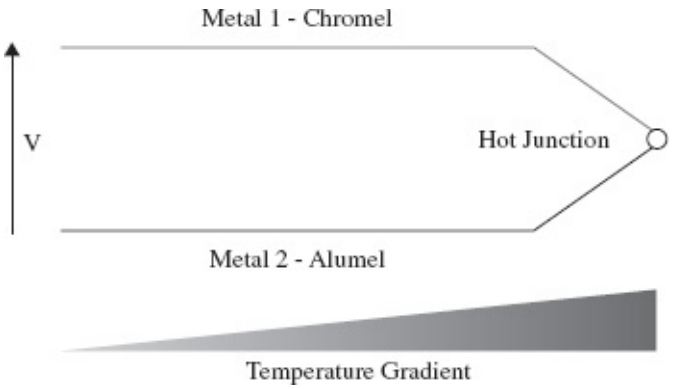


FIGURE 6.7 Thermocouples

Any conductor subjected to a thermal gradient will experience something called the Seebeck effect; that is, it will generate a small voltage. The magnitude of this voltage is dependent on the type of metal, so if two different metals are joined, the temperature of that junction can be measured by measuring the voltage across the metal leads, at the far end from the junction. It is also necessary to measure the temperature at the other end of the thermocouple leads (often using a thermistor, since this will probably be at room temperature). This second temperature is called the "cold-junction" temperature.

Lookup tables are often used to calculate the absolute temperature at the junction, based on the voltage and the cold-junction temperature, as the relationship is not completely linear and needs a fifth-order polynomial to model it accurately. Manufacturer's data sheets for thermocouples will normally contain large tables for calculating the temperature.

The most commonly used metals for a thermocouple are the alloys chromel (90 percent nickel and 10

percent chromium) and alumel (95 percent nickel, 2 percent manganese, 2 percent aluminum, and 1 percent silicon). A thermocouple made with these materials will typically be able to measure temperatures over the range –200°C to +1350°C. The sensitivity is 41 μV / °C for these metals.

6.2.3 *Resistive Temperature Detectors*

Resistive temperature detectors (RTDs) are perhaps the simplest temperature sensors to understand. Like thermistors, RTDs rely on their resistance changing with temperature. However, rather than use a special material that is sensitive to temperature changes (like a thermistor), they simply use a coil of wire (normally platinum) around a glass or ceramic core. The resistance of the core is often contrived to be 100 Ω at 0°C.

RTDs are much less sensitive than thermistors and can therefore be used over a much wider range of temperatures. The resistance of platinum changes in a relatively linear manner, and can be assumed to be linear for a range of 100°C or so. For the range 0 to 100°C, the resistance of a platinum RTD will vary by 0.003925 Ω/Ω/°C.

So, a 100 Ω (at 0°C) platinum RTD will, at a temperature of 100°C, have a resistance of:

$$100\Omega + 100°C \times 100\ \Omega \times 0.003925\ \Omega/\Omega/°C = 139.25\ \Omega$$

The first 100 Ω in the equation above is for the base resistance of the RTD at 0°C.

This can be arranged in a potential divider in the same way as a thermistor.

6.2.4 *Analog Output Thermometer ICs*

An alternative to using a thermistor and fixed-value resistor in a potential divider arrangement is to use a special-purpose temperature measurement IC.

Devices such as the TMP36 come in a three-pin package and are used as shown in Fig. 6.8.

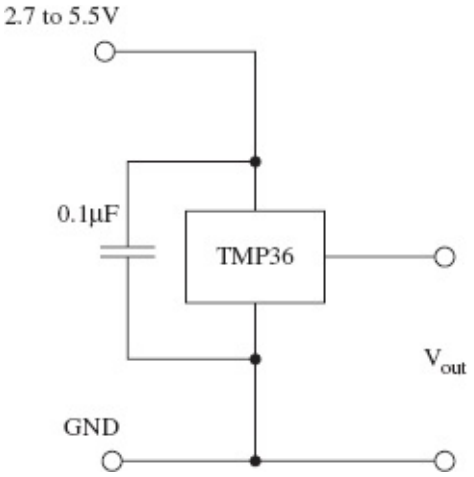


FIGURE 6.8 TMP36 temperature sensor

Unlike a thermistor, the sensor's output voltage is almost linear at 10 mV / °C for a temperature range of –40 to +125°C. The accuracy is only ±2°C over the temperature range.

The temperature in °C is simply calculated from V_{out} using the formula:

$$T = 100V_{out} - 50$$

The constant 50 is specified in the data sheet for the TMP36.

These devices are considerably more expensive than thermistors and series resistors, but they are very easy and convenient to use.

6.2.5 Digital Thermometer ICs

An even higher technology approach to temperature measurement is to use a digital thermometer IC. These devices have a serial interface for use by microcontrollers.

A typical digital thermometer IC is the DS18B20. It uses a serial bus standard called 1-Wire that can allow multiple sensors to share the same data line (see Fig. 6.9).

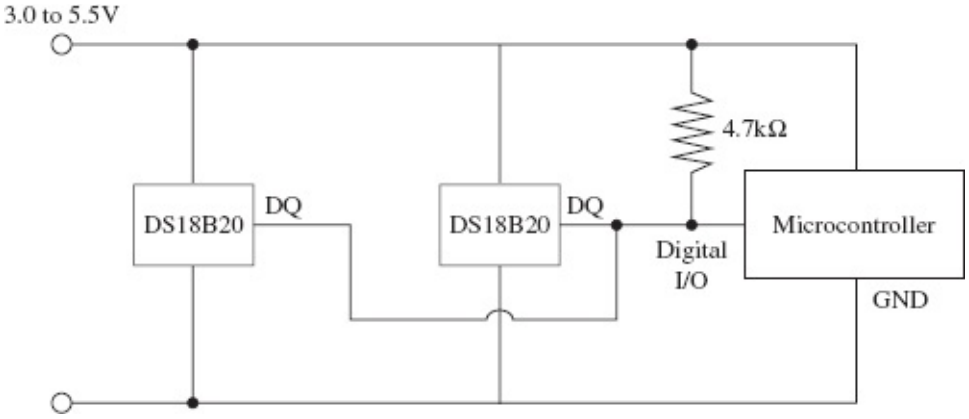


FIGURE 6.9 DS18B20 temperature sensor IC

These devices are more accurate than the linear devices such as the TMP36. The digital thermometer IC has a stated accuracy of ±0.5°C and a temperature range of –55°C to +125°C.

Since digital thermometer ICs transmit their data digitally, they are very suitable for remote sensing, as lead length and electrical interference will have less effect than with an analog device. In fact, they will either work accurately or not at all in an electrically noisy environment.

Digital thermometer ICs can also be configured to work in “parasitic” power mode, where they harvest power from the data line, allowing them to be connected with just two wires. The GND and positive supply to the DS18B20 are both grounded, and the microcontroller must control a MOSFET transistor that pulls the data line up to the supply voltage under strict timing conditions.

Chapter 13 contains more information about using the 1-Wire interface with microcontrollers.

6.2.6 Infrared Thermometers/Pyrometers

If you have been to the doctor recently and have had your temperature taken, you will probably have seen an infrared thermometer that is placed near your ear and measures the temperature just on the inside surface of the ear without any actual contact. The ear is chosen because it is effectively a hole into the body that is free from the influence of external radiation, so that the back surface of the ear acts as a “black-body” radiator.

Pyrometers, or more specifically, broadband pyrometers, measure the radiation intensity, which has dimensions of energy per second per unit of area and use the Stefan-Boltzmann law to determine the temperature:

$$j^* = \sigma T^4$$

The radiation intensity (j^*) is proportional to the fourth power of the temperature, where σ is the Stefan-Boltzmann constant ($5.6704 \times 10^{-8}\ Js^{-1}m^{-2}K^{-4}$).

These devices use an infrared sensor such as the MLX90614 coupled with optics that focus the infrared radiation from the subject onto the sensor. When used for other measurement applications, the infrared thermometer will often have a low-power laser align with the sensor so that it can be aimed to take spot readings of temperature.

Devices such as the MLX90614 are more than just a sensor, containing a low-noise amplifier, high-precision analog-to-digital converter, and all the associated electronics to produce a digital output.

As with other IC sensors, the advantages of including all the electronics in the same package as the sensor itself are the reduction of noise and simplicity of interfacing.

Other sensors are available for high temperatures, such as those found in industrial furnaces, and the term "pyrometer," with its overtones of fire, is usually reserved for such high-temperature applications.

6.2.7 Summary

Table 6.1 shows a number of temperature sensors and their characteristics.

TABLE 6.1 Temperature Sensors

SENSOR	TYPICAL TEMP RANGE (°C)	ACCURACY (±°C)	PROS	CONS	APPLICATIONS
Thermistor	–40 to 125	1	Low cost		Ambient temperature measurement
Thermocouple	–200 to 1350	3	Low cost	The lead is part of the sensor and it measures temperature difference	Industrial Domestic furnace
RTD	–260 to 800	1	Accurate Good linearity	Expensive Slow response time	
Analog IC	–40 to 125 (TMP36)	2	Simple to interface	More expensive than thermistor	Domestic thermostat Digital thermometer
Digital IC	–55 to 125 (DS18B20)	0.5	Simple to use with microcontrollers Accurate	More expensive than thermistor	Domestic thermostat Digital thermometer
Infrared thermometer/ pyrometer	–70 to 380 (MLX90614) Up to 1030 (MLX90616)	0.5	Does not need contact	More expensive than contact solutions	Medical Industrial control, especially where contact sensing is difficult

The figures for accuracy in Table 6.1 are somewhat arbitrary, as for most sensors, high accuracy can be obtained by individual calibration. It is also the accuracy of the overall system—including the electronics that use the sensor value—that counts. The figures assume the use of off-the-shelf components using data sheet parameters without any individual calibration and are meant as an indication of the likely accuracy of measurement that can be achieved with such a sensor.

6.3 Proximity and Touch

This section deals with detecting objects or measuring their distance from the sensor. Figure 6.10 shows a selection of such sensors



FIGURE 6.10 Proximity and touch sensors

6.3.1 Touch Screens

Touch screens are commonly used in cellular phones and tablet computers. There are many different technologies used in touch screens. The most commonly used are resistive touch screens.

Resistive touch screens are one of the older and more readily used touch screen technologies. They rely on a transparent sheet on top of the display. This sheet is flexible and also conductive. Figure 6.11 shows a typical arrangement for a four-wire resistive touch screen.

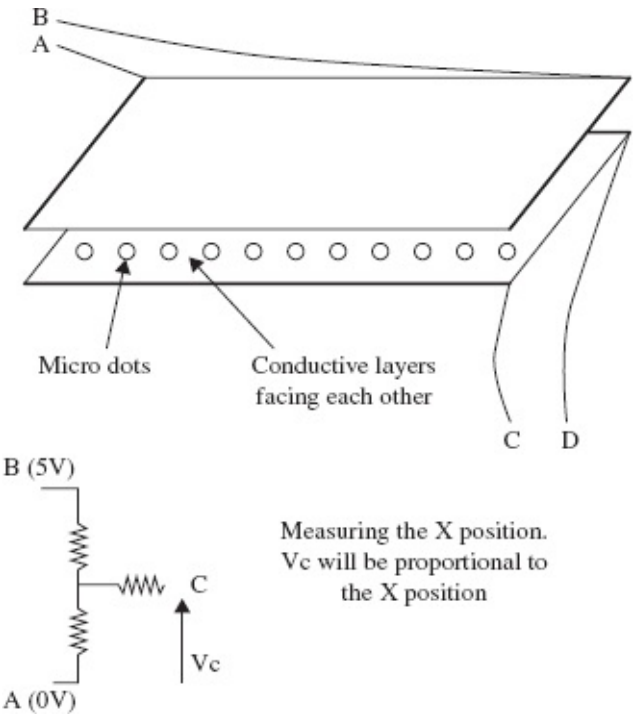


FIGURE 6.11 Resistive touch screen

Both the top and bottom surfaces are coated with a conductive layer. Insulating dots are evenly sprayed onto the rigid bottom surface to keep the layers apart, except when they are pressed together.

To determine the X position of a touch, A is set to 0 V and B to 5 V, establishing a voltage gradient across the top surface. The voltage measured at C, or for that matter D, will be proportional to the X position. This is converted into coordinates using an analog-to-digital converter.

The conductive layer acts just like a potentiometer with C as the slider. If the device measuring the voltage at C has a very high input impedance, then the resistance of the track from the surface to the terminal C can be ignored. Most microcontrollers will have an analog-to-digital converter with a high input resistance, typically several MΩ. So, the voltage at C will be between 0 and 5 V in direct proportion to the distance from A of the touch.

When it comes to reading the Y position, some crafty footwork has to take place. Now C will be set to 0V, D to 5V, and the voltage measured at A or B.

All this processing will be carried out by either a special-purpose controller chip or a microcontroller.

6.3.2 Ultrasonic Distance

Ultrasonic distance-measuring devices are much loved by the developers of hobby robots. They are also found in commercial equivalents to the tape measure.

These devices send out a pulse of ultrasound and time how long it takes for the reflection to come back. A simple calculation involving the speed of sound will determine the distance to the object (see Fig. 6.12).

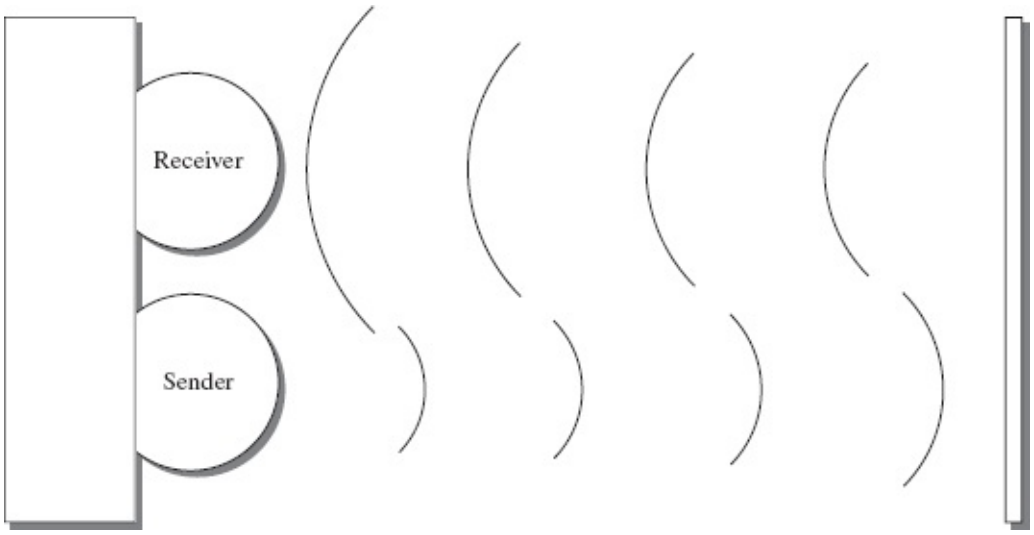


FIGURE 6.12 Ultrasonic distance measurement

These devices have a measurement range of up to around 5 meters (15 ft), but this depends very much on the size and sound reflecting properties of the object. They also suffer limits of accuracy due to variations in the speed of sound, which is dependent on many factors, including atmospheric pressure, humidity, and temperature. For example, at 0°C, the speed of sound in dry air is 331 meters/seconds, but this rises to 346 meters/second at 25°C—a difference of 4.5 percent.

Sometimes separate transducers are used for sending and receiving the pulses of ultrasound, and sometimes a single transducer is used in both roles.

Some devices rely on a microcontroller to initiate the pulse and time the echo. Other, more expensive, devices will have their own digital processing that generates an output that can be read from a microcontroller as one of the following:

- An analog output voltage proportional to the distance
- Serial data
- A train of varying pulse lengths dependent on the distance

In some cases, all three outputs are provided, such as with the SparkFun device SKU: SEN-00639.

Example: Assuming that in dry air at 20°C, the speed of sound is 340 meters/second, if the time period between a pulse of ultrasound being sent is 10 milliseconds, how far away is the reflecting object?

Answer: *distance* = *velocity* × *time* = 340 × 0.01 = 3.4m

However, that is the distance for the entire round-trip, so the actual distance to the object is half of that, or 1.7 meters.

6.3.3 Optical Distance

Another useful device for measuring distance at close range is the infrared optical sensor (see Fig. 6.13). This type of device uses the amount of reflection of a modulated infrared pulse to determine the distance to the object. The Sharp GP2Y0A21YK (SparkFun SKU: SEN-00242) is a typical sensor of this type.

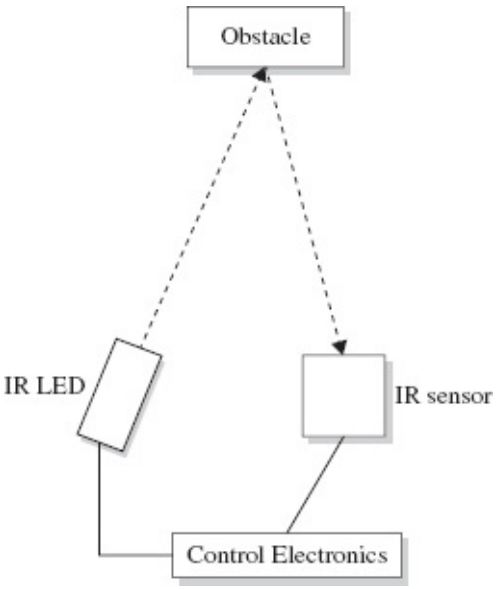


FIGURE 6.13 Optical distance measurement

Infrared optical sensors are used for closer range measurements than the ultrasonic devices and produce an analog output. This is not linear or particularly accurate, so these devices are most useful for simple proximity detection rather than actual measurement.

At close range, the measurements become ambiguous, as shown in the distance/output voltage plot of Fig. 6.14. Therefore, they are usually mounted recessed so that no object can get closer than the ambiguous distance, which is about 5 cm for the sensor in Fig. 6.14.

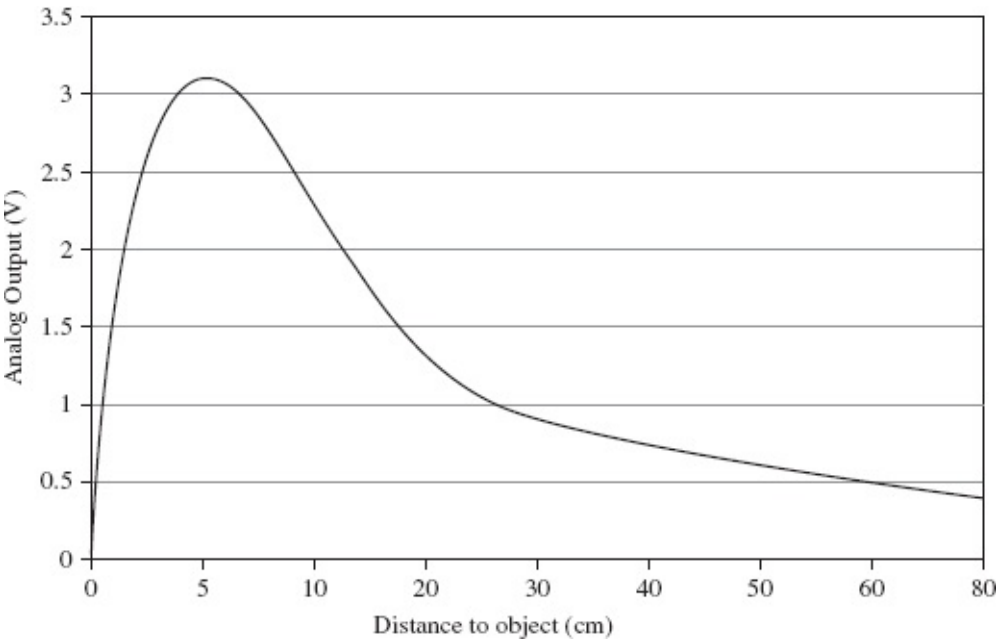


FIGURE 6.14 Output against distance for a Sharp GP2Y0A21YK

When this type of sensor is used with a microcontroller, a lookup table is used to convert the measured voltage to a distance reading.

In some circumstances, you just need to know whether or not something is present. This can be detected with a slotted optical sensor, where the infrared source and sensor are aligned with a slot in between. When the beam is broken, the object is detected.

6.3.4 Capacitive Sensors

Capacitive sensors are frequently used as proximity or touch sensors as a replacement for mechanical push switches. They sense conductors and therefore are ideal for sensing proximity of a hand or finger.

There are various different approaches to capacitive sensing, but they all follow the basic principals shown in Fig. 6.15.

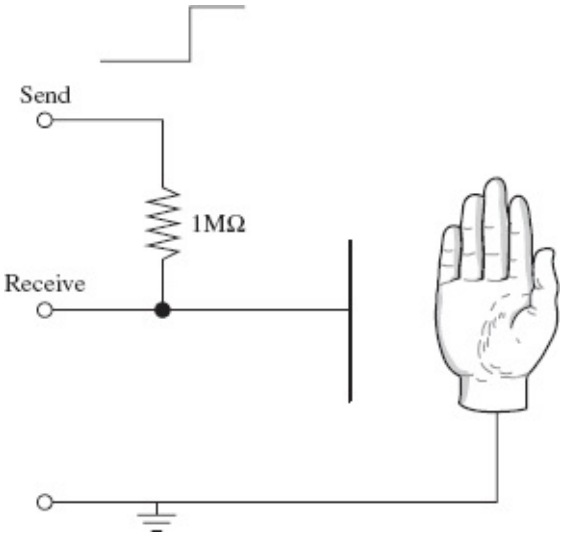


FIGURE 6.15 Capacitive sensing

In this case, the sensing is accomplished using two general-purpose input/output (GPIO) pins of a microcontroller. The Send pin is configured as an output, and the Receive pin as an input. A single fixed resistor between the two pins forms one half of an RC arrangement, with the capacitor formed by the object being sensed and a sensor plate forming the other half of the capacitor.

When a hand moves close to the plate, the capacitance increases. This is sensed as the control software toggles the state of the Send pin and times how long it takes for the Receive pin to catch up with the changed state. In this way, it effectively measures the capacitance and hence the proximity of the object being sensed. The longer it takes for the Receive pin to change to the same state as the Send pin, the closer the object is to the plate.

This type of sensing has the advantages that it requires very little special hardware and can sense through glass, plastic, and other insulators. More advanced versions of this approach can sense in two dimensions to make a capacitative touch screen.

6.3.5 Summary

Table 6.2 compares the relative merits of the different types of distance sensors.

TABLE 6.2 Distance Sensors

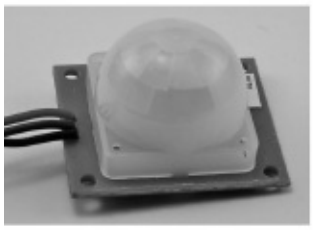
SENSOR	DISTANCE RANGE	ACCURACY	PROS	CONS	APPLICATIONS
Ultrasonic	150 mm to 6 m	25 mm	Simple to interface to	Relatively high cost	Industrial control Security Robotics
Optical reflective	100 to 800 mm	10 mm	Simple to interface to Low cost	Measurement sensitive to IR reflectivity of object	Industrial control Security Robotics
Optical slotted	N/A	N/A	Low cost	Detects only presence or absence, not distance	Industrial control
Capacitive	0 to 30 cm	N/A	Low cost	Detects conductive object with resistive link to earth only	Push switch replacement Touch screens

Another type of sensor that is used in industrial settings is the inductive sensor. This is in essence a mini-metal detector.

Proximity of a magnet can also be detected using magnetic sensors such as Hall effect sensors (discussed in Section 6.6.3) and even the humble read switch, which consists of a pair of contacts in a glass envelope that close when near a magnet.

6.4 Movement, Force, and Pressure

Many different types of sensors can tell us about how things are moving. This has become a common occurrence in smartphones, where an accelerometer is used to detect the orientation of the device for automatic switching of the screen between landscape and portrait format, as well as for game playing. Figure 6.16 shows a variety of different movement-detecting devices.



PIR



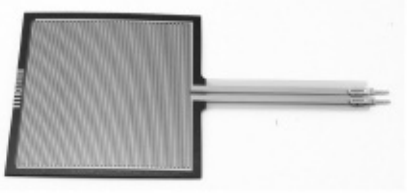
Accelerometer Module



Potentiometer



Rotary Encoder



Force Sensitive Resistor



Vibration Sensor

FIGURE 6.16 Movement sensors

6.4.1 Passive Infrared

Passive infrared (PIR) detectors are most commonly used in intruder alarms. They detect changes in

infrared heat. Because they require a plastic lens, they are usually sold as a detector module on a printed circuit board.

More advanced movement detectors will have multiple detectors that respond to a rapid change in the level of infrared heat. When this is detected, the module may activate a relay, or in some modules, switch a transistor with an open collector arrangement.

Figure 6.17 shows how to use an open collector device such as the PIR sensor from SparkFun (SKU: SEN-08630) to give a digital output.

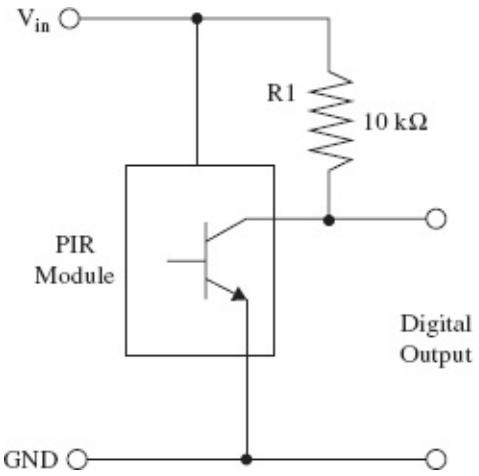


FIGURE 6.17 Using a passive infrared sensor

6.4.2 Acceleration

The ready use of accelerometers in cellular phones has led to them becoming a low-cost and readily available component. The most common of these use a technology called microelectromechanical systems (MEMS). The basic principal for these devices is to measure the position of a mass attached to a spring that stretches under acceleration (see Fig. 6.18). The neat trick is that all of this is fabricated onto an IC.

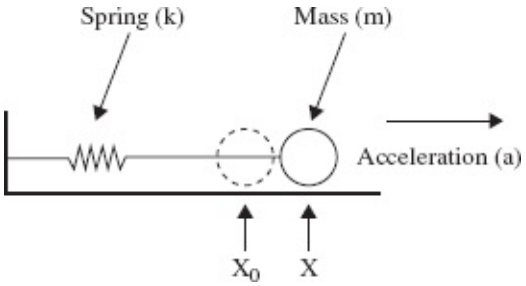


FIGURE 6.18 Mass and spring accelerometer

This arrangement is basically the same as a spring balance of the sort favored by fishermen and travelers worried about exceeding their baggage allowance. In fact, turning it through 90 degrees, you would be measuring the gravitational constant whose units are those of an acceleration. This is how an accelerometer can determine its orientation on the vertical axis.

The force acting on the mass due to the acceleration is:

$$F = k\,(x - x_0) = m\,a$$

So:

$$a = \frac{k(x - x_0)}{m}$$

Since the manufacturer will know *k* and *m*, we just need to be able to measure the displacement of the mass *m*. Measuring this movement often uses the capacitive effect of the mass, in relation to one or two plates. A common arrangement is for the mass to move between two plates making a capacitive divider (see Fig. 6.19 and Section 2.23.14 in Chap. 2).

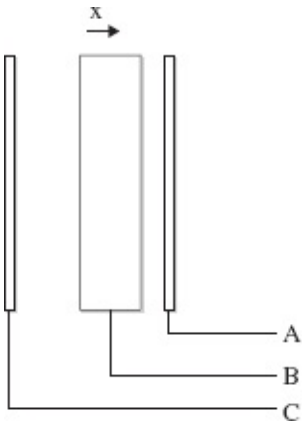


FIGURE 6.19 Measuring the offset of the mass

If the mass, shown as a rectangle in between two plates in Fig. 6.19, is halfway between the two plates A and C, then the capacitance between A and B and between B and C will be equal. An acceleration to the left of the diagram will move the weight to the right, changing the ratio of the capacitance.

Most accelerometers come in a little IC package that not only includes three accelerometers—one for each dimension—but also all the necessary electronics, analog-to-digital converters, and digital electronics to provide serial communication of the data back to a microcontroller using an I2C serial interface.

A much used low-cost accelerometer is the MMA8452Q (SparkFun SKU: COM-10953).

6.4.3 Rotation

Measuring rotation can be as simple as using a potentiometer as a voltage divider, where the voltage at the slider is proportional to the angle of rotation (see Fig. 6.20). However, the potentiometer will not measure a full 360 degrees, and it has stops at both ends of the track that prevent it from turning all the way around.

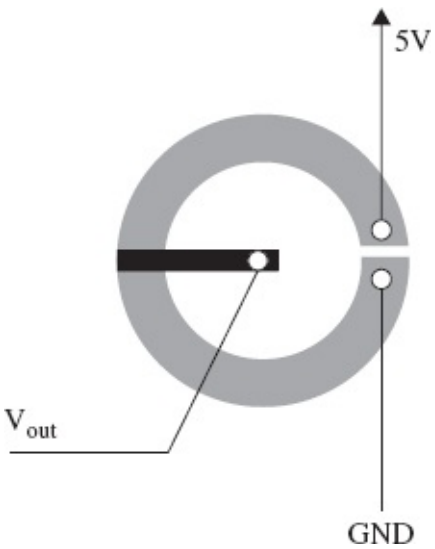


FIGURE 6.20 Measuring rotation with a potentiometer

A more flexible alternative to the potentiometer is the quadrature encoder. This type of device rotates

without any stops, so you can just keep turning it. These are often used as an alternative to potentiometers in electrical appliances. They actually just measure steps to the left or steps to the right, rather than the absolute position.

A quadrature encoder is really just a pair of concentric tracks that each has a contact. They act as a pair of switches producing pulses as the shaft is rotated (see Fig. 6.21).

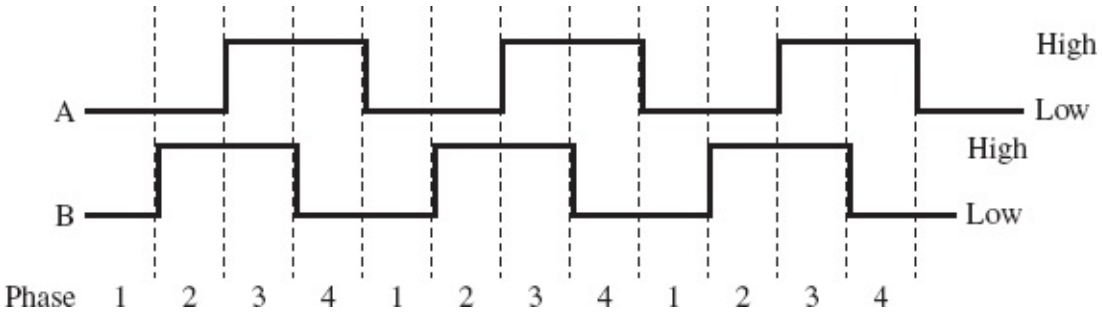


FIGURE 6.21 Outputs from a quadrature encoder

The transitions of A and B are used to determine the direction in which the shaft is being rotated. For example, a clockwise rotation would be read as going from phase 1 to 4 and a counterclockwise from 4 down to 1.

Clockwise (AB): LH HL HH HL

Counterclockwise (AB): HL HH LH LL

As you can tell, this is a digital device that will be connected directly to a microcontroller.

Quadrature encoders come in various resolutions. Low-cost ones may have only 12 pulses per revolution. Better-quality devices can have up to 200 pulses per revolution and be designed for high-speed operation of up to 30,000 rpm. Such devices will use optical sensors rather than contacts.

A variation on the quadrature encoders that measure relative rotation are absolute rotary encoders. This type of device is essentially a rotary switch with a number of tracks that then indicate the angle by providing a binary output.

Rather than count in "normal" binary, the binary numbers for each position are arranged so that only one bit changes between one position of the switch and the next. This type of encoding is called a Gray code. For example, the 3-bit Gray code looks like this: 000, 001, 011, 010, 110, 111, 101, 100.

Table 6.3 shows some of the pros and cons of rotation sensors.

TABLE 6.3 Rotation Sensors

SENSOR	ACCURACY	PROS	CONS	APPLICATIONS
Potentiometer	2 to 5°	Cheap Measures absolute angle	End stops limited to 300°	Volume control, etc.
Quadrature encoder (low cost)	30°	Cheap	Measures changes in angle, not absolute angle	Volume control, etc.
Quadrature encoder (optical)	2°	Accurate High speed	Expensive	Industrial control
Absolute rotary encoder	22° (4-bit encoder)	Measures absolute angle	Not very accurate	Industrial control

6.4.4 Flow

The simplest sensors for measuring flow are variations on some kind of paddle wheel or turbine that is

placed in the flow and spins at a speed proportional to the flow. The problem of measuring the rate of rotation remains. An anemometer that measures the wind speed is a typical example of this (see Fig. 6.22).

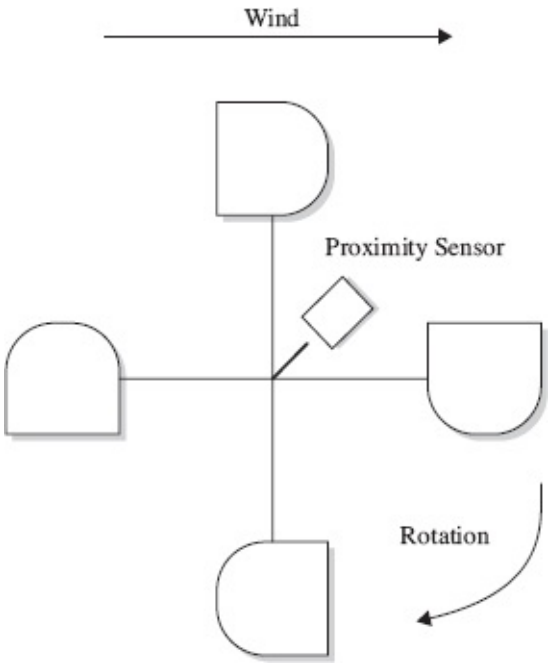


FIGURE 6.22 A cup anemometer

The wind will cause the arrangement of cups to spin at a rate of rotation proportional to the wind strength. Some fixed point on the cup assembly passes close to a proximity sensor that produces a pulse for each full rotation. In this way, you get a stream of pulses at a frequency proportional to the wind speed. Measuring the speed is then just a matter of using a microcontroller to count the number of pulses in a given period of time.

The proximity sensor may be a fixed magnet on the cup assembly and a magnetic sensor such as the Hall effect sensor (described in Section 6.6.3), or may be an optical sensor where a beam is interrupted.

Various types of fluid flow sensors are summarized in Table 6.4.

TABLE 6.4 Flow Sensors

SENSOR	DESCRIPTION	PROS	CONS
Hotwire	A current is passed through the wire to keep it at a constant temperature. Fluid flow draws heat away from the wire, and the current required to maintain the temperature is an indication of the rate of flow.	Simple	Hot
Ultrasonic (transit time)	Ultrasound is injected into the fluid from an upstream transducer, and the time to reach a downstream transducer is measured. The process is then reversed, and the fluid velocity can be determined from the difference in times.	No moving parts	Very dependent on geometry of the arrangement
Ultrasonic (Doppler)	A transmitter and receiver are used side by side to measure the difference in frequency between the transmission and ultrasound reflected from particles in the fluid.	No moving parts	Requires particles in the fluid
Laser (Doppler)	Similar to ultrasonic Doppler, monochromatic light from a laser is used. The difference in wavelength of the reflected light is measured.	Accurate	Requires particles in the fluid Expensive

6.4.5 Force

Force can be sensed using a force-sensitive resistor. These devices operate in a similar manner to a resistive touch screen. They are two layers, with conducting tracks and a microdot insulating layer, but the harder you press, the more of the top conductor comes into contact with the bottom conductor and the lower the resistance. These devices are not usually very accurate.

Strain gauges operate on the principal of deforming the geometry of a resistor as it is put under tension or compression (see Fig. 6.23).

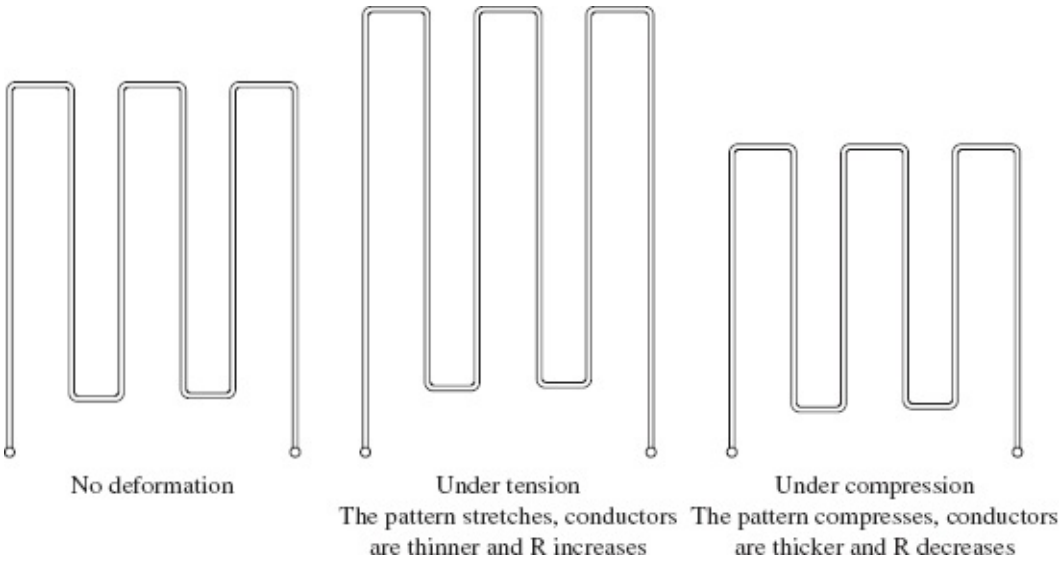


FIGURE 6.23 Strain gauge

For a more accurate sensing of force, a load cell is used. This arranges two or more strain gauges bonded onto a deformable metal block. They will often also include temperature compensation.

6.4.6 Tilt

Although an accelerometer can be used to measure the angle of tilt, or just detect tilt from the horizontal, if all that is required is to know that the sensor has tilted from the horizontal, then a simpler sensor can be used. The simplest of these use a small metal ball in a case, which sits on contacts when horizontal but rolls off when the device is tipped.

6.4.7 Vibration and Mechanical Shock

Piezoelectric materials are good for detecting shock and vibration. The vibration sensor in Fig. 6.16 uses a small weight (just a rivet) on the end of a flexible strip of piezoelectric material. When the weight wobbles due to vibration, a voltage is generated across the terminals. The output of these voltages will drift, so for practical use, place a 10 MΩ resistor in parallel with them. This same approach will work for detecting mechanical shock.

6.4.8 Pressure

Unless you plan to work in process control or build a weather station, sensing pressure is unlikely to be something that you will need to do. Except for when you have very particular sensing requirements, the solution is nearly always to buy a digital sensor on a chip, such as the two listed in Table 6.5.

TABLE 6.5 Pressure Sensors

SENSOR	RANGE	DESCRIPTION	FEATURES	APPLICATIONS
Silicon (general purpose) (MPX2010)	0 to 10 kPa	A sensor on a chip with nozzles that measures pressure differential	Temperature compensated Good linearity Factory-calibrated Low cost	Respiratory diagnostics (medical) Air-movement control
Silicon (Atmospheric) (KP125)	40kPa to 115 kPa	A sensor on a chip top surface sensor that measures atmospheric pressure	Accurate (±1.2 kPa) Factory-calibrated	Altimeter Barometer

Other types of pressure sensors are on a larger scale, usually based on traditional pressure measurement, such as bellows or a Bourdon tube. These create a physical displacement that is then measured by one of the techniques discussed earlier, such as a potentiometer, a strain gauge, or capacitative position sensing like that used in an IC accelerometer.

6.5 Chemical

There are many sensors available for detecting chemicals of various kinds—far too many to list here. The selection shown in Fig. 6.24 illustrates a range of the sensors available, and we'll look at their principals of operation in the following sections.

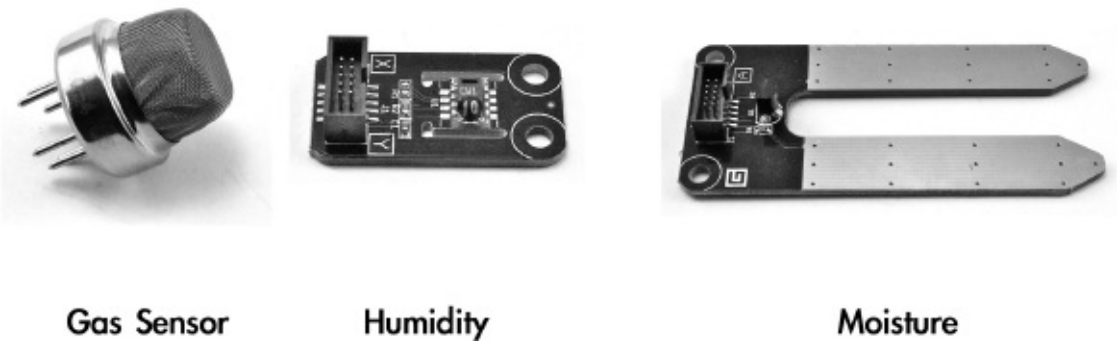


FIGURE 6.24 Chemical sensors

6.5.1 Smoke

Domestic smoke detectors are (or should be) a feature of every home. The most common devices are actually quite sophisticated in their sensing (see Fig. 6.25).

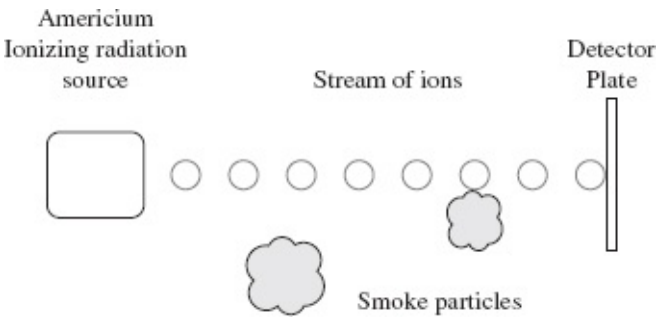


FIGURE 6.25 An ionizing smoke detector

They use a radioisotope (Americium) that generates a stream of ionized particles in a chamber that is

open to the air. This allows a small current to flow between the metal casing of the radiation source and a detecting plate at the far side of the chamber.

If smoke particles enter the chamber—even microscopic particles—they attach themselves to the ions, neutralizing them and reducing the current flowing. This is detected by the control electronics.

The other common type of smoke detector (optical) uses a focused infrared LED with a photodiode, off-axis from the infrared LED. When smoke enters the sensor, the smoke particles scatter light, which can be detected by the photodiode.

6.5.2 Gas

Gas detectors, such as the MQ-4 from Hanwei Electronics (SparkFun SKU: SEN-09404) shown in Fig. 6.24, contain a small heating element and a catalytic detector to detect concentrations of methane as low as 200 parts per million. Other types of sensors use different catalysts to make them sensitive to other gases.

Using such a device involves supplying a voltage (often 5 V) to the heating element (which typically draws a few tens of milliamps) and putting the sensor pins in a voltage divider arrangement with a fixed resistor to create a measurable output voltage.

6.5.3 Humidity

Although highly accurate techniques are available for measuring humidity, in many applications, an accuracy of 1 or 2 percent is acceptable and can be achieved with a low-cost capacitative sensor. These sensors are often combined with a temperature sensor, control electronics, and a serial interface for use by a microcontroller.

Capacitative sensing relies on humidity changing the dielectric constant of a polymer separating the two plates. These devices achieve better accuracy by laser calibration; that is, they are calibrated during manufacture by using a laser to write into digital memory on the device, setting a calibration parameter.

Unless extreme accuracy is required in a very specialized application, there is little reason to use anything other than an IC humidity sensor.

6.6 Light, Radiation, Magnetism, and Sound

Figure 6.26 shows a selection of various types of sensors associated with detecting radiation or magnetism.



FIGURE 6.26 Light, radiation, and magnetic sensors

6.6.1 Light

Light detection is used in many different ways, from optical transmission of data to proximity detection. There are many types of sensors for this purpose, including some we have already discussed, such as the photoresistor. For information about the use of photoresistors, photodiodes, and phototransistors, see Chapter 5.

6.6.2 Ionizing Radiation

To sense ionizing radiation, a Geiger–Müller tube is usually used. These tubes contain a low-pressure inert gas such as neon. A high voltage (400 to 500 V) is applied between the outer conducting cathode and a wire anode running down the middle of the tube (see Fig. 6.27).

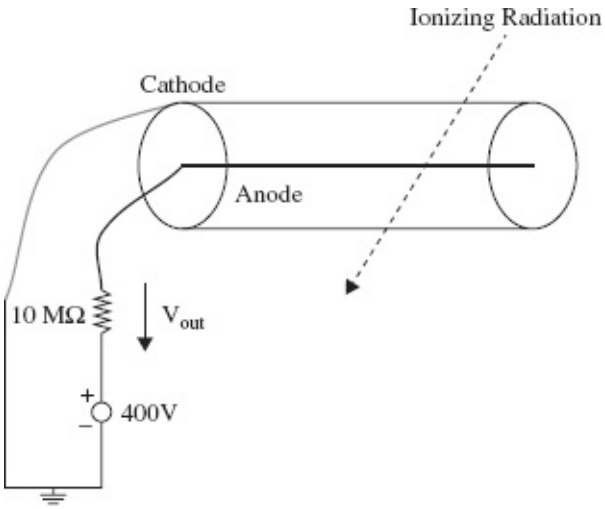


FIGURE 6.27 A Geiger–Müller tube

When ionizing radiation passes through the tube, the gas is ionized and conducts, producing a pulse at V_{out}. These pulses are counted using a Geiger counter, traditionally producing a click sound with each event.

Different tubes are designed to be more sensitive to different types of radiation. To detect alpha particles, tubes with a very thin mica end will allow the easily stopped alpha particles to pass through and be detected.

6.6.3 Magnetic Fields

In discussing sensing rotational speed in a flow meter, I suggested detecting each rotation by sensing when a magnet passed a sensor. This sensor could just be a coil of wire, in which a current would be induced. However, the slower the magnet moves past the coil, the smaller the signal available. An alternative to this is to measure the static magnetic field and just detect the presence or absence of the magnet. A common device for doing this is called a Hall effect sensor.

The Hall effect is an electrical effect that occurs when a current passes through a conductor in the presence of a magnetic field. Referring to Fig. 6.28, when current is flowing along a bar of thickness d, in the presence of a magnetic field B, a voltage V_H will appear between the top and bottom surfaces of the conductor.

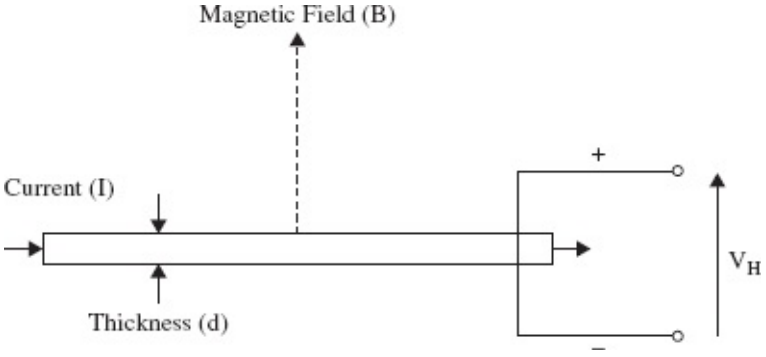


FIGURE 6.28 Hall effect sensor

The voltage from the Hall effect is given by the equation:

$$V_H = \frac{-IB}{ned}$$

where n is the charge carrier density, e is the charge of an electron, d is the length of the conductor, I is the current flowing through it, and B is the magnetic field density.

Practical Hall effect sensors usually combine the sensor itself with high-gain amplification in a single IC. Others take this a step further and actually incorporate a serial interface to send the data to a microcontroller.

There are two types of Hall effect sensors. The most common type also acts as a switch and just provides an on-off indication of the presence of a nearby magnet. A more flexible type of Hall effect sensor (a linear sensor) provides a linear output voltage proportional to the strength of the magnetic field.

6.6.4 Sound

Sensing sound relies on using a microphone and amplifier (see Chap. 15). If you are sensing the sound level, then you will also need to use a low-pass filter and rectification to determine an indication of the level of sound (see Fig. 6.29).

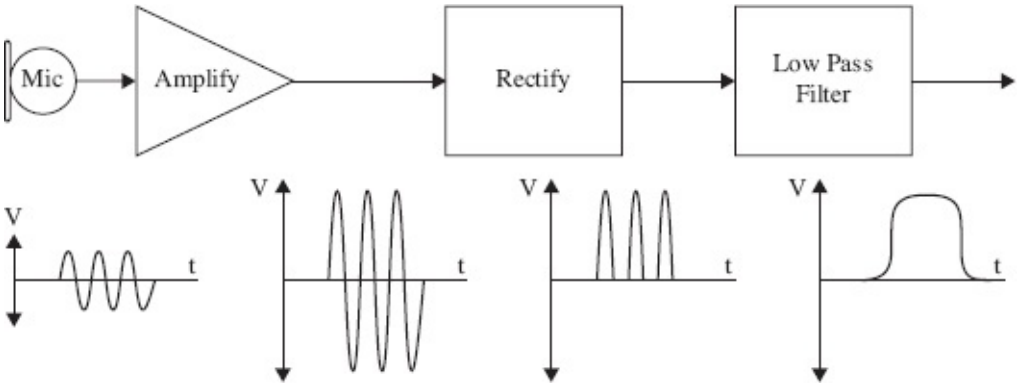


FIGURE 6.29 Sensing sound level

6.7 GPS

A side effect of consumer devices, such as smartphones and dedicated satellite navigation systems, is that GPS modules are readily available at relatively low cost.

GPS relies on a constellation of satellites (see Fig. 6.30). Each satellite contains a highly accurate clock that is synchronized with all the other satellites in the constellation. The satellites then broadcast this time signal.

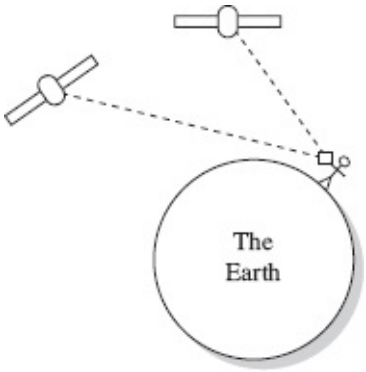


FIGURE 6.30 Global positioning system

A GPS receiver on the ground will attempt to receive time signals from as many satellites as possible and use the differences in those times, due to the time for the signals to travel, to calculate the position of the receiver.

GPS modules are generally a single IC, but require a special antenna that is usually the largest part of the module. Figure 6.31 shows the Venus 638 FLPx GPS module from SparkFun with its antenna attached. This module and modules like it have a serial interface that transmits a stream of updates giving the position of the receiver and various other information, such as the time and number of satellites currently being received.

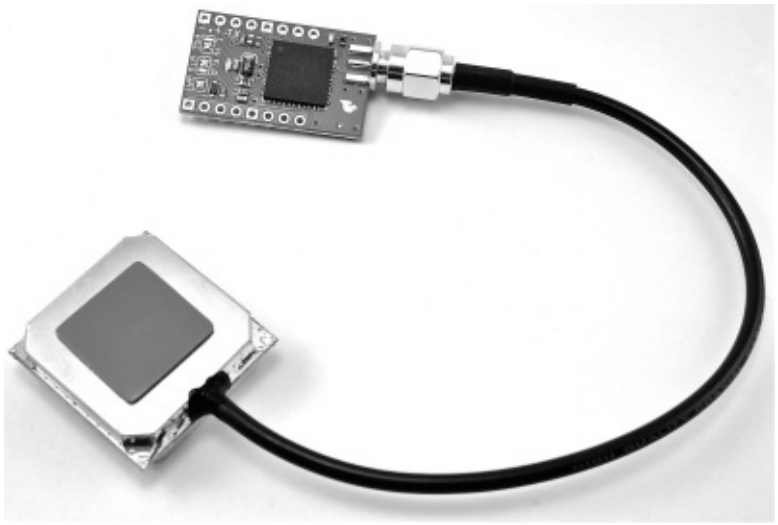


FIGURE 6.31 The SparkFun Venus638FLPx GPS module

Hands-on Electronics

7.1 Safety

7.1.1 Lecture on Safety

Probably the most hazardous thing in electronics is your household line voltage, around 120 V at 60 Hz in the United States (240 V at 50 Hz in many other countries). What often happens is an ungrounded metal object comes in contact with a hot wire, and then you touch the "hot object," discovering in the process that your body acts as a conducting medium for current to flow to ground. (The ground and ground wires are, in effect, at the same potential as the neutral wire—they are tied together at the utility box.) The frequency and amplitude of the line voltage tend to be at the perfect levels for inducing involuntary muscle contractions that prevent you from letting go of a hot or energized object. This freezing effect is extremely dangerous. The longer you're stuck, the more damage current will inflict upon your internal tissues (tissue heating), and the more likely you'll suffer from cardiac and/or respiratory arrest.

Voltage and frequency levels below household line levels can cause cardiac/respiratory arrest, too (especially if your internal resistance is low, such as with a wet body), but are less likely to cause the dreaded freezing effect. Frequencies above household line levels actually are less likely to cause severe muscle contractions, so the chances of fatality decrease somewhat. At very high voltages, there is an arcing effect that can occur, which can give you a nasty jolt even if you're not touching an object directly. The electric potential between you and the energized object is so great that the air becomes a conducting medium. Ironically, the initial jolt you receive from an arc will often be large enough to throw you clear of danger. However, besides being extremely painful, such jolts may induce cardiac/respiratory arrest, and depending on where you're standing, the fall may be what kills you.

In more specific terms, what kills you or what causes greatest tissue damage is the amperage. When you come in contact with a live wire or energized object, the amount of current that passes through your body to ground depends on the voltage level (assuming an ideal voltage source) and your internal resistance. A heavily callused dry palm may present a resistance of 1 MΩ, while a thin, wet palm may present only 100 Ω. Resistance tends to be lower in children, and different body tissues exhibit a range of resistances. Nerves, arteries, and muscles are low in resistance. Bone, fat, and tendon are relatively high in resistance. Across the chest of an average adult, the resistance is about 70 to 100 Ω. The low resistance across the chest and the higher resistance through the air-filled lungs provide a path of least resistance through the heart and spinal cord regions—the critical regions for life support. A 100- to 1000-mA current is sufficient to induce cardiac/respiratory arrest. Thermal burns due to I^2R losses through body parts can also be significant, resulting in loss of life or limb, even long after the initial exposure. Since the thermal heating effects increase with the square of the current, higher current levels truly cause nastier burns, whether they are internal or external.

As a rough guideline, a 10-mA, 50- to 60-Hz cycle line current, from hand to foot, will merely give you a tingly sensation. However, above 10 mA, at the same frequency, it's possible you'll freeze to the circuit

you're touching. Outlets using ground fault circuit interrupters (GFCIs) will cut power if they sense sudden ground current flow in this current range—the tripping point around 5 to 10 mA. Currents from 20 to 100 mA may be fatal, but it's the 100-mA to 1-A range that's the deadliest. Above 1 A, the heart is thrown into a single contraction, and internal heating becomes significant. You may be thrown clear of the power source, but then cardiac and/or respiratory arrest may ensue.

The most lethal form of electrocution occurs via hand-to-hand current flow; current passes right across the heart, lungs, and spinal cord. Hand-to-foot electrocution is less fatal, carrying a mortality rate of around 20 percent. For this reason, when working on line voltages, make sure to keep one hand in your pocket. Another safety practice is to use the back of your hand when touching unknown entities. This prevents the "grip of death."

Consumer electronics equipment—such as TVs, computer monitors, microwave ovens, and electronic flash units—use voltage at power levels that are potentially lethal. Normally, the hazardous circuitry is safely enclosed to prevent accidental contact. However, during servicing, the cabinet or enclosure will likely be open and safety interlocks may be defeated. Depending on the type of circuitry and your general state of health, there is a wide variation of voltage, current, and total energy that can kill.

Microwave ovens are probably the most dangerous household appliance to mess around with. They operate at several thousand volts (5000 V or more), at higher current levels (more than an amp may be available momentarily). This is instantly a lethal combination.

Old-fashioned cathode-ray tube TVs and monitors may have 35 kV on the cathode-ray tube, with current lower than a couple of milliamps. However, the cathode-ray tube capacitance can hold a painful charge for quite some time. In addition, portions of the circuitry within TVs and monitors, as well as all other devices that plug into the wall socket (such as switching power supplies), are line connected. These circuits' internal grounds may be several hundred volts higher than the utility earth ground. Without isolation between the 120 V and the circuit, a real potential shock hazard exists if you come in contact with the circuit's floating ground (see the discussion of isolation transformers in Section 7.5.12). Switch-mode power supplies, electronic flash units, and strobe lights have large energy storage capacitors that can deliver a lethal discharge, long after the power has been removed. This even applies to disposable cameras with a flash. Even portions of apparently harmless devices like VCRs and CD players, or vacuum cleaners and toasters, can be hazardous (though the live parts may be insulated or protected, you shouldn't count on it).

The following list of safety tips will help you prolong your life.

ELECTRONICS SAFETY TIPS

1. Do not attempt repairs on line-powered circuits with the power on. Always remove the main power first.
2. Use one hand when taking measurements, and keep your other hand at your side or in your pocket. If you do get shocked, it's less likely that current will pass through your heart.
3. If you need to probe, solder, or otherwise touch circuits with the power off, discharge (across) a large power supply filter or an energy-storage capacitor with a 2-W or greater resistor of 100 to 500 Ω/V; for example, for a 200-V capacitor, use a 20- to 100-KΩ resistor. (A screwdriver's metal tip is often used to discharge a capacitor, but the sudden discharge may not be well tolerated by the capacitor.) Monitor while discharging and verify that there is no residual charge with a suitable voltmeter. Large capacitors can store a lethal amount of charge and may retain the charge for a number of days. Even capacitors rated at voltages as low as 5 or 10 V can be dangerous.
4. Perform as many tests as possible with the power off and the equipment unplugged. For

example, the semiconductors in the power supply section can be tested for short circuits with an ohmmeter.

5. Avoid standing in a position that could be dangerous if you were to lose muscle control due to shock. Often the fall alone is more dangerous than the initial shock.
6. When working on high-power circuits, bring someone else along who can assist you if something goes wrong. If you see that someone cannot let go of a "hot" object, do not grab onto that person. Instead, use a stick or insulated object to push the person away from the source. It's not a bad idea to learn CPR, too.
7. All high-voltage test instruments (such as power supplies, signal generators, and oscilloscopes) that are operated on a 120 or 240 V ac supply should use three-wire line cable.
8. Use only shielded (insulated) lead probes when testing circuits. Never allow your fingers to slip down to the metal probe tip when testing a "hot" circuit. Also, make sure to remove power from a circuit when making a connection with wires and cables.
9. Connect/disconnect any test leads with the equipment unpowered or unplugged. Use clip leads or solder temporary wires to reach cramped or difficult-to-access locations. If you must probe live, put electrical tape over all but the last 1/16 of the test probe to avoid the possibility of an accidental short that could cause damage to various components. Clip the reference end of the meter or scope to the appropriate ground return so that you need to probe with only one hand.
10. If circuit boards need to be removed from their mounting, put insulating material between the boards and anything they may short to. Hold them in place with string or electrical tape. Prop them up with insulation sticks made of plastic or wood.
11. Set up your work area away from possible grounds that you may accidentally contact.
12. When working on ac-line circuits, wearing rubber-bottomed shoes or sneakers, or standing on a sheet of rubber or wood, can reduce possible shocks.
13. The use of a GFCI-protected outlet is a good idea, but it will not protect you from shock from many points in a line-connected device, such as a TV, monitor, or the high-voltage side of a microwave oven. (Note that a GFCI may be a nuisance by tripping at power-on or at other random times due to leakage paths, like your scope probe ground, or due to the highly capacitive or inductive input characteristics of line-powered equipment.)
14. A fuse or circuit breaker is too slow and insensitive to provide any protection for you or, in many cases, your equipment. However, these devices may save your scope probe ground wire should you accidentally connect it to a live chassis.
15. Know your equipment. TVs and monitors may use parts of the metal chassis as the ground return, yet the chassis may be electrically live with respect to the earth ground of the ac line. Microwave ovens use the chassis as the ground return for the high voltage. In addition, do not assume that the chassis is a suitable ground for your test equipment! Use an isolation transformer if there is any chance of contacting line-connected circuits. A Variac is not isolated, so you need a combination of a Variac and an isolation transformer for safety.
16. Make sure all components that are connected to ac power lines meet required power ratings.
17. When building power supplies and other instruments, make sure that all wires and components are enclosed in a metal box or insulated plastic enclosure. If you use a metal box, it is important that you ground the conductive shell (attach a wire from the inner surface of the box to the ground wire of the power cable, preferably with a screw along with solder). Grounding a metal box eliminates shock incurred as a result of a hot wire coming loose and

falling onto the box, thereby making the whole box hot.

18. When making holes in metal boxes through which to insert an ac line cable, place a rubber grommet around the inner edge of the hole to eliminate the chance of fraying the cable.
19. Wear eye protection: large plastic-lens eyeglasses or safety goggles.
20. Don't wear any jewelry or other articles that could accidentally contact circuitry and conduct current.

7.1.2 Damaging Components with Electrostatic Discharge

Scuffling across a carpet while wearing sneakers on a dry day can result in a transfer of electrons from the carpet to your body. In such a case, it is entirely possible that you will assume a potential of 1000 V relative to ground. Handling a polyethylene bag can result in static voltages of 300 V or more, whereas combing your hair can result in voltages as high as 2500 V. The drier the conditions (lower the humidity), the greater the chance is for these large voltages to form. Now, the amount of electrostatic discharge (ESD) that can result from an electrostatically charged body coming in contact with a grounded object is not of much concern in terms of human standards. However, the situation is entirely different when subjecting certain types of semiconductor devices to similar discharges.

Devices that are particularly vulnerable to damage include field-effect transistors, such as MOSFETs and JFETs. For example, a MOSFET, with its delicate gate-channel oxide insulator, can be destroyed easily if an electrostatically charged individual touches its gate. The gate-channel breakdown voltage will be exceeded, and a hole will be blown through the insulator, which will destroy the transistor. Here is a rundown of the vulnerable devices out there:

- *Extremely vulnerable:* MOS transistors, MOS ICs, JFETs, laser diodes, microwave transistors, and metal film resistors
- *Moderately vulnerable:* CMOS ICs, LS TTL ICs, Schottky TTL ICs, Schottky diodes, and linear ICs
- *Somewhat vulnerable:* TTL ICs, small signal diodes and transistors, and piezoelectric crystals
- *Not vulnerable:* Capacitors, carbon-composite resistors, inductors, and many other analog devices

7.1.3 Component Handling Precautions

Devices that are highly vulnerable to damage are often marked with "Caution, components subject to damage by static electricity." If you see a label like this, use the following precautions:

- Store components in their original packages, in electrically conductive containers (such as a metal sheet or aluminum foil), or in conductive foam packages.
- Do not touch leads of ESD-sensitive components.
- Discharge the static electricity on your body before touching components by touching a grounded metal surface such as a water pipe or large appliance.
- Never allow your clothing to make contact with components.
- Ground tabletops and soldering irons (or use a battery-powered soldering iron). You also should ground yourself with a conductive wrist guard that's connected with a wire to ground.

Never install or remove an ESD-sensitive component into or from a circuit when power is applied. Once the component is installed, the chances for damage are greatly reduced.

While these risks are real, and you will no doubt occasionally lose a component to static, it is also important to keep a sense of perspective and take the necessary precautions around devices that you know are expensive or vulnerable. You certainly do not need to wear a grounded wrist strap all the time.

7.2 Constructing Circuits

This section briefly discusses the steps involved in building a working circuit: drawing a circuit schematic, building a prototype, making a permanent circuit, finding a suitable enclosure for the circuit, and applying a sequence of troubleshooting steps to fix improperly functioning circuits.

7.2.1 Drawing a Circuit Schematic

The *circuit schematic*, or *circuit diagram*, is a blueprint of a circuit. For a schematic to be effective, it must include all the information necessary so that you or anyone else reading it can figure out what parts to buy, how to assemble the parts, and possibly what kind of output behavior to expect. The following are guidelines for making an easy-to-read, unambiguous schematic:

- The standard convention used when drawing a schematic is to place inputs on the left, outputs on the right, positive supply terminals on the top, and negative supply or ground terminals on the bottom of the drawing.
- Keep functional groups—such as amplifiers, input stages, filters, and so on—separated within the schematic. This will make it easier to isolate problems during the testing phase.
- Give all circuit components symbol designations (such as R_1, C_3, Q_1, and *IC4*), and provide the exact size or type of component information (such as 100 k, 0.1 µF, 2N2222, or 741). It is also important to include the power rating for certain devices, such as resistors, capacitors, relays, speakers, and so on.
- Use abbreviations for large-valued components (for example, 100 kΩ instead of 100,000 Ω, or 100 pF instead of 100×10^{-12} F). Common unit prefixes include p = 10^{-12}, n = 10^{-9}, µ = 10^{-6}, k = 10^{3}, and M = 10^{6}.
- When labeling ICs, place lead designations (such as pin numbers) on the outside of the device symbol, and place the name of the device on the inside.
- In certain circuits where the exact shape of a waveform is of importance (such as logic circuits and inverting circuits), it is helpful to place a picture of the expected waveforms at locations of interest on your circuit diagram. This will help isolate problems later during the testing phase.
- Power-supply connections to op amps and digital ICs are usually assumed and typically left out of the schematic. However, if you anticipate confusion later on, include these supply voltages in your drawing.
- To indicate joined wires, place a dot at the junction point. Unjoined wires simply cross (do not include a dot in such cases).
- Include a title area near the bottom of the page that contains the circuit name, the designer's name, and the date. It is also useful to leave room for a list of revisions.

Figure 7.1 shows a sample schematic.

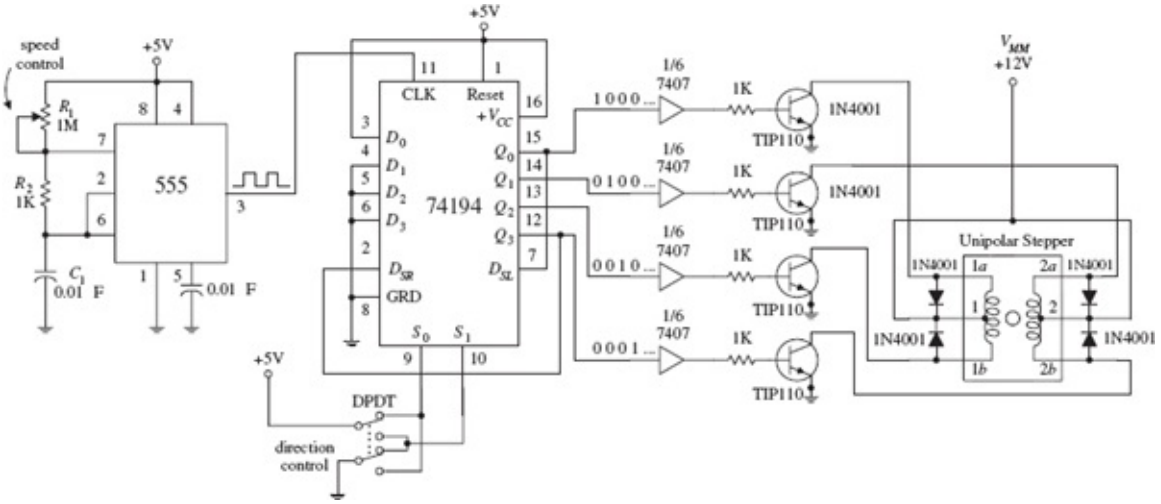


FIGURE 7.1

Once you have completed the circuit schematic, see if anything looks fishy. Are there missing connections or missing component values? Are component polarities indicated? Have you considered the power ratings of the components? Have you made connections as simple as possible? It is better to check things now; finding an error when you are soldering things together is much more annoying than erasing a few lines on a drawing.

You can draw the schematic with a pencil on paper, and often this is good enough while you are developing the project, especially if it is a simple one. However, at some stage, you will probably want to draw the circuit using an electronic computer-aided design (CAD) tool. The advantages of using a CAD tool are as follows:

- When you modify your schematic, the diagram stays clear in a way that a schematic with a lot of erasing doesn't.
- A good CAD system will apply electrical design rules to your schematic—at the very least, warning you about unconnected leads.
- A CAD system provides two views of a project: the schematic and the printed circuit board (PCB) layout. You can usually automatically route and lay out a PCB from the schematic.
- To be able to lay out a PCB, the CAD package needs to know about the geometry of the component. A catalog of components is supplied that contains information about the component package and pin assignments. Manufacturers will often provide component libraries for CAD packages.

There are many electronic CAD packages, including some that are very expensive. One of the most popular is EAGLE CAD. A free version of this product is available that restricts use to relatively small PCBs and a maximum of two layers of copper. Options for CAD packages are listed in Section 7.5.22. Most require a reasonable investment in time to learn to use properly.

If a full-blown CAD system is more than you need (or are prepared to learn), an alternative is to use a general-purpose drawing tool such as Microsoft Visio or OmniGraffle, both of which have libraries of electronics component symbols available.

7.2.2 A Note on Circuit Simulator Programs

Now, before you build a circuit, or even before you finish the schematic, you might consider using a circuit simulator program to test your idea to see if it works. Circuit simulator programs allow you to construct a computer model of your circuit and then test it (measure voltages, currents, wave patterns, logic states, and so on) without ever needing to touch a real component. A typical simulator program

contains a library of analog and digital devices, both discrete and integrated in form. If you wish to model an oscillator circuit—one that is built from a few bipolar transistors, some resistors, a capacitor, and a dc power supply—all you do is select the parts from the library, set the values of these parts, and then arrange the parts to form an oscillator circuit.

To test the circuit, simply choose one of the simulator's test instruments, and then attach the test instrument's probes to the desired test points within the circuit. For example, if you are interested in what the output waveform of the oscillator looks like, choose the simulator's oscilloscope, and then attach the test leads and measure the output. The computer screen will then display a voltage versus time graph of the output. Some other instruments found in simulator programs include multimeters, logic analyzers, function generators, and bode plotters.

Why make a computer simulation of a circuit before building a real one? For several reasons:

- When you make a computer simulation, you do not need to worry about working with faulty components.
- There is no need to worry about destroying components with excessive current; the computer program does not "care about the numbers."
- A simulator program does all the mathematical work for you. The simulator allows you to fiddle around with component values until the circuit is working as desired.
- Using a simulator program can make learning electronics an intuitive process and saves time spent at the workbench.

Some popular simulator programs include Electronics Workbench, CircuitMaker, and MicroSim/Pspice. Electronics Workbench and CircuitMaker are relatively easy to use, while Pspice is a bit more technical.

7.2.3 Making a Prototype of Your Circuit

Once you are satisfied with the schematic, the next step is to make a prototype of your circuit. The most common tool used during the prototype phase is a solderless modular breadboard. A breadboard acts as a temporary assembly board on which electrical parts such as resistors, transistors, and ICs are placed and joined together by wires or built-in conductive pathways hidden underneath the surface of the breadboard (see Fig. 7.2).

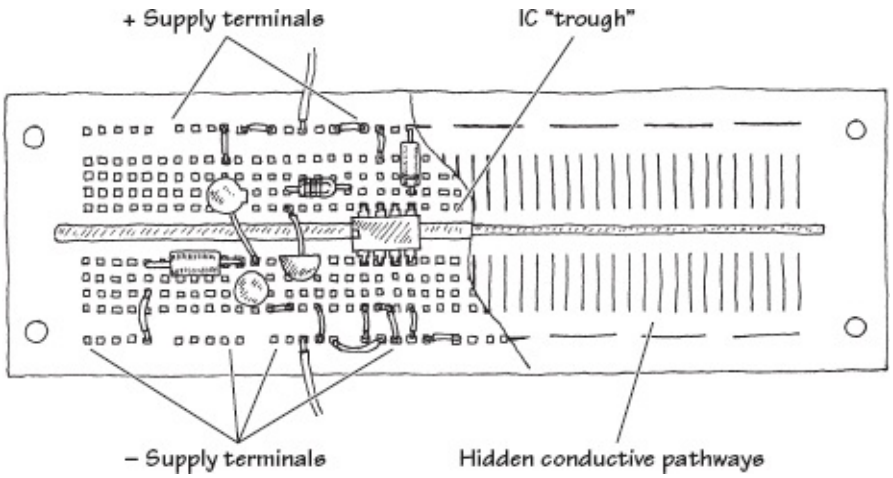


FIGURE 7.2

Breadboards come with an array of small square sockets spaced about 0.100 in from center to center. When a wire or component lead is inserted into one of these sockets, a spring-like metal sleeve built into

the socket acts to hold the wire or lead in place. Breadboard sockets are designed to accept 22-gauge wire but can expand to fit wire diameters between 0.015 and 0.032 in (0.38 and 0.81 mm). The upper and lower rows of sockets of a breadboard typically are reserved for power-supply connections, while the sockets between the central gap region are reserved for DIP ICs.

See Section 7.5.17 for more on prototyping boards.

7.2.4 The Final Circuit

Once you are finished making a successful prototype, the next step is to construct a more permanent circuit. At this point, you must choose the type of mounting board on which to place your circuit. Your choices include a perforated board, a wire-wrap board, a pre-etched board, or a custom-etched PCB. Let's take a closer look at each of these boards.

Perforated Board

A *perforated board* is an insulated board with an array of holes drilled into it (see Fig. 7.3). To join a lead from one electrical component to another, each of the components' leads is placed through neighboring holes. The lead ends sticking out the back side of the board are then twisted together (and possibly soldered).

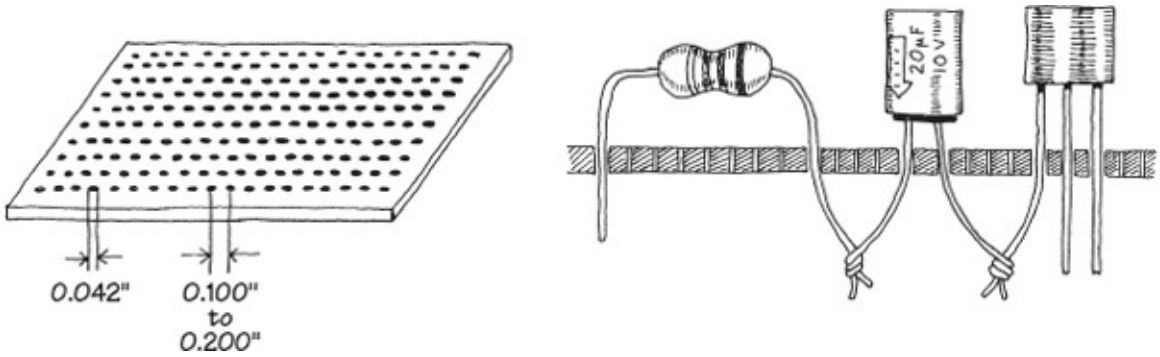


FIGURE 7.3

Constructing a circuit on a perforated board is easy. Few supplies are needed, and making connections does not require much skill. However, what you get in the end is a fairly bulky circuit that is liable to fall apart over time and may pick up noise inadvertently (jumper wires will act like little antennas). In general, perforated boards are used for constructing simple, noncritical types of circuits.

Wire-Wrap

Using a *wire-wrap board* is another way to assemble moderately complex circuits containing ICs. Every wire-wrap board is made up of an array of sockets, each of which has a corresponding pin-like extension sticking out of the opposite side of the board (see Fig. 7.4).

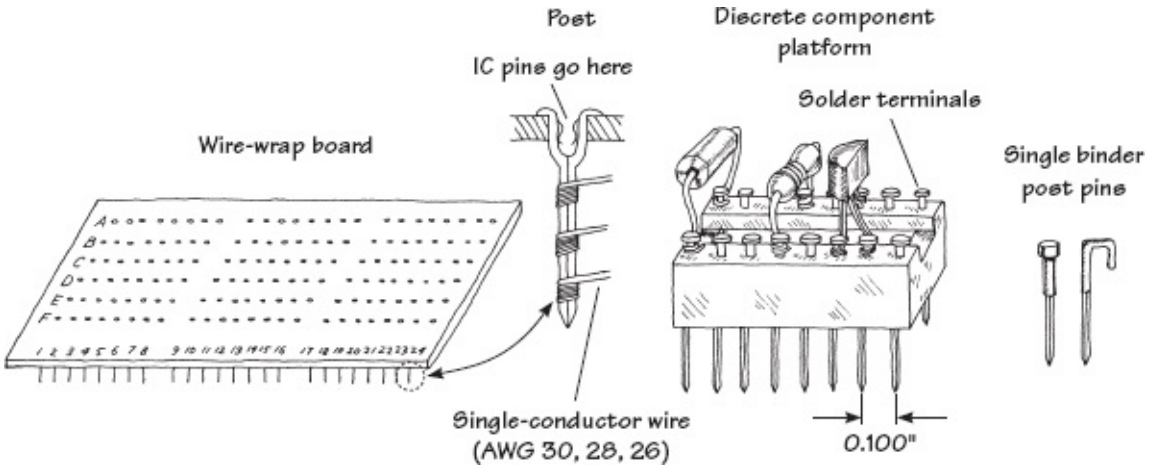


FIGURE 7.4

IC leads are inserted directly into the wire-wrap's sockets, while discrete components, such as resistors, capacitors, and transistors, must be mounted on special platforms like blocks or single post-like pins (see Fig. 7.4b). Each of these platforms contains a number of nail-like heads on which discrete component leads are attached, either by coiling the leads around the nail head or by fastening the leads to the nail head with solder. The nail-like tips of the platforms are then inserted into the board's sockets.

To connect components together, the pins on the back side of the wire-wrap board are joined together with wires (typically 30-, 28-, or 26-gauge single-conductor wires). In order to fasten the joining wires to the pins, a special wire-wrap tool is used (see Fig. 7.5). This tool wraps the joining wire around the pin by means of a bit section that rotates around a hollow core. The wire is inserted into the bit, the hollow core is placed over the pin, and then the wire-wrap tool is twisted several times (usually around seven times) and removed.

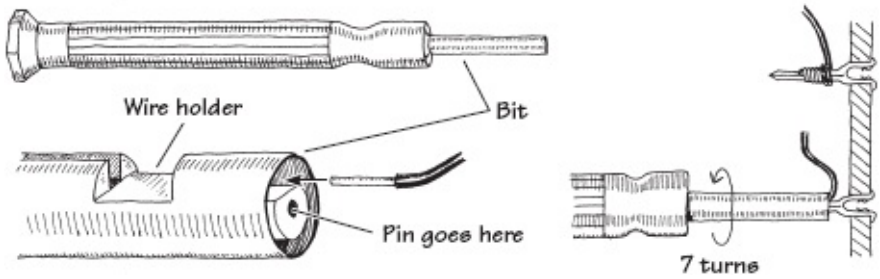


FIGURE 7.5

In practice, to save time and avoid making mistakes, it is desirable to do all the wrapping in a single pass. To follow this approach, some bookkeeping is required. Notice in Fig. 7.4a that each socket/pin is given a row/column designation. For example, a pin located three rows down and two rows in from the left is given a C2 designation, and a pin that is located five rows down and seven rows in is given an E7 designation.

To figure out how to arrange electrical components on the board, making a simple sketch of the wire-wrap board is helpful. On the sketch, draw in all the components, fixing each component lead to a specific row/column coordinate. Once the sketch is complete, simply grab your wire-wrap tool and start making wire connections between the pins, using the sketch as a guide.

Wire-wrap boards are suitable platforms for circuits that contain a number of ICs, such as logic circuits. However, because the sockets of these boards are not designed for linear component leads (you must use platforms in such cases), it may be easier to use a pre-etched, or custom-etched PCB for building analog circuits.

Pre-etched Perforated Boards

A *pre-etched perforated board* is made of an insulating material that is coated with a pre-etched copper pattern and has a number of holes drilled into it. To join electrical components, simply place the leads of the components in the appropriate holes that are joined by a copper-etched strip, and then apply solder.

Pre-etched boards come with a variety of different etched patterns. Figure 7.6 shows samples of the kinds of patterns available.

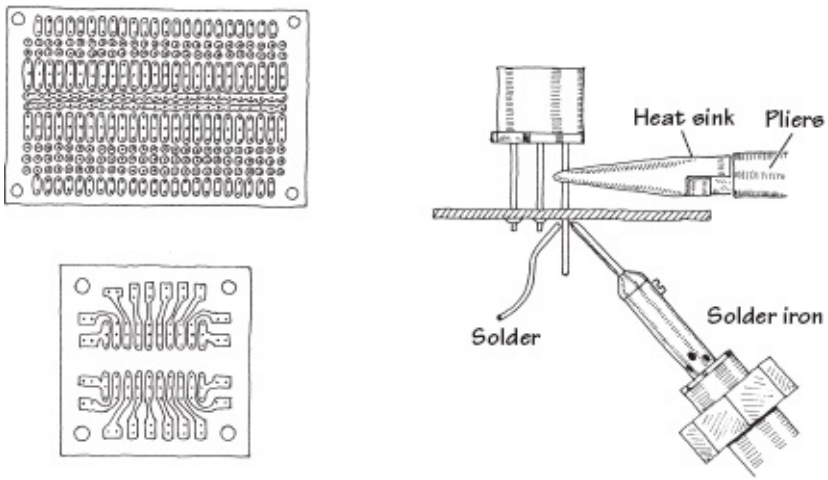


FIGURE 7.6

Custom-Etched Boards (PCBs)

If you are looking to build a circuit with a professional appearance, designing a *custom-etched circuit board*, or *PCB*, is the ticket, and the next section describes how to do this.

Designing custom-etched circuit boards takes a bit longer, but the time spent is often well worth the effort. There are times when making a custom-etched board is essential to ensure proper circuit operation, especially when dealing with circuits that contain components whose properties are greatly influenced by the length of the leads. For example, very high-speed logic circuits require the unique microstrip line geometries and precise placement of components to achieve fast rise times and at the same time avoid crosstalk among circuit elements. Sensitive low-level amplifier circuits also benefit from well-placed microstrip interconnections; the short and direct interconnections help eliminate noise pickup.

7.2.5 Making a PCB

To create the PCB layout, you can either draw the design by hand or use a CAD tool. Once you have the design, you have a number of choices for actually creating the board:

- Using an etch-resistant pen
- Printing the layout onto overhead projector (OHP) transparency film and using photosensitive board
- Using laser printer toner transfer
- Using CAD and a desktop router to cut out the pattern on the copper-clad board
- Sending the Gerber design files away to a PCB shop

Each of these approaches has pros and cons. The last option has started to become cost-effective, with prices as low as a dollar a board for a run of ten boards, and will produce professional-quality results. However, you do need to wait, sometimes for a week or two.

We will look at each of these approaches in turn.

Etch-Resistant Pen

Custom etching involves using graphic and chemical techniques to convert a copper-covered board into a

custom-etched one. By doing your own custom etching, you can construct highly reliable, tightly compact circuits that require few jumper wires (see Fig. 7.7).

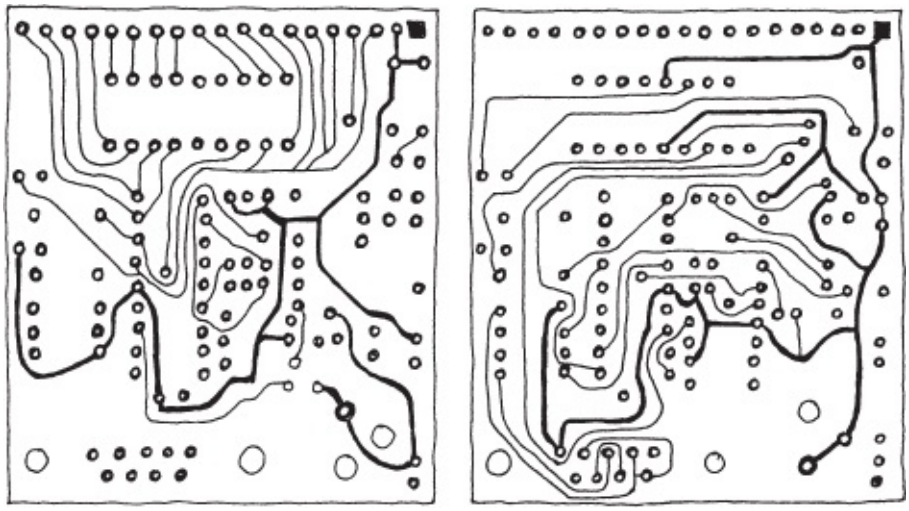


FIGURE 7.7

To design a custom-etched PCB, you first need an insulated board (usually 1/16 in. thick and made from a fire-resistant epoxy-bonded fiberglass) that is completely covered on one or both sides with a very thin copper coating. Next, you must transform your circuit schematic into a PCB layout. This involves rearranging components in such ways as to make all conductive pathways short and direct. The layout should also eliminate any wire crossing, if possible.

Once you feel that your hardwired sketch is complete, the next step is to transfer it onto the copper-coated board. Afterward, the trick is to etch out all the undesired copper-coated sections while leaving the conductive pathways intact.

At this point, there are a number of different transferring/etching techniques from which you can choose. Perhaps the simplest technique involves using a PCB kit, which you can buy for a few bucks from a store such as RadioShack. A typical kit comes with a single or dual copper-coated board, a bottle of etching solvent, a permanent marker, a bottle of rubbing alcohol, and a drill bit.

To make your custom-etched board, you first transfer your hardwired sketch onto the surface of the board with, say, a pencil. Next, you drill in the appropriate holes where component leads are to go. Now, with the etch-resistant pen, you trace over the pencil sketch, making sure to encircle the drilled-out holes. After that, you place the board in a tub of etching solvent (typically ferric chloride) and wait until the copper dissolves away from the sections of the board that are not coated with magic-marker ink (the ink does not dissolve in the solvent; it acts to protect the underlying copper). After the board is removed from the solvent bath, it is washed off with water, and the magic-marker ink is then removed with a rag doused in rubbing alcohol.

Using a PCB kit is great for simple, single-run productions. PCB kits are easy to use, inexpensive, and require practically no special equipment other than what is provided in the kit itself. However, one problem with these kits is that you can construct only one circuit board at a time. Another problem with these kits is the limited precision you get by using a magic marker to create conductive pathways. If you are interested in making multiple copies of a circuit, and if you are looking for greater line precision, a more sophisticated technique that involves photochemical processes is required.

Photo-etching with OHP Transparency Film

Only the simplest of designs are really suited to hand-drawn PCB design. If you invest the time to learn a CAD system, it is unlikely that you will go back to the manual approach.

A great advantage of having your PCB design in a CAD system is that you can print it. For this

technique, you print it onto transparent film, of the type that used to be used on overhead projectors. This film is available in letter and A4 sizes and comes in two types: one for laser printers and one for inkjet printers. Don't put the inkjet film into a laser printer (it is likely to melt onto the drum).

Using EAGLE PCB, as shown in Fig. 7.8, you can see the schematic (a), the PCB with all the component positions and labels (b), and, finally, just the copper (c).

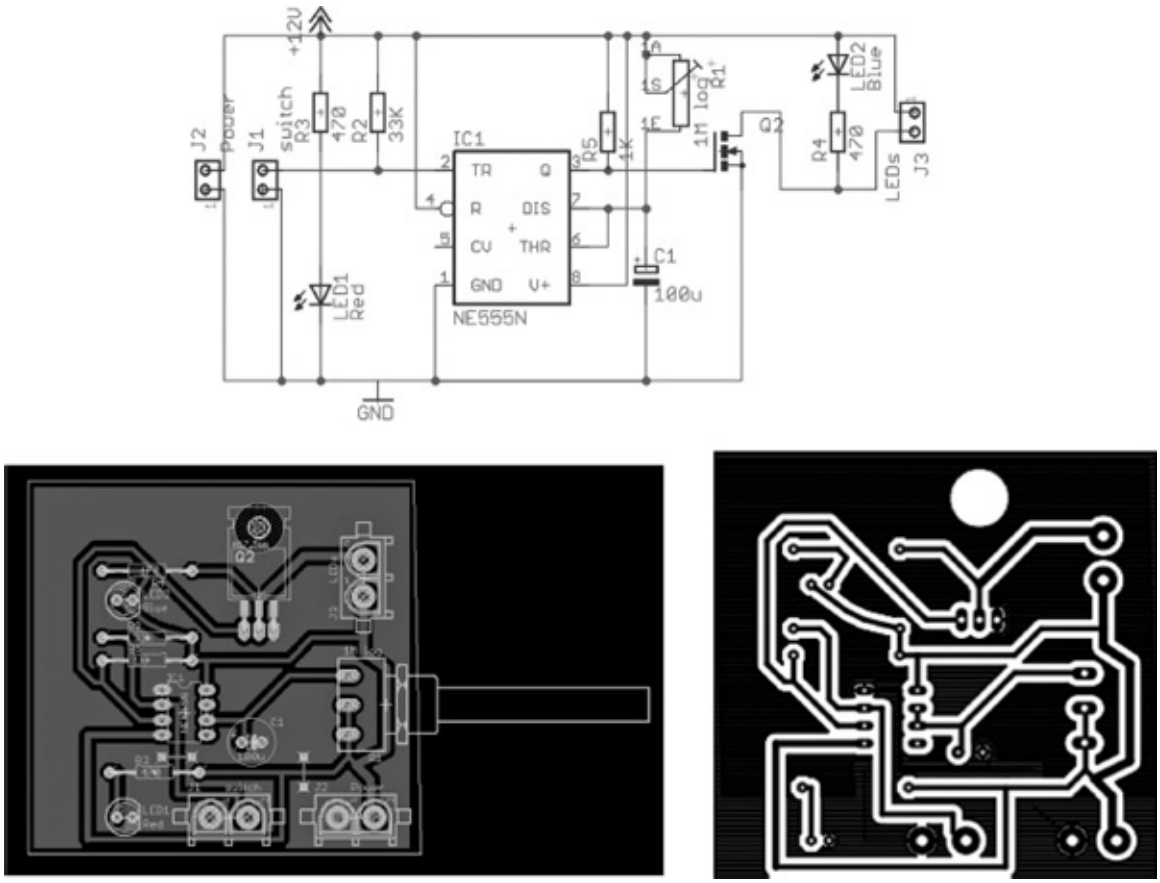


FIGURE 7.8 EAGLE PCB, schematic and copper layer

The process followed here is to draw the schematic and then switch to the board view. Here, you drag the components onto the board, leaving airwires connecting the components according to the schematic. You can then load a set of design rules and click a button, and a PCB layout will be generated automatically. You will probably want to intervene, and rip up some of the wires and route them your own way, but generally, the automated layout isn't a bad start.

Once the design is printed onto the film, it is placed on top of presensitized copper-clad board. This board is PCB with a photosensitive layer that hardens when exposed to UV light. This is available from most hobby electronic stores. While short exposure to room lighting will do it no harm, you should keep it in a light-proof container.

The board-film sandwich is often held in a frame (a clip photo frame will do) while an exposure is made using a UV light box. The board is then put into a tray of developer, and the image of the PCB tracks will become visible on the board, just like an old-fashioned photograph being developed.

Next, the board is etched in a chemical that dissolves the copper except where it is protected by the photographic image of the PCB tracks. This part of the process is the same as for a pen-drawn PCB.

Figure 7.9 shows a homemade photo-etching kit that uses an array of UV LEDs inside an index card box to provide the UV exposure. Commercial kits are available, of course.



FIGURE 7.9 A photo-etching kit

Note that the black areas will end up as copper. Quite large areas of black can be seen on the film. This reduces the quantity of copper that needs to be dissolved and hence extends the life of the chemical. The large areas of copper are generally connected to GND and are termed a *ground plane*.

Laser Printer Toner Transfer

A variation on the transparency film option is to print the PCB layout onto special paper using a laser printer, and then use a domestic iron to transfer the toner onto the copper-clad board. In this case, the image must be printed reversed.

With practice, good results can be achieved with this technique, which has the advantage of not requiring special tools. There are a number of Internet videos and tutorials describing this approach.

Creating a PCB with a Router

Low-cost desktop CNC routers offer a chemical-free method of producing PCBs by using normal copper-clad PCB, but then using a computer-controlled CNC router to cut away the unwanted copper (see Fig. 7.10).

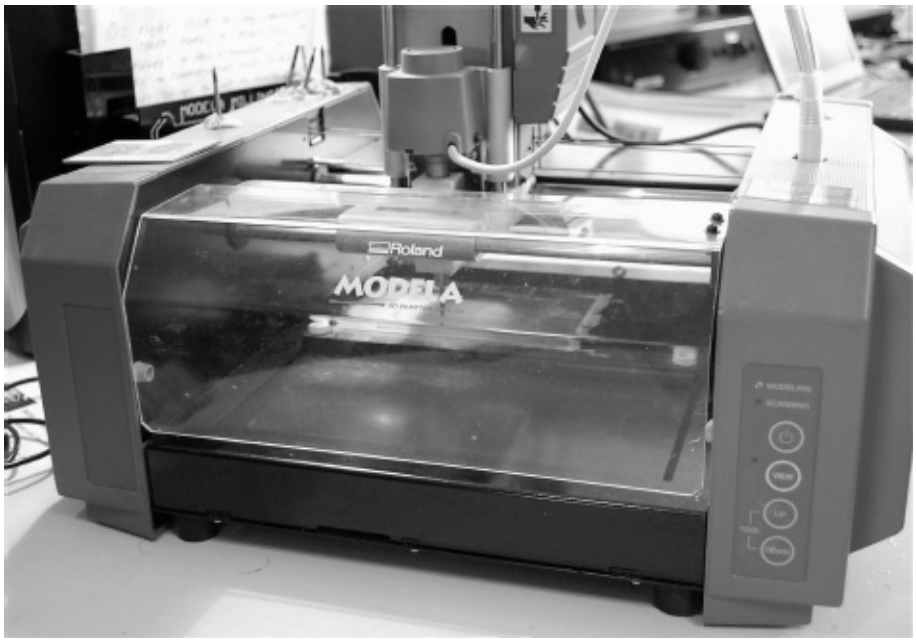


FIGURE 7.10 A CNC router cutting a PCB

The process is similar to the photo-etching method. Once the PCB artwork is done, the copper layer is dispatched to the router as if it were a printer.

Using a PCB Service

It used to be that designing a PCB was a very specialized skill, and in some ways, it still is. For very complex projects heading toward quantity production, the services of a specialist PCB design company are worth considering.

However, programs like EAGLE PCB will take a lot of the difficulty away, both by automating the track layout on the PCB and enforcing design rules such as minimum track widths, separations, and so on. This makes it entirely feasible to design your own PCBs and submit the finished design files to a PCB service. This service is likely to be almost entirely automated, so you will need to make sure your design is right before you send it, because these services will not check your design.

A PCB service will produce a great-looking result. All of these services generally allow a two-layer board, which provides a top and bottom copper layer. What is great is that they also do the following:

- Create the vias that link a track on one layer to one beneath it.
- Provide a silk-screen layer that can label the components and mark their positions.
- Provide a mask layer that covers all the copper not actually intended to be soldered.

Figure 7.11 shows a typical PCB produced by such a service. The cost for ten of these boards can be as low as $10 plus shipping.

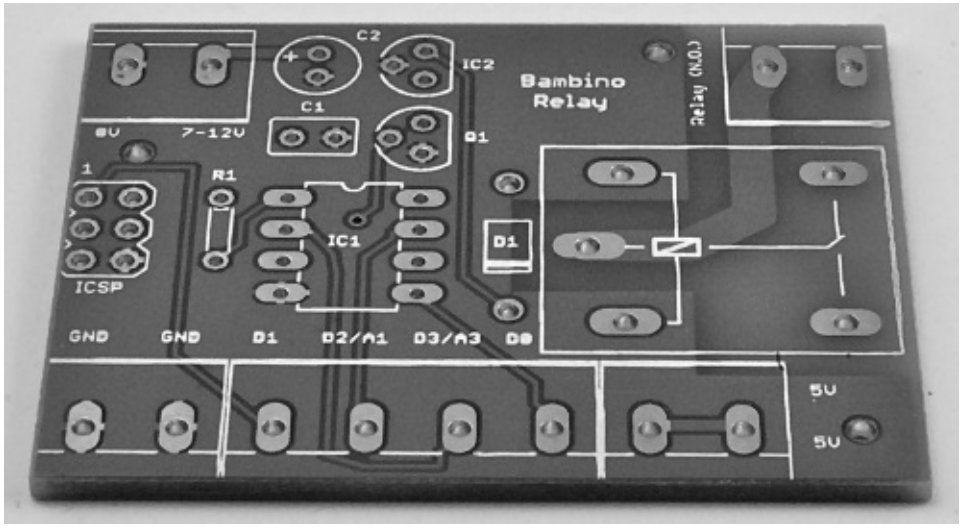


FIGURE 7.11 A PCB created by a PCB services

The design files that you send to a PCB service are called Gerber files. Table 7.1 shows the files that you should submit. Each of the files has a different extension that indicates its contents. The computer-aided manufacturing (CAM) job feature of EAGLE CAD will produce these files for you automatically, using a CAM job file that you will find in the instructions on the PCB service's web site.

TABLE 7.1 Gerber File Types for CAM

FILE	CONTENTS
Myboard.GTL	Top layer (copper)
Myboard.GBL	Bottom layer (copper)
Myboard.GTS	Solder stop mark (top)
Myboard.GBS	Solder stop mark (bottom)
Myboard.GTO	Silk screen (top)
Myboard.GBO	Silk screen (bottom)
Myboard.TXT	Drill

All that remains to do is to pay, create a zip file containing the files, and e-mail them to the PCB service. Your boards will come back in the post days or weeks later, depending on the service level you chose.

Some PCB services will also accept EAGLE CAD .brd files and other CAD formats without requiring the generation of Gerber files.

When schematics get even a little complicated, you rapidly find that trying to route everything on one layer of copper becomes impossible, and a two-layer board becomes necessary. The top and bottom layers can either connect where a component lead naturally passes through both layers or using vias that are holes in the board just to allow a trace to jump layers.

New and better PCB deals come along all the time, so check the newsgroups to see what people are using. Two useful providers are Itead Studio (www.iteadstudio.com) and Elektor PCB Service (www.elektor.com).

A Note About Board Layout

When arranging components on a circuit board, ICs and resistors should be placed in rows and should all be pointing in the same direction. Also, make sure to leave about a 2mm border around the circuit board to allow room for card lifters, guides, and standoffs.

Bring power supply leads or other I/O leads to the edge of the board, connecting them through an edge connector, D-connector, barrier-strip connector, or single-binder posts fixed to the edge of the board. Avoid mounting heavy components on circuit boards to prevent damage in case of a fall.

It is also a good idea to place polarity marking on the board next to devices such as diodes and electrolytic capacitors. Placing labels next to IC pins is also helpful. Consider labeling test points, trimmer functions (such as zero adjustment), inputs and outputs, indicator light functions, and power-supply terminals as well.

7.2.6 Special Pieces of Hardware Used in Circuit Construction

Three pieces of hardware are used frequently during the construction phase: prototyping boards with I/O interfacing gold-plated fingers, IC and transistor socket holders, and heat sinks.

Prototyping boards with gold-plated fingers typically are inserted into a card cage along with a number of other boards. Each board is inserted through a plastic guide and into an edge connector. Separate boards can be linked by means of a flat multiple-conductor cable (see Fig. 7.12a). The nice thing about these boards is that you can easily remove them from a cage to work on them, without making a mess of things in the process. When designing multiple-board devices, it is wise to use a separate board for each functional group of a circuit (for example, amplifier sections, memory-chip sections, and so on). This will make it easy to find and fix problems later on.

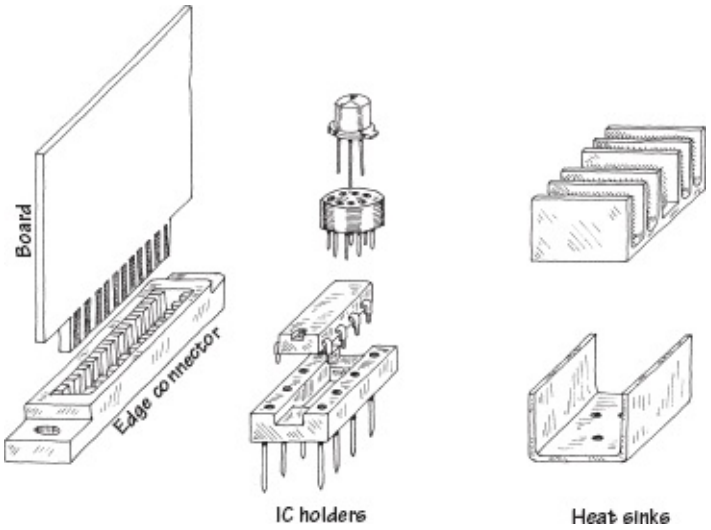


FIGURE 7.12

IC sockets are used in situations where there is a good chance that the device they house will need replacing (see Fig. 7.12b). It is tempting to use such holders everywhere within a circuit; however, placing too many IC sockets in a circuit board can lead to headaches later on. Often, the socket sections of these holders are poorly designed and may prove unreliable over time.

Heat sinks are metal devices with large surface areas that are connected to heat-generating devices (such as power diodes and transistors) to help dissipate heat energy. A heat sink is usually connected to a component by means of a screw and washer fastener (see Fig. 7.12c). Silicon grease placed between the washer and heat sink is often used to enhance the thermal conductivity between the electrical component and heat sink.

7.2.7 Soldering

Solder is a tin-lead (these days, mostly without the lead) alloy used to join component leads together. Electrical solder often comes with a rosin flux mixed in, which is used to dissolve oxides that are present on the metal surfaces to be joined. Recent legislation in Europe (known as the Restriction of Hazardous Substances Directive, or RoHS) has outlawed the use of lead in consumer electronics. In the United States, there are tax incentives to reduce the use of lead. This has led (no pun intended) to the development of lead-free solders.

Before solder can be applied with a soldering iron, all metal surfaces must be cleansed of oils, silicones, waxes, or grease. Use a solvent, steel wool, or fine sandpaper to remove undesirable residue.

When soldering PCBs, use a low-wattage iron (25 to 40 W). To ensure good soldering connections, a thin, bright coating of molten solder should be present on the tip of the iron. With time, this coat becomes contaminated with oxides and should be renewed by wiping its surface with a sponge and then reapplying fresh solder. (Applying a fresh coat of solder to the tip of an iron is referred to as *tinning* an iron.)

The trick to making good soldering connections is to first heat the two metal pieces to be joined. Do not melt the solder first; otherwise, you will not be able to control the placement of the molten solder. Solder likes to flow toward hot spots.

Also when working on a circuit, make sure not to splatter solder on your board. If a small piece of solder lands between two separate conductive lines, you will end up shorting them. When you are finished soldering, inspect your work carefully for stray solder spatters and for a good sound joint.

To protect sensitive components from the heat of the soldering iron’s tip, heat-sink components by gripping the component lead with a needle-nose pliers. Special heat-sink clips are also available for this purpose.

See Section 7.5.16 for more on soldering equipment.

7.2.8 Desoldering

If you make a bad connection or need to replace a component, you must melt the solder and start over again. But simply melting the solder and then attempting to yank the part while the solder is still wet is not always easy. This is especially true when it comes to dealing with ICs.

The trick to freeing a component from the solder's hold is to first melt the solder and then remove the solder with an aspiration tool, or "sucker." An aspiration tool typically resembles a turkey baster or a large syringe-like device. Another method that can be used to remove solder is to use a solder wick. This device resembles a braided copper wire and acts to draw solder away from a connection by means of capillary action.

7.2.9 Enclosing the Circuit

Circuits typically are enclosed in either an aluminum or plastic box. Aluminum enclosures are often used when designing high-voltage devices, whereas plastic containers typically are used for lower-voltage applications. If you design a high-voltage circuit housed in an aluminum box, make sure to ground the box to avoid getting shocked.

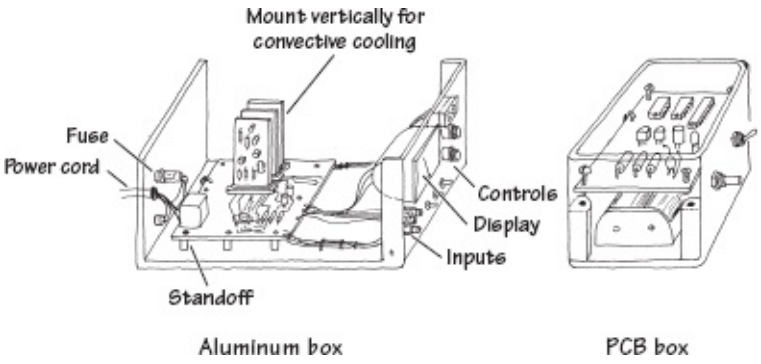


FIGURE 7.13

Circuit boards that are placed within an aluminum box should be supported off the ground with standoffs. If the circuit is ac-powered, drill a hole through the back side of the box, and insert a strain relief that grips the cable and a grommet around the edges of the hole. Place frequently used switches, dials, and indicators on the front panel, and place seldom-used switches and fuses on the back panel.

If you expect that your circuit will be generating a lot of heat (running more than 10 W or so), consider installing a blower fan. For circuits running on moderately low power, simple perforated holes placed on the top and bottom of the box will aid in conductive cooling.

Place major heat-producing components, such as power resistors and transistors, toward the back of the box, connecting them through the back panel to heat sinks. Make sure to orient heat sinks with their fins in the vertical direction. Also, if you are building a multiple-board cage device, align all boards vertically to allow for efficient ventilation.

Plastic boxes usually come with built-in standoffs on which to rest the circuit board. Typically, these boxes allow for extra room underneath the board for items such as batteries and speakers.

7.2.10 Useful Items to Keep Handy

The following items are worth having at your workbench:

- Needle-nose pliers
- Snippers
- Solder
- Soldering iron
- Solder sucker
- IC inserter
- Lead bender
- Solvent
- Clip-on heat sinks
- Circuit-board holder
- Screws (flathead and roundhead)
- Nuts
- Flat and lock washers (4-40, 6-32, 10-24)
- Binding posts
- Grommets
- Standoffs
- Cable clamps
- Line cord
- Hookup wire
- Shrink tubing in assorted sizes
- Eyelets
- Fuse holders

Sections 7.5.18 through 7.5.21 provide a more complete list.

7.2.11 Troubleshooting the Circuits You Build

If your circuit is malfunctioning, see if you have overlooked any suggestion in the troubleshooting flowchart in Fig. 7.14.

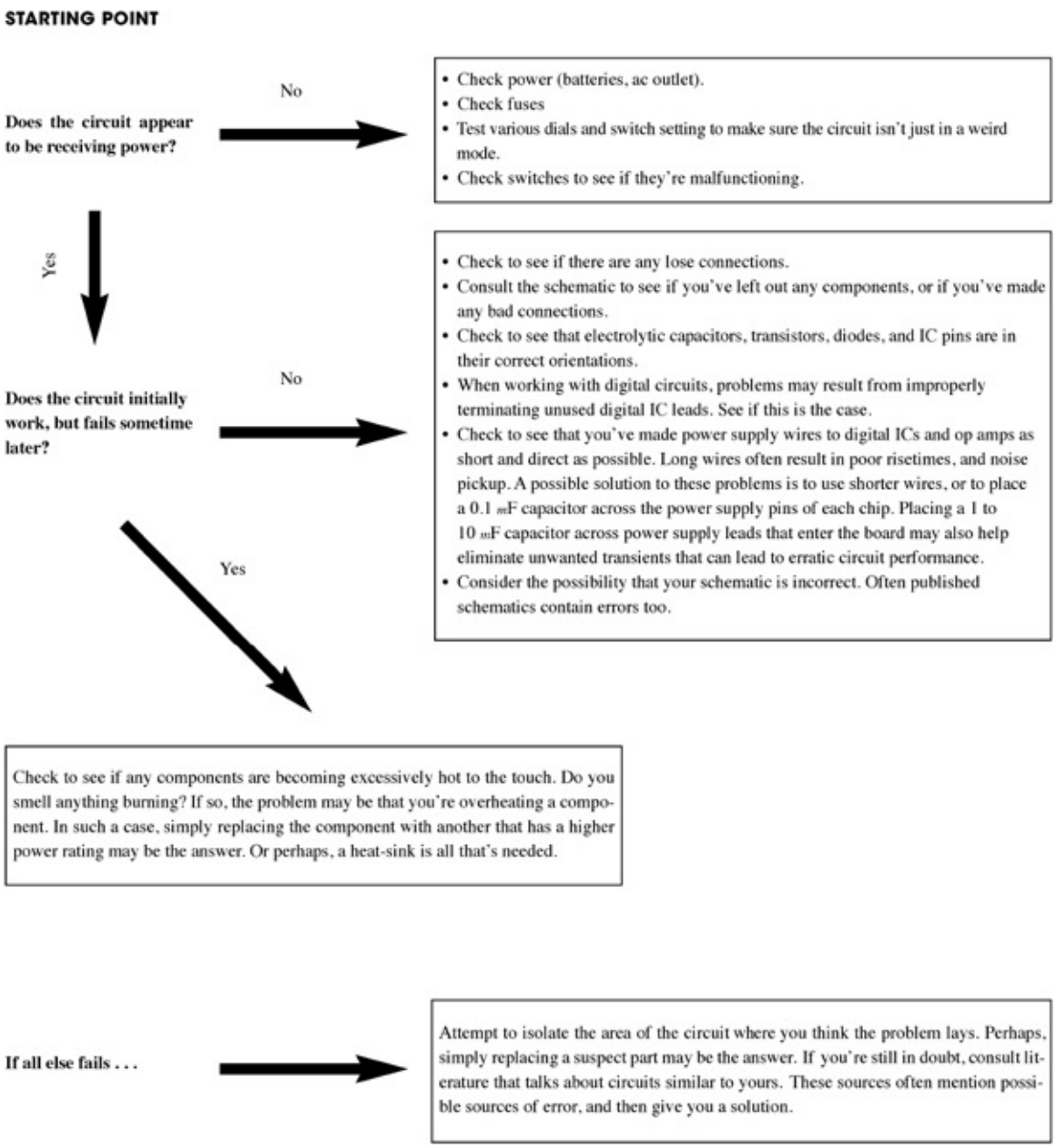


FIGURE 7.14

7.3 Multimeters

A *multimeter*, or volt-ohm-milliammeter (VOM), is an instrument used to measure current, voltage, and resistance. The two most common types of multimeters include the analog VOM and the digital VOM, as shown in Fig. 7.15.

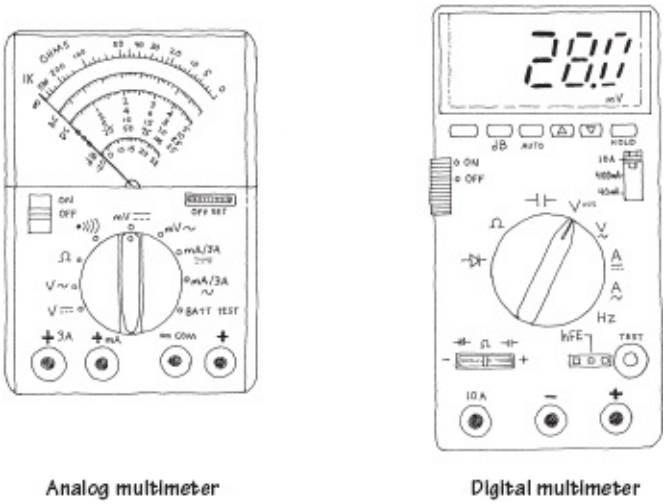


FIGURE 7.15

The obvious difference between the two types of VOMs is that an analog VOM uses a moving-pointer mechanism that swings along a calibrated scale, while the digital VOM uses some complex digital circuitry to convert input measurements into a digitally displayed reading. Technically speaking, analog VOMs are somewhat less accurate than digital VOMs (they typically have a 3 percent higher error in reading than a digital VOM), and multiple scales make them harder to read. Also, the resolution (displayable accuracy) for an analog VOM is roughly 1 part in 100, as compared with a 1 part in 1000 resolution for a digital VOM. Despite these limitations, analog VOMs are superior to digital VOMs when it comes to testing circuits that contain considerable electrical noise. Unlike digital VOMs, which may go blank when noise is present, analog VOMs are relatively immune to such disturbances.

Here, we'll look at how these devices work. Also see Section 7.5.3, which discusses which type of multimeter you may need for your electronics laboratory.

7.3.1 Basic Operation

Measuring Voltages

The trick to measuring voltages with a VOM is to turn the selector knob to the voltage setting. If you want to measure a dc voltage, the knob is turned to the appropriate dc voltage-level setting. If you wish to measure an ac voltage, the knob is turned to the ac voltage setting (V_{ac}, or V_{rms}). Note that the displayed voltage in the V_{ac} setting is the RMS voltage ($V_{rms} = 0.707\ V_{peak\text{-}to\text{-}peak}$). Once the VOM is set correctly, the voltage between two points in a circuit can be measured by touching the VOM's probes to these points (the VOM is placed in parallel). For example, Fig. 7.16 shows the procedure used to measure the voltage drop across a resistor.

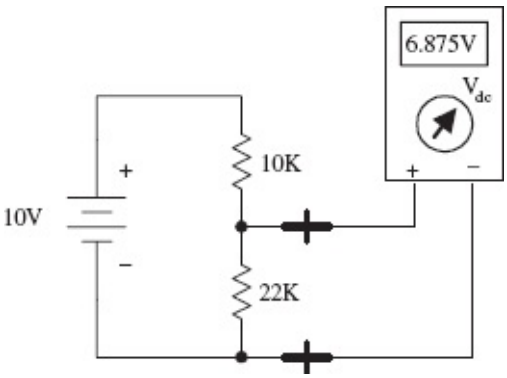


FIGURE 7.16

Measuring Currents

Measuring currents with a VOM is almost as easy as measuring voltages. The only difference (besides changing the setting) is that you must break the test circuit at the location where you wish to make a current reading. Once the circuit is open, the two probes of the VOM are placed across the break to complete the circuit (VOM is placed in series). Fig. 7.17 shows how this is

done. When measuring ac currents, the VOM must be set to the RMS current setting.

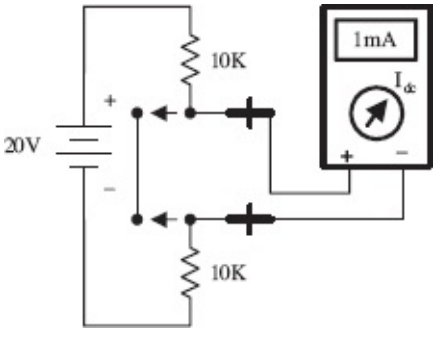


FIGURE 7.17

Measuring Resistances

Measuring resistances with a VOM is simple enough: remove the power to the resistive section of interest, and then place the VOM's probes across this section. Of course, make sure to turn the VOM selector knob to the ohms setting beforehand.

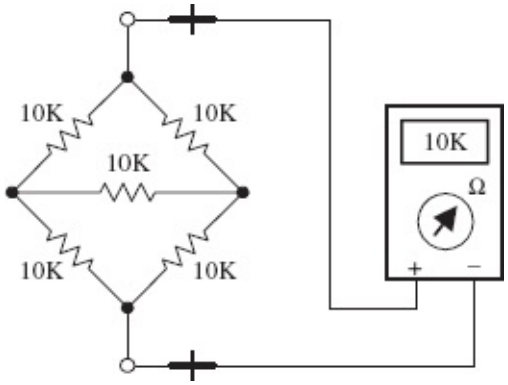
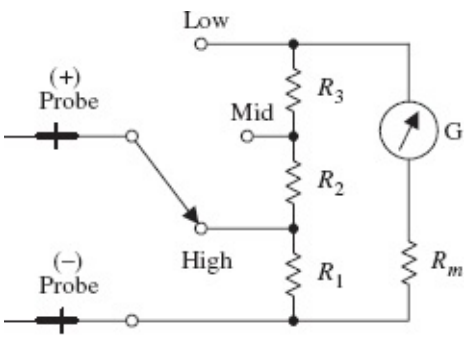


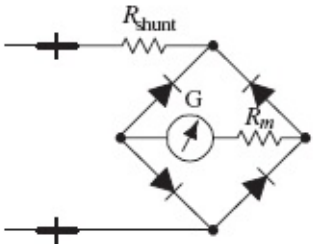
FIGURE 7.18

7.3.2 How Analog VOMs Work

An analog VOM contains an ammeter, voltmeter, and ohmmeter all in one. In principle, understanding how each one of these meters works individually will help to explain how an analog VOM works as a whole.

Ammeter





An ammeter uses a D-Arsonval galvanometer that consists of a current-controlled electromagnetic that imparts a torque on a

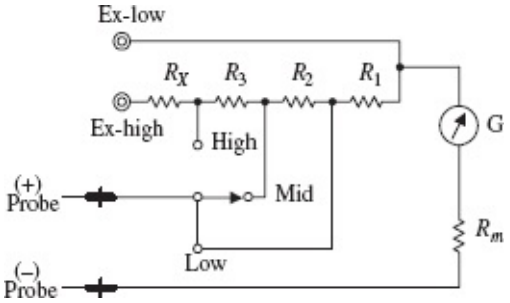
spring-loaded rotatable needle. The amount of deflection of the needle is proportional to the current flow through the electromagnet. The electromagnetic coil has some resistance built in, which means you need to throw R_m into the circuit, as shown in Fig. 7.19. (R_m is typically around 2 k or so.)

Now, a galvanometer alone could be used to measure currents directly; however, if the input current is excessively large, it will force the needle beyond the viewable scale. To avoid this effect, a number of *shunt resistors* placed in parallel with the galvanometer make up a current divider capable of diverting some of the "needle-bending" current away from the galvanometer. The current value read from the display must be read from the appropriate ruler marking on the display that corresponds to the shunt resistance chosen.

To make this device capable of measuring ac currents, a bridge rectifier can be incorporated into the design to provide a dc current to the galvanometer (see the lower circuit). The dc current will produce a needle swing that is proportional to the alternating voltage measured. A typical ammeter has about a 2-k input resistance. Ideally, an ammeter should have zero input resistance.

FIGURE 7.19

Voltmeter

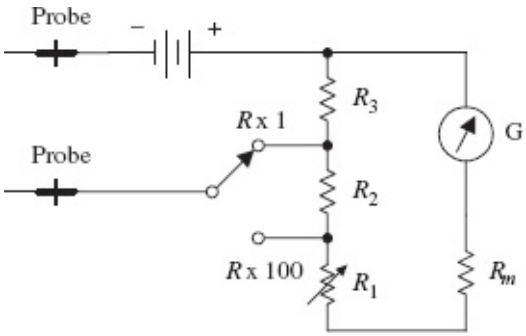


An analog voltmeter, like the ammeter, also uses a D-Arsonval galvanometer. Again, the galvanometer has some internal resistance (R_m). When the voltmeter's leads are placed across a voltage difference, a current will flow from the higher potential to lower potential, going through the galvanometer in the process. The current flow and the needle deflection are proportional to the voltage difference.

Again, as with the ammeter, shunt resistors are used to calibrate and control the amount of needle deflection. To make ac voltage measurements, a bridge rectifier, like the one shown in the previous example, can be incorporated into the meter's design. A typical voltmeter has an input resistance of 100 k. An ideal voltmeter should have infinite input resistance.

FIGURE 7.20

Ohmmeter

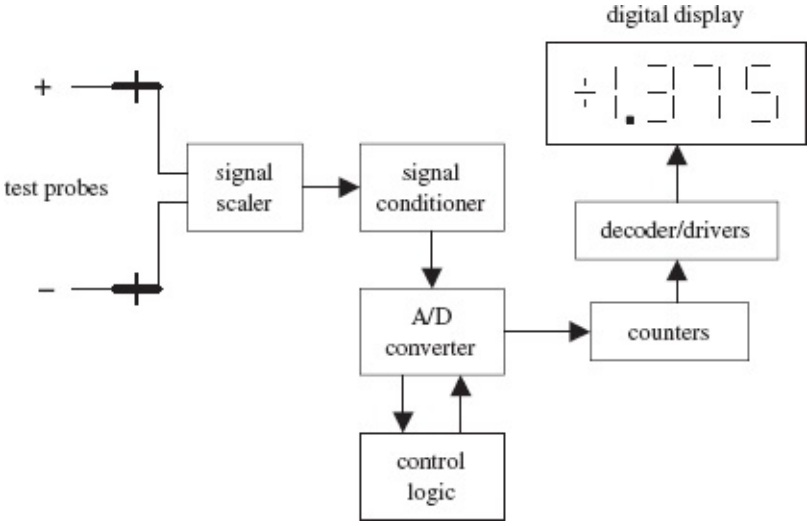


To measure resistance, an ohmmeter uses an internal battery to supply a current through the measured load and through a galvanometer (the load and galvanometer are in series). If the tested load is small, a large current will flow through the galvanometer, and a large deflection will occur. However, if the tested load resistance is large, the current flow and the deflection will be small. (In a VOM, the ohmmeter calibration markings are backward: 0 W is set to the right of the scale.)

The amount of current flow through the galvanometer is proportional to the load resistance. The ohmmeter is first calibrated by shorting the probe leads together and then zeroing the needle. Like the other meters, an ohmmeter uses a number of shunt resistors to control and calibrate the needle deflections. A typical ohmmeter has an input resistance of about 50 Ω. An ideal ohmmeter should have zero input resistance.

FIGURE 7.21

7.3.3 How Digital Multimeters Work



A digital multimeter is composed of a number of functional groups, as shown in the block diagram. The signal scaler is an attenuator amplifier that acts as a range selector. The signal conditioner converts the scaled input signal to a dc voltage within the range of the analog-to-digital converter (A/D converter). In the case of ac voltage measurements, the ac voltage is converted into a dc voltage via a precision rectifier-filter combination. The gain of the active filter is set to provide a dc level equal to the RMS value of the ac input voltage or current.

The signal conditioner also contains circuits to convert current and/or resistance into proportional dc voltages. The A/D converter converts the dc analog input voltage into a digital output voltage. The digital display provides a digital readout of the measured input. Control logic is used to synchronize the operation of the A/D converter and digital display.

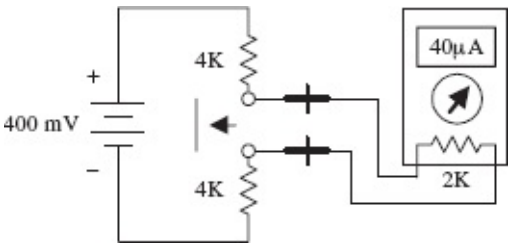
FIGURE 7.22

7.3.4 A Note on Measurement Errors

When measuring the current through (or voltage/resistance across) a load, the reading obtained from the VOM will always be different when compared with the true value present before the meter was connected. This error comes from the internal resistance of the VOM.

For each setting (ammeter, voltmeter, and ohmmeter), there will be a different internal resistance. A real ammeter typically will have a small input resistance of around 2 k, while a voltmeter may have an internal resistance of 100 k or more. For an ohmmeter, the internal resistance is usually around 50 Ω. It is crucial to know these internal resistances in order to make accurate measurements. The following examples show how large the percentage error in readings can be for meters with corresponding input resistances.

Current-Measurement Error



If an ammeter has an input resistance of 2 k, calculate the percentage error in reading for the circuit shown here.

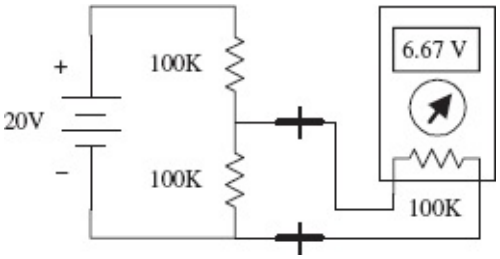
$$I_{true} = \frac{400\,\text{mV}}{4\text{k} + 4\text{k}} = 50\,\mu A$$

$$I_{measured} = \frac{400\,\text{mV}}{4\text{k} + 4\text{k} + 2\text{k}} = 40\,\mu\text{A}$$

$$\%\ \text{error} = \frac{50\ \mu\text{A} - 40\ \mu\text{A}}{50\ \mu\text{A}} \times 100\% = 20\%$$

FIGURE 7.23

Voltage-Measurement Error



If a voltmeter has an input resistance of 100 k, calculate the percentage error in reading for the circuit shown here.

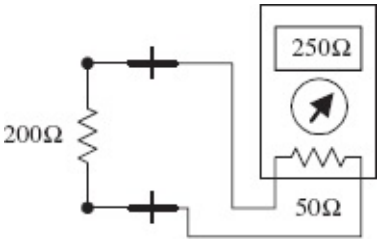
$$V_{true} = \frac{100\text{ k}}{100\text{ k} + 100\text{ k}}(20\text{V}) = 10\text{V}$$

$$V_{measured} = \frac{100\text{ k}}{100\text{ k} + (100\text{ k} \times 100\text{ k}) / (100\text{ k} \times 100\text{ k})} = 6.67\text{ V}$$

$$\%\text{ error} = \frac{10\text{V} - 6.67\text{ V}}{10\text{V}} \times 100\% = 33\%$$

FIGURE 7.24

Resistance-Measurement Error



If an ohmmeter has an input resistance of 50 Ω, calculate the percentage error in reading for the circuit shown here.

$$R_{true} = 200\ \Omega$$

$$R_{measured} = 200\ \Omega + 50\ \Omega = 250\ \Omega$$

$$\%\text{ error} = \left|\frac{200\ \Omega - 250\ \Omega}{200\ \Omega}\right| \times 100\% = 25\%$$

FIGURE 7.25

To minimize the percentage error, an ammeter's input resistance should be less than the Thevenin resistance of the original circuit by 20 times or more. Conversely, a voltmeter should have an input resistance that is larger than the Thevenin resistance of the original circuit by 20 times or more. The same goes for the ohmmeter: it should have an input resistance that is at least 20 times the Thevenin resistance of the original circuit. By following these simple rules, it is possible to reduce the error to below 5 percent.

Another approach (perhaps a bit more tedious) is to look up the internal resistances of your VOM, make your measurements, and then add or subtract the internal resistances afterward. (See Appendix B for information about detailed error analysis.)

7.4 Oscilloscopes

Oscilloscopes measure voltages, not currents or resistances—just voltages. This is an important point to get straight from the start. An oscilloscope is an extremely fast xy plotter capable of plotting an input

signal versus time or versus another input signal.

As a signal is supplied to the input of a scope, a luminous spot appears on the screen. As changes in the input voltage occur, the luminous spot responds by moving up or down, or left or right. In most applications, the oscilloscope's vertical-axis (y axis) input receives the voltage part of an incoming signal and then moves the spot up or down depending on the value of the voltage at a particular instant in time. The horizontal axis (x axis) is usually used as a time axis, where an internally generated linear ramp voltage is used to move the spot across the screen from left to right at a rate that can be controlled by the operator.

If the signal is repetitive, such as a sinusoidal wave, the oscilloscope can make the sinusoidal pattern appear to stand still. This makes a scope a useful tool for analyzing time-varying voltages.

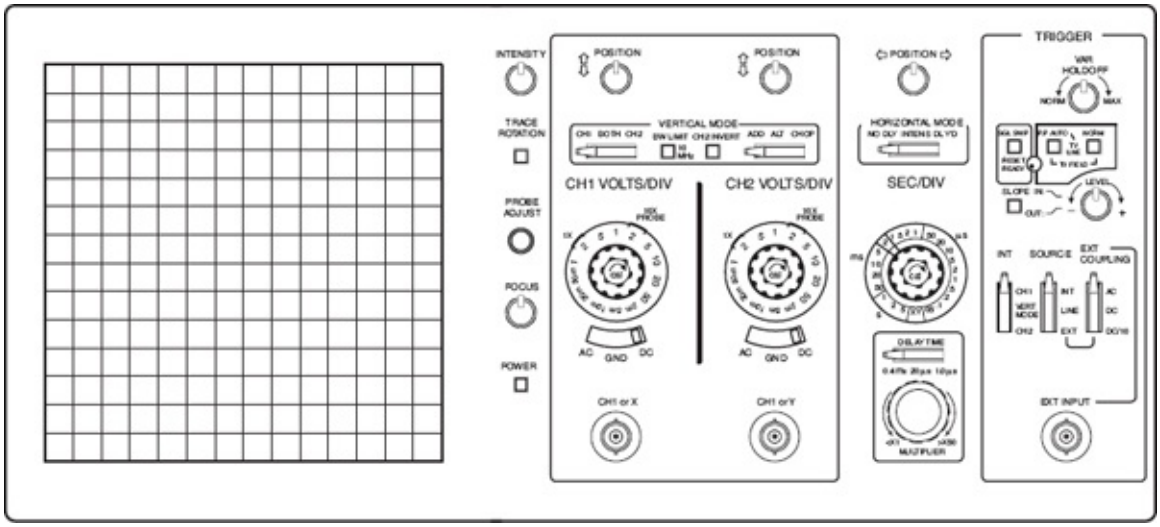


FIGURE 7.26

Even though oscilloscopes measure only voltage, it is possible to convert quantities such as current, strain, acceleration, pressure, and so on into voltages that the scope can use. To convert a current into a voltage, a resistor of known resistance is used; the current is measured indirectly by measuring the voltage drop across the resistor and then applying Ohm's law. To convert strain, movement, and so on into voltage requires the use of transducers (electromechanical devices). By applying some calibrating tricks, the magnitude of, say, a pressure applied to a pressure transducer can be measured accurately.

Here, we'll look at how oscilloscopes work. Also see Section 7.5.5, which discusses which type of scope you may need for your electronics laboratory.

7.4.1 How Scopes Work

While it is still possible to buy an analog oscilloscope that uses a cathode-ray tube, the majority of modern oscilloscopes are effectively a computer with fast analog-to-digital converters and an LCD or OLED display. As well as reducing the size and weight of the oscilloscope, the use of digital signal processing allows the scope to do things that an analog scope cannot do, such as generate clearer color-coded signals, provide the memory to pan left and right, allow you to export the display image to your computer, and show a frequency domain display. Digital scopes are comparable in price to their analog counterparts.

Interestingly, a veteran user of an analog oscilloscope would recognize the controls of a digital oscilloscope. The basic procedure for operating a digital oscilloscope is exactly the same as for an analog oscilloscope. It is actually easier to understand what an oscilloscope does by studying the workings of an analog oscilloscope than a digital one.

An analog oscilloscope is built around a cathode-ray tube. All the circuits inside the scope are designed to take an input signal and modify it into a set of electrical instructions that supply the tube's electron gun with aiming instructions (the location where to focus the beam). Most of the knobs and switches on the face of a scope that are connected to the interior circuitry are designed to help modify the instructions sent to the cathode-ray tube. For example, these controls set voltage scale, time scale, intensity of beam, focus of beam, channel selection, triggering, and so on.

Cathode-Ray Tube

A cathode-ray tube consists of an electron gun (filament, cathode, control grid, and anode), a second anode, vertical deflection plates, horizontal deflection plates, and a phosphor-coated screen. When current flows through the filament, the filament heats the cathode to a point where electrons are emitted. The control grid controls the amount of electrons that flow through the electron gun, and thus controls the intensity of the beam. If this grid is made more negative in voltage, more electrons will be repelled away from the grid, thus reducing the electron flow.

The electron beam is focused into a sharp point-like beam by applying a controlling voltage or focus voltage to anode 1. The second anode is supplied with a large voltage that is used to give the electrons within the beam the additional momentum needed to cause a photon emission once they collide with the phosphor screen. The beam-focusing section of the tube is referred to as the *electron gun*.

There are two sets of electrostatic deflection plates (vertical and horizontal) that are set between the second anode and the inner face of the phosphor screen. One set of plates is used to deflect the electron beam vertically; the other set is used to deflect the beam horizontally. For example, when one of the plates of a pair of plates is made more negative in charge than the other, the electron beam will bend away from the negative plate and veer toward the more positive plate. (The electrons in the beam are usually moving with sufficient forward velocity that they never actually come in contact with the plates.) When a sawtooth voltage is applied to the horizontal plates, the gradually rising potential across the plates pulls the electron beam from the negative plate to the positive plate, causing the beam to scan across the phosphor screen. The vertical plates cause the electron beam to move up and down.

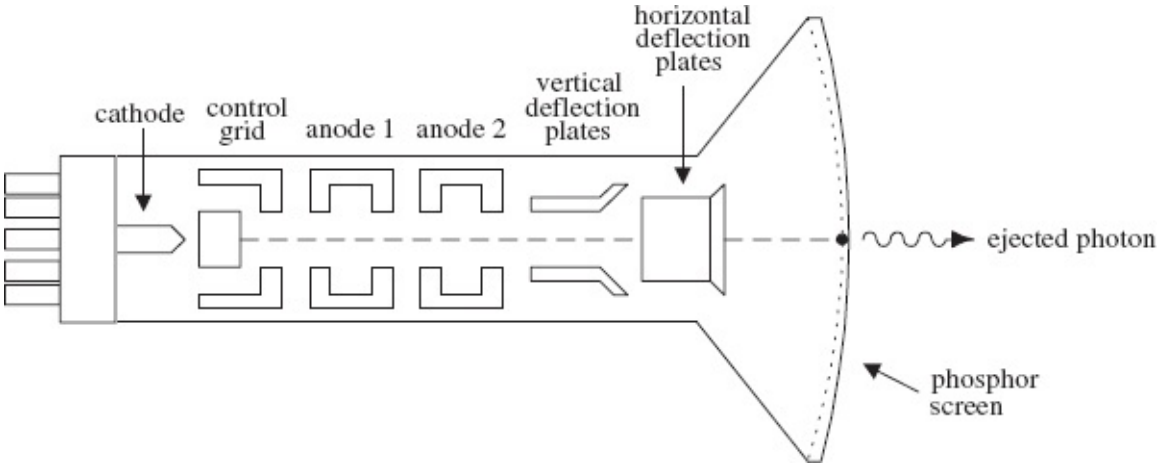


FIGURE 7.27

The next step in understanding how an oscilloscope works is knowing how an incoming signal is converted into a set of electrical signals or applied voltages that control the beam-aiming mechanisms of the cathode-ray tube. This is where the interior circuitry comes in.

7.4.2 Interior Circuitry of a Scope

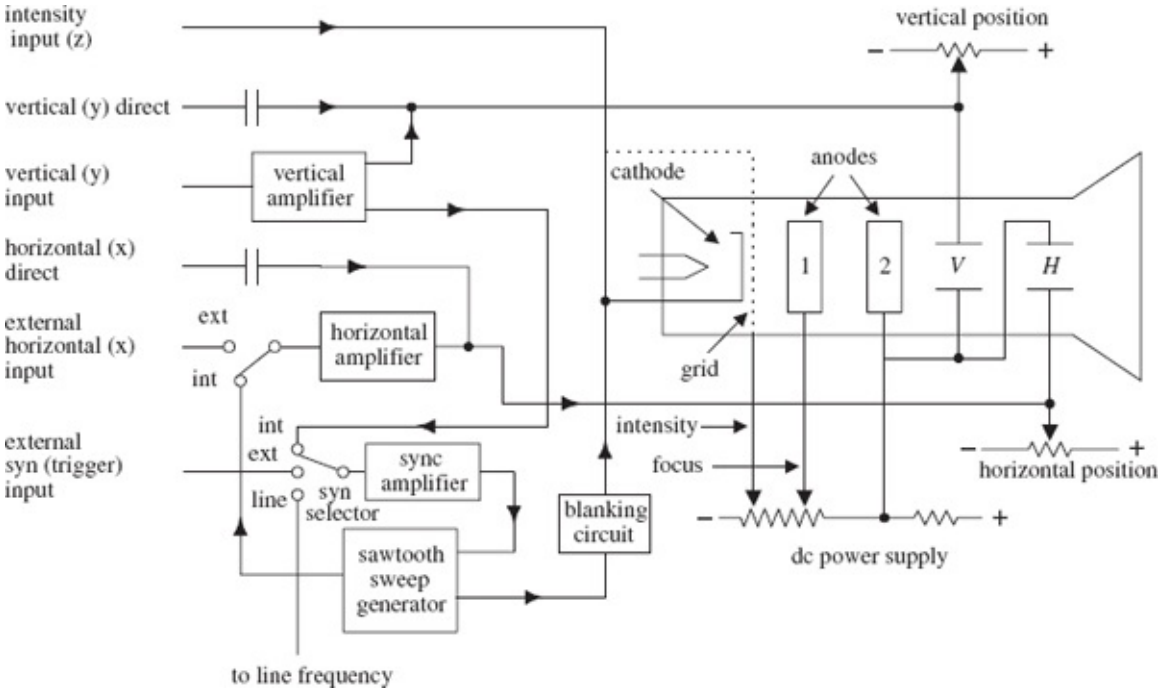


FIGURE 7.28

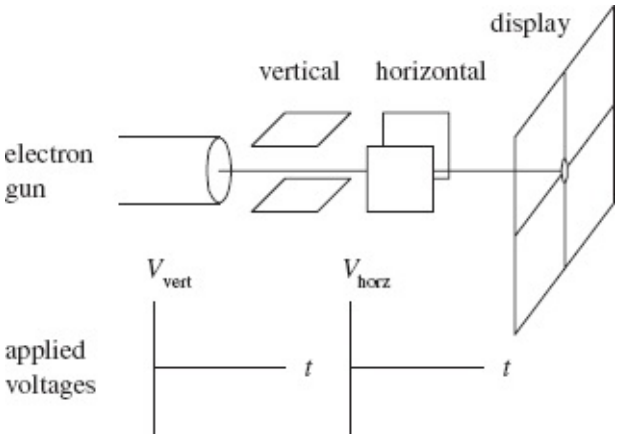
Let's take a sinusoidal signal and see how the interior circuits of a scope convert it into something you can see on a display. The first thing you do is apply the sinusoidal signal to the vertical input. From the vertical input, the sinusoidal signal is sent to a vertical amplifier, where it is amplified so that it can supply enough voltage to deflect the electron beam. The vertical amplifier then sends a signal to the sweep selector. When the sweep selector is switched to the internal position (the other positions are explained in Section 7.4.5), the signal from the vertical amplifier will enter the sync amplifier.

The sync amplifier is used to synchronize the horizontal sweep (sawtooth in this case) with the signal under test. Without the sync amplifier, the display pattern would drift across the screen in a random fashion. The sync amplifier then sends a signal to the sawtooth sweep generator, telling it to start a cycle. The sawtooth sweep generator then sends a sawtooth signal to a horizontal amplifier (when horizontal input is set to internal). At the same time, a signal is sent from the sawtooth sweep generator to the blanking circuit. The blanking circuit creates a high negative voltage on the control grid (or high positive voltage on the cathode-ray tube cathode), which turns off the beam as it snaps back to the starting point. Finally, voltages from the vertical and horizontal amplifiers (sawtooth) are sent to the vertical and horizontal plates in a synchronized fashion. The final result is a sinusoidal pattern displayed on the scope's screen.

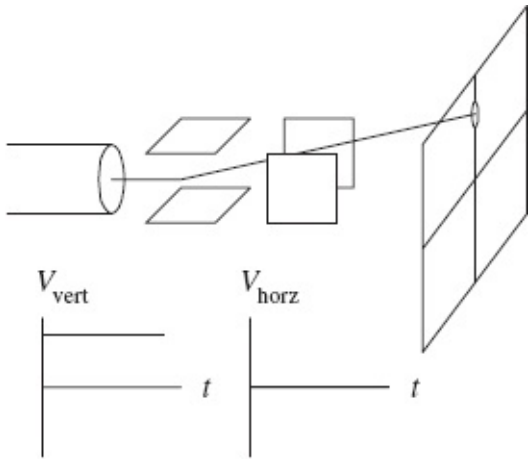
The other features—such as vertical direct and horizontal direct inputs, external horizontal input, external trigger, line frequency, and *xy* mode—are described in Section 7.4.5.

It is important to note that the scope does not always use a sawtooth voltage applied to the horizontal plates. You can change the knobs and inputs and use another input signal for the horizontal axis. Controls such as intensity, focus, and horizontal and vertical position of the beam can be understood by looking at the oscilloscope circuit diagram.

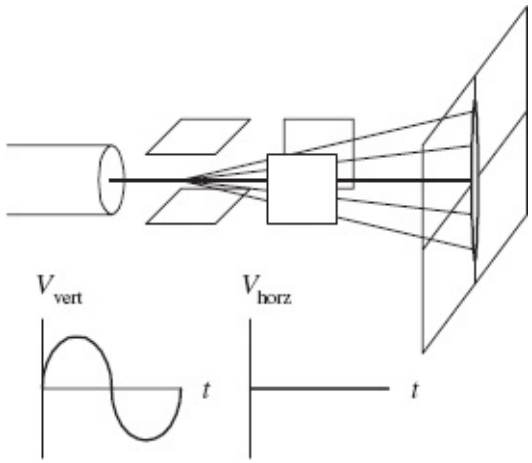
7.4.3 Aiming the Beam



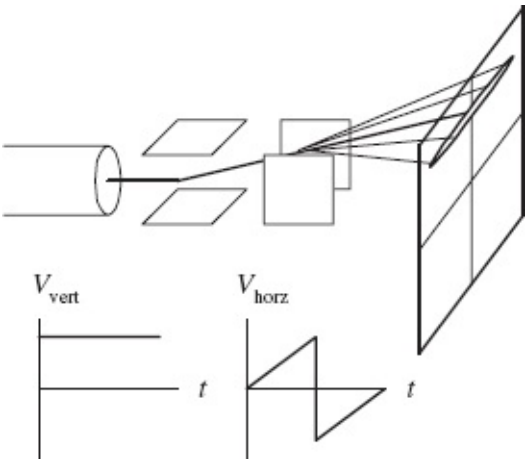
When no voltages are applied to the horizontal and vertical plates, the electron beam is focused at the center of the scope's display.



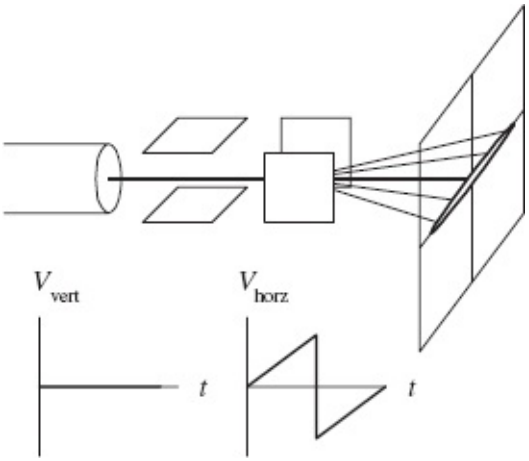
When a dc voltage is applied to the vertical plates, while no voltage is applied to the horizontal plates, the electron beam shifts up or down depending on the sign of the applied voltage.



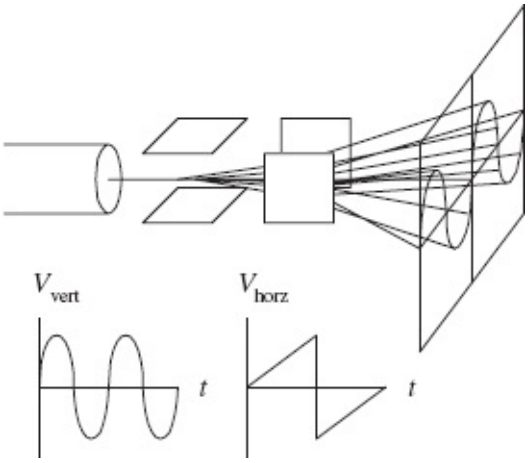
When a sinusoidal voltage is applied to the vertical plates, while no voltage is applied to the horizontal plates, a vertical line is traced on the y axis.



When a sawtooth voltage is applied to the horizontal plates, while no voltage is applied to the vertical plate, the electron beam traces a horizontal line from left to right. After each sawtooth, the beam jumps back to the left and repeats its left-to-right sweep.



When a dc voltage is applied to the vertical plates, while a sawtooth voltage is applied to the horizontal plates, a horizontal line is created that is shifted up or down depending on the sign of the dc vertical plate voltage (+ or −).

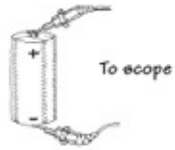


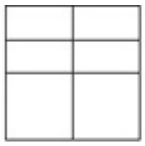
When a sinusoidal voltage is applied to the vertical plates and a sawtooth voltage is applied to the horizontal plates, the electron beam moves up as the signal voltage increases and at the same time moves to the left as the sawtooth voltage is applied to the horizontal plates. The display gives a sinusoidal graph. If the applied sinusoidal frequency is twice that of the sawtooth frequency, two cycles appear on the display.

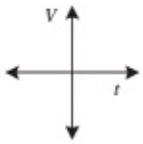
FIGURE 7.29

7.4.4 Scope Usage

DC Voltmeter

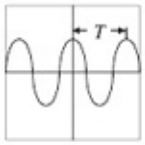






AC Voltmeter/AC Frequency Meter





T = period

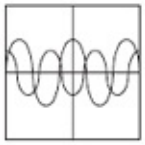
f = frequency

$f = 1/T$

$$V_{rms} = \frac{1}{\sqrt{2}} V_{max}$$

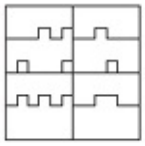
Phase Relationships Between Two Signals

source 1→chan 1
source 2→chan 2



The scope can be used to compare two source signals (e.g., measure phase shifts, voltage and frequency differences, etc.).

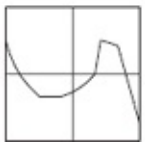
Digital Measurements



A scope can be used to create timing diagrams for digital circuits.

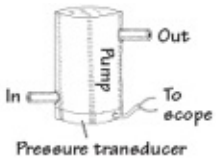
xy Graphics (xy Mode)

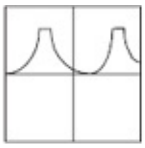
chan 1 input→x axis
chan 2 input→y axis



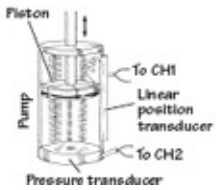
The scope no longer uses the x axis as the time axis, but uses signal voltage from another external source.

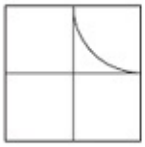
Measurements Using Transducers





By using transducers to convert an input quantity, such as pressure, into a voltage, the scope can be transformed into a pressure meter.
y axis → pressure
x axis → time





Here, a scope is used in xy mode, where
y axis → pressure
x axis → piston

FIGURE 7.30

7.4.5 What All the Little Knobs and Switches Do

Figure 7.31 shows a typical layout of an oscilloscope control panel. The control panel of the scope you use may look slightly different (knob positions, digital display, number of input channels, and so on), but the basic ingredients are the same. If you do not find what you need in this section, refer to the oscilloscope user manual that comes with the scope.

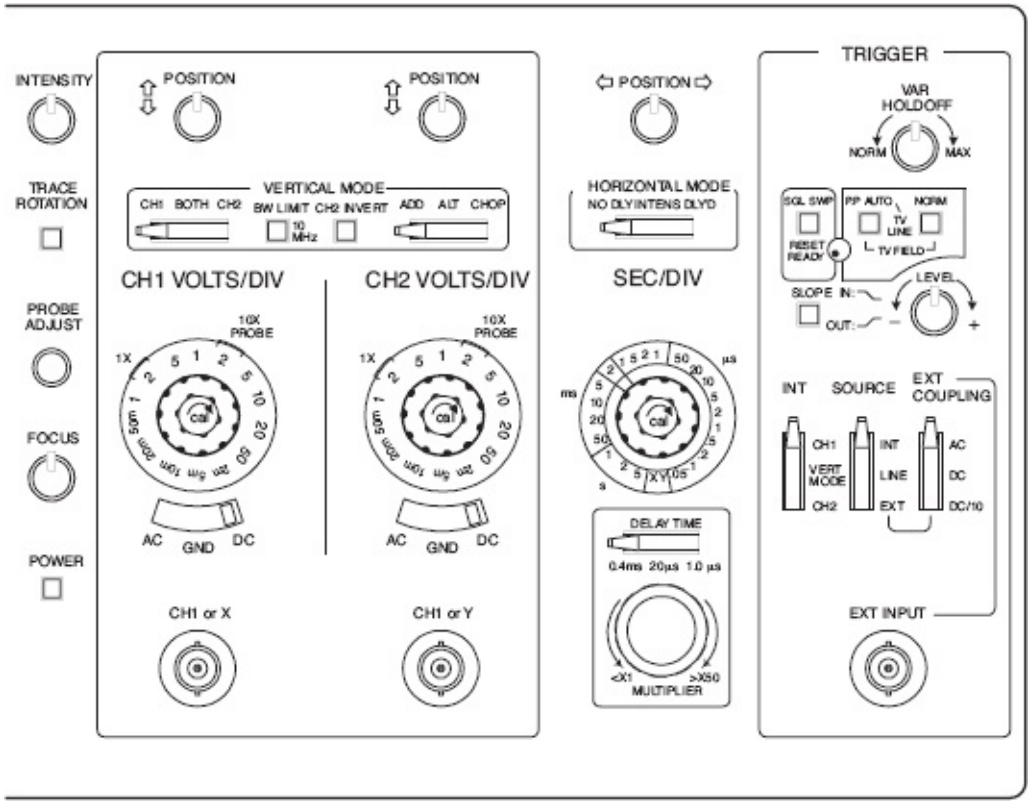


FIGURE 7.31

The control panel of an oscilloscope is divided into the following sections:

Vertical mode: This section of the scope contains all the knobs, buttons, and so on that control the vertical graphics of the scope. Most of these are associated with the voltage amplitude of an incoming signal.

Horizontal mode: This section of the scope contains all the knobs, buttons, and so on that control the horizontal portion of the graphics display. These usually are associated with the time base for the scope.

Trigger mode: This section of the scope contains all the knobs, buttons, and so on that control the way in which the scope "reads" an incoming signal. This section of the scope is probably the most technical. To understand triggering, read the upcoming "Trigger Mode" section.

Vertical Mode

The following are the vertical mode controls:

CH1 and CH2 coaxial inputs: Where input signals enter the scope.

AC, GRD, and DC switches:

- *AC:* Blocks the dc component of the signal, passing only the ac part of the signal.
- *DC:* Measures direct input of both ac and dc components of the input signal.
- *GRD:* Grounds input, causing vertical plates in cathode-ray tubes to become uncharged, thus eliminating electron beam deflection. Used to recalibrate the vertical component of the electron beam to a reference position on the display after altering the vertical position of the knob.

CH1 VOLTS/DIV and CH2 VOLTS/DIV knobs: Used to set the voltage scale on the display. For example, 5 V/div means that each division (1 cm) on the display is 5 V high.

MODE switches:

- *CH1, BOTH (DUAL), CH2 switch:* This switch allows you to pick between displaying a signal from

channel 1 or channel 2, or display both channels at the same time.

NORM, INVERT: This switch lets you choose to display a signal normally or inverted.

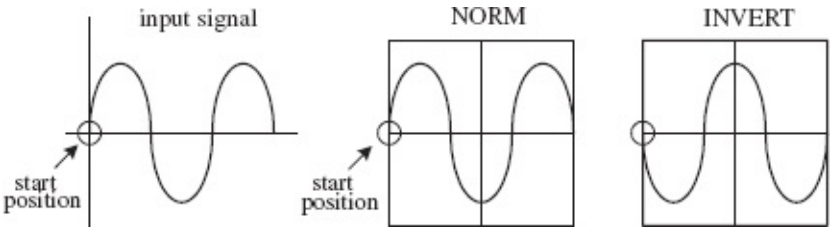


FIGURE 7.32

ADD, ALT, and CHOP:

- *ADD:* Adds signals from channel 1 and channel 2 together arithmetically.

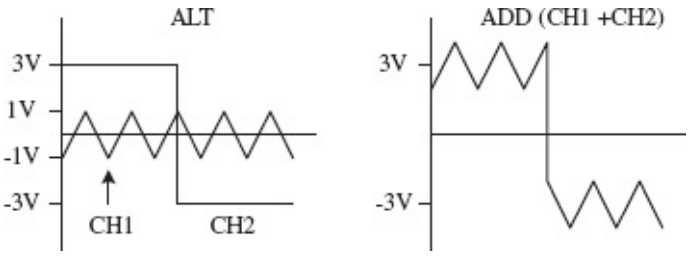


FIGURE 7.33

- *ALT:* Alternate sweep is selected regardless of sweep time, and the NORM-CHOP switch has no effect.
- *CHOP:* Operates the triggering SOURCE switch, providing automatic or manual selection of the alternate or chop method of dual-trace sweep generation.

POSITION knob: This knob allows you to move the displayed image up or down on the screen.

XY mode: When selected, the sweep rate (time base) supplied by the scope is switched off, and an external signal voltage applied to the channel 2 input replaces it.

Horizontal Mode

The following are the horizontal mode controls:

SEC/DIV knob: This knob sets the sweep speed or the scale for the horizontal time display. For example, 0.5 ms/DIV means each division (1 cm) on the display is 0.5 ms wide.

MODE switches:

- *NO DLY:* This setting takes the horizontal signal and presents it to display immediately.
- *DLY'D:* This setting delays the horizontal signal for a time you specify on the delay time section of scope. Use this to set the delay time of a signal.

SWEEP-TIME variable control: Sometimes known as the sweep frequency control, fine frequency control, or frequency vernier. Used as a fine sweep-time adjustment. In the extreme clockwise (CAL) position, the sweep time is calibrated by using the SWEEP TIME/CM switch. In the other positions, the variable control provides a continuously variable sweep rate.

POSITION knob: Moves the horizontal display left or right. This feature is useful when comparing two

input signals. It allows you to align the wave patterns for comparison.

Trigger Mode

The following are the trigger mode controls:

EXT TRIG jack: Input terminal for external trigger signals.

CAL terminal: Provides a calibrated 1-kHz, 0.1-V peak-to-peak squarewave signal. The signal can be used to calibrate the vertical-amplifier attenuators and to check frequency compensation of probes used with the scope.

HOLDOFF control: Used to adjust holdoff time (ignore triggers until holdoff time has expired).

TRIGGERING mode switch:

- *SINGLE*: When a signal is nonrepetitive, or if it varies in amplitude, shape, or time, a conventional repetitive display can produce an unstable presentation. SINGLE enables the RESET switch for triggered single-sweep operation. The signal sweep can be used to photograph a nonrepetitive signal. Pushing the RESET button initiates a single sweep that begins when the next sync trigger occurs.
- *NORM*: Used for triggered sweep operation. The triggering threshold is adjustable by means of the triggering LEVEL control. No sweep is generated in the absence of the triggering signal or if the LEVEL control is set in such a way as to allow the threshold to exceed the amplitude of the triggering signal (see Fig. 7.35).

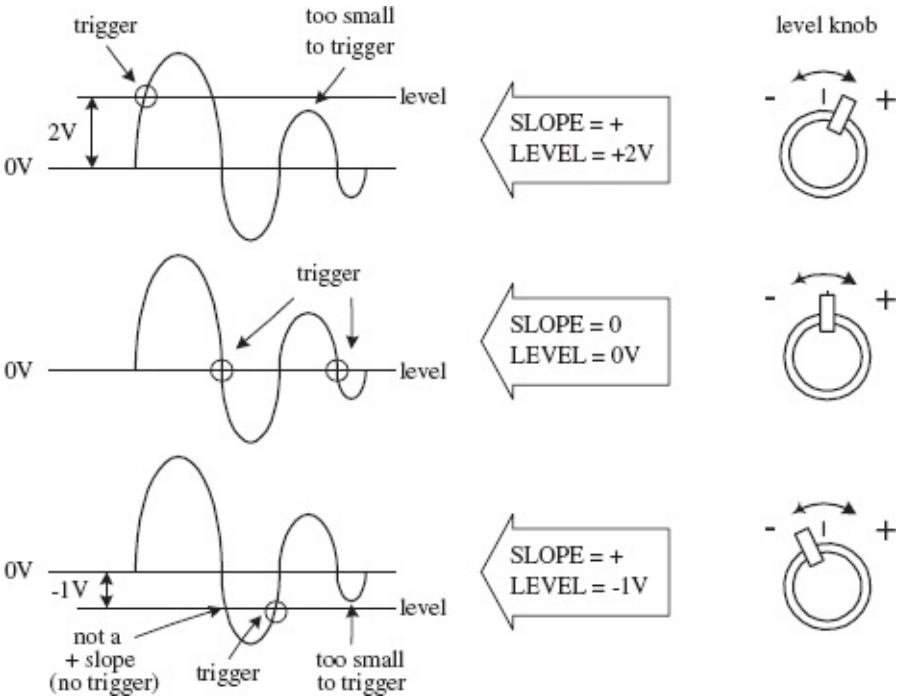


FIGURE 7.35

- *AUTO*: Selects automatic sweep operation, where the sweep generator free-runs and generates a sweep without a trigger signal (this is often referred to as a *recurrent sweep operation*). In AUTO mode, the sweep generator automatically switches to triggered sweep operation if an acceptable trigger signal is present. The AUTO position is useful when first setting up the scope in order to observe a waveform. It provides a sweep for waveform operation until other controls can be set properly. DC AUTO sweep must be used for dc measurements and for signals of such low amplitude

that the sweep is not triggered.

- *FIX*: This is the same as the AUTO mode, except that triggering always occurs at the center of the sync trigger waveform regardless of the LEVEL control setting.

SLOPE button: Selects the point at which a scope will trigger. When positive slope is selected, the scope will begin a sweep only when the signal voltage crosses the LEVEL voltage during a positive sloping rise (see the description of the LEVEL knob). A negative slope setting initiates a sweep to occur when the signal crosses the LEVEL voltage during a negative sloping fall (see Fig. 7.34).

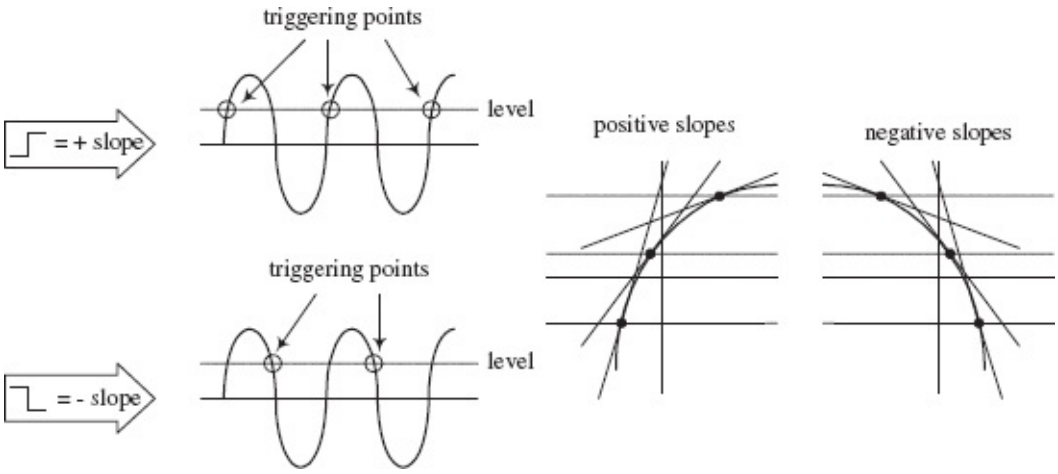


FIGURE 7.34

RESET button: When the triggering MODE switch is set to SINGLE, pushing the RESET button initiates a single sweep that begins when the next sync trigger occurs.

READY/TRIGGER indicator: In SINGLE trigger mode, an indicator light turns on when the RESET button is pushed in, indicating that a sweep is beginning. The light turns off when the sweep is completed. In the NORM, AUTO, and FIX triggering modes, the indicator turns on for the duration of the triggered sweep. The indicator also shows when the LEVEL control is set properly to obtain triggering.

LEVEL knob: Used to trigger a sweep. LEVEL sets the point when the scope will trigger based on the amplitude of the applied signal. The level can be shifted up or down. The READY/TRIGGER indicator turns on when the sweep is triggered, indicating that the triggering LEVEL control is within the proper range (see Fig. 7.35).

COUPLING switch: Used to select input coupling for the sync trigger signal.

- *AC:* This is the most commonly used position. AC position permits triggering from 10 Hz to typically over 35 MHz (depending on your scope) and blocks any dc component of the sync trigger signal.
- *LF REJ*: The dc signals are rejected, and signals below 10 kHz are attenuated; the sweep is triggered by only the higher-frequency components of the signal. This is useful for providing stable triggering when the trigger signal contains line-frequency components, such as 60-Hz hum.

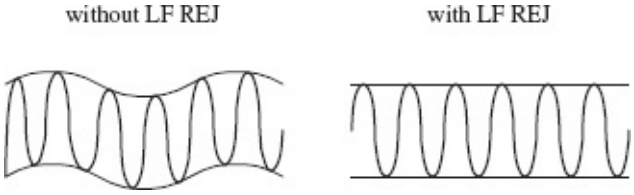


FIGURE 7.36

- *HF REJ*: Attenuates signals above 100 kHz. This is used to reduce high-frequency noise or to initiate

a trigger from the amplitude of a modulated envelope rather than the carrier frequency.

- *VIDEO*: Used to view composite video signals.
- *DC*: Permits triggering from dc to typically over 35 MHz. The DC coupling can be used to provide stable triggering for low-frequency signals that would otherwise be attenuated if measured in the AC setting. The LEVEL control can be adjusted to provide triggering at the desired dc level on the waveform.



FIGURE 7.37

7.4.6 Measuring Things with Scopes

A scope's buttons and knobs must be set properly to obtain accurate measurements. If just one of these buttons or switches is set wrong, things can go haywire. You must make sure every button is set correctly.

This section covers a number of applications for oscilloscopes, such as making phase measurements between two signals. For each application, the procedure is to set the scope to the initial settings listed next, and then adjust particular buttons and knobs to put the scope in the proper configuration needed for that particular application.

Initial Scope Settings

First, start with the following settings:

- *Power switch:* Off
- *Internal recurrent sweep (TRIGGER mode switch):* Off (NORM or AUTO position)
- *Focus:* Lowest setting
- *Gain:* Lowest setting
- *Intensity:* Lowest setting
- *Sync controls* (LEVEL, HOLDOFF): Lowest settings
- *Sweep selector:* External (EXT)
- *Vertical position control:* Midpoint
- *Horizontal position control:* Midpoint

Next, adjust the settings as follows:

- *Power switch:* On
- *Focus:* Until beam is in focus
- *Intensity:* Desired luminosity
- *Sweep selector:* Internal (use the linear internal sweep if more than one sweep is available)
- *Vertical position control:* Until beam is centered on display
- *Horizontal position control:* Until beam is centered on display
- *Internal recurrent sweep:* On; set the sweep frequency to any frequency above 100 Hz
- *Horizontal gain control:* Check that the luminous spot has expanded into a horizontal trace or line; return horizontal gain control to zero or lowest setting

- *Internal recurrent sweep:* Off
- *Vertical gain control:* To midpoint; touch the vertical input with a finger; the stray signal pickup should cause the spot to be deflected vertically into a trace or line; check that the line length is controllable by adjusting the vertical gain control; return the vertical gain control to zero or the lowest setting
- *Internal recurrent sweep:* On; advance the horizontal gain control to expand the spot into a horizontal line

Measuring a Sinusoidal Voltage Signal

1. Connect the equipment as shown in Fig. 7.38.

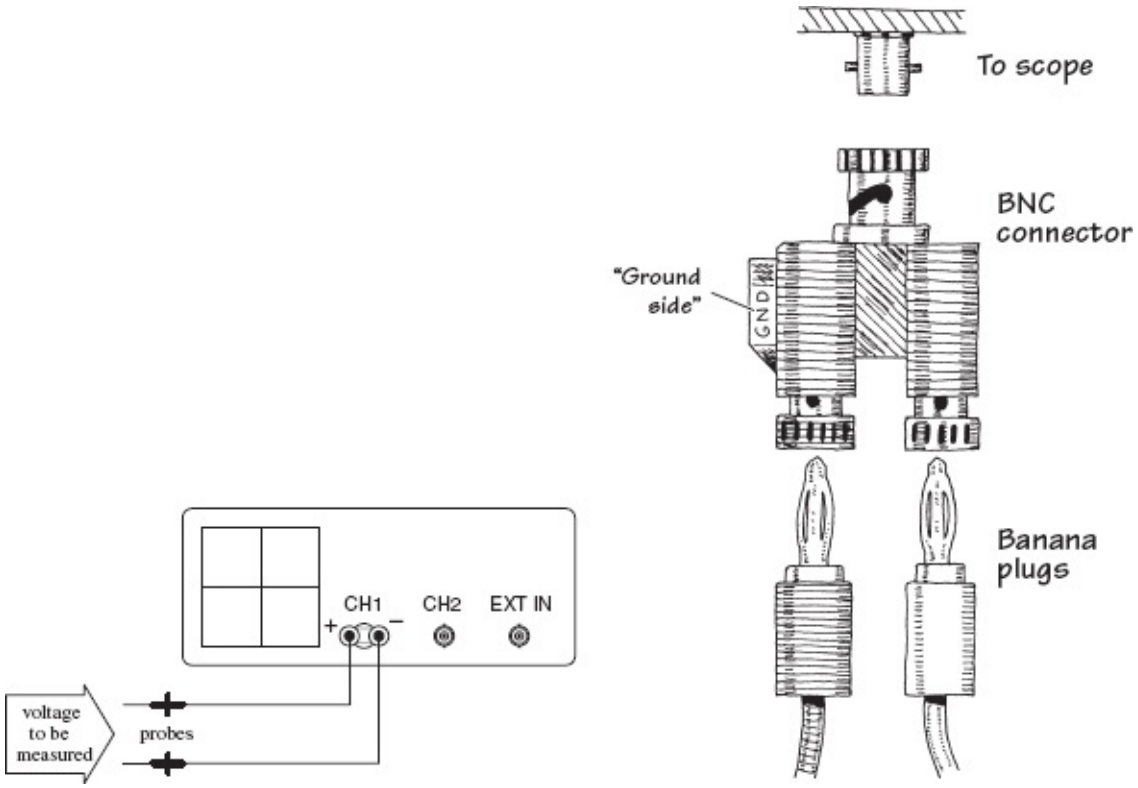
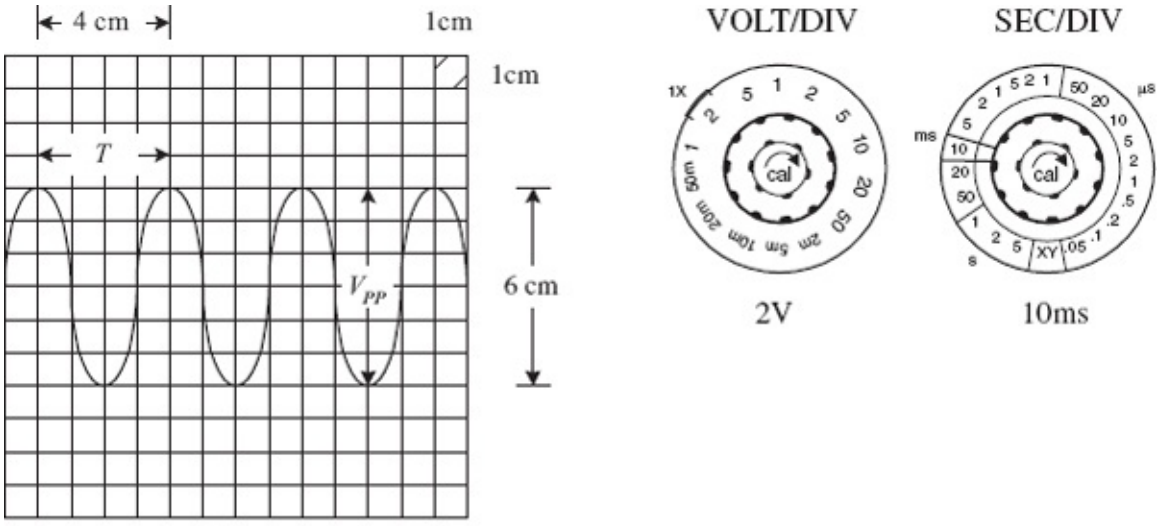


FIGURE 7.38

2. Set the scope to the initial settings listed in the previous section.
3. Fiddle with the vertical VOLT/DIV knob until the signal comes into view.
4. Set the input selector (AC/GRD/DC) to ground (GRD).
5. Switch the scope to the internal recurrent sweep. Fiddle with the SEC/DIV knob until the electron beam is tracing out a desired path of the screen.
6. Now you should have a horizontal line in view. Center this line on the x axis or some desired reference position by adjusting the vertical position knob. Make sure you do not fiddle around with the vertical position knob after it has been set to the desired reference point. If you do, your measurements will be offset. If you think you have accidentally moved the vertical position line, set the input selector to GRD and recalibrate.
7. Set the input selector switch (AC/GRD/DC) to DC. Connect the probe to the signal being measured.
8. Fiddle with the vertical and horizontal VOLT/DIV knob and SEC/DIV knob to get the signal into view.
9. Once you have an image of your signal on the screen, take a look at your VOLT/DIV and SEC/DIV knobs and record these settings. Now, visually measure the period, peak-to-peak voltage, and so on of the displayed image, using the centimeter grid lines on the scope's screen as a ruler. To find the actual voltages and times, multiply the measurement made in centimeters by the VOLT/DIV (or VOLTS/cm)

and SEC/DIV (SEC/cm) recorded set values. The example in Fig. 7.39 shows how to calculate the peak-to-peak voltage, root-mean-square voltage, period, and frequency of a sinusoidal waveform.



Peak-to-peak voltage (V_{pp}):

$$V_{pp} = (6\text{ cm})\frac{(2\text{ V})}{(1\text{cm})} = 12\text{ V}$$

Root-mean-square voltage (V_{rms})

$$V_{rms} = \frac{1}{\sqrt{2}} V_{pp} = 8.5\text{ V}$$

Period (T):

$$T = (4\text{ cm})\frac{(10\text{ cm})}{(1\text{ cm})} = 40\text{ ms}$$

Frequency (f):

$$f = \frac{1}{T} = \frac{1}{40\text{ ms}} = 25\text{ Hz}$$

FIGURE 7.39

Measuring Current

As mentioned earlier, oscilloscopes can measure only voltages; they do not directly measure currents. However, with the help of a resistor and Ohm's law, you can trick the scope into making current measurements. You simply measure the voltage drop across a resistor of known resistance and let $I = V/R$ do the rest. Typically, the resistor's resistance must be small to avoid disturbing the operating conditions within the circuit that is being measured. A high-precision 1-Ω resistor is often used for such instances.

Let's now take a look at the specifics of how to measure currents with a scope.

1. Set up your equipment as shown in Fig. 7.40.

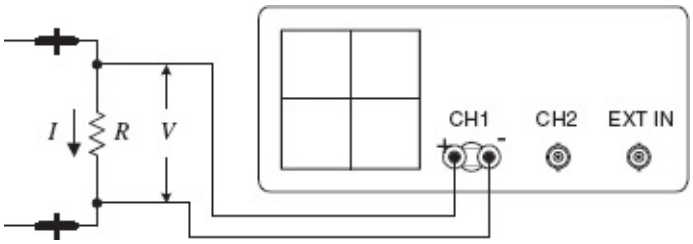


FIGURE 7.40

2. Set the scope to the initial settings listed earlier in the "Initial Scope Settings" section.
3. Apply a dc current to be measured through the resistor. Here, we'll use a 1-Ω resistor to make the calculations simple and to avoid altering the dynamics of the circuit being tested. The wattage of the resistor must be at least 2 Ω times the square of the maximum current (expressed in amps). For example, if the maximum anticipated current is 0.5 A, the minimum wattage of the resistor should be $2\ \Omega \times (0.5\text{ A})^2 = 1/2\text{ W}$.
4. Measure the voltage drop across the resistor using the scope. The unknown current will equal the

magnitude of the voltage measured, provided that you stick with the 1-W resistor. Figure 7.41 shows some example measurements, two of which describe how to measure RMS and total (dc + ac) effective currents.

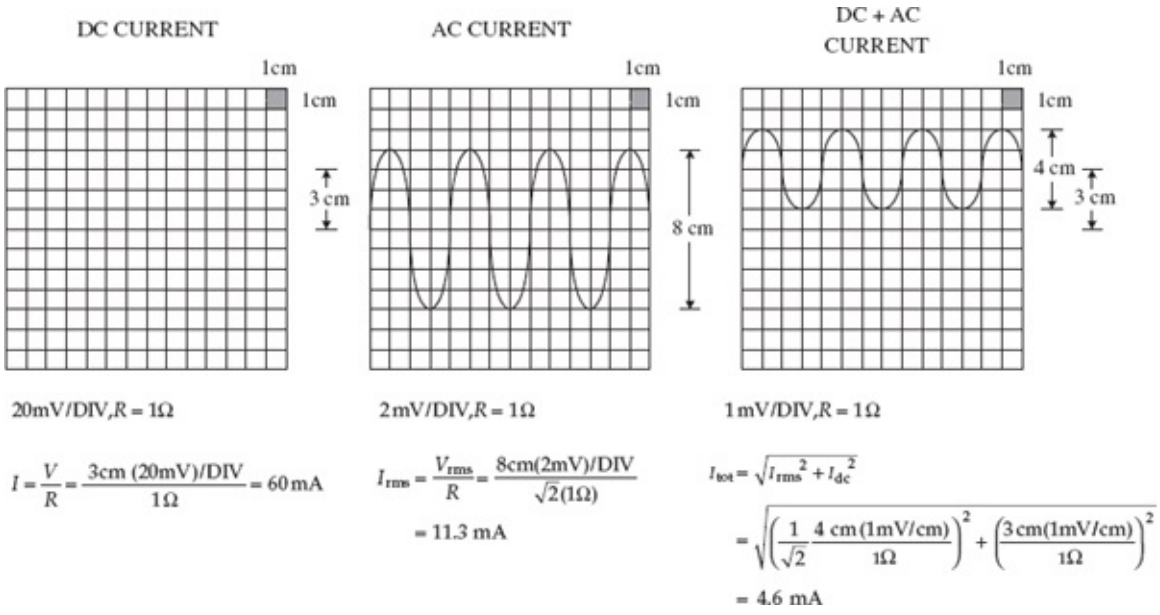


FIGURE 7.41

Phase Measurements Between Two Signals

Suppose that you wish to compare the phase relationship between two voltage signals. To do so, apply one of the signals to CH1 and the other to CH2. Then, using the DUAL setting (or BOTH setting), you can display both signals at the same time and align them side by side to compare the phased difference between them. Here's how to view both signals:

1. Step up your equipment as shown in Fig. 7.42.

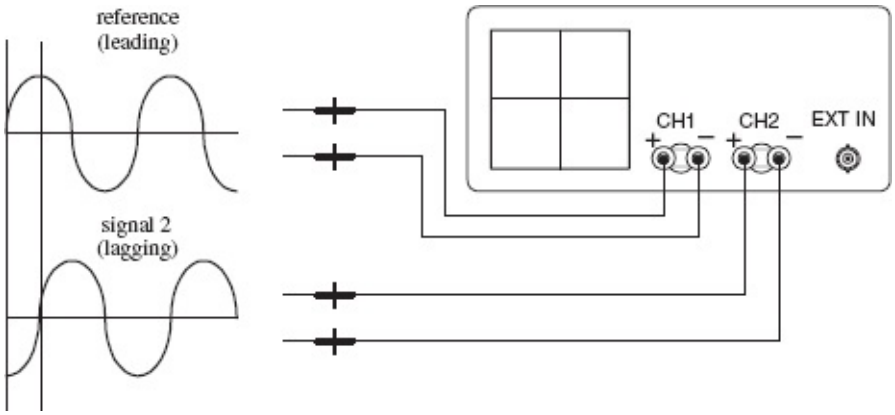


FIGURE 7.42

2. Set the scope to the initial settings listed earlier in the "Initial Scope Settings" section. Note that cables should be short, of the same length, and have similar electrical characteristics. At high frequencies, a difference in cable length or a difference in electrical characteristics between cables can introduce improper phase shifts.
3. Switch the scope's internal recurrent sweep to on.
4. Set the scope to dual trace (DUAL) mode.
5. Fiddle with the CH1 and CH2 VOLT/DIV settings until both signals are of similar amplitudes. This makes measuring phase differences easier.
6. Determine the phase factor of the reference signal. If one period (360°) of a signal is 8 cm, then 1 cm

equals one-eighth of 360°, or 45°. The 45° value represents the phase factor (see Fig. 7.43).

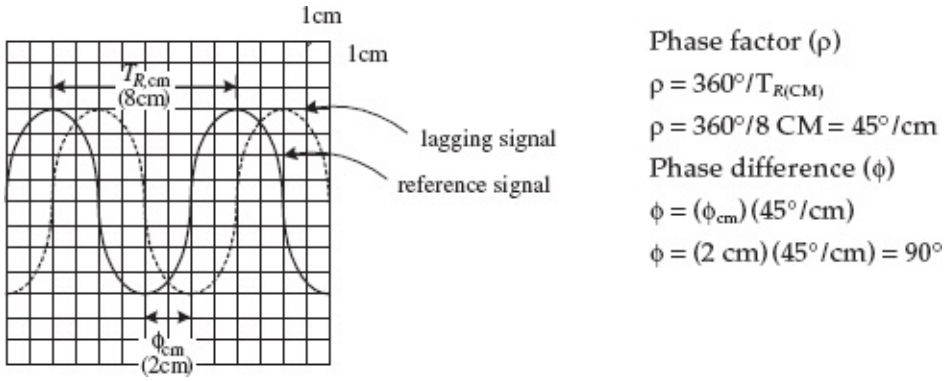


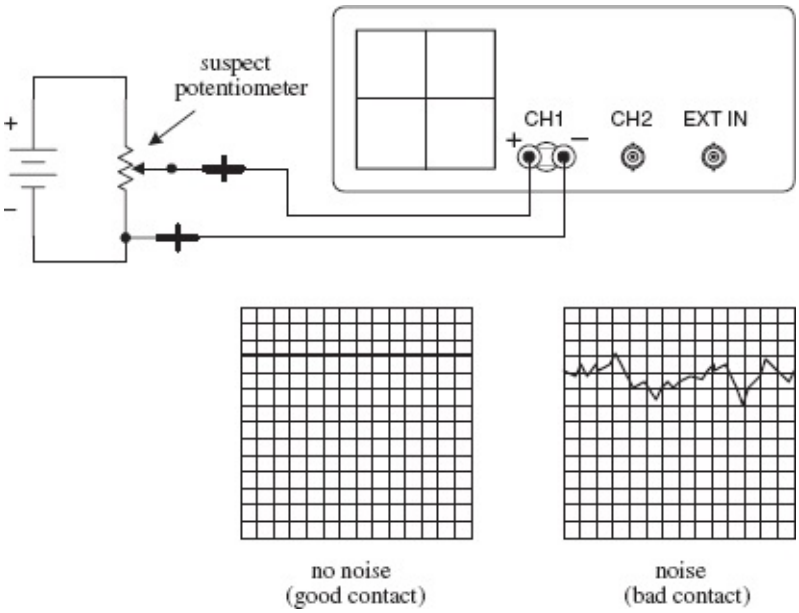
FIGURE 7.43

7. Measure the horizontal distance between corresponding points (for example, corresponding peaks or troughs) of the two waveforms. Multiply this measured distance by the phase factor to get the phase difference (see Fig. 7.43). For example, if the measured difference between the two signals is 2 cm, then the phase difference is 2 × 45°, or 90°.

7.4.7 Scope Applications

The ability for an oscilloscope to "freeze" a high-frequency waveform makes it an incredibly useful instrument for testing electronic components and circuits whose response curves, transient characteristics, phase relationships, and timing relationships are of fundamental importance. For example, scopes are used to study the shape of particular waveforms (such as squarewave, sawtooth, and so on). They are used to measure static noise (current variation caused by poor connections between components), pulse delays, impedances, digital signals, and other values. The list goes on. This section presents a few example scope applications.

Checking Potentiometers for Static Noise



Here, a scope is used to determine if the sliding contact of a potentiometer is faulty. A good potentiometer will present a solid voltage line on the scope's screen, whereas a bad potentiometer will present a noisy pattern on the display. Before concluding that a potentiometer is bad, make sure that noise was not present beforehand. For example, the cables used in this test may have been at fault.

FIGURE 7.44

Pulse Measurements

Scopes are often used to
definitions, shows some c

Rise time (t_r): **The time**
value

Fall time (tf): **The time**
value

Pulse width (t_w): **The ti**

Time delay (t_d): **The tir**
output pulse

Tilt: **A measure of the f**

$$\text{Percent tilt} = \frac{A}{B} \times 100\%$$

Overshoot: **A measure**

$$\text{Percent overshoot} = \frac{C}{D} \times$$

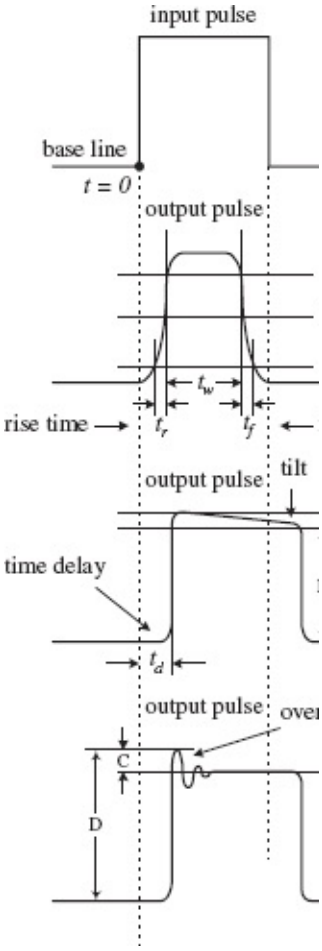


FIGURE 7.45

7.4.8 Measuring

The method of measu
output pulse. When t
and sent back along
impedance. The line
impedance (thing be
than the source impe

CHAPTER 8

Operational Amplifiers

Operational amplifiers (op amps) are incredibly useful high-performance differential amplifiers that can be employed in a number of amazing ways. A typical op amp is an integrated device with a noninverting input, an inverting input, two dc power supply leads (positive and negative), an output terminal, and a few other specialized leads used for fine-tuning. The positive and negative supply leads, as well as the fine-tuning leads, are often omitted from circuit schematics. If you do not see any supply leads, assume that a dual supply is being used.

Note that we have labeled the supply voltages $+V_s$ and $-V_s$, as they are usually the same. However, they do not need to be, as you will see when we look at single-supply op amps in this chapter.

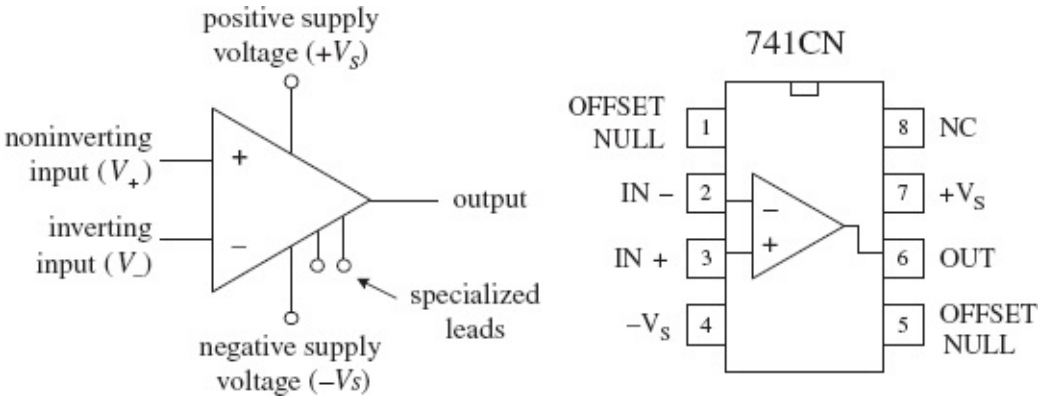


FIGURE 8.1

By itself, an op amp's operation is simple. If the voltage applied to the *inverting terminal* V_- is more positive than the voltage applied to the *noninverting terminal* V_+, the output saturates toward the *negative supply voltage* $-V_S$. Conversely, if $V_+ > V_-$, the output saturates toward the positive supply voltage $+V_S$ (see Fig. 8.2). This "maxing out" effect occurs with the slightest difference in voltage between the input terminals.

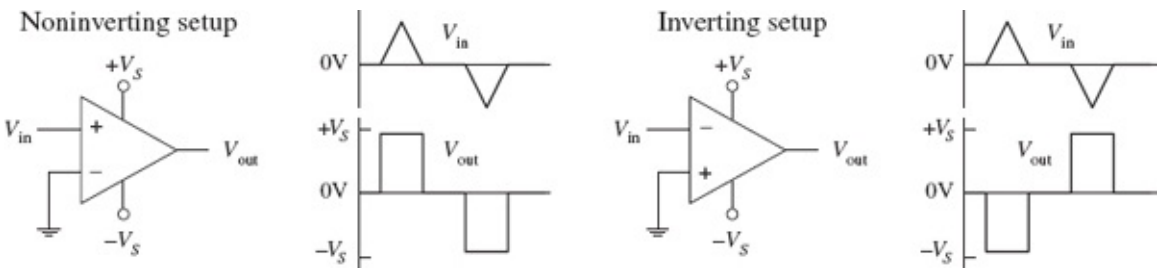


FIGURE 8.2

At first glance, it may appear that an op amp is not a very impressive device—it switches from one maximum output state to another whenever there's a voltage difference between its inputs. Big deal, right? By itself, it does indeed have limited applications. The trick to making op amps useful devices involves applying what is called *negative feedback*.

When voltage is “fed” back from the output terminal to the inverting terminal (this is referred to as *negative feedback*), the gain of an op amp can be controlled—the op amp’s output is prevented from saturating. For example, a feedback resistor R_F placed between the output and the inverting input, as shown in Fig. 8.3, acts to convey the state of the output back to the op amp’s input. This feedback information basically tells the op amp to readjust its output voltage to a value determined by the resistance of the feedback resistor. The circuit in Fig. 8.3, called an *inverting amplifier*, has an output equal to $-V_{in}/(R_F/R_{in})$ (you will learn how to derive this formula later in this chapter). The negative sign means that the output is inverted relative to the input. The gain is then simply the output voltage divided by the input voltage, or $-R_F/R_{in}$ (the negative sign indicates that the output is inverted relative to the input). As you can see from this equation, if you increase the resistance of the feedback resistor, there is an increase in the voltage gain. On the other hand, if you decrease the resistance of the feedback resistor, there is a decrease in the voltage gain.

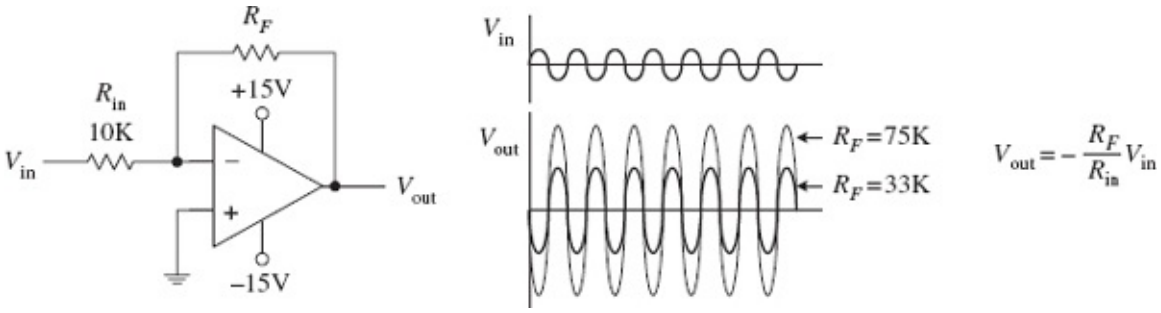


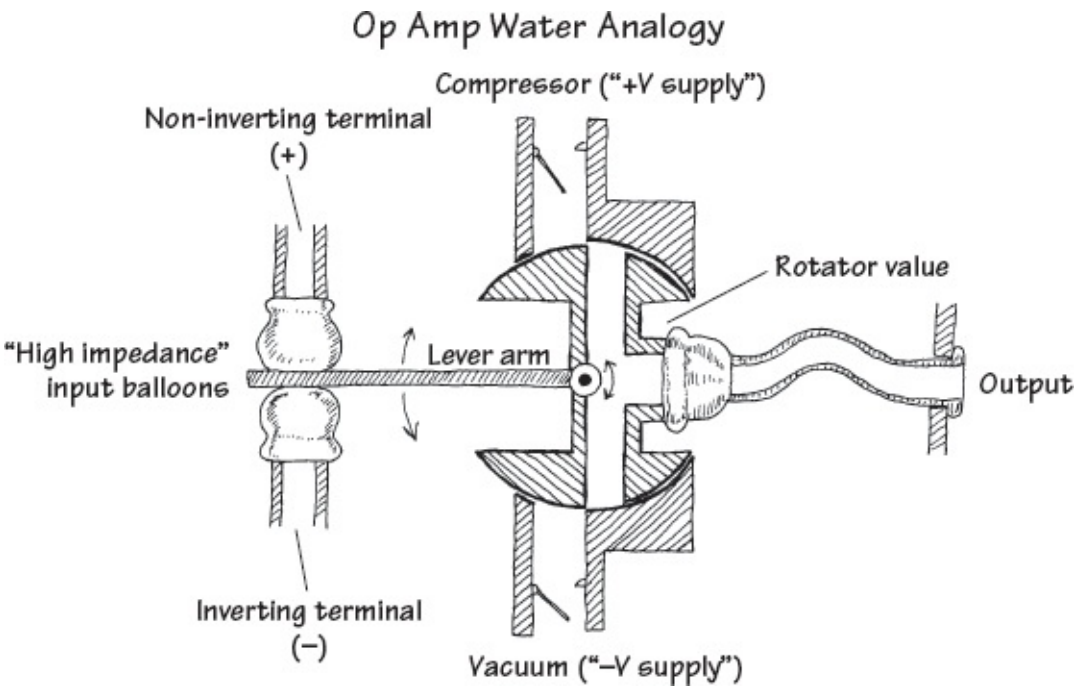
FIGURE 8.3

By adding other components to the negative-feedback circuit, an op amp can be made to do a number of interesting things besides pure amplification. Other interesting op amp circuits include voltage-regulator circuits, current-to-voltage converters, voltage-to-current converters, oscillator circuits, mathematical circuits (adders, subtractors, multipliers, differentiators, integrators, etc.), waveform generators, active filter circuits, active rectifiers, peak detectors, sample-and-hold circuits, etc. Most of these circuits will be covered in this chapter.

Besides negative feedback, there’s positive feedback, where the output is linked through a network to the noninverting input. Positive feedback has the opposite effect as negative feedback; it drives the op amp harder toward saturation. Although positive feedback is seldom used, it finds applications in special comparator circuits that are often used in oscillator circuits. Positive feedback also will be discussed in detail in this chapter.

8.1 Operational Amplifier Water Analogy

This is the closest thing I could come up with in terms of a water analogy for an op amp. To make the analogy work, you have to pretend that water pressure is analogous to voltage and water flow is analogous to current flow.



The inverting and noninverting terminals of the water op amp are represented by the two tubes with elastic balloon ends. When the water pressure applied to both input tubes is equal, the lever arm is centered. However, if the water pressure applied to the noninverting tube is made larger than the pressure applied to the inverting tube, the noninverting balloon expands and forces the lever arm downward. The lever arm then rotates the rotator valve counterclockwise, thus opening a canal from the compressor tube (analogous to the positive supply voltage) to the output tube. (This is analogous to an op amp saturating in the positive direction whenever the noninverting input is more positive in voltage than the inverting input.) Now, if the pressure applied at the non-inverting tube becomes less than the pressure applied at the inverting tube, the lever arm is pushed upward by the inverting balloon. This causes the rotator valve to rotate clockwise, thus opening the canal from the vacuum tube (analogous to the negative supply voltage) to the output. (This is analogous to an op amp saturating in the negative direction whenever the inverting input is made more positive in voltage than the noninverting input.) See what you can do with the analogy in terms of explaining negative feedback. Also note that in the analogy there is an infinite "input water impedance" at the input tubes, while there is a zero "output water impedance" at the output tube. As you will see, ideal op amps also have similar input and output impedance. In real op amps, there are always some leakage currents.

FIGURE 8.4

8.2 How Op Amps Work (The "Cop-Out" Explanation)

An op amp is an integrated device that contains a large number of transistors, several resistors, and a few capacitors. Figure 8.5 shows a schematic diagram of a typical low-cost general-purpose bipolar operational amplifier.

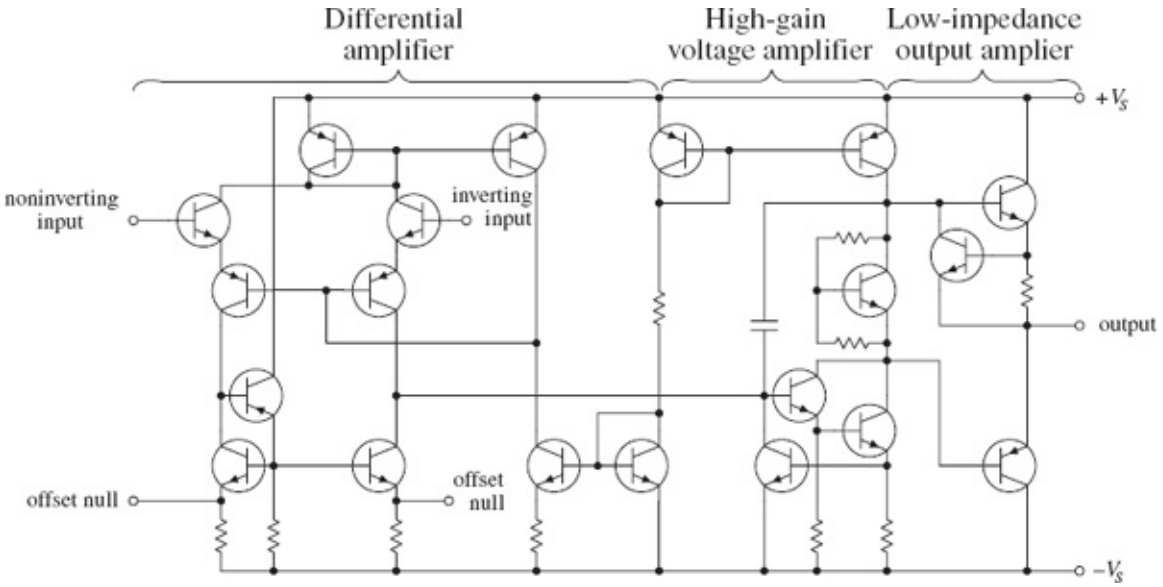


FIGURE 8.5

This op amp basically consists of three stages: a high-input-impedance differential amplifier, a high-gain voltage amplifier with a level shifter (permitting the output to swing positive and negative), and a low-impedance output amplifier. However, realizing that an op amp is composed of various stages does not help you much in terms of figuring out what will happen between the input and output leads. That is, if you attempt to figure out what the currents and voltages are doing within the complex system, you will be asking for trouble. It is just too difficult a task. What is important here is not to focus on understanding the op amp's internal circuitry but instead to focus on memorizing some rules that individuals came up with that require only working with the input and output leads. This approach seems like a "cop-out," but it works.

8.3 Theory

There is essentially only one formula you will need to know for solving op amp circuit problems. This formula is the foundation on which everything else rests. It is the expression for an op amp's output voltage as a function of its input voltages V_+ (noninverting) and V_- (inverting) and of its *open-loop voltage gain* A_o*:*

$$V_{out} = A_o(V_+ - V_-)$$

This expression says that an *ideal op amp* acts like an ideal voltage source that supplies an output voltage equal to $A_o(V_+ - V_-)$ (see Fig. 8.6). Things can get a little more complex when we start talking about *real op amps*, but generally, the open-loop voltage expression above pretty much remains the same, except now we have to make some slight modifications to our equivalent circuit. These modifications must take into account the nonideal features of an op amp, such as its input resistance R_{in} and output resistance R_{out}. Figure 8.6 *right* shows a more realistic equivalent circuit for an op amp.

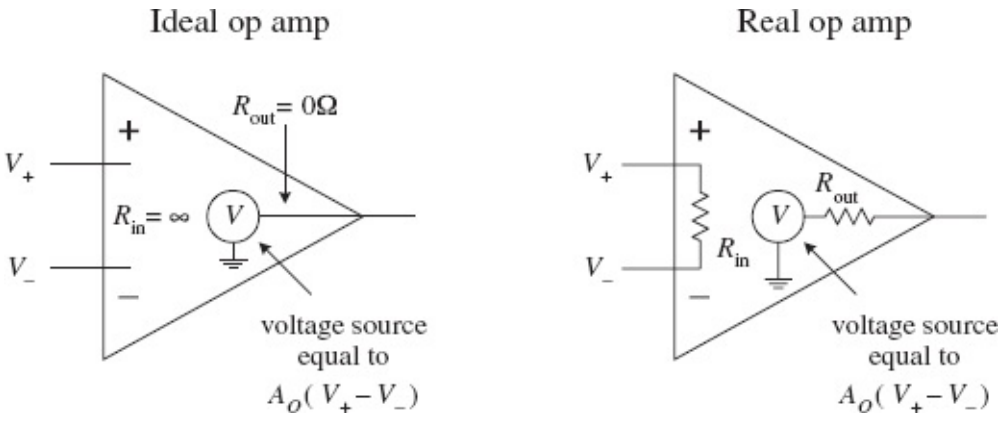


FIGURE 8.6

To give meaning to the open-loop voltage gain expression and to the ideal and real equivalent circuits, the values of A_o, R_{in}, and R_{out} are defined within the following rules:

Rule 1: For an ideal op amp, the open-loop voltage gain is infinite ($A_o = \infty$). For a real op amp, the gain is a finite value, typically between 10^4 to 10^6.

Rule 2: For an ideal op amp, the input impedance is infinite ($R_{in} = \infty$). For a real op amp, the input impedance is finite, typically between 10^6 (e.g., typical bipolar op amp) to 10^{12} Ω (e.g., typical JFET

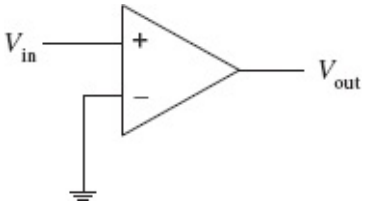
op amp). The output impedance for an ideal op amp is zero ($R_{out} = 0$). For a real op amp, R_{out} is typically between 10 to 1000 Ω.

Rule 3: The input terminals of an ideal op amp draw no current. Practically speaking, this is true for a real op amp as well—the actual amount of input current is usually (but not always) insignificantly small, typically within the picoamps (e.g., typical JFET op amp) to nanoamps (e.g., typical bipolar op amp) range.

Now that you are armed with $V_{out} = A_o(V_+ - V_-)$ and rules 1 through 3, let's apply them to a few simple example problems.

EXAMPLE 1

Solve for the gain (V_{out}/V_{in}) of the circuit below.



Since V_- is grounded (0 V) and V_+ is simply V_{in}, you can plug these values into the open-loop voltage gain expression:

$$V_{out} = A_o(V_+ - V_-)$$

$$= A_o(V_{in} - 0\ V) = A_o V_{in}$$

Rearranging this equation, you get the expression for the gain:

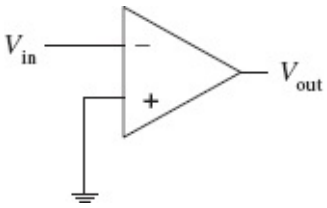
$$\text{Gain} = \frac{V_{out}}{V_{in}} = A_o$$

If you treat the op amp as ideal, A_o would be infinite. However, if you treat the op amp as real, A_o is finite (around 10^4 to 10^6). This circuit acts as a simple noninverting comparator that uses ground as a reference. If $V_{in} > 0$ V, the output ideally goes to $+\infty$ V; if $V_{in} < 0$V, the output ideally goes to $-\infty$ V. With a real op amp, the output is limited by the supply voltages (which are not shown in the drawing but assumed). The exact value of the output voltage is slightly below and above the positive and negative supply voltages, respectively. These maximum output voltages are called the *positive* and *negative saturation voltages*.

FIGURE 8.7

EXAMPLE 2

Solve for the gain (V_{out}/V_{in}) of the circuit below.



Since V_+ is grounded (0V) and V_- is simply V_{in}, you can substitute these values into the open-loop voltage gain expression:

$$V_{out} = A_o(V_+ - V_-)$$

$$= A_o(0\ V - V_{in}) = -A_o V_{in}$$

Rearranging this equation, you get the expression for the gain:

$$\text{Gain} = \frac{V_{out}}{V_{in}} = -A_o$$

If you treat the op amp as ideal, $-A_o$ is negatively infinite. However, if you treat the op amp as real, $-A_o$ is finite (around -10^4 to -10^6). This circuit acts as a simple inverting comparator that uses ground as a reference. If $V_{in} > 0$ V, the output ideally goes to $-\infty$ V; if $V_{in} < 0$ V, the output ideally goes to $+\infty$ V. With a real op amp, the output swings are limited to the saturation voltages.

FIGURE 8.8

8.4 Negative Feedback

Negative feedback is a wiring technique where some of the output voltage is sent back to the inverting terminal. This voltage can be "sent" back through a resistor, capacitor, or complex circuit or simply can be sent back through a wire. So exactly what kind of formulas do you use now? Well, that depends on the feedback circuit, but in reality, there is nothing all that new to learn. In fact, there is really only one formula you need to know for negative-feedback circuits (you still have to use the rules, however). This formula looks a lot like our old friend $V_{out} = A_o(V_+ - V_-)$. There is, however, the V_- in the formula—this you must reconsider. V_- in the formula changes because now the output voltage from the op amp is "giving" extra voltage (positive or negative) back to the inverting terminal. What this means is that you must replace V_- with fV_{out}, where f is a fraction of the voltage "sent" back from V_{out}. That's the trick!

There are two basic kinds of negative feedback, voltage feedback and operational feedback, as shown in Fig. 8.9.

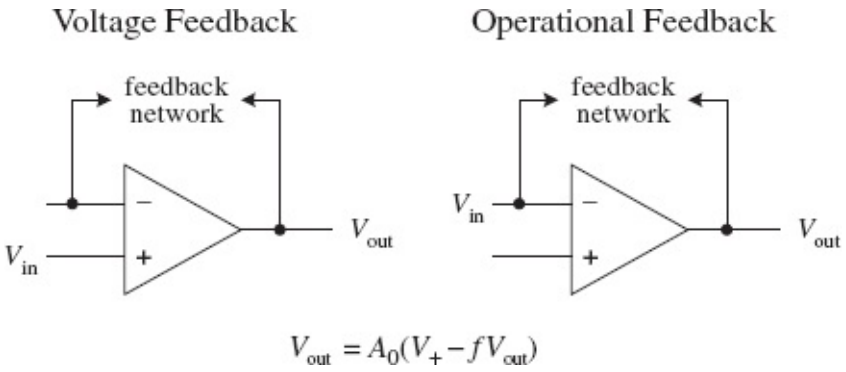


FIGURE 8.9

Now, in practice, figuring out what the fraction f should be is not important. That is, you do not have to calculate it explicitly. The reason why I have introduced it in the open-loop voltage expression is to provide you with a bit of basic understanding as to how negative feedback works in theory. As it turns out, there is a simple trick for making op amp circuits with negative feedback easy to calculate. The trick is as follows: If you treat an op amp as an ideal device, you will notice that if you rearrange the open-loop voltage expression into $V_{out}/A_o = (V_+ - V_-)$, the left side of the equation goes to zero—A_o is infinite for an ideal op amp. What you get in the end is then simply $V_+ - V_- = 0$. This result is incredibly important in terms of simplifying op amp circuits with negative feedback—so important that the result receives its own rule (the fourth and final rule).

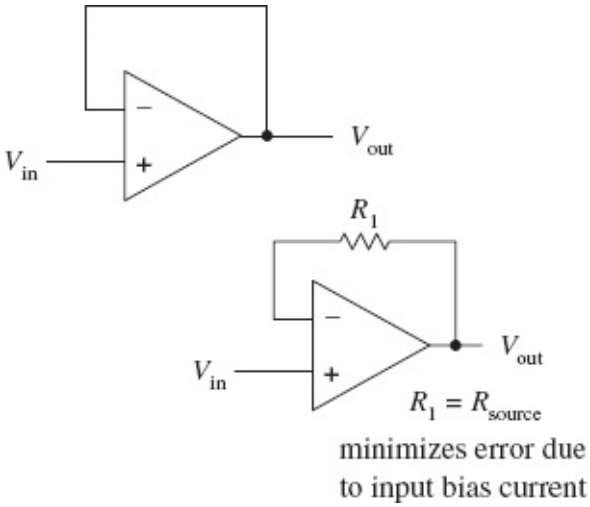
Rule 4: Whenever an op amp senses a voltage difference between its inverting and noninverting inputs, it responds by feeding back as much current/voltage through the feedback network as is necessary to keep this difference equal to zero ($V_+ - V_- = 0$). This rule only applies for negative feedback.

The following sample problems are designed to show you how to apply rule 4 (and the other rules) to op amp circuit problems with negative feedback.

Negative Feedback Example Problems

BUFFER (UNITY GAIN AMPLIFIER)

Solve for the gain (V_{out}/V_{in}) of the circuit below.



Since you are dealing with negative feedback, you can apply rule 4, which says that the output will attempt to make $V_+ - V_- = 0$. By examining the simple connections, notice that $V_{in} = V_+$ and $V_- = V_{out}$. This means that $V_{in} - V_{out} = 0$. Rearranging this expression, you get the gain:

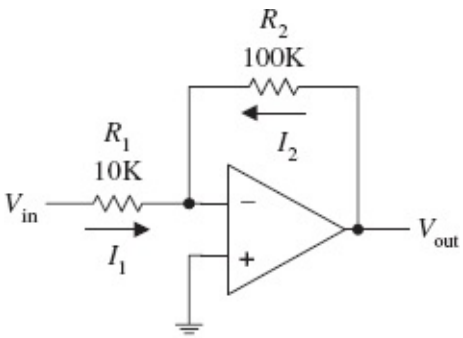
$$\text{Gain} = \frac{V_{out}}{V_{in}} = 1$$

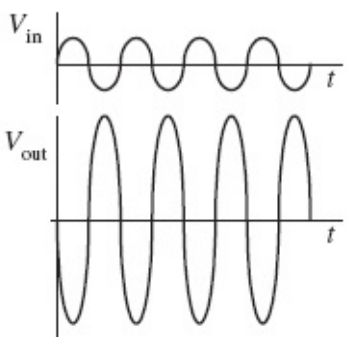
A gain of 1 means that there is no amplification; the op amp's output follows its input. At first glance, it may appear that this circuit is useless. However, it is important to recall that an op amp's input impedance is huge, while its output impedance is extremely small (rule 2). This feature makes this circuit useful for circuit-isolation applications. In other words, the circuit acts as a buffer. With real op amps, it may be necessary to throw in a resistor in the feedback loop (lower circuit). The resistor acts to minimize voltage offset errors caused by input bias currents (leakage). The resistance of the feedback resistor should be equal to the source resistance. I will discuss input bias currents later in this chapter.

FIGURE 8.10

INVERTING AMPLIFIER

Solve for the gain (V_{out}/V_{in}) of the circuit below.





Because you have negative feedback, you know the output will attempt to make the difference between V_+ and V_- zero. Since V_+

is grounded (0 V), this means that V_- also will be 0 V (rule 4). To figure out the gain, you must find currents I_1 and I_2 so you can come up with an expression containing V_{out} in terms of V_{in}. Using Ohm's law, you find I_1 and I_2 to be

$$I_1 = \frac{V_{in} - V_-}{R_1} = \frac{V_{in} - 0\,V}{R_1} = \frac{V_{in}}{R_1}$$

$$I_2 = \frac{V_{out} - V_-}{R_2} = \frac{V_{out} - 0\,V}{R_2} = \frac{V_{out}}{R_2}$$

Because an ideal op amp has infinite input impedance, no current will enter its inverting terminal (rule 3). Therefore, you can apply Kirchhoff's junction rule to get $I_2 = -I_1$. Substituting the calculated values of I_1 and I_2 into this expression, you get $V_{out}/R_2 = -V_{in}/R_1$. Rearranging this expression, you find the gain:

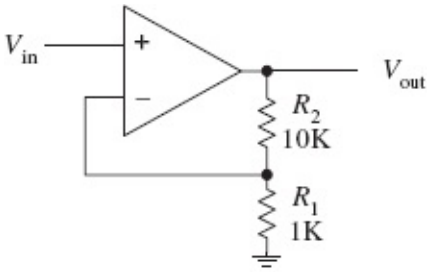
$$\text{Gain} = \frac{V_{out}}{V_{in}} = -\frac{R_2}{R_1}$$

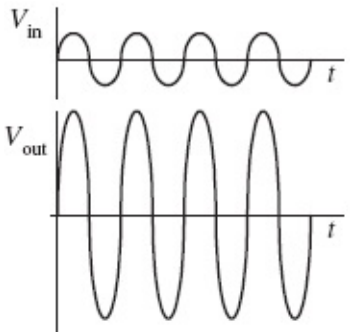
The negative sign tells you that the signal that enters the input will be inverted (shifted 180°). Notice that if $R_1 = R_2$, the gain is −1 (the negative sign simply means the output is inverted). In this case you get what's called a *unity-gain inverter*, or an *inverting buffer*. When using real op amps that have relatively high input bias currents (e.g., bipolar op amps), it may be necessary to place a resistor with a resistance equal to $R_1 \| R_2$ between the noninverting input and ground to minimize voltage offset errors.

FIGURE 8.11

NONINVERTING AMPLIFER

Solve for the gain (V_{out}/V_{in}) of the circuit below.





By inspection, you know that $V_+ = V_{in}$. By applying rule 4, you then can say that $V_- = V_+$. This means that $V_- = V_{in}$. To come up with an expression relating V_{in} and V_{out} (so that you can find the gain), the voltage divider relation is used:

$$V_- = \frac{R_1}{R_1 + R_2} V_{out} = V_{in}$$

Rearranging this equation, you find the gain:

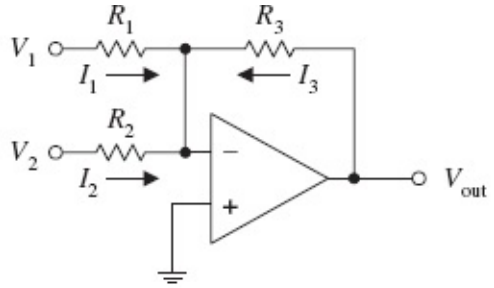
$$\text{Gain} = \frac{V_{out}}{V_{in}} = \frac{R_1 + R_2}{R_1} = 1 + \frac{R_2}{R_1}$$

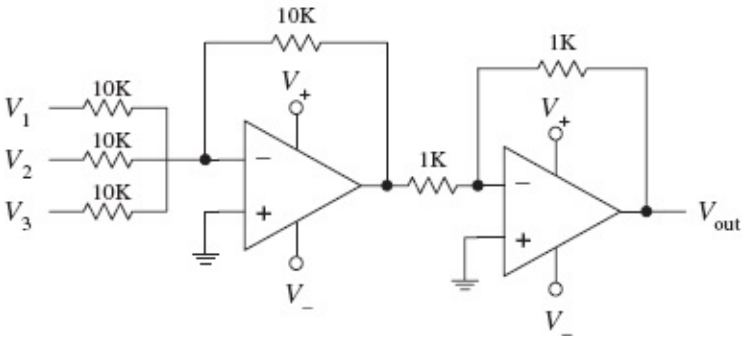
Unlike the inverting amplifier, this circuit's output is in phase with its input—the output is "noninverted." With real op amps, to minimize voltage offset errors due to input bias current, set $R_1 \| R_2 = R_{source}$.

FIGURE 8.12

SUMMING AMPLIFIER

Solve for V_{out} in terms of V_1 and V_2.





Since you know that V_+ is grounded (0 V), and since you have feedback in the circuit, you can say that $V_+ = V_- = 0$ V (rule 4). Now that you know V_+, solve for I_1, I_2, and I_3 in order to come up with an expression relating V_{out} with V_1 and V_2. The currents are found by applying Ohm's law:

$$I_1 = \frac{V_1 - V_-}{R_1} = \frac{V_1 - 0\,\text{V}}{R_1} = \frac{V_1}{R_1}$$

$$I_2 = \frac{V_2 - V_-}{R_2} = \frac{V_2 - 0\,\text{V}}{R_2} = \frac{V_2}{R_2}$$

$$I_3 = \frac{V_{out} - V_-}{R_3} = \frac{V_{out} - 0\,\text{V}}{R_3} = \frac{V_{out}}{R_3}$$

Like the last problem, assume that no current enters the op amp's inverting terminal (rule 3). This means that you can apply Kirchhoff's junction rule to combine I_1, I_2, and I_3 into one expression: $I_3 = -(I_1 + I_2) = -I_1 - I_2$. Plugging the results above into this expression gives the answer:

$$V_{out} = -\frac{R_3}{R_1}V_1 - \frac{R_3}{R_2}V_2 = -\left(\frac{R_3}{R_1}V_1 + \frac{R_3}{R_2}V_2\right)$$

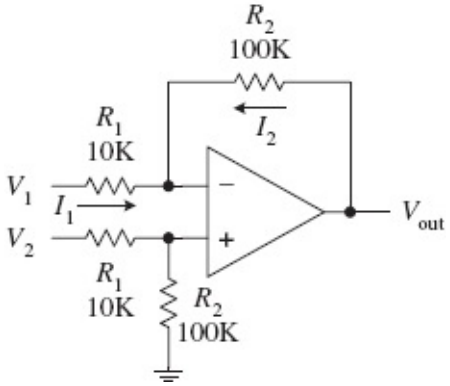
If you make $R_1 = R_2 = R_3$, $V_{out} = -(V_1 + V_2)$. Notice that the sum is negative. To get a positive sum, you can add an inverting stage, as shown in the lower circuit. Here, three inputs are added together to yield the following output: $V_{out} = V_1 + V_2 + V_3$. Again, for some real op amps, an additional input-bias compensation resistor placed between the noninverting input and ground

may be needed to avoid offset error caused by input bias current. Its value should be equal to the parallel resistance of all the input resistors.

FIGURE 8.13

DIFFERENCE AMPLIFIER

Determine V_{out}.



First, you determine the voltage at the noninverting input by using the voltage divider relation (again, assume that no current enters the inputs):

$$V_+ = \frac{R_2}{R_1 + R_2} V_2$$

Next, apply Kirchhoff's current junction law to the inverting input ($I_1 = I_2$):

$$\frac{V_1 - V_-}{R_1} = \frac{V_- - V_{out}}{R_2}$$

Using rule 4 ($V_+ = V_-$), substitute the V_+ term in for V_- in the last equation to get

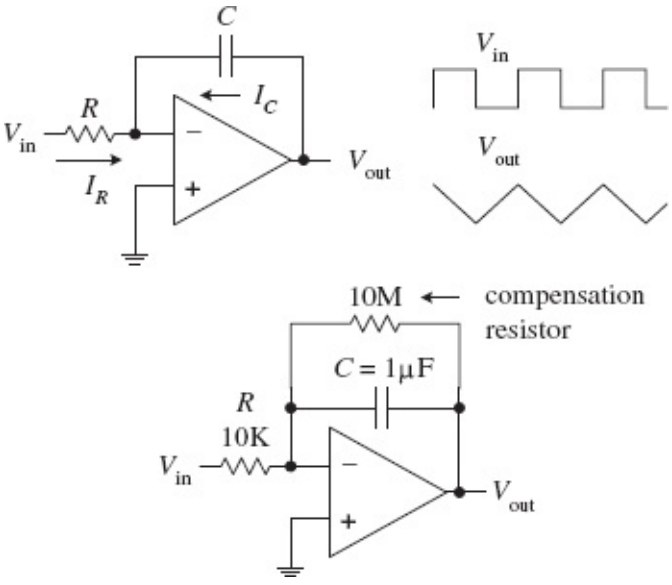
$$V_{out} = \frac{R_2}{R_1}(V_2 - V_1)$$

If you set $R_1 = R_2$, then $V_{out} = V_2 - V_1$.

FIGURE 8.14

INTEGRATOR

Solve for V_{out} in terms of V_{in}.



Because you have feedback, and because V_+ = 0 V, you can say that V_- is 0 V as well (rule 4). Now that you know V_-, solve for I_R and I_C so that you can come up with an expression relating V_{out} with V_{in}. Since no current enters the input of an op amp (rule 3), the displacement current I_C through the capacitor and the current I_R through the resistor must be related by $I_R + I_C = 0$. To find I_R, use Ohm's law:

$$I_R = \frac{V_{in} - V_-}{R} = \frac{V_{in} - 0\,\text{V}}{R} = \frac{V_{in}}{R}$$

I_C is found by using the displacement current relation:

$$I_C = C\frac{dV}{dt} = C\frac{d(V_{out} - V_-)}{dt} = C\frac{d(V_{out} - 0\,\text{V})}{dt} = C\frac{dV_{out}}{dt}$$

Placing these values of I_C and I_R into $I_R + I_C = 0$ and rearranging, you get the answer:

$$dV_{out} = -\frac{1}{RC}V_{in}\,dt$$

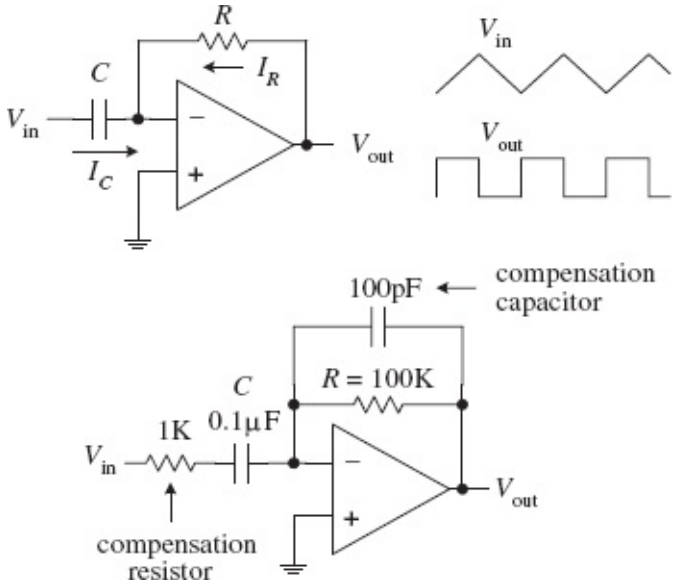
$$V_{out} = -\frac{1}{RC}V_{in}t$$

Such a circuit is called an *integrator;* the input signal is integrated at the output. Now, one problem with the first circuit is that the output tends to drift, even with the input grounded, due to nonideal characteristics of real op amps such as voltage offsets and bias current. A large resistor placed across the capacitor can provide dc feedback for stable biasing. Also, a compensation resistor may be needed between the noninverting terminal and ground to correct voltage offset errors caused by input bias currents. The size of this resistor should be equal to the parallel resistance of the input resistor and the feedback compensation resistor.

FIGURE 8.15

DIFFERENTIATOR

Solve for V_{out} in terms of V_{in}.



Since you know that V_+ is grounded (0 V), and since you have feedback in the circuit, you can say that V_- = V_+ = 0 V (rule 4). Now that you know V_-, solve for I_R and I_C so that you can come up with an expression relating V_{out} with V_{in}. Since no current enters the input of an op amp (rule 3), the displacement current I_C through the capacitor and the current I_R through the resistor must be related by $I_R + I_C = 0$. To find I_C, use the displacement current equation:

$$I_C = C\frac{dV}{dt} = C\frac{d(V_{in} - V_-)}{dt} = C\frac{d(V_{in} - 0\,V)}{dt} = C\frac{dV_{in}}{dt}$$

The current I_R is found using Ohm's law:

$$I_R = \frac{V_{out} - V_-}{R} = \frac{V_{out} - 0\,V}{R} = \frac{V_{out}}{R}$$

Placing these values of I_C and I_R into $I_R + I_C = 0$ and rearranging, you get the answer:

$$V_{out} = -RC\frac{dV_{in}}{dt}$$

Such a circuit is called a *differentiator;* the input signal is differentiated at the output. The first differentiator circuit shown is not in practical form. It is extremely susceptible to noise due to the op amp's high ac gain. Also, the feedback network of the differentiator acts as an *RC* low-pass filter that contributes a 90° phase lag within the loop and may cause stability problems. A more practical differentiator is shown below the first circuit. Here, both stability and noise problems are corrected with the addition of a feedback capacitor and input resistor. The additional components provide high-frequency rolloff to reduce high-frequency noise. These components also introduce a 90° lead to cancel the 90° phase lag. The effect of the additional components, however, limits the maximum frequency of operation—at very high frequencies, the differentiator becomes an integrator. Finally, an additional input-bias compensation resistor placed between the noninverting input and ground may be needed to avoid offset error caused by input bias current. Its value should be equal to the resistance of the feedback resistor.

FIGURE 8.16

8.5 Positive Feedback

Positive feedback involves sending output voltage back to the noninverting input. In terms of the theory, if you look at our old friend $V_{out} = A_o(V_+ - V_-)$, the V_+ term changes to fV_{out} (f is a fraction of the voltage sent back), so you get $V_{out} = A_o(fV_{out} - V_-)$. Now, an important thing to notice about this equation (and about positive feedback in general) is that the voltage fed back to the noninverting input will act to drive the op amp "harder" in the direction the output is going (toward saturation). This makes sense in terms of the equation; fV_{out} adds to the expression. Recall that negative feedback acted in the opposite way; the fV_{out} (= V_-) term subtracted from the expression, preventing the output from "maxing out." In electronics, positive feedback is usually a bad thing, whereas negative feedback is a good thing. For most applications, it is desirable to control the gain (negative feedback), while it is undesirable to go to the extremes (positive feedback).

There is, however, an important use for positive feedback. When using an op amp to make a comparator, positive feedback can make output swings more pronounced. Also, by adjusting the size of the feedback resistor, a comparator can be made to experience what is called *hysteresis*. In effect, hysteresis gives the comparator two thresholds. The voltage between the two thresholds is called the *hysteresis voltage*. By obtaining two thresholds (instead of merely one), the comparator circuit becomes more immune to noise that can trigger unwanted output swings. To better understand hysteresis, let's take a look at the following comparator circuit that incorporates positive feedback.

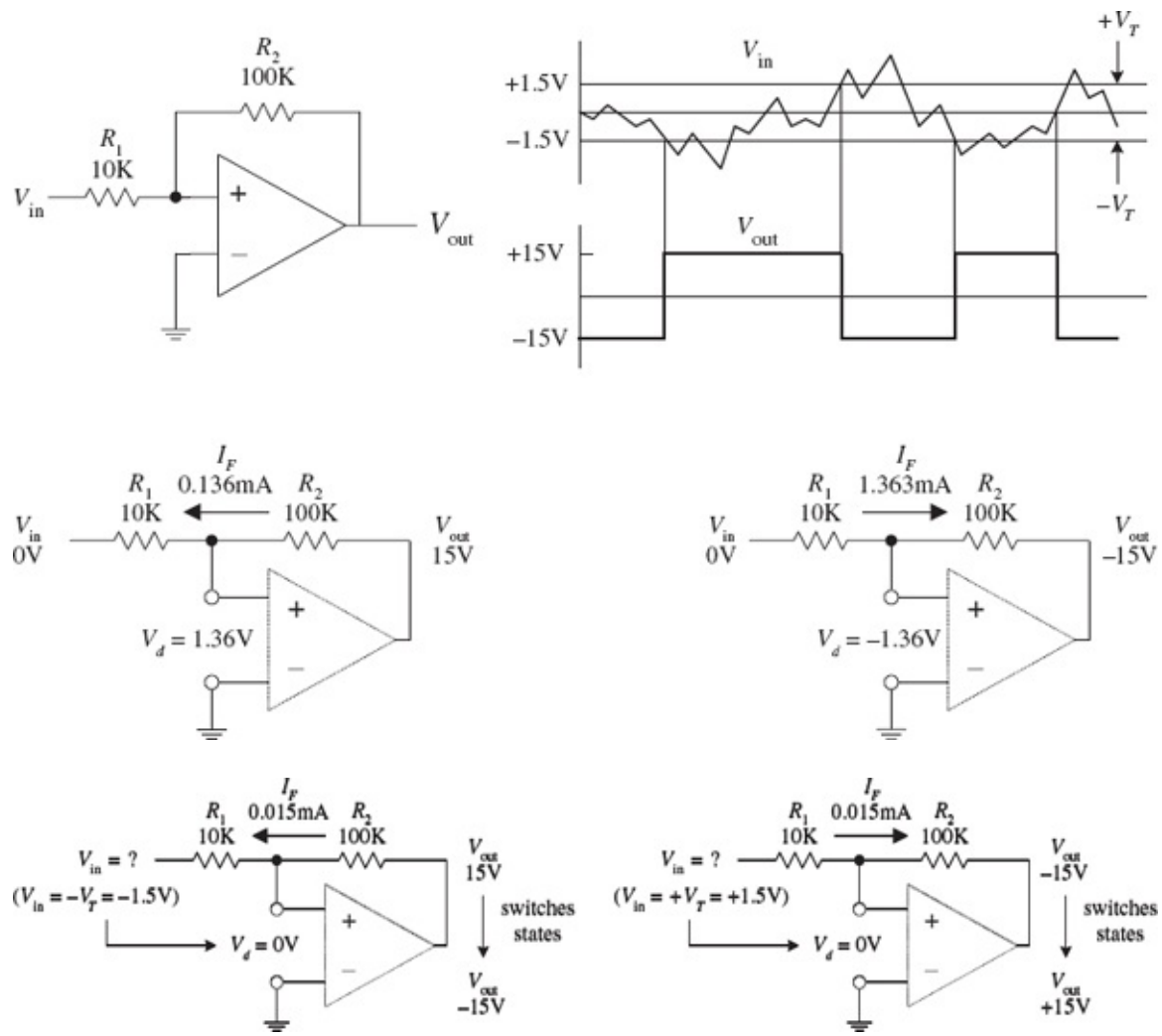


FIGURE 8.17

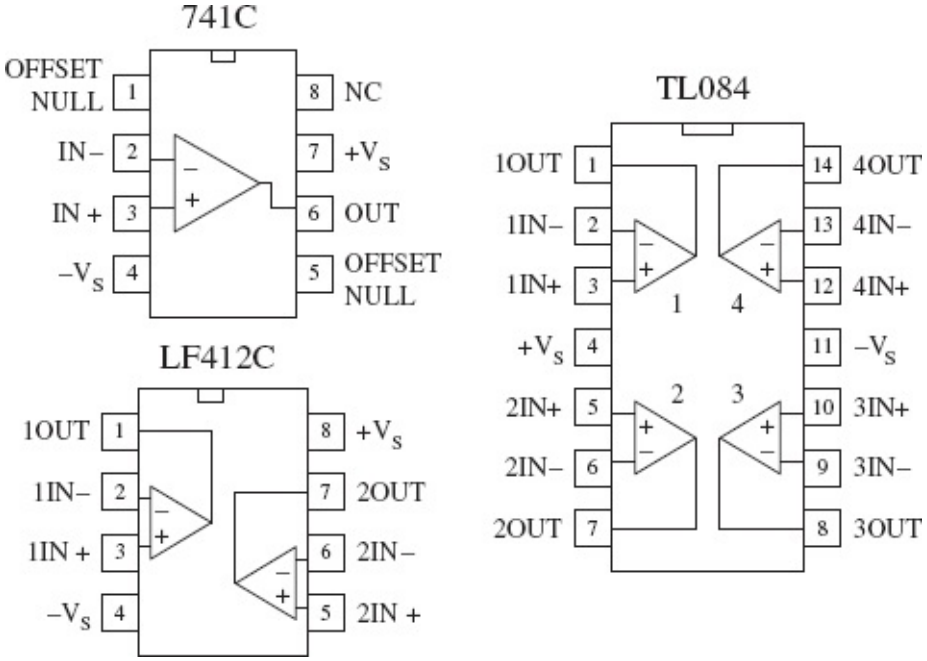
Assume that the op amp's output is at positive saturation, say, +15 V. If V_{in} is 0 V, the voltage difference between the inverting input and noninverting input (V_d) will be 1.36 V. You get this by using Ohm's law:

$$I_F = (V_{out} - V_{in})/(R_1 + R_2)$$
$$V_d = I_F R_1$$

This does not do anything to the output; it remains at +15 V. However, if you reduce V_{in}, there is a point when V_d goes to 0 V, at which time the output switches states. This voltage is called the *negative threshold voltage* ($-V_T$). The negative threshold voltage can be determined by using the previous two equations—the end result being $-V_T = -V_{out}/(R_2/R_1)$. In the example, $-V_T = -1.5$ V. Now, if the output is at negative saturation (−15 V) and 0 V is applied to the input, $V_d = -1.36$ V. The output remains at −15 V. However, if the input voltage is increased, there is a point where V_d goes to zero and the output switches states. This point is called the *positive threshold voltage* ($+V_T$), which is equal to $+V_{out}/(R_2/R_1)$. In the example, $+V_T = +1.5$ V. Now the difference between the two saturation voltages is the hysteresis voltage: $V_h = +V_T - (-V_T)$. In the example, $V_h = 3$ V.

8.6 Real Kinds of Op Amps

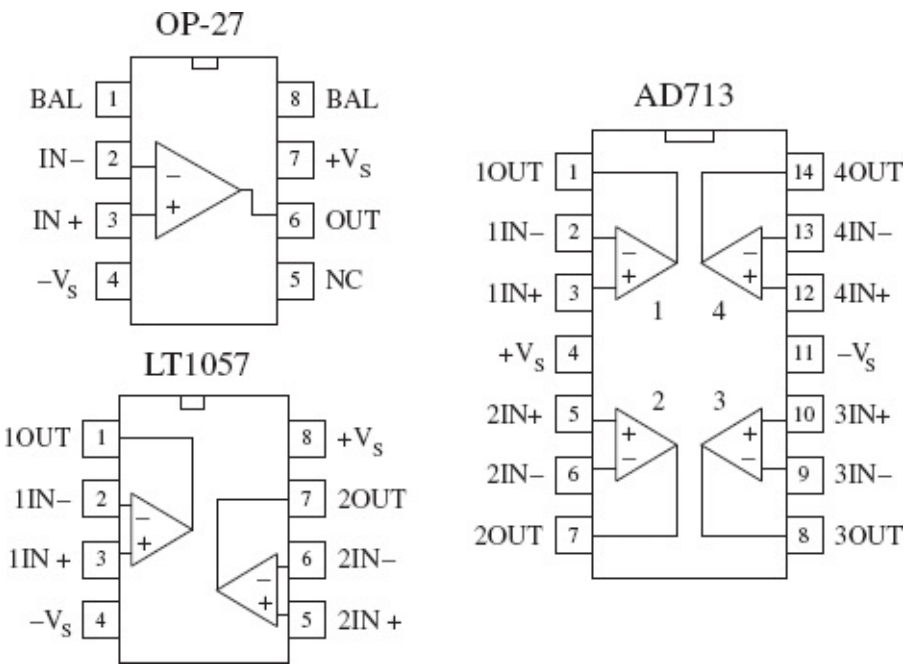
General Purpose



There is a huge selection of general-purpose and precision op amps to choose from. Precision op amps are specifically designed for high stability, low offset voltages, low bias currents, and low drift parameters. Because the selection of op amps is so incredibly large, I will leave it to you to check out the electronics catalogs to see what devices are available. When checking out these catalogs, you will find that op amps (not just general-purpose and precision) fall into one of the following categories (based on input circuitry): bipolar, JFET, MOSFET, or some hybrid thereof (e.g., BiFET). In general, bipolar op amps, like the 741 (industry standard), have higher input bias currents than either JFET or MOSFET types. This means that their input terminals have a greater tendency to "leak in" current. Input bias current results in voltage drops across resistors of feedback networks, biasing networks, or source impedances, which in turn can offset the output voltage. The amount of offset a circuit can tolerate ultimately depends on the application. Now, as I briefly mentioned earlier in this chapter, a compensation resistor placed between the noninverting terminal and ground (e.g., bipolar inverting amplifier circuit) can reduce these offset errors. (More on this in a minute.)

FIGURE 8.18

Precision



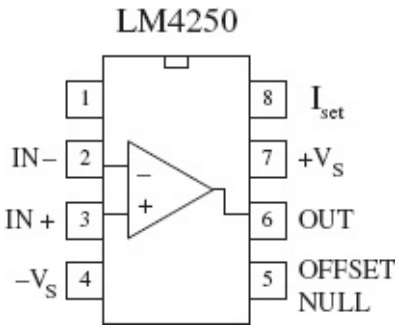
A simple way to avoid problems associated with input bias current is to use a FET op amp. A typical JFET op amp has a very low input bias current, typically within the lower picoamp range as compared with the nanoamp range for a typical bipolar op amp. Some MOSFET op amps come with even lower input bias currents, often as low as a few tenths of a picoamp. Though FET op amps have lower input bias current than bipolar op amps, there are other features they have that are not quite as desirable. For example, JFET op amps often experience an undesired effect called *phase inversion*. If the input common-mode voltage of the JFET approaches the negative supply too closely, the inverting and noninverting input terminals may reverse directions—negative feedback becomes positive feedback, causing the op amp to latch up. This problem can be avoided by using a bipolar

op amp or by restricting the common-mode range of the signal. Here are some other general comments about bipolar and FET op amps: offset voltage (low for bipolar, medium for JFET, medium to high for MOSFET), offset drift (low for bipolar, medium for FET), bias matching (excellent for bipolar, fair for FET), bias/temperature variation (low for bipolar, fair for FET).

To avoid getting confused by the differences between the various op amp technologies, it is often easier to simply concentrate on the specifications listed in the electronics catalogs. Characteristics to look for include speed/slew rate, noise, input offset voltages and their drift, bias currents and their drift, common-mode range, gain, bandwidth, input impedance, output impedance, maximum supply voltages, supply current, power dissipation, and temperature range. Another feature to look for when purchasing an op amp is whether the op amp is internally or externally frequency compensated. An externally compensated op amp requires external components to prevent the gain from dropping too quickly at high frequencies, which can lead to phase inversions and oscillations. Internally compensated op amps take care of these problems with internal circuitry. All the terms listed in this paragraph will be explained in greater detail in a minute.

FIGURE 8.19

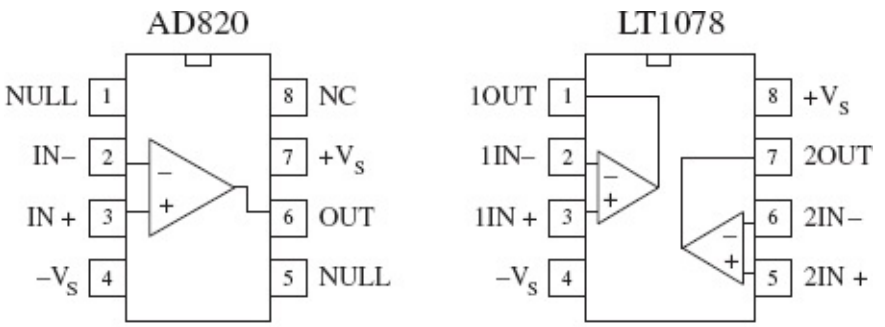
Programmable Op Amp

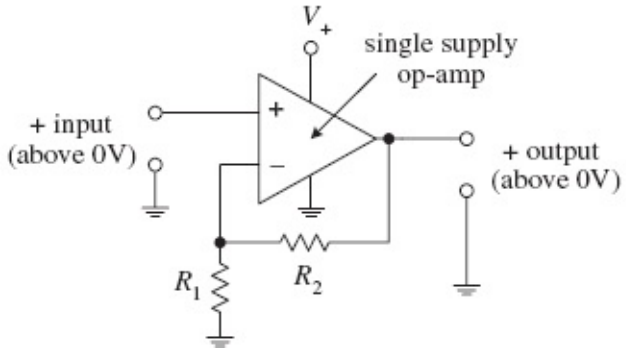


A programmable op amp is a versatile device that is used primarily in low-power applications (e.g., battery-powered circuits). These devices can be programmed with an external current for desired characteristics. Some of the characteristics that can be altered by applying a programming current include quiescent power dissipation, input offset and bias currents, slew rate, gain-bandwidth product, and input noise characteristics—all of which are roughly proportional to the programming current. The programming current is typically drawn from the programming pin (e.g., pin 8 of the LM4250) through a resistor and into ground. The programming current allows the op amp to be operated over a wide range of supply currents, typically from around a few microamps to a few millamps. Because a programmable op amp can be altered so as to appear as a completely different op amp for different programming currents, it is possible to use a single device for a variety of circuit functions within a system. These devices typically can operate with very low supply voltages (e.g., 1 V for the LM4250). A number of different manufacturers make programmable op amps, so check the catalogs. To learn more about how to use these devices, check out the manufacturers' literature (e.g., for National Semiconductor's LM4250, go to *www.national.com*).

FIGURE 8.20

Single-Supply Op Amps

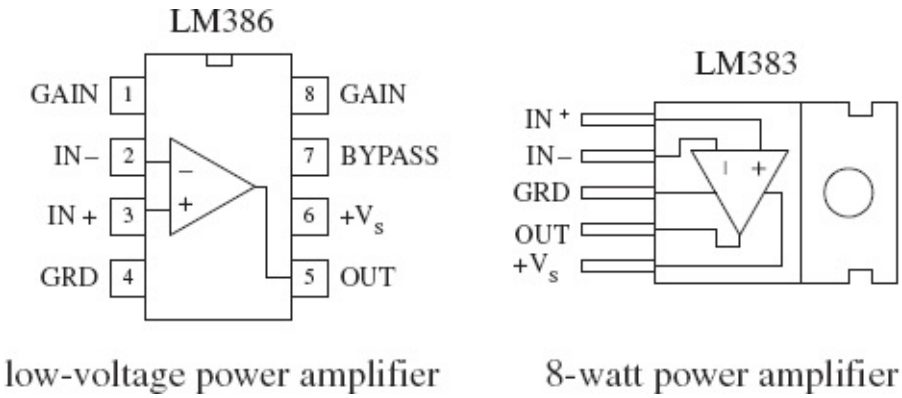




These op amps are designed to be operated from a single positive supply (e.g., +12 V) and allow input voltages all the way down to the negative rail (normally tied to ground). Figure 8.21 shows a simple dc amplifier that uses a single-supply op amp. It is important to note that the output of the amplifier shown cannot go negative; thus it cannot be used for, say, ac-coupled audio signals. These op amps are frequently used in battery-operated devices.

FIGURE 8.21

Audio Amplifiers



These are closely related to conventional op amps but designed specifically to operate best (low audio band noise, crossover distortion, etc.) within the audio-frequency spectrum (20 to 20,000 Hz). These devices are used mainly in sensitive preamplifiers, audio systems, AM-FM radio receivers, servo amplifiers, and intercom and automotive circuits. There are a number of audio amplifiers to choose from. Some of these devices contain unique features that differ when compared with those of conventional op amps. For example, the popular LM386 low-voltage audio amplifier has a gain that is internally fixed at 20 but which can be increased to up to 200 with an external capacitor and resistor placed across its gain leads (pins 1 and 8). This device is also designed to drive low-impedance loads, such as an 8-Ω speaker, and runs off a single supply from +4 to +12 V—an ideal range for battery-powered applications. The LM383 is another audio amplifier designed as a power amplifier. It is a high-current device (3.5 A) designed to drive a 4-Ω load (e.g., one 4-Ω speaker or two 8-Ω speakers in parallel). This device also comes with thermal shutdown circuitry and a heat sink. We'll take a closer look at audio amplifiers in Chap. 15.

FIGURE 8.22

8.7 Op Amp Specifications

Common-mode rejection ratio (CMRR). The input to a difference amplifier, in general, contains two components: a common-mode and a difference-mode signal. The common-mode signal voltage is the

average of the two inputs, whereas the difference-mode signal is the difference between the two inputs. Ideally, an amplifier affects the difference-mode signals only. However, the common-mode signal is also amplified to some degree. The common-mode rejection ratio (CMRR), which is defined as the ratio of the difference signal voltage gain to the common-mode signal voltage gain provides an indication of how well an op amp does at rejecting a signal applied simultaneously to both inputs. The greater the value of the CMRR, the better is the performance of the op amp.

Differential-input voltage range. Range of voltage that may be applied between input terminals without forcing the op amp to operate outside its specifications. If the inputs go beyond this range, the gain of the op amp may change drastically.

Differential input impedance. Impedance measured between the noninverting and inverting input terminals.

Input offset voltage. In theory, the output voltage of an op amp should be zero when both inputs are zero. In reality, however, a slight circuit imbalance within the internal circuitry can result in an output voltage. The input offset voltage is the amount of voltage that must be applied to one of the inputs to zero the output.

Input bias current. Theoretically, an op amp should have an infinite input impedance and therefore no input current. In reality, however, small currents, typically within the nanoamp to picoamp range, may be drawn by the inputs. The average of the two input currents is referred to as the *input bias current*. This current can result in a voltage drop across resistors in the feedback network, the bias network, or source impedance, which in turn can lead to error in the output voltage. Input bias currents depend on the input circuitry of an op amp. With FET op amps, input bias currents are usually small enough not to cause serious offset voltages. Bipolar op amps, on the other hand, may cause problems. With bipolar op amps, a compensation resistor is often required to center the output. I will discuss how this is done in a minute.

Input offset current. This represents the difference in the input currents into the two input terminals when the output is zero. What does this mean? Well, the input terminals of a real op amp tend to draw in different amounts of leakage current, even when the same voltage is applied to them. This occurs because there is always a slight difference in resistance within the input circuitry for the two terminals that originates during the manufacturing process. Therefore, if an op amp's two terminals are both connected to the same input voltage, different amounts of input current will result, causing the output to be offset. Op amps typically come with offset terminals that can be wired to a potentiometer to correct the offset current. I will discuss how this is done in a minute.

Voltage gain (A_V). A typical op amp has a voltage gain of 10^4 to 10^6 (or 80 to 120 dB; gain in dB = 20 $\log_{10}A_0$11) at dc. However, the gain drops to 1 at a frequency called the *unity-gain frequency f_T*, typically from 1 to 10 MHz—a result of high-frequency limitations in the op amp's internal circuitry. I will talk more about high-frequency behavior in op amps in a minute.

Output voltage swing. This is the peak output voltage swing, referenced to zero, that can be obtained without clipping.

Slew rate. This represents the maximum rate of change of an op amp's output voltage with time. The limitation of output change with time results from internal or external frequency compensation capacitors slowing things down, which in turn results in delayed output changes with input changes (propagation delay). At high frequencies, the magnitude of an op amp's slew rate becomes more critical. A general-purpose op amp like the 741 has a 0.5 V/μs slew rate—a relatively small value when compared with the high-speed HA2539's slew rate of 600 V/μs.

Supply current. This represents the current that is required from the power supply to operate the op amp with no load present and with an output voltage of zero.

Table 8.1 is a sample op amp specifications table.

TABLE 8.1 Sample Op Amp Specifications

TYPE	TOTAL SUPPLY VOLTAGE MAX (V)	TOTAL SUPPLY VOLTAGE MIN (V)	SUPPLY CURRENT (mV)	OFFSET VOLTAGE TYPICAL (mV)	OFFSET VOLTAGE MAX (mV)	CURRENT BIAS MAX (nA)	CURRENT OFFSET MAX (nA)	SLEW RATE TYPICAL (V/μS)	f_t TYPICAL (MHz)	CMRR MIN (dB)	GAIN MIN (mA)	OUTPUT CURRENT MAX (mA)
Bipolar 741C	10	36	2.8	2	6	500	200	0.5	1.2	70	86	20
MOSFET CA3420A	2	22	1	2	5	0.005	0.004	0.5	0.5	60	86	2
JFET LF411	10	36	3.4	0.8	2	0.2	0.1	15	4	70	88	30
Bipolar, precision LM10	1	45	0.4	0.3	2	20	0.7	0.12	0.1	93	102	20

8.8 Powering Op Amps

Most op amp applications require a dual-polarity power supply. A simple split ±15-V supply that uses a tapped transformer is presented in Chap. 11. If you are using batteries to power an op amp, one of the following arrangements can be used.

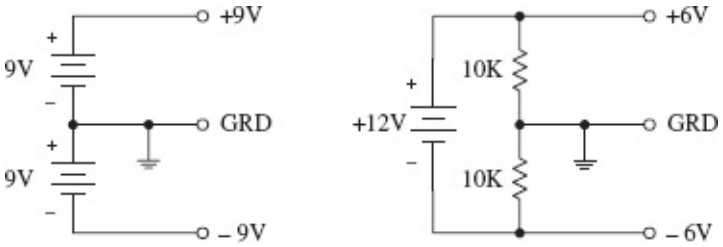


FIGURE 8.23

Now, it is often desirable to avoid split-supplies, especially with small battery-powered applications. One option in such a case is to use a single-supply op amp. However, as I pointed out a second ago, these devices will clip the output if the input attempts to go negative, making them unsuitable for ac-coupled applications. To avoid clipping while still using a single supply, it is possible to take a conventional op amp and apply a dc level to one of the inputs using a voltage-divider network. This, in turn, provides a dc offset level at the output. Both input and output offset levels are referenced to ground (the negative terminal of the battery). With the input offset voltage in place, when an input signal goes negative, the voltage applied to the input of the op amp will dip below the offset voltage but will not go below ground (provided you have set the bias voltage large enough, and provided the input signal is not too large; otherwise, clipping occurs). The output, in turn, will fluctuate about its offset level. To allow for input and output coupling, input and output capacitors are needed. The two circuits in Fig. 8.24 show noninverting and inverting ac-coupled amplifiers (designed for audio) that use conventional op amps that run off a single supply voltage.

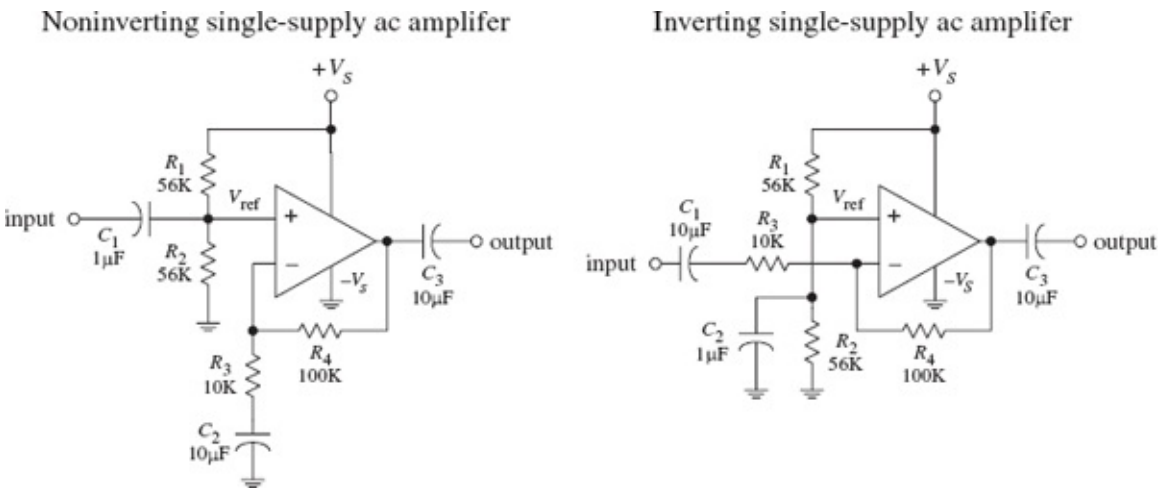


FIGURE 8.24

In the noninverting circuit, the dc offset level is set to one-half the supply voltage by R_1 and R_2 to allow for maximum symmetrical swing. C_1 (and R_2) and C_3 (and R_{load}) act as ac coupling (filtering) capacitors that block unwanted dc components and low-level frequencies. C_1 should be equal to $1/(2\pi f_{3dB}R_1)$, while C_3 should be equal to $1/(2\pi f_{3dB}R_{load})$, where f_{3dB} is the cutoff frequency (see Chaps. 9 and 15).

When using conventional op amps with single supply voltages, make sure to stay within the minimum supply voltage rating of the op amp, and also make sure to account for maximum output swing limitations and maximum common-mode input range.

8.9 Some Practical Notes

As a note of caution, never reverse an op amp's power supply leads. Doing so can result in a zapped op amp IC. One way to avoid this fate is to place a diode between the op amp's negative supply terminal and the negative supply, as shown in Fig. 8.25.

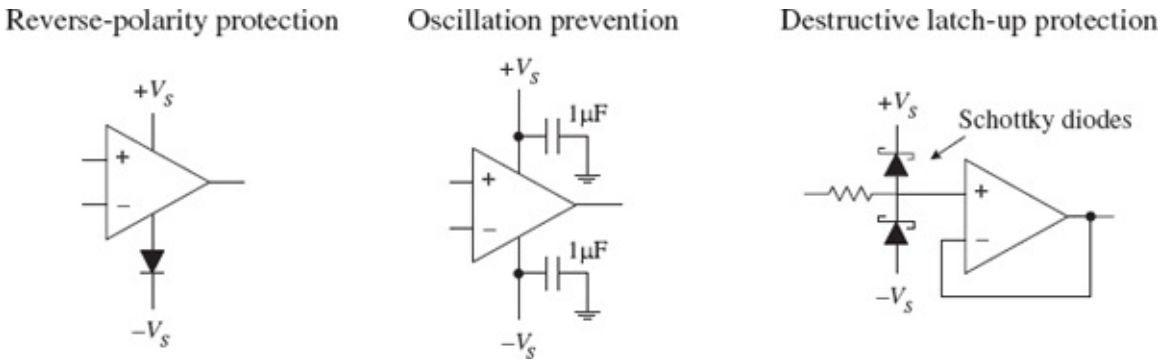


FIGURE 8.25

Also keep wires running from the power supply to the op amp's supply terminals short and direct. This helps prevent unwanted oscillations/noise from arising in the output. Disturbances also may arise from variations in supply voltage. To eliminate these effects, place bypass capacitors from the supply terminals to ground as shown in Fig. 8.25. A 0.1-μF disk capacitor or a 1.0-μF tantalum capacitor should do the trick.

Both bipolar and JFET op amps can experience a serious form of latch if the input signal becomes more positive or negative than the respective op amp power supplies. If the input terminals go more positive than $+V_S$ + 0.7 V or more negative than $-V_s$ − 0.7 V, current may flow in the wrong direction within the internal circuitry, short-circuiting the power supplies and destroying the device. To avoid this potentially fatal latch-up, it is important to prevent the input terminals of op amps from exceeding the

power supplies. This feature has vital consequences during device turn-on; if a signal is applied to an op amp before it is powered, it may be destroyed at the moment power is applied. A "hard wire" solution to this problem involves clamping the input terminals at risk with diodes (preferably fast low-forward-voltage Schottky diodes; see Fig. 8.25). Current-limiting resistors also may be needed to prevent the diode current from becoming excessive. This protection circuitry has some problems, however. Leakage current in the diodes may increase the error. See manufacturers' literature for more information.

8.10 Voltage and Current Offset Compensation

In theory, the output voltage of an op amp should be zero when both inputs are zero. In reality, however, a slight circuit imbalance within the internal circuitry can result in an output voltage (typically within the microvolt to millivolt range). The input offset voltage is the amount of voltage that must be applied to one of the inputs to zero the output—this was discussed earlier. To zero the input offset voltage, manufacturers usually include a pair of offset null terminals. A potentiometer is placed between these two terminals, while the pot's wiper is connected to the more negative supply terminal, as shown in Fig. 8.26. To center the output, the two inputs can be shorted together and an input voltage applied. If the output saturates, the input offset needs trimming. Adjust the pot until the output approaches zero.

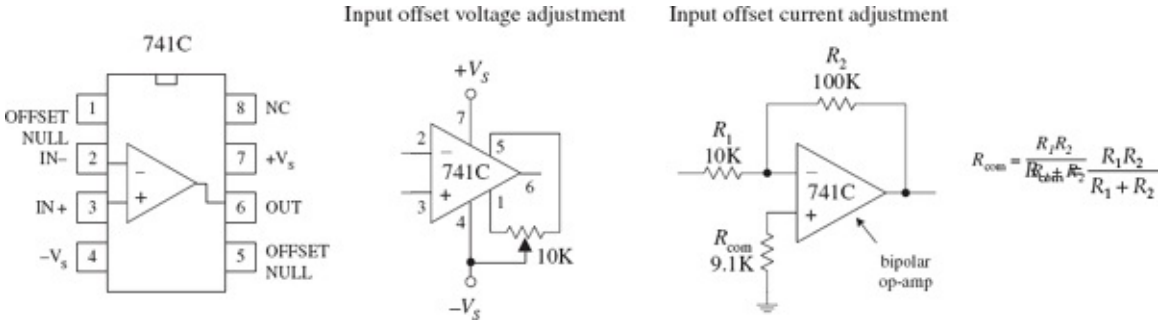


FIGURE 8.26

Notice the resistor placed between the noninverting terminal and ground within the inverting amplifier circuit shown in Fig. 8.26 *right*. What is the resistor used for? It is used to compensate for output voltage errors caused by a voltage drops across R_1 and R_2 as a result of input bias current. As discussed earlier, bipolar op amps tend to have larger input bias currents than FET op amps. With FET op amps, the input bias errors are usually so small (in the picoamp range) that the output voltage error is insignificant, and the compensation resistor is not needed. However, with bipolar op amps, this is not the case (input bias currents in the nanoamp range), and compensation is often necessary. Now, in the inverting amplifier, the bias current—assuming for now that the compensation resistor is missing—introduces a voltage drop equal to $V_{in} = I_{bias}(R_1 \| R_2)$, which is amplified by a factor of $-R_2/R_1$. In order to correct this problem, a compensation resistor with a resistance equal the $R_1 \| R_2$ is placed between the noninverting terminal and ground. This resistor makes the op amp "feel" the same input driving resistance.

8.11 Frequency Compensation

For a typical op amp, the open-loop gain is typically between 10^4 and 10^6 (80 to 120 dB). However, at a certain low frequency, called the *breakover frequency* f_B, the gain drops by 3 dB, or drops to 70.7 percent of the open-loop gain (maximum gain). As the frequency increases, the gain drops further until it reaches 1 (or 0 dB) at a frequency called the *unity-gain frequency* f_T. The unity-gain frequency for an op amp is typically around 1 MHz and is given in the manufacturer's specifications (see Fig. 8.27 *left*). The rolloff

in gain as the frequency increases is caused by low-pass filter-like characteristics inherently built into the op amp's inner circuitry. If negative feedback is used, an improvement in bandwidth results; the response is flatter over a wider range of frequencies, as shown in the far-left graph in Fig. 8.27. Now op amps that exhibit an open-loop gain drop of more than 60 dB per decade at f_T are unstable due to phase shifts incurred in the filter-like regions within the interior circuitry. If these phase shifts reach 180° at some frequency at which the gain is greater than 1, negative feedback becomes positive feedback, which results in undesired oscillations (see center and right grafts in Fig. 8.27). To prevent these oscillations, frequency compensation is required. *Uncompensated* op amps can be frequency compensated by connecting an *RC* network between the op amp's frequency-compensation terminals. The network, especially the capacitor, influences the shape of the response curve. Manufacturers will supply you with the response curves, along with the component values of the compensation network for a particular desired response.

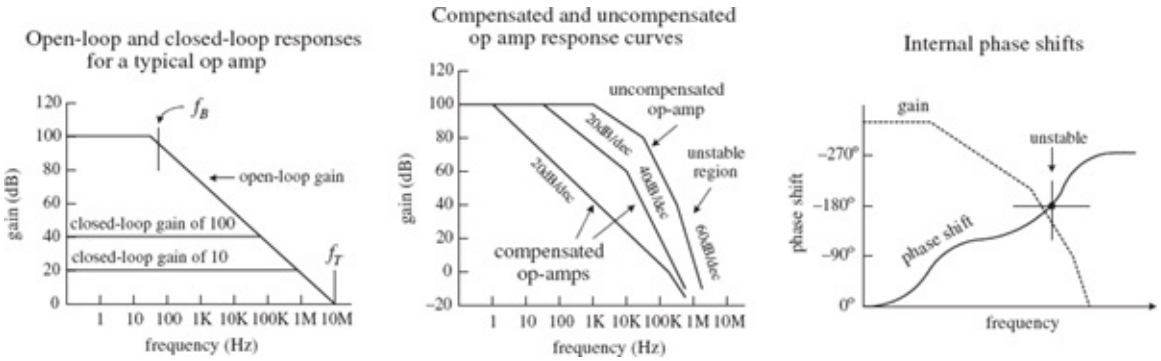


FIGURE 8.27

Perhaps the easiest way to avoid dealing with frequency compensation is to buy an op amp that is internally compensated.

8.12 Comparators

In many situations, it is desirable to know which of two signals is bigger or to know when a signal exceeds a predetermined voltage. Simple circuits that do just this can be constructed with op amps, as shown in Fig. 8.28. In the noninverting comparator circuit, the output switches from low (0 V) to high (positive saturation) when the input voltage exceeds a reference voltage applied to the inverting input. In the inverting comparator circuit, the output switches from high to low when the input exceeds the reference voltage applied to the noninverting terminal. In the far-right circuit, a voltage divider (pot) is used to set the reference voltage.

Note that not all op amps are able to operate with a grounded negative supply, and a special-purpose comparator IC is a better approach.

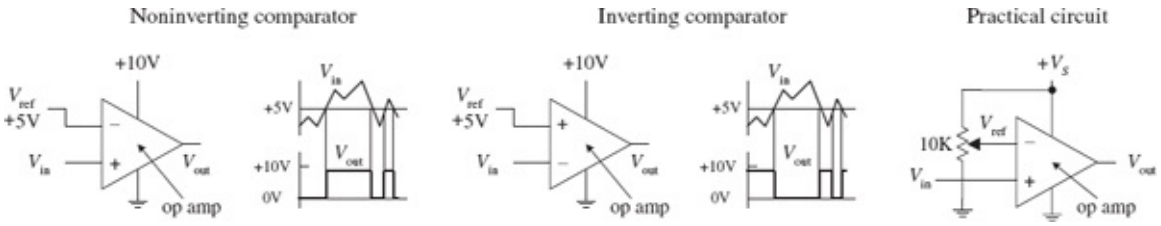


FIGURE 8.28

Another, more popular method for comparing two voltages is to use a special IC called a *comparator*. A comparator, like an op amp, has an inverting input, a non-inverting input, an output, and power supply leads—its schematic looks like an op amp, too. However, unlike op amps, comparators are not frequency

compensated and therefore cannot be used as linear amplifiers. In fact, comparators never use negative feedback (they often use positive feedback, as you will see). If negative feedback were used with a comparator, its output characteristics would be unstable. Comparators are specifically designed for high-speed switching—they have much larger slew rates and smaller propagation delays than op amps. Another important difference between a comparator and an op amp has to do with the output circuitry. Unlike an op amp, which typically has a push-pull output stage, a comparator uses an internal transistor whose collector is connected to the output and whose emitter is grounded. When the comparator's noninverting terminal is less positive in voltage than the inverting terminal, the output transistor turns on, grounding the output. When the noninverting terminal is made more positive than the inverting terminal, the output transistor is off. In order to give the comparator a high output state when the transistor is off ($V_- < V_+$), an external *pull-up resistor* connected from a positive voltage source to the output is used. The pull-up resistor acts like the collector resistor in a transistor amplifier. The size of the pull-up resistor should be large enough to avoid excessive power dissipation yet small enough to supply enough drive to switch whatever load circuitry is used on the comparator's output. The typical resistance for a pull-up resistor is anywhere from a few hundred to a few thousand ohms. Figure 8.29 shows a simple noninverting and inverting comparator circuit with pull-up resistors included. Both circuits have an output swing from 0 to +5 V.

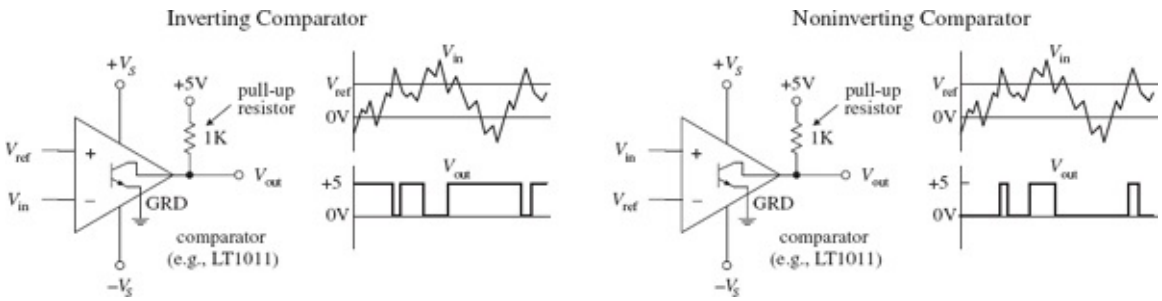


FIGURE 8.29

Comparators are commonly used in analog-to-digital conversion. A typical application might be to connect a magnetic tape sensor or photodiode to an input of a comparator (with reference voltage set at the other input) and allow the sensor to drive the comparator's output to a low or high state suitable for driving logic circuits. Analog-to-digital conversion is discussed in greater detail in Chap. 12.

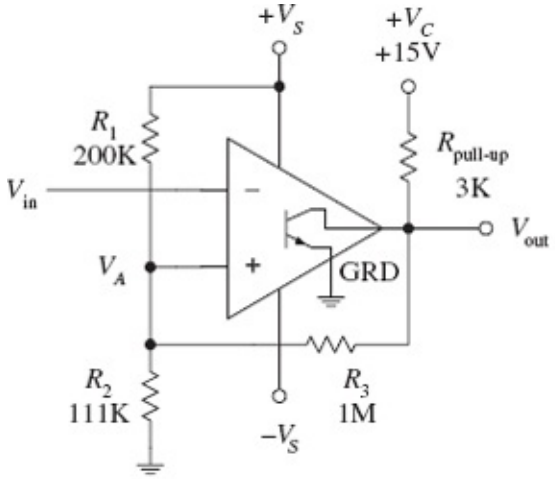
8.13 Comparators with Hysteresis

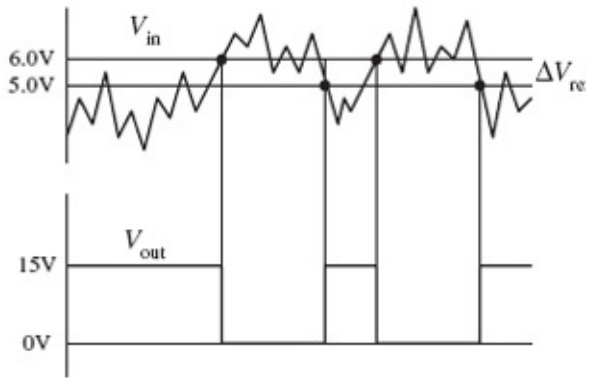
Now there is a basic problem with both comparator circuits shown in Fig. 8.29. When a slowly varying signal is present that has a level near the reference voltage, the output will "jitter" or flip back and forth between high and low output states. In many situations having such a "finicky" response is undesirable. Instead, what is usually desired is a small "cushion" region that ignores small signal deviations. To provide the cushion, positive feedback can be added to the comparator to provide hysteresis, which amounts to creating two different threshold voltages or triggering points. The details of how hysteresis works within comparator circuits are given in the following two examples.

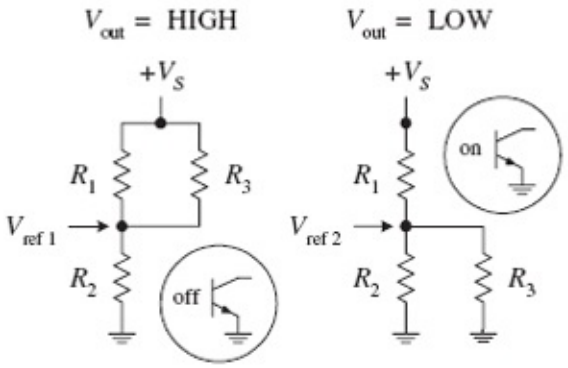
8.13.1 Inverting Comparator with Hysteresis

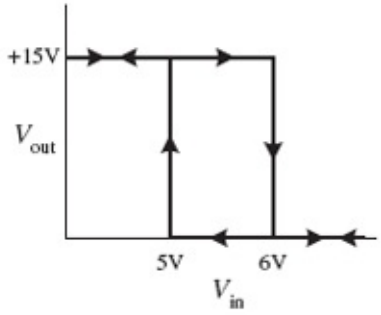
In the inverting comparator circuit shown in Fig. 8.30, positive feedback through R_3 provides the comparator with two threshold voltages or triggering points. The two threshold voltages result from the fact that the reference voltage applied to the noninverting terminal is different when V_{out} is high (+15 V)

and when V_{out} is low (0 V)—a result of feedback current. Let's call the reference voltage when the output is high V_{ref1}, and call the reference voltage when V_{out} is low V_{ref2}. Now let's assume that the output is high (transistor is off) and $V_{in} > V_{ref1}$. In order for the output to switch high, V_{in} must be greater than V_{ref1}. However, what is V_{ref1}? It is simply the reference voltage that pops up at the noninverting terminal when the output transistor is off and the output is high (+15 V).









$$1)\ \frac{\Delta V_{ref}}{V_{ref2}} = n$$

$$2)\ R_1 = nR_3$$

$$3)\ R_2 = \frac{R_1 \| R_3}{(+V_S/V_{ref1}) - 1}$$

To calculate V_{ref1}, simply use the basic resistor network shown below the main circuit and to the left:

$$V_{ref1} = \frac{+V_S R_2}{(R_1 \| R_3) + R_2} = \frac{+V_S R_2 (R_1 + R_2)}{R_1 R_2 + R_1 R_3 + R_2 R_3}$$

If $V_{in} > V_{ref1}$ when the output is already high, the output suddenly goes low—transistor turns on. With the output now low, a new reference voltage V_{ref2} is in place. To calculate V_{ref2}, use the resistor network shown to the right of the first:

$$V_{ref2} = \frac{+V_S R_2 \| R_3}{R_1 + (R_2 \| R_3)} = \frac{+V_S R_2 R_3}{R_1 R_2 + R_1 R_3 + R_2 R_3}$$

When the input voltage decreases to V_{ref2} or lower, the output suddenly goes high. The difference in the reference voltages is called the *hysteresis voltage*, or ΔV_{ref}:

$$\Delta V_{ref} = V_{refl} - V_{ref2} = \frac{+V_S R_1 R_2}{R_1 R_2 + R_1 R_3 + R_2 R_3}$$

Now, let's try the theory out on a real-life design example.

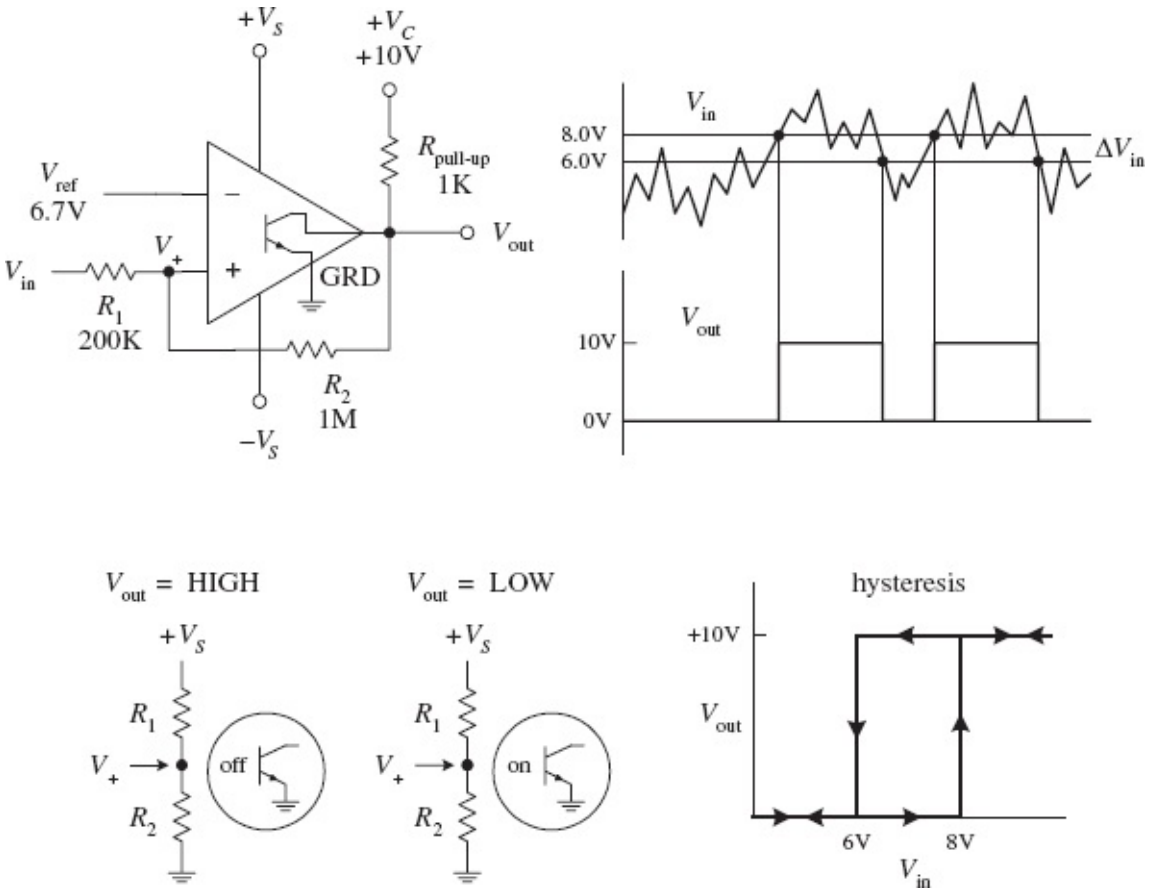
Say we want to design a comparator circuit with a V_{ref1} = +6 V, a V_{ref2} = +5 V, and $+V_C$ = +15 V that drives a 100-kΩ load. The first thing to do is pick a pull-up resistor. As a rule of thumb,

$R_{pull\text{-}up} < R_{load}$
$R_3 > R_{pull\text{-}up}$

Why? Because heavier loading on $R_{pull\text{-}up}$ (smaller values of R_3 and R_{load}) reduce the maximum output voltage, thereby reducing the amount of hysteresis by lowering the value of V_{ref1}. Pick $R_{pull\text{-}up}$ = 3 kΩ, and choose R_3 equal to 1 MΩ. Combining equations above gives us the practical formulas below the diagrams. With equation (1), calculate n, which is (6 V − 5 V)/5 V = 0.20. Next, using equation (2), find R_1, which is simply (0.2)(1 M) = 200 kΩ. Using equation (3), find R_3, which is 166 k/(15 V/6 V − 1) = 111 kΩ. These are the values presented in the circuit.

FIGURE 8.30

8.13.2 Noninverting Comparator with Hysteresis



What's desired: V_{in1} = 8 V, V_{in2} = 6 V, given $+V_C$ = 10 V and load of 100 kΩ. Question: What should V_{ref}, $R_{pull\text{-}up}$, R_2, and R_1 be?

First, choose $R_{pull\text{-}up} < R_{load}$, and $R_2 > R_{pull\text{-}up}$ to minimize the effects of loading. For $R_{pull\text{-}up}$, choose 1 kΩ, and for R_2, choose 1 M. Next, using the equations to the right, we find R_1 and V_{ref}.

$$\frac{R_1}{R_2} = \frac{\Delta V_{in}}{V_C} = \frac{10-8}{10} = 0.20$$

$$R_1 = 0.20R_2 = 0.20(1\text{ M}) = 200\text{ k}\Omega$$

$$V_{ref} = \frac{V_{in1}}{1 + R_1/R_2} = \frac{8\text{ V}}{1\text{ V} - 0.20\text{ V}} = 6.7\text{ V}$$

Unlike the inverting comparator, the noninverting comparator only requires two resistors for hysteresis to occur. (Extra resistors are needed if you wish to use a voltage divider to set the reference voltage. However, these resistors do not have a direct role in developing the hysteresis voltage.) Also, the terminal to which the input signal is applied is the same location where the threshold shifting occurs—a result of positive feedback. The threshold level applied to the noninverting terminal is shifted about the reference voltage as the output changes from high ($+V_C$) to low (0 V). For example, assume that V_{in} is at a low enough level to keep V_{out} low. For the output to switch high, V_{in} must rise to a triggering voltage, call it V_{in1}, which is found

simply by using the resistor network shown to the far left:

$$V_{in1} = \frac{V_{ref}(R_1 + R_2)}{R_2}$$

As soon as V_{out} switches high, the voltage at the noninverting terminal will be shifted to a value that's greater than V_{ref} by:

$$\Delta V_+ = V_{in} + \frac{(V_{CC} - V_{in1})R_1}{R_1 + R_2}$$

To make the comparator switch back to its low state, V_{in} must go below ΔV_+. In other words, the applied input voltage must drop below what is called the lower trip point, V_{in2}:

$$V_{in2} = \frac{V_{ref}(R_1 + R_2) - V_{CC}R_1}{R_2}$$

The hysteresis is then simply the difference between V_{in1} and V_{in2}:

$$\Delta V_{in} = V_{in1} - V_{in2} = \frac{V_{CC}R_1}{R_2}$$

A practical design example is presented to the left.

FIGURE 8.31

8.14 Using Single-Supply Comparators

Like op amps, comparator ICs come in both dual- and single-supply forms. With single-supply comparators, the emitter and the "negative supply" are joined internally and grounded, whereas the dual-supply comparator has separate emitter (ground) and negative supply leads. A few sample comparator ICs, along with two single-supply comparator circuits are shown below.

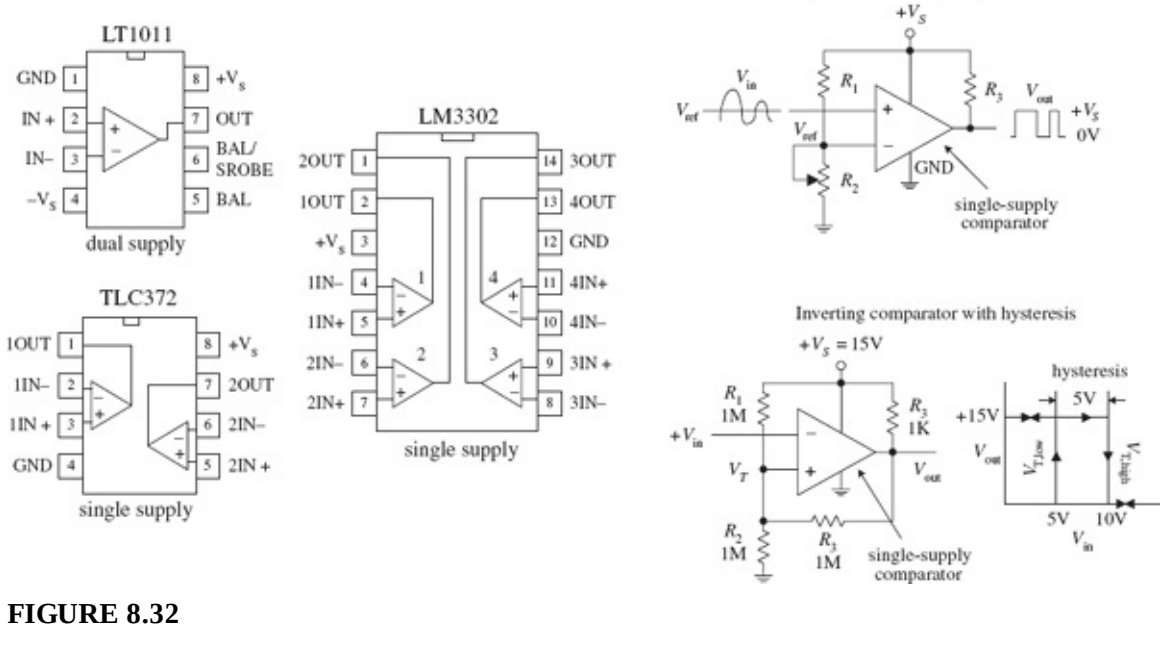


FIGURE 8.32

8.15 Window Comparator

A window comparator is a very useful circuit that changes its output state whenever the input voltage is anywhere between predetermined high and low reference voltages. The region between these two

reference voltages is called the *window*. Figure 8.33 shows a simple window comparator built with two comparators (op amps also can be used). In the left-most circuit, the window is set between +3.5 V ($V_{ref,high}$) and +6.5 V($V_{ref,low}$). If V_{in} is below +3.5 V, the lower comparator's output is grounded, while the upper comparator's output floats. Only one ground is needed, however, to make V_{out} = 0 V. If V_{in} is above +6.5 V, the upper comparator's output is grounded, while the lower comparator's output floats—again V_{out} goes to 0 V. Only when V_{in} is between +3.5 and +6.5 V will the output go high (+5 V). The right-most circuit uses a voltage-divider network to set the reference voltages.

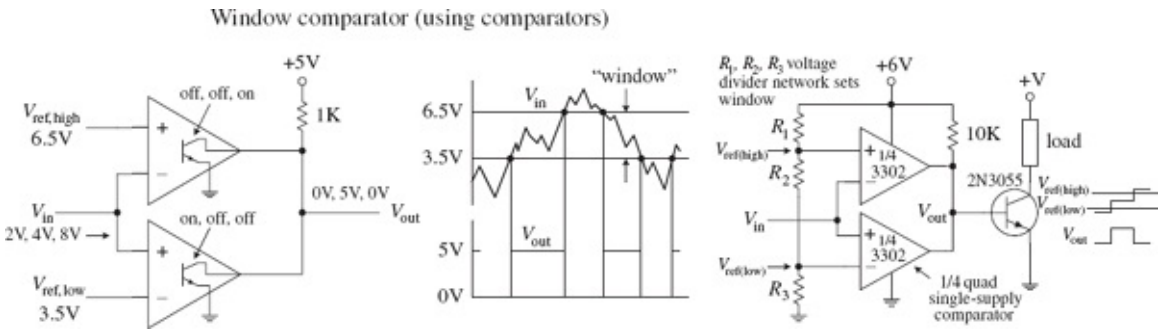


FIGURE 8.33

8.16 Voltage-Level Indicator

A simple way to make a voltage-level indicator is to take a number of comparators that share a common input and then supply each comparator with a different reference or triggering voltage, as shown in Fig. 8.34. In this circuit, the reference voltage applied to a comparator increases as you move up the chain of comparators (a result of the voltage-divider network). As the input voltage increases, the lower comparator's output is grounded first (diode turns on), followed in succession by the comparators (LEDs) above it. The potentiometer provides proportional control over all the reference voltages.

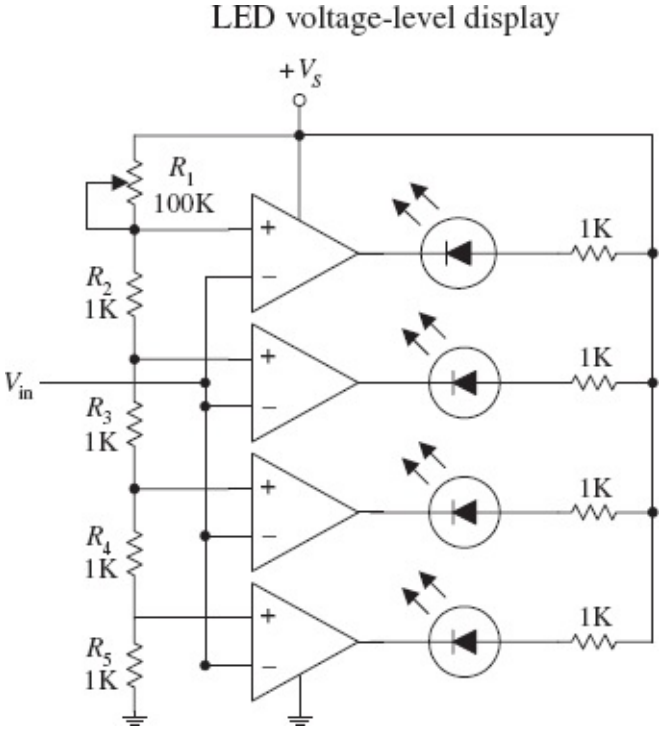


FIGURE 8.34

8.17 Applications

Op Amp Output Drivers (for Loads That Are Either On or Off)

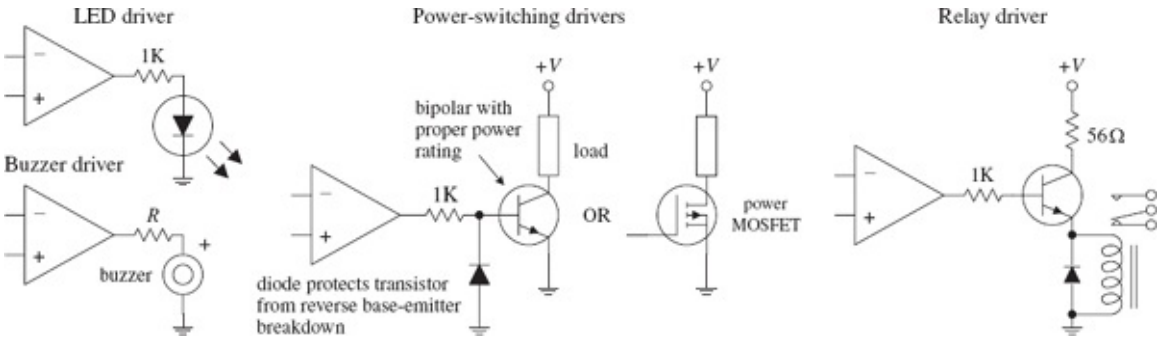


FIGURE 8.35

Comparator Output Drivers

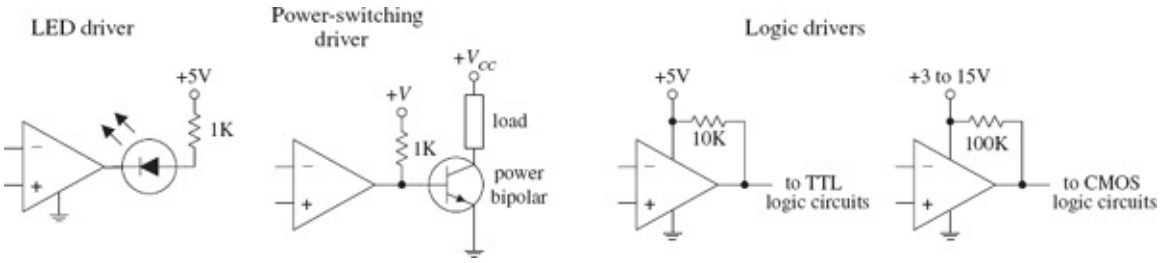
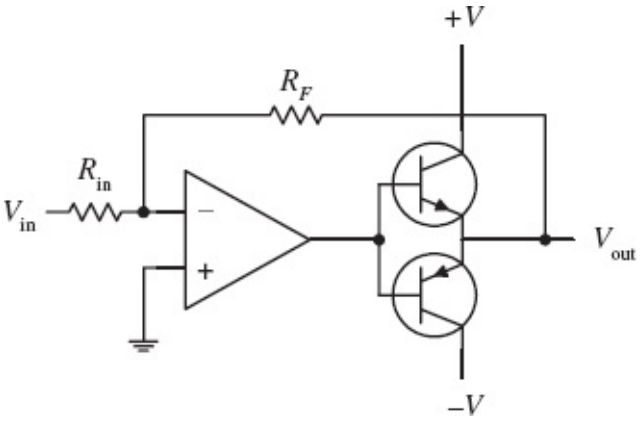


FIGURE 8.36

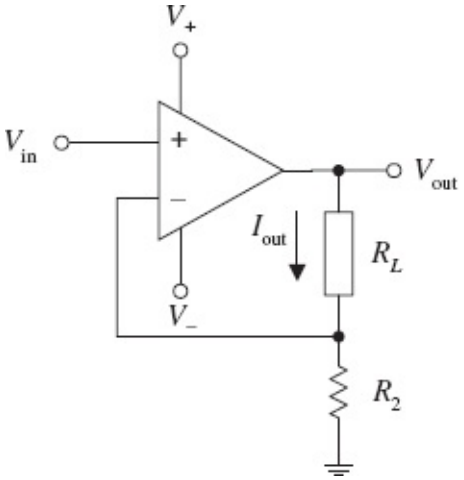
Op Amp Power Booster (AC Signals)



There are times when an op amp needs a boost in output power-handling capacity while at the same time maintaining both positive and negative output swings. A simple way to increase the output power while maintaining the swing integrity is to attach a complementary transistor push-pull circuit to the op amp's output, as shown in this circuit. At high speeds, additional biasing resistors and capacitors are needed to limit crossover distortion. At low speeds, negative feedback helps eliminate much of the crossover distortion.

FIGURE 8.37

Voltage-to-Current Converter



Here is a simple current source whose output current is determined by an input voltage applied to the noninverting terminal of the op amp. The output current and voltage are determined by the following expressions:

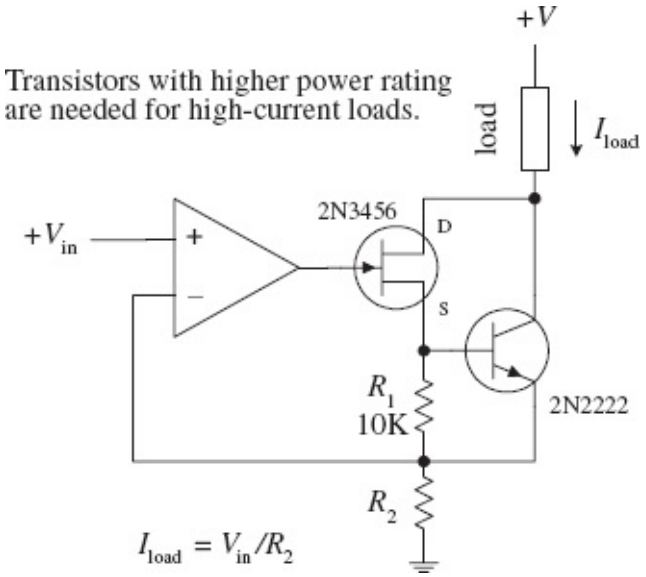
$$V_{out} = \frac{(R_L + R_2)}{R_2} V_{in}$$

$$I_{out} = \frac{V_{out}}{R_1 + R_2} = \frac{V_{in}}{R_2}$$

V_{in} can be set with a voltage divider.

FIGURE 8.38

Precision Current Source



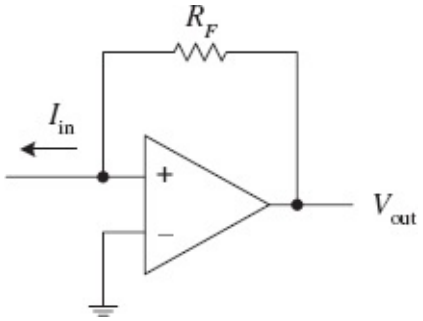
Here, a precision current uses a JFET to drive a bipolar output transistor used to sink current through a load. Unlike the preceding current source, this circuit is less susceptible to output drift. Use of the JFET helps achieve essentially zero bias current error (a single bipolar output stage will leak base current). This circuit is accurate for output currents larger than the JFET's $I_{DS(on)}$ and provided V_{in} is greater than 0 V. For large currents, the FET-bipolar combination can be replaced with a Darlington transistor, provided its base current does not introduce significant error. The output current or load current is determined by

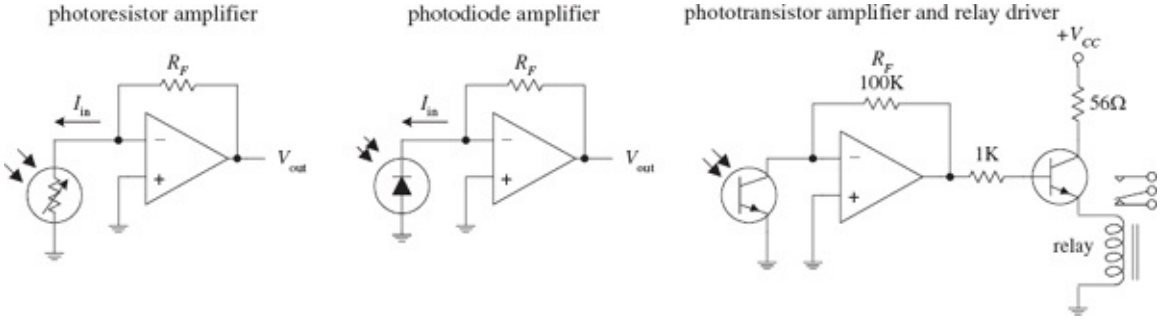
$I_{load} = V_{in}/R_2$

R_2 acts as an adjustment control. Additional compensation may be required depending on the load reactance and the transistors' parameters. Make sure to use transistors of sufficient power ratings to handle the load current in question.

FIGURE 8.39

Current-to-Voltage Converter





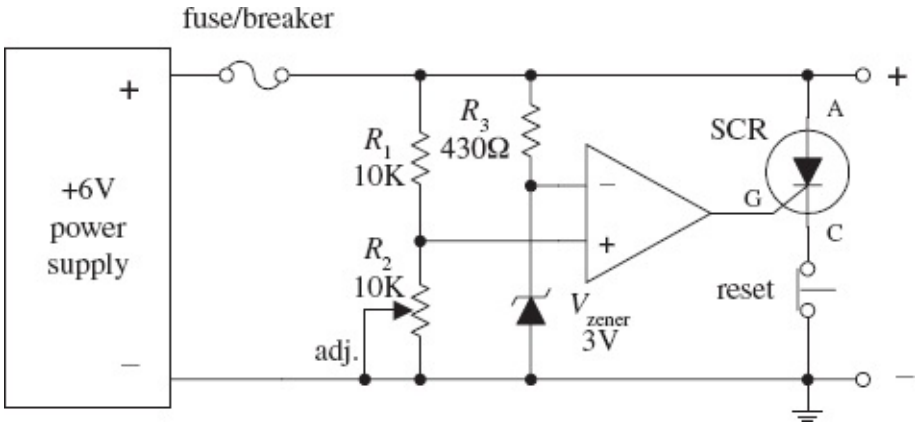
The circuit shown here transforms a current into a voltage. The feedback resistor R_F helps establish a voltage at the inverting input and controls the swing of the output. The output voltage for this circuit is given by

$$V_{out} = I_{in}R_F$$

The light-activated circuits shown below use this principle to generate an output voltage that is proportional to the amount of input current drawn through the light sensor.

FIGURE 8.40

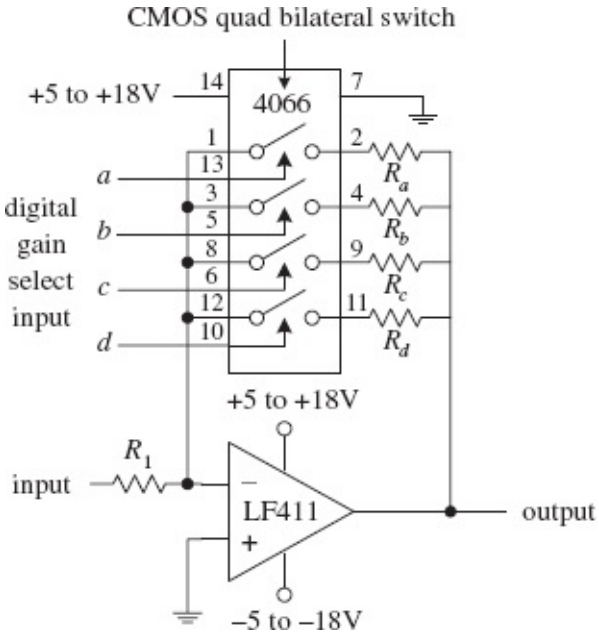
Overvoltage Protection (Crowbar)



This circuit acts as a fast-acting overvoltage protection control used to protect sensitive loads from voltage surges generated in the power supply. Initially, let's assume that the supply is doing what it should—generating a constant +6 V. In this case, the voltage applied to the op amp's noninverting input is set to 3 V (by means of the R_1,R_2 voltage divider—the pot provides fine-tuning adjustment). At the same time, the inverting input is set to 3 V by means of the 3-V zener diode. The op amp's differential input voltage in this case is therefore zero, making the op amp's output zero (op amp acts as a comparator). With the op amp's output zero, the SCR is off, and no current will pass from anode (A) to cathode to ground. Now, let's say there's a sudden surge in the supply voltage. When this occurs, the voltage at the noninverting input increases, while the voltage at the inverting input remain at 3 V (due to the 3-V zener). This causes the op amp's output to go high, triggering the SCR on and diverting all current from the load to ground in the process. As a result, the fuse (breaker) blows and the load is saved. The switch is opened to reset the SCR.

FIGURE 8.41

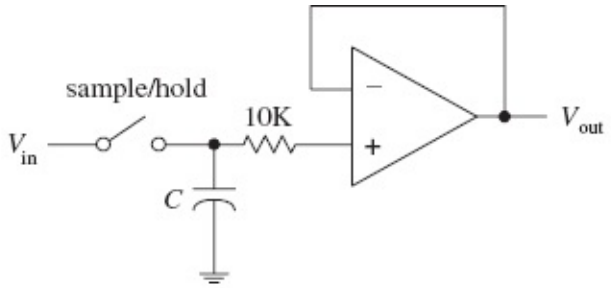
Programmable-Gain Op Amp

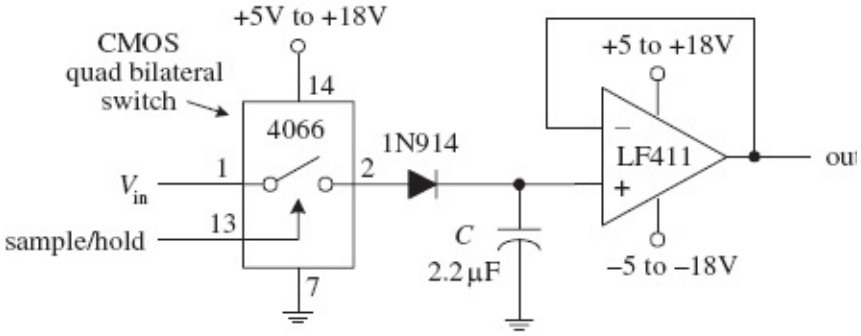


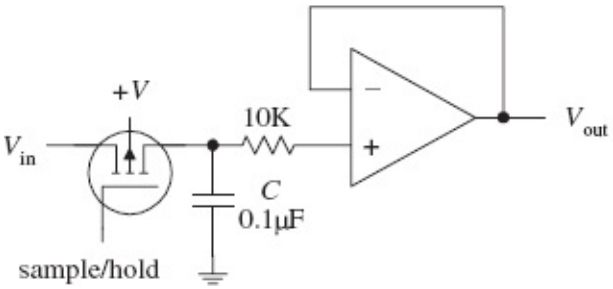
This circuit is simply an inverting amplifier whose feedback resistance (gain) is selected by means of a digitally controlled bilateral switch (e.g., CMOS 4066). For example, if the bilateral switch's input *a* is set high (+5 to +18 V) while its *b* through *d* inputs are set low (0 V), only resistor R_a will be present in the feedback loop. If you make inputs *a* through *d* high, then the effective feedback resistance is equal to the parallel resistance of resistors R_a through R_d. Bilateral switches are discussed in greater detail in Chap. 12.

FIGURE 8.42

Sample-and-Hold Circuits



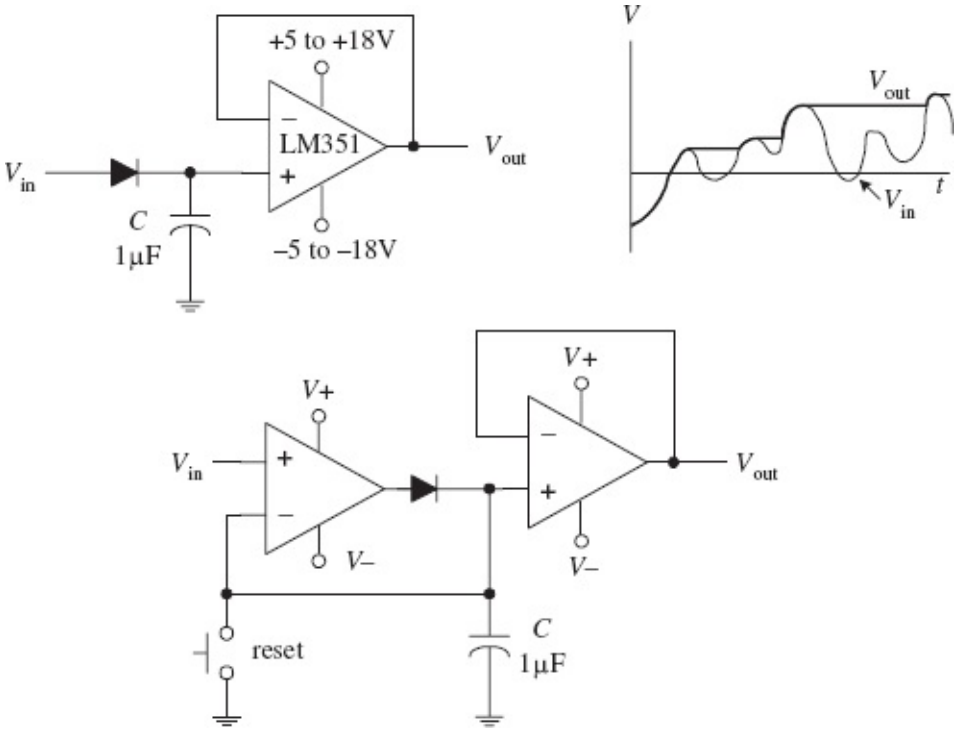




Sample-and-hold circuits are used to sample an analog signal and hold it so that it can be analyzed or converted into, say, a digital signal at one's leisure. In the first circuit, a switch acts as a sample/ hold control. Sampling begins when the switch is closed and ends when the switch is opened. When the switch is opened, the input voltage present at that exact moment will be stored in *C*. The op amp acts as a unity-gain amplifier (buffer), relaying the capacitor's voltage to the output but preventing the capacitor from discharging (recall that ideally, no current enters the inputs of an op amp). The length of time a sample voltage can be held varies depending on how much current leaks out of the capacitor. To minimize leakage currents, use op amps with low input-bias currents (e.g., FET op amps). In the other two circuits, the sample/hold manual switch is replaced with an electrically controlled switch—the left-most circuit uses a bilateral switch, while the right-most circuit uses a MOSFET. Capacitors best suited for sample/hold applications include Teflon, polyethylene, and polycarbonate dielectric capacitors.

FIGURE 8.43

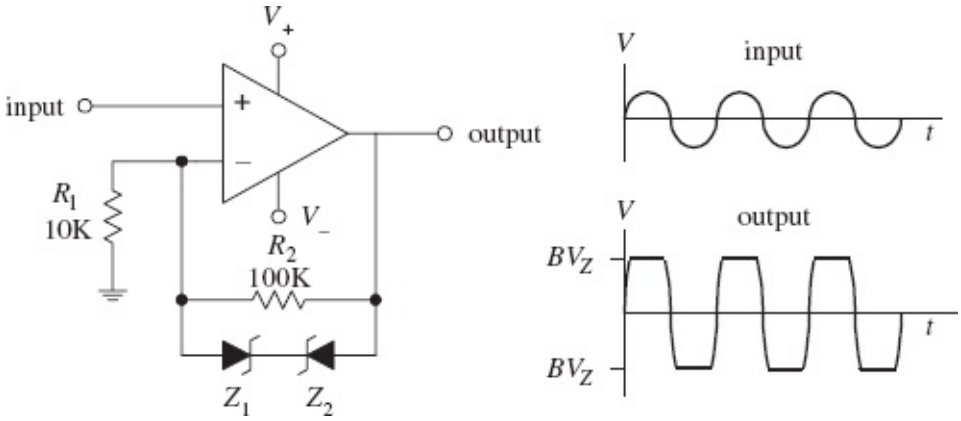
Peak Detectors



The circuits shown here act as peak detectors—they follow an incoming voltage signal and store its maximum voltage within *C* (see graph). The op amp in the upper circuit acts as a buffer—it "measures" the voltage in *C*, outputs that voltage, and prevents *C* from discharging. The diode also prevents the capacitor from discharging when the input drops below the peak voltage stored on *C*. The second circuit is a more practical peak detector. The additional op amp makes the detector more sensitive; it compensates for the diode voltage drop (around 0.6 V) by feeding back *C*'s voltage to the inverting terminal. In other words, it acts as an active rectifier. Also, this circuit incorporates a switch to reset the detector. Often, peak detectors use a FET in place of the diode and use the FET's gate as a reset switch. Reducing the capacitance of *C* promotes faster response times for changes in V_{in}.

FIGURE 8.44

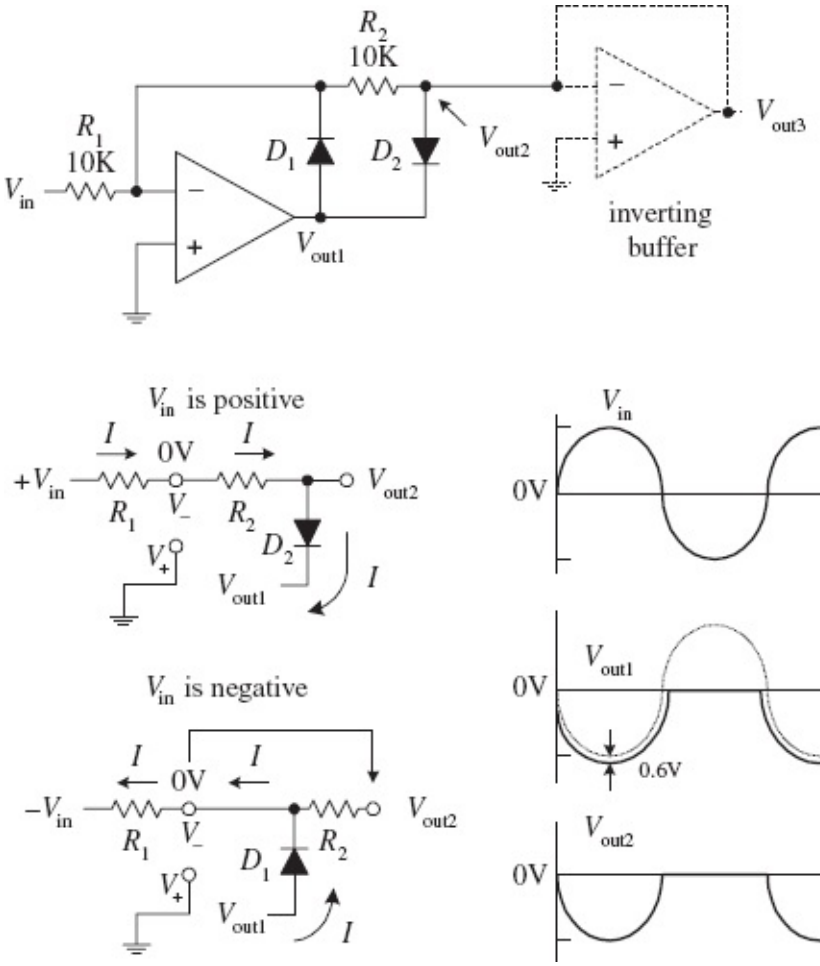
Noninverting Clipper Amplifier



This simple amplifier circuit acts to clip both positive and negative going portions of the output signal. The clipping occurs in the feedback network whenever the feedback voltage exceeds a zener diode's breakdown voltage. Removing one of the zener diodes results in partial clipping (either positive or negative going, depending on which zener diode is removed). This circuit can be used to limit overloads within audio amplifiers and as a simple sinewave-to-squarewave converter.

FIGURE 8.45

Active Rectifiers



A single diode acts to rectify a signal—however, at the cost of a diode drop (e.g., 0.6 V). Not only does this voltage drop lower the level of the output, but it also makes it impossible to rectify low-level signals below 0.6 V. A simple solution to this problem is to construct an active rectifier like the one shown here. This circuit acts as an ideal rectifier, in so far as it rectifies signals all the way down to 0 V. To figure out how this circuit works, let's apply the rules we have learned. If V_{in} is positive, current I will flow in the direction shown in the simplified network shown below the main circuit. Since V_+ is grounded, and since we have feedback, $V_- = V_+ = 0$ V (rule 4), we can use Kirchhoff's voltage law to find V_{out1}, V_{out2}, and finally, V_{out3}: $0\text{ V} - IR_2 - 0.6\text{ V} - V_{out1} = 0$

$$V_{out1} = 0\,\text{V} - \frac{V_{in}R_2}{R_1} - 0.6\,\text{V} = -V_{in} - 0.6\,\text{V}$$

$$V_{out2} = V_{out1} + 0.6\,\text{V} = -V_{in}$$

$$V_{out3} = V_{out2} = -V_{in}$$

Notice that there is no 0.6-V drop present at the final output; however, the output is inverted relative to the input. Now, if V_{in} is negative, the output will source current through D_1 to bring V_- to 0 V (rule 4). But since no current will pass through R_2 (due to buffer), the 0 V at V_- is present at V_{out2} and likewise present at V_{out3}. The buffer stage is used to provide low output impedance for the next stage without loading down the rectifier stage. To preserve the polarity of the input at the output, an inverting buffer (unity-gain inverter) can be attached to the output.

FIGURE 8.46

CHAPTER 9

Filters

A *filter* is a circuit that is capable of passing a specific range of frequencies while blocking other frequencies. As you discovered in Chap. 2, the four major types of filters include *low-pass filters, high-pass filters, bandpass filters,* and *notch filters* (or *band-reject filters*). A low-pass filter passes low-frequency components of an input signal, while a high-pass filter passes high-frequency components. A bandpass filter passes a narrow range of frequencies centered around the filter's resonant frequency, while a notch filter passes all frequencies except those within a narrow band centered around the filter's resonant frequency.

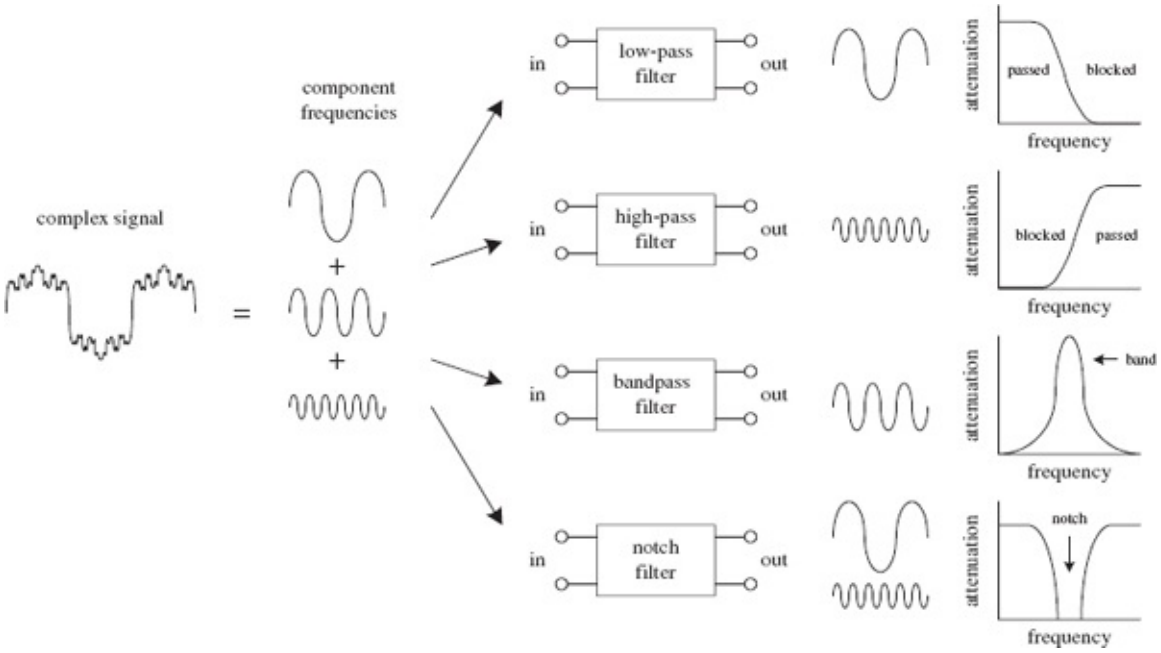


FIGURE 9.1

Filters have many practical applications in electronics. For example, within a dc power supply, filters can be used to eliminate unwanted high-frequency noise present within the ac line voltage, and they act to flatten out pulsing dc voltages generated by the supply's rectifier section. In radio communications, filters make it possible for a radio receiver to provide the listener with only the desired signal while rejecting all others. Likewise, filters allow a radio transmitter to generate only one signal while attenuating other signals that might interfere with different radio transmitters' signals. In audio electronics, filter networks called *crossover networks* are used to divert low audio signals to woofers, middle-range frequencies to midrange speakers, and high frequencies to tweeters. A high-pass filter is often used to eliminate 60 Hz mains hum from audio circuits. The list of filter applications is extensive.

There are two filter types covered in this chapter, namely, *passive filters* and *active filters*. Passive filters are designed using passive elements (e.g., resistors, capacitors, and inductors) and are most responsive to frequencies between around 100 Hz and 300 MHz. (The lower frequency limit results from the fact that at low frequencies the capacitance and inductance values become exceedingly large, meaning

prohibitively large components are needed. The upper frequency limit results from the fact that at high frequencies parasitic capacitances and inductances wreak havoc.) When designing passive filters with very steep attenuation falloff responses, the number of inductor and capacitor sections increases. As more sections are added to get the desired response, greater is the chance for signal loss to occur. Also, source and load impedances must be taken into consideration when designing passive filters.

Active filters, unlike passive filters, are constructed from op amps, resistors, and capacitors—no inductors are needed. Active filters are capable of handling very low frequency signals (approaching 0 Hz), and they can provide voltage gain if needed (unlike passive filters). Active filters can be designed to offer comparable performance to *LC* filters, and they are typically easier to make, less finicky, and can be designed without the need for large-sized components. Also, with active filters, a desired input and output impedance can be provided that is independent of frequency. One major drawback with active filters is a relatively limited high-frequency range. Above around 100 kHz or so, active filters can become unreliable (a result of the op amp's bandwidth and slew-rate requirements). At radiofrequencies, it is best to use a passive filter.

9.1 Things to Know Before You Start Designing Filters

When describing how a filter behaves, a response curve is used, which is simply an attenuation (V_{out}/V_{in}) versus frequency graph (see Fig. 9.2). As you discovered in Chap. 2, attenuation is often expressed in decibels (dB), while frequency may be expressed in either angular form ω (expressed in rad/s) or conventional form f (expressed in Hz). The two forms are related by $\omega = 2\pi f$. Filter response curves may be plotted on linear-linear, log-linear, or log-log paper. In the case of log-linear graphs, the attenuation need not be specified in decibels.

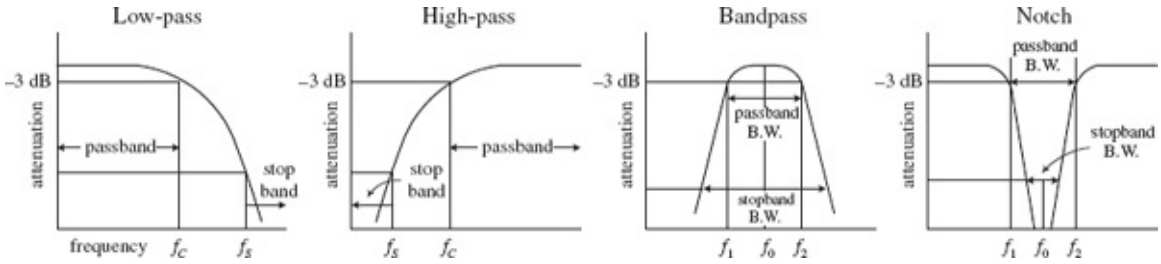


FIGURE 9.2

Here are some terms that are commonly used when describing filter response:

−3-dB Frequency (f_{3dB}). This represents the input frequency that causes the output signal to drop to −3 dB relative to the input signal. The −3-dB frequency is equivalent to the cutoff frequency—the point where the input-to-output power is reduced by one-half or the point where the input-to-output voltage is reduced by $1/\sqrt{2}$. For low-pass and high-pass filters, there is only one −3-dB frequency. However, for bandpass and notch filters, there are two −3-dB frequencies, typically referred to as f_1 and f_2.

Center frequency (f_0). On a linear-log graph, bandpass filters are geometrically symmetrical around the filter's resonant frequency or center frequency—provided the response is plotted on linear-log graph paper (the logarithmic axis representing the frequency). On linear-log paper, the central frequency is related to the −3-dB frequencies by the following expression:

$$f_0 = \sqrt{f_1 f_2}$$

For narrow-band bandpass filters, where the ratio of f_2 to f_1 is less than 1.1, the response shape approaches arithmetic symmetry. In this case, we can approximate f_0 by taking the average of −3-dB frequencies:

$$f_0 = \frac{f_1 + f_2}{2}$$

Passband. This represents those frequency signals that reach the output with no more than −3 dB worth of attenuation.

Stop-band frequency (f_s). This is a specific frequency where the attenuation reaches a specified value set by the designer. For low-pass and high-pass filters, the frequencies beyond the stop-band frequency are referred to as the *stop band*. For bandpass and notch filters, there are two stop-band frequencies, and the frequencies between the stop bands are also collectively called the *stop band*.

Quality factor (Q). This represents the ratio of the center frequency of a bandpass filter to the −3-dB bandwidth (distance between −3-dB points f_1 and f_2):

$$Q = \frac{f_0}{f_2 - f_1}$$

For a notch filter, use $Q = (f_2 - f_1)/f_0$, where f_0 is often referred to as the *null frequency*.

9.2 Basic Filters

In Chap. 2 you discovered that by using the reactive properties of capacitors and inductors, along with the resonant behavior of *LC* series and parallel networks, you could create simple low-pass, high-pass, bandpass, and notch filters. Here's a quick look at the basic filters covered in Chap. 2:

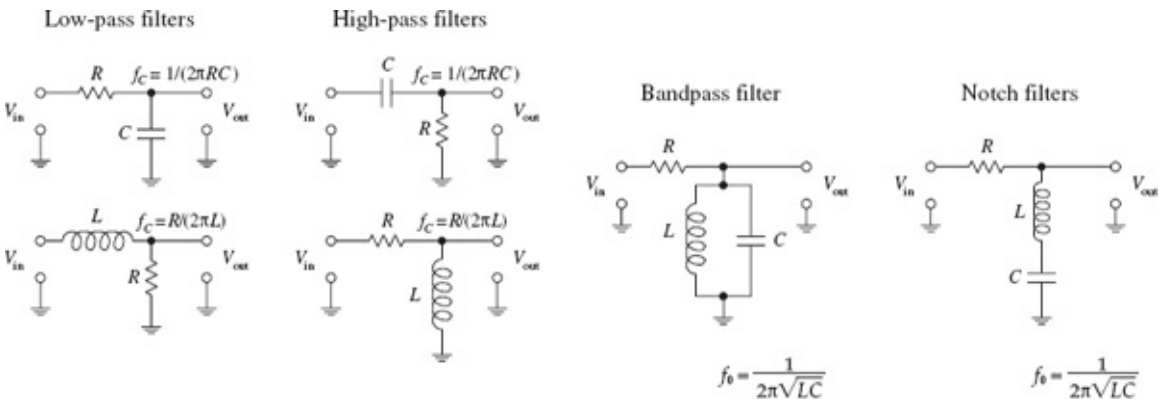


FIGURE 9.3

Now, all the filters shown in this figure have a common limiting characteristic, namely, a shallow 6-dB per octave falloff response beyond the −3-dB point(s). (You can prove this to yourself by going back to Chap. 2 and fiddling with the equations.) In certain noncritical applications, a 6-dB per octave falloff works fine, especially in cases where the signals you want to remove are set well beyond the −3-dB point. However, in situations where greater frequency selectivity is needed (e.g., steeper falloffs and flatter passbands), 6-dB per octave filters will not work. What is needed is a new way to design filters.

Making Filters with Sharper Falloff and Flatter Passband Responses

One approach used for getting a sharper falloff would be to combine a number of 6-dB per octave filters

together. Each new section would act to filter the output of the preceding section. However, connecting one filter with another for the purpose of increasing the "dB per octave" slope is not as easy as it seems and in fact becomes impractical in certain instances (e.g., narrow-band bandpass filter design). For example, you have to contend with transient responses, phase-shift problems, signal degradation, winding capacitances, internal resistances, magnetic noise pickup, etc. Things can get nasty.

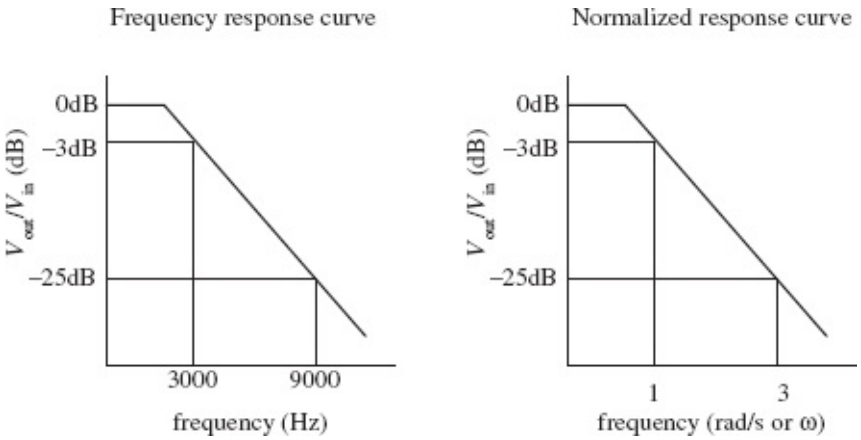
To keep things practical, what I will do is skip the hard-core filter theory (which can indeed get very nasty) and simply apply some design tricks that use basic response graphs and filter design tables. To truly understand the finer points of filter theory is by no means trivial. If you want in-depth coverage of filter theory, refer to a filter design handbook. (A comprehensive handbook written by Zverck covers almost everything you would want to know about filters.)

Let's begin by jumping straight into some practical filter design examples that require varying degrees of falloff response beyond 6 dB per octave. As you go through these examples, important new concepts will surface. First, I will discuss passive filters and then move on to active filters.

9.3 Passive Low-Pass Filter Design

Suppose that you want to design a low-pass filter that has a f_{3dB} = 3000 Hz (attenuation is −3 dB at 3000 Hz) and an attenuation of −25 dB at a frequency of 9000 Hz—which will be called the *stop frequency* f_s. Also, let's assume that both the signal-source impedance R_s and the load impedance R_L are equal to 50 Ω. How do you design the filter?

Step 1 (Normalization)



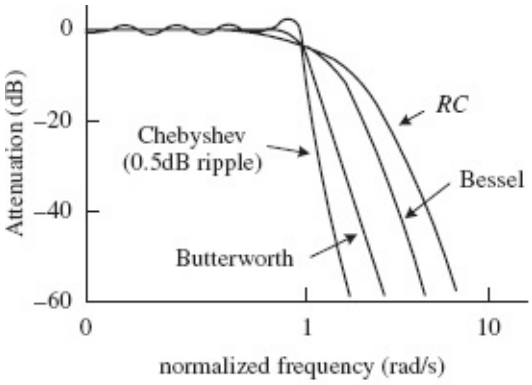
First, make a rough attenuation versus frequency graph to give yourself a general idea of what the response looks like (see far-left figure). Next, you must normalize the graph. This means that you set the −3-dB frequency f_{3dB} to 1 rad/s. The figure to the near left shows the normalized graph. (The reason for normalizing becomes important later on when you start applying design tricks that use normalized response curves and tables.) In order to determine the normalized stop frequency, simply use the following relation, which is also referred to as the *steepness factor:*

$$A_s = \frac{f_s}{f_{3dB}} = \frac{9000\ \text{Hz}}{3000\ \text{Hz}} = 3$$

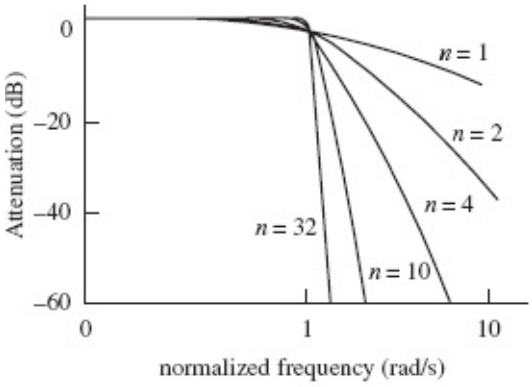
This expression tells you that the normalized stop frequency is three times larger than the normalized −3-dB point of 1 rad/s. Therefore, the normalized stop frequency is 3 rad/s.

FIGURE 9.4

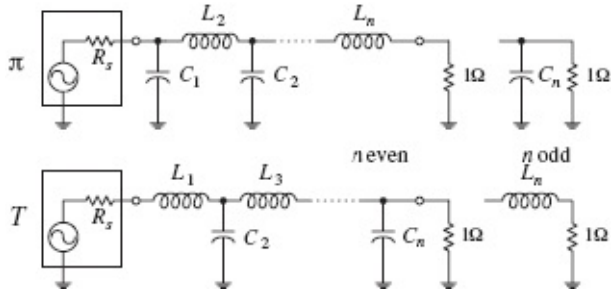
Step 2 (Pick Response Curve)



Normalized low-pass Butterworth filter response curves



LC low-pass filter networks



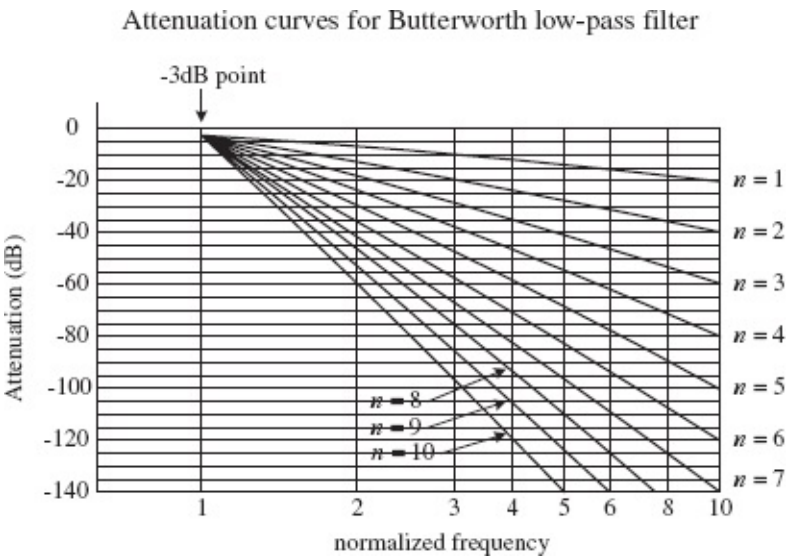
Next, you must pick a filter response type. Three of the major kinds to choose from include the Butterworth, Chebyshev, and Bessel. Without getting too technical here, what is going on is this: Butterworth, Chebyshev, and Bessel response curves are named after individuals who were able to model *LC* filter networks after a mathematical function called the *transfer function*, given here:

$$T(S) = \frac{V_{out}}{V_{in}} = \frac{N_m S^m + N_{m-1} S^{m-1} + \cdots + N_1 S + N_0}{D_n S^n + D_{n-1} S^{n-1} + \cdots + D_1 S + D_0}$$

The *N*'s in the equation are the numerator's coefficients, the *D*'s are the denominator's coefficients, and $S = j\omega$ ($j = \sqrt{-1}$, $\omega = 2\pi f$). The highest power *n* in the denominator is referred to as the order of the filter or the number of poles. The highest power *m* in the numerator is referred to as the number of zeros. Now, by manipulating this function, individuals (e.g., Butterworth, Chebyshev, and Bessel) were able to generate unique graphs of the transfer function that resembled the attenuation response curves of cascaded *LC* filter networks. What is important to know, for practical purposes, is that the number of poles within the transfer function correlates with the number of *LC* sections present within the cascaded filter network and determines the overall steepness of the response curve (the decibels per octave). As the number of poles increases (number of *LC* sections increases), the falloff response becomes steeper. The coefficients of the transfer function influence the overall shape of the response curve and correlate with the specific capacitor and inductor values found within the filter network. Butterworth, Chebyshev, and Bessel came up with their own transfer functions and figured out what values to place in the coefficients and how to influence the slope of the falloff by manipulating the order of the transfer function. Butterworth figured out a way to manipulate the function to give a maximally flat passband response at the expense of steepness in the transition region between the passband and the stop band. Chebyshev figured out a way to get a very steep transition between the passband and stop band at the expense of ripples present in the passband, while Bessel figured out a way to minimize phase shifts at the expense of both flat passbands and steep falloffs. Later I will discuss the pros and cons of Butterworth, Chebyshev, and Bessell filters. For now, however, let's concentrate on Butterworth filters.

FIGURE 9.5

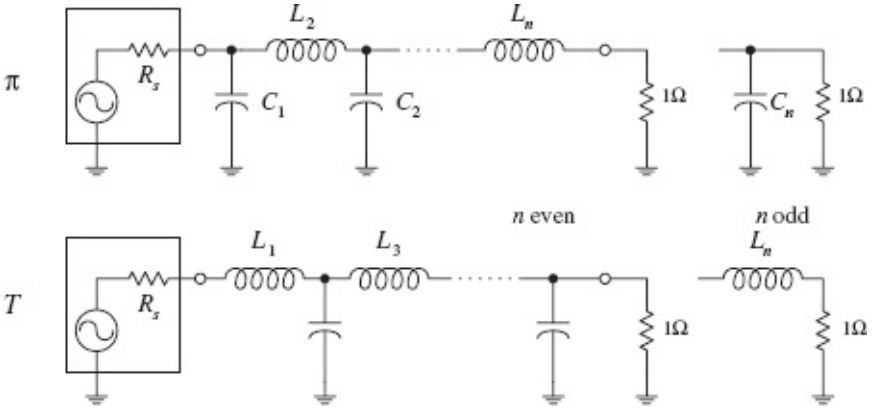
Step 3 (Determine the Number of Poles Needed)



Continuing on with our low-pass filter problem, lef's choose the Butterworth design approach, since it is one of the more popular designs used. The next step is to use a graph of attenuation versus normalized frequency curves for Butterworth low-pass filters, shown in the figure. (Response curves like this are provided in filter handbooks, along with response curves for Chebyshev and Bessel filters.) Next, pick out the single response curve from the graph that provides the desired −25 dB at 3 rad/s, as stated in the problem. If you move your finger along the curves, you will find that the *n* = 3 curve provides sufficient attenuation at 3 rad/s. Now, the filter that is needed will be a third-order low-pass filter, since there are three poles. This means that the actual filter that you will construct will have three *LC* sections.

FIGURE 9.6

Step 4 (Create a Normalized Filter)



Now that you have determined the order of the filter, move on to the next step—creating a normalized *LC* filter circuit. (This circuit will not be the final filter circuit you will use—it will need to be altered.) The circuit networks that are used in this step take on either a π or the *T* configuration, as shown in the figure. If the source and load impedances match, either configuration can be used—though a π network is more attractive because fewer inductors are needed. However, if the load impedance is greater than the source impedance, it is better to use *T* configuration. If the load impedance is smaller than the source impedance, it is better to use the π configuration. Since the initial problem stated that the source and load impedances were both 50 Ω, choose the π configuration. The values of the inductors and capacitors are given in Table 9.1. (Filter handbooks will provide such tables, along with tables for Chebyshev and Bessel filters.) Since you need a third-order filter, use the values listed in the *n* = 3 row. The normalized filter circuit you get in this case is shown in Fig. 9.8.

FIGURE 9.7

TABLE 9.1 Butterworth Active Filter Low-Pass Values

n (π)	R_s ($1/R_s$)	C_1 (L_1)	L_2 (C_2)	C_3 (L_3)	L_4 (C_4)	C_5 (L_5)	L_6 (C_6)	C_7 (L_7)
2	1.000	1.4142	1.4142					
3	1.000	1.0000	2.0000	1.0000				
4	1.000	0.7654	1.8478	1.8478	0.7654			
5	1.000	0.6180	1.6180	2.0000	1.6180	0.6180		
6	1.000	0.5176	1.4142	1.9319	1.9319	1.4142	0.5176	
7	1.000	0.4450	1.2470	1.8019	2.0000	1.8019	1.2470	0.4450

Note: Values of L_n and C_n are for a 1-Ω load and −3-dB frequency of 1 rad/s and have units of H and F. These values must be scaled down. See text.

As mentioned a moment ago, this circuit is not the final circuit that we'll use. That is, the component values listed here will not work! This is so because the graphs and tables you used to get to this point used the normalized frequency within them. Also, you haven't considered the effects of the source and load impedances. In order to construct the final working circuit, you must frequency and impedance scale the component values listed in the circuit in Fig. 9.8. This leads us to the next step.

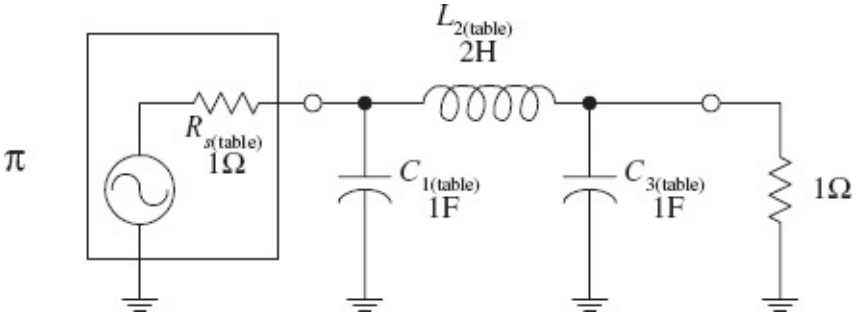


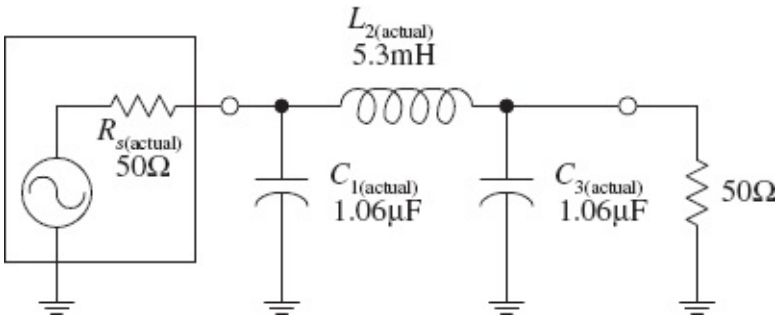
FIGURE 9.8

Step 5 (Frequency and Impedance Scaling)

$$L_{2(actual)} = \frac{R_L L_{2(table)}}{2\pi f_{3dB}} = \frac{(50\Omega)(2\text{ H})}{2\pi(3000\text{ Hz})} = 5.3\text{ mH}$$

$$C_{1(actual)} = \frac{C_{1(table)}}{2\pi f_{3dB} R_L} = \frac{1\text{ F}}{2\pi(3000\text{ Hz})(50\ \Omega)} = 1.06\mu\text{F}$$

$$C_{3(actual)} = \frac{C_{3(table)}}{2\pi f_{3dB} R_L} = \frac{1\text{ F}}{2\pi(3000\text{ Hz})(50\ \Omega)} = 1.06\mu\text{F}$$



To account for impedance matching of the source and load, as well as getting rid of the normalized frequency, apply the following frequency and impedance scaling rules. To frequency scale, divide the capacitor and inductor values that you got from the table by $\omega = 2\pi f_c$. To impedance scale, multiply resistor and inductor values by the load impedance and divide the capacitor values by the load impedance. In other words, use the following two equations to get the actual component values needed:

$$L_{n(\text{actual})} = \frac{R_L L_{n(\text{table})}}{2\pi f_{3\text{dB}}}$$

$$C_{n(\text{actual})} = \frac{C_{n(\text{table})}}{2\pi f_{3dB} R_L}$$

The calculations and the final low-pass circuit are shown in the figure.

FIGURE 9.9

9.4 A Note on Filter Types

It was briefly mentioned earlier that Chebyshev and Bessel filters could be used instead of Butterworth filters. To design Chebyshev and Bessel filters, you take the same approach you used to design Butterworth filters. However, you need to use different low-pass attenuation graphs and tables to come up with the component values placed in the π and *T LC* networks. If you are interested in designing Chebyshev and Bessel filters, consult a filter design handbook. Now, to give you a better understanding of the differences between the various filter types, the following few paragraphs should help.

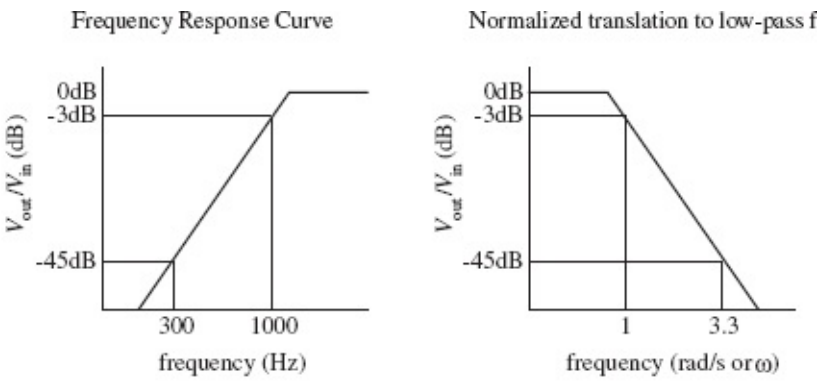
Butterworth filters are perhaps the most popular filters used. They have very flat frequency response in the middle passband region, although they have somewhat rounded bends in the region near the −3-dB point. Beyond the −3-dB point, the rate of attenuation increases and eventually reaches $n \times 6$ dB per octave (e.g., $n = 3$, attenuation = 18 dB/octave). Butterworth filters are relatively easy to construct, and the components needed tend not to require as strict tolerances as those of the other filters.

Chebyshev filters (e.g., 0.5-dB ripple, 0.1-dB ripple Chebyshev filter) provide a sharper rate of descent in attenuation beyond the −3-dB point than Butterworth and Bessel filters. However, there is a price to pay for the steep descent—the cost is a ripple voltage within the passband, referred to as the *passband ripple*. The size of the passband ripple increases with order of the filter. Also, Chebyshev filters are more sensitive to component tolerances than Butterworth filters.

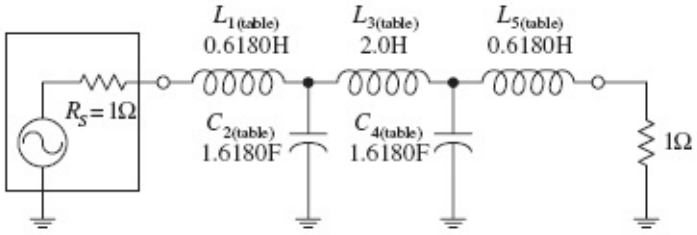
Now, there is a problem with Butterworth and Chebyshev filters—they both introduce varying amounts of delay time on signals of different frequencies. In other words, if an input signal consists of a multiple-frequency waveform (e.g., a modulated signal), the output signal will become distorted because different frequencies will be displaced by different delay times. The delay-time variation over the passband is called *delay distortion*, and it increases as the order of the Butterworth and Chebyshev filters increases. To avoid this effect, a Bessel filter can be used. Bessel filters, unlike Butterworth and Chebyshev filters, provide a constant delay over the passband. However, unlike the other two filters, Bessel filters do not have as sharp an attenuation falloff. Having a sharp falloff, however, is not always as important as good signal reproduction at the output. In situations where actual signal reproduction is needed, Bessel filters are more reliable.

9.5 Passive High-Pass Filter Design

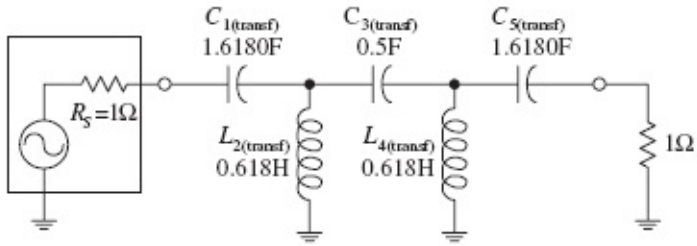
Suppose that you want to design a high-pass filter that has an $f_{3\text{dB}} = 1000$ Hz and an attenuation of at least −45 dB at 300 Hz—which we call the *stop frequency* f_s. Assume that the filter is hooked up to a source and load that both have impedances of 50 Ω and that a Butterworth response is desired. How do you design the filter? The trick, as you will see in a second, involves treating the high-pass response as an inverted low-pass response, then designing a normalized low-pass filter, applying some conversion tricks on the low-pass filter's components to get a normalized high-pass filter, and then frequency and impedance scaling the normalized high-pass filter.



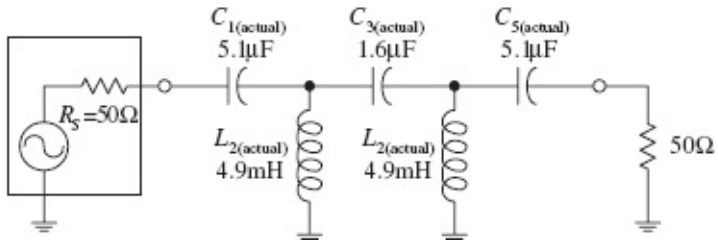
Start with a "*T*" low-pass filter...



Transform low-pass filter into a high-pass filter...



Impedance and frequency scale high-pass filter to get final circuit



First, make a simple sketch of the response curve for the high-pass filter, as shown in the far-left graph. Next, take the high-pass curve and flip it around in the horizontal direction to get a low-pass response. Then normalize the low-pass response. (This allows you to use the low-pass design techniques. Later you will need to apply a transformation trick on the normalized component values of the low-pass filter to get the desired high-pass filter.) To find the steepness factor A_s and normalized stop-band frequency f_s, follow the same basic procedure as you used in the low-filter example, except now you must take f_{3dB} over *f:*

$$A_s = \frac{f_{3dB}}{f_s} = \frac{1000 \text{ Hz}}{300 \text{ Hz}} = 3.3$$

This expression tells us that the normalized stop-band frequency is 3.3 times larger than the normalized −3-dB frequency. Since the normalized graph sets f_{3dB} to 1 rad/s, f_s becomes 3.3 rad/s.

Next, take the low-pass filter response from the preceding step and determine which response curve in Fig. 9.6 provides an attenuation of at least −45 dB at 3.3 rad/s. The n = 5 curve does the trick, so you create a fifth-order *LC* network. Now, the question to ask is, Do you use the π or the *T* network? Initially, you might assume the π network would be best, since the load and source impedances are equal and since fewer inductors are needed. However, when you apply the transformational trick to get the low-pass filter back to a high-pass filter, you will need to interchange inductors for capacitors and capacitors for inductors. Therefore, if you choose the low-pass *T* network now, you will get fewer inductors in the final high-pass circuit. The fifth-order normalized low-pass filter network is shown in the figure.

To convert the low-pass into a high-pass filter, replace the inductors with capacitors that have value of 1/*L*, and replace the capacitors with inductors that have values of 1/*C*. In other words, do the following:

FIGURE 9.10

$$C_{1(\text{transf})} = 1/L_{1(\text{table})} = 1/0.6180 = 1.6180 \text{ F}$$

$$C_{3(\text{transf})} = 1/L_{3(\text{table})} = 1/2.0 = 0.5 \text{ F}$$

$$C_{5(\text{transf})} = 1/L_{5(\text{table})} = 1/0.6180 = 1.6180 \text{ F}$$

$$L_{2(\text{transf})} = 1/C_{2(\text{table})} = 1/1.6180 = 0.6180 \text{ H}$$

$$L_{4(\text{transf})} = 1/C_{4(\text{table})} = 1/1.6180 = 0.6180 \text{ H}$$

Next, frequency and impedance scale to get the actual component values:

$$C_{1(\text{actual})} = \frac{C_{1(\text{trans})}}{2\pi f_{3\text{dB}} R_L} = \frac{1.618 \text{ H}}{2\pi(1000 \text{ Hz})(50\ \Omega)} = 5.1\ \mu\text{F}$$

$$C_{3(\text{actual})} = \frac{C_{3(\text{trans})}}{2\pi f_{3\text{dB}} R_L} = \frac{0.5 \text{ H}}{2\pi(1000 \text{ Hz})(50\ \Omega)} = 1.6\ \mu\text{F}$$

$$C_{5(\text{actual})} = \frac{C_{5(\text{trans})}}{2\pi f_{3\text{dB}} R_L} = \frac{1.618 \text{ H}}{2\pi(1000 \text{ Hz})(50\ \Omega)} = 5.1\ \mu\text{F}$$

$$L_{2(\text{actual})} = \frac{L_{2(\text{trans})} R_L}{2\pi f_{3\text{dB}}} = \frac{(0.6180 \text{ F})(50\ \Omega)}{2\pi(1000 \text{ Hz})} = 4.9 \text{ mH}$$

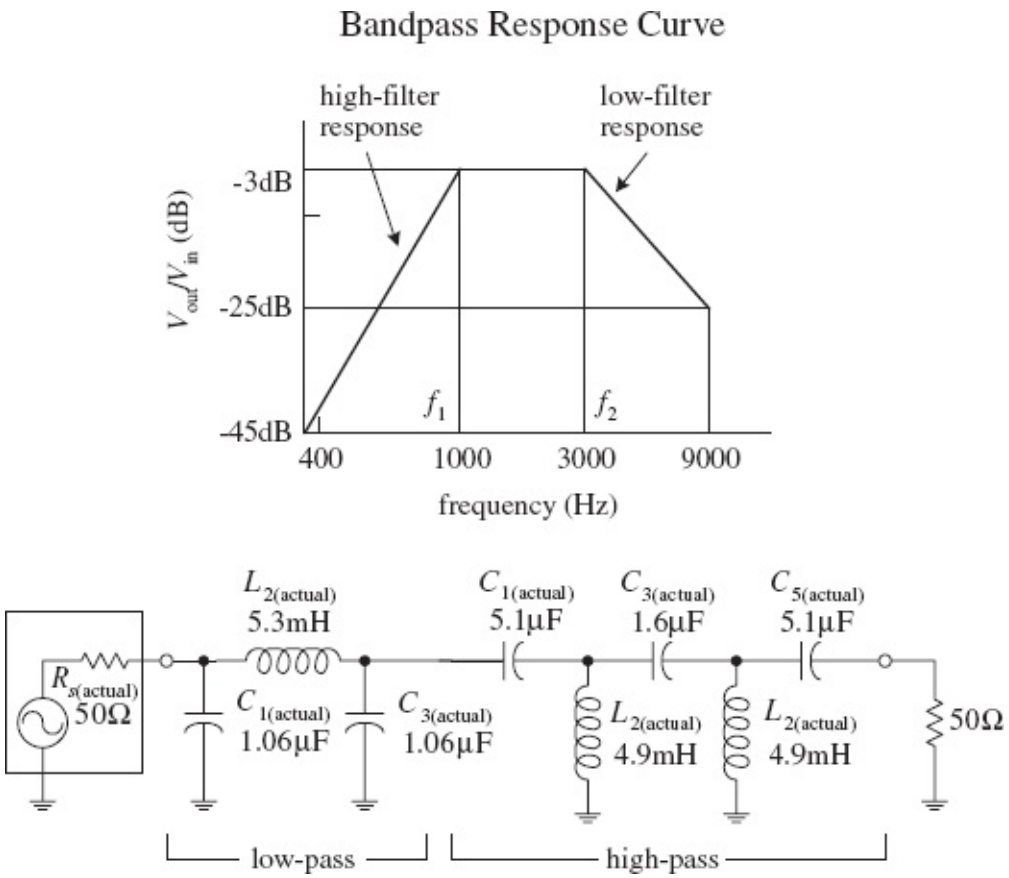
$$L_{4(\text{actual})} = \frac{L_{4(\text{trans})} R_L}{2\pi f_{3\text{dB}}} = \frac{(0.6180 \text{ F})(50\ \Omega)}{2\pi(1000 \text{ Hz})} = 4.9 \text{ mH}$$

9.6 Passive Bandpass Filter Design

Bandpass filters can be broken down into narrow-band and wide-band types. The defining difference between the two is the ratio between the upper −3-dB frequency f_1 and lower −3-dB frequency f_2. If f_2/f_1 is greater than 1.5, the bandpass filter is placed in the wide-type category. Below 1.5, the bandpass filter is placed in the narrow-band category. As you will see in a moment, the procedure used to design a wide-band bandpass filter differs from that used to design a narrow-band filter.

Wide-Band Design

The basic approach used to design wide-band bandpass filters is simply to combine a low-pass and high-pass filter together. The following example will cover the details. Suppose that you want to design a bandpass filter that has −3-dB points at f_1 = 1000 Hz and f_2 = 3000 Hz and at least −45 dB at 300 Hz and more than −25 dB at 9000 Hz. Also, again assume that the source and load impedances are both 50 Ω and a Butterworth design is desired.



The basic sketch shown here points out the basic response desired. The ratio f_2/f_1 = 3, which is larger than 1.5, so you do indeed have a wide-band situation. Notice how the sketch resembles low-pass and high-pass response curves placed together on the same graph. If you break up the response into low- and high-phase curves, you get the following results:

Low-pass	–3 dB at 3000 Hz
	–25 dB at 9000 Hz
High-pass	–3 dB at 1000 Hz
	–45 dB at 300 Hz

Now, to design the wide-band bandpass filter, construct a low-pass and high-pass filter using the values above and the design technique used in the preceding two example problems. Once you have done this, simply cascade the low-pass and high-pass filters together. The nice thing about this problem is the low-pass and high-pass filters that are needed are simply the filters used in the preceding low-pass and high-pass examples. The final cascade network is shown at the bottom of the figure.

FIGURE 9.11

Narrow-Band Design

Narrow-band filters ($f_2/f_1 < 1.5$), unlike wide-band filters, cannot be made simply by cascading low-pass and high-pass filters together. Instead, you must use a new, slightly tricky procedure. This procedure involves transforming the −3-dB band-width ($\Delta f_{BW} = f_2 - f_1$) of a bandpass filter into the −3-dB frequency f_{3dB} of a low-pass filter. At the same time, the stop-band bandwidth of the bandpass filter is transformed into the corresponding stop-band frequency of a low-pass filter. Once this is done, a normalized low-pass filter is created. After the normalized low-pass filter is created, the filter must be frequency scaled in a special way to get the desired bandpass filter. (The normalized circuit also must be impedance scaled, as before.) When frequency scaling the components of the normalized low-pass filter, do not divide by $\omega = 2\pi f_{3dB}$, as you would do with low-pass scaling. Instead, the normalized low-pass filter's components are divided by $2\pi(\Delta f_{BW})$. Next, the scaled circuit's branches must be resonated to the bandpass filter's center frequency f_0 by placing additional inductors in parallel with the capacitors and placing additional capacitors in series with the inductors. This creates *LC* resonant circuit sections. The values of the additional inductors and capacitors are determined by using the *LC* resonant equation (see Chap. 2 for the

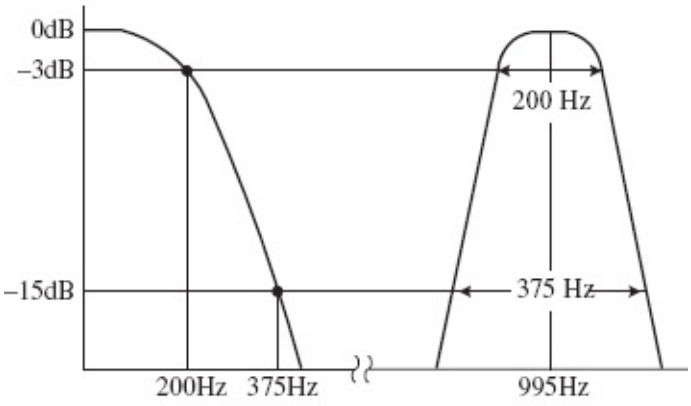
details):

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

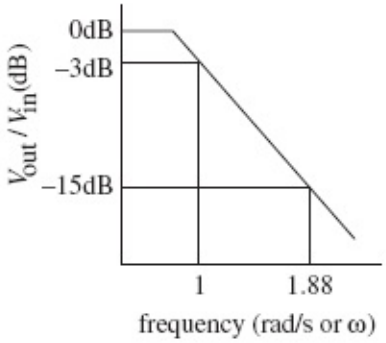
NARROW-BANDWIDTH BANDPASS FILTER EXAMPLE

Suppose that you want to design a bandpass filter with −3-dB points at f_1 = 900 Hz and f_2 = 1100 Hz and at least −20 dB worth of attenuation at 800 and 1200 Hz. Assume that both the source and load impedances are 50 Ω and that a Butterworth design is desired.

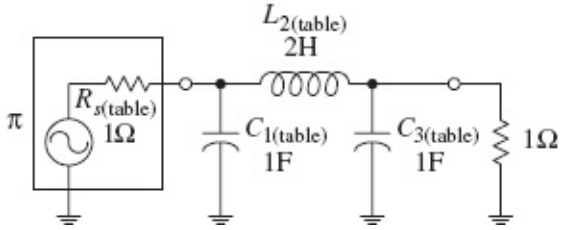
Low-pass bandpass relationship



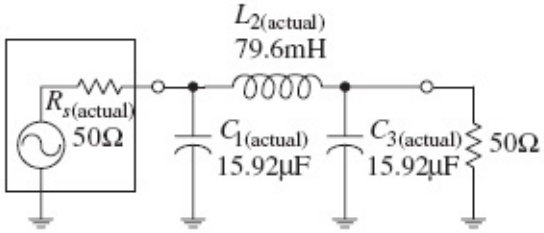
Normalized low-pass response



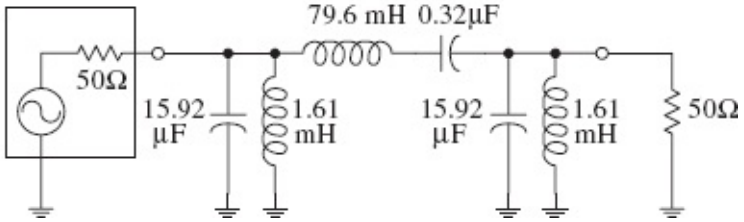
Normalized low-pass filter



Impedance and frequency scaled low-pass filter



Final bandpass filter



Since f_2/f_1 = 1.2, which is less than 1.5, a narrow-band filter is needed. The initial step in designing a narrow-band bandpass filter is to normalize the bandpass requirements. First, the geometric center frequency is determined:

$$f_0 = \sqrt{f_1 f_2} = \sqrt{(900\text{ Hz})(1100)} = 995\text{ Hz}$$

Next, compute the two pair of geometrically related stop-band frequencies by using

$$f_a f_b = f_0^2$$

$$f_a = 800\text{ Hz} \qquad f_b = \frac{f_0^2}{f_a} = \frac{(995\text{ Hz})^2}{800\text{ Hz}} = 1237\text{ Hz} \qquad f_b - f_a = 437\text{ Hz}$$

$$f_b = 1200\text{ Hz} \qquad f_a = \frac{f_0^2}{f_b} = \frac{(995\text{ Hz})^2}{1200\text{ Hz}} = 825\text{ Hz} \qquad f_b - f_a = 375\text{ Hz}$$

Notice how things are a bit confusing. For each pair of stop-band frequencies, you get two new pairs—a result of making things "geometrical" with respect to f_0. Choose the pair having the least separation, which represents the more severe requirement −375 Hz.

The steepness factor for the bandpass filter is given by

$$A_s = \frac{\text{stop-band bandwidth}}{\text{3-dB bandwidth}} = \frac{375\text{ Hz}}{200\text{ Hz}} = 1.88$$

Now choose a low-pass Butterworth response that provides at least −20 dB at 1.88 rad/s. According Fig. 9.6, the n = 3 curve does the trick. The next step is to create a third-order normalized low-pass filter using the π configuration and Table 9.1.

Next, impedance and frequency scale the normalized low-pass filter to require an impedance level of 50 Ω. and a −3-dB frequency equal to the desired bandpass filter's bandwidth ($\Delta f_{BW} = f_2 - f_1$)—which in this example equals 200 Hz. Notice the frequency-scaling trick! The results follow:

$$C_{1(\text{actual})} = \frac{C_{1(\text{table})}}{2\pi(\Delta f_{BW})R_L} = \frac{1\text{ F}}{2\pi(200\text{ Hz})(50\ \Omega)} = 15.92\ \mu\text{F}$$

$$C_{3(\text{actual})} = \frac{C_{3(\text{table})}}{2\pi(\Delta f_{BW})R_L} = \frac{1\text{ F}}{2\pi(200\text{ Hz})(50\ \Omega)} = 15.92\ \mu\text{F}$$

$$L_{2(\text{actual})} = \frac{L_{2(\text{table})}R_L}{2\pi(\Delta f_{BW})} = \frac{(2\text{ H})(50\ \Omega)}{2\pi(200\text{ Hz})} = 79.6\text{ mH}$$

The important part comes now. Each circuit branch of the low-pass filter must be resonated to f_0 by adding a series capacitor to each inductor and a parallel inductor to each capacitor. The *LC* resonant equation is used to determine the additional component values:

$$L_{(\text{parallel with } C1)} = \frac{1}{(2\pi f_0)^2 C_{1(\text{actual})}} = \frac{1}{(2\pi \cdot 995\ \text{Hz})^2 (15.92\ \mu\text{F})} = 1.61\ \text{mH}$$

$$L_{(\text{parallel with } C3)} = \frac{1}{(2\pi f_0)^2 C_{3(\text{actual})}} = \frac{1}{(2\pi \cdot 995\ \text{Hz})^2 (15.92\ \mu\text{F})} = 1.61\ \text{mH}$$

$$C_{(\text{series with } L2)} = \frac{1}{(2\pi f_0)^2 L_{2(\text{actual})}} = \frac{1}{(2\pi \cdot 995\ \text{Hz})^2 (79.6\ \text{mH})} = 0.32\ \mu\text{F}$$

The final bandpass circuit is shown at the bottom of the figure.

FIGURE 9.12

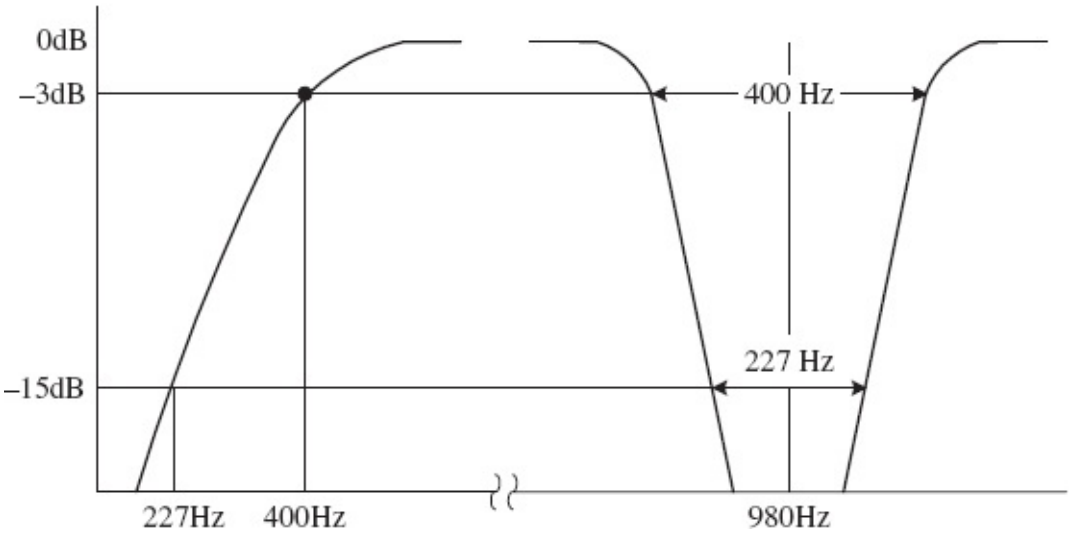
9.7 Passive Notch Filter Design

To design a notch filter, you can apply a technique similar to the one you used in the narrow-band bandpass example. However, now you use a high-pass filter instead of a low-pass filter as the basic building block. The idea here is to relate the notch filter's −3-dB bandwidth ($\Delta f_{BW} = f_1 - f_2$) to the −3-dB frequency of a high-pass filter and relate the notch filter's stop-band bandwidth to the stop-band frequency of a high-pass filter. After that, a normalized high-pass filter is created. This filter is then frequency scaled in a special way—all its components are divided by $2\pi\Delta f_{BW}$. (This circuit also must be impedance scaled, as before.) As with the narrow-band bandpass filter example, the scaled high-pass filter's branches must be resonated to the notch filter's center frequency f_0 by inserting additional series capacitors with existing inductors and inserting additional parallel inductors with existing capacitors.

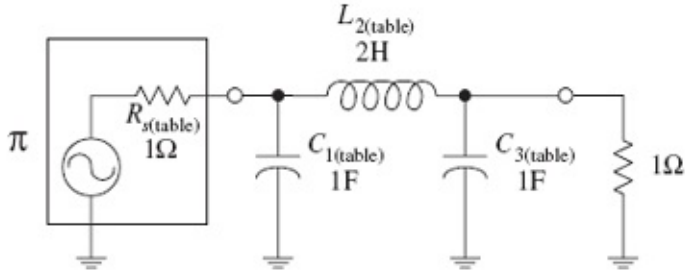
EXAMPLE

Suppose that you want to design a notch filter with −3-dB points at f_1 = 800 Hz and f_2 = 1200 Hz and at least −20 dB at 900 and 1100 Hz. Let's assume that both the source and load impedances are 600 Ω and that a Butterworth design is desired.

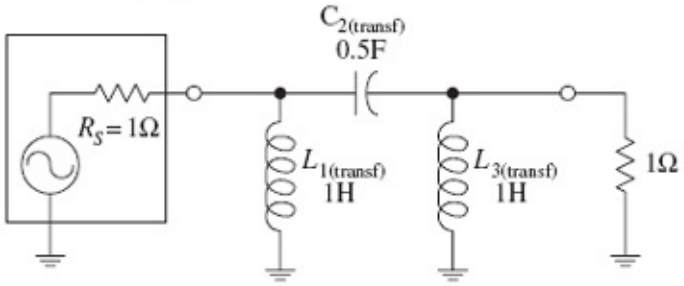
High-pass bandpass relationship



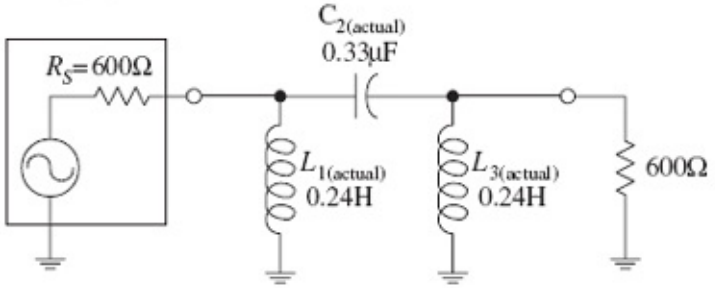
Normalized low-pass filter



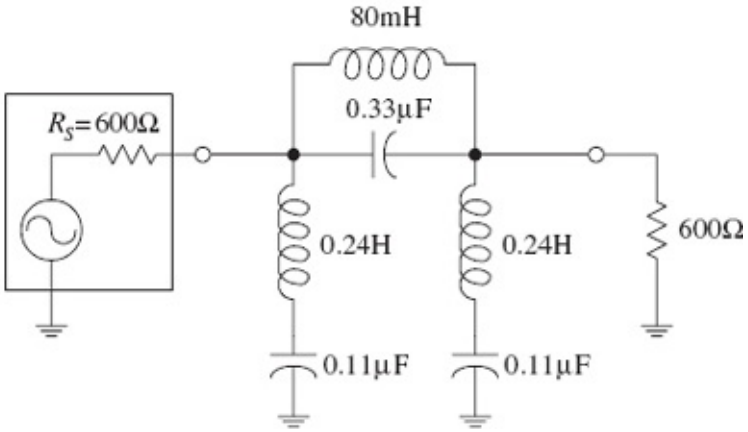
Normalized high-pass filter



Actual high-pass filter



Final bandpass filter



First, you find the geometric center frequency:

$$f_0 = \sqrt{f_1 f_2} = \sqrt{(800\text{ Hz})(1200\text{ Hz})} = 980\text{ Hz}$$

Next, compute the two pairs of geometrically related stop-band frequencies:

$$f_a = 900Hz \qquad f_b = \frac{f_0^2}{f_a} = \frac{(980\text{ Hz})^2}{900\text{ Hz}} = 1067\text{ Hz}$$

$$f_b - f_a = 1067\text{ Hz} - 900\text{ Hz} = 167\text{ Hz}$$

$$f_b = 1100\text{ Hz} \qquad f_a = \frac{f_0^2}{f_b} = \frac{(980\text{ Hz})^2}{1100\text{ Hz}} = 873\text{ Hz}$$

$$f_b - f_a = 1100\text{ Hz} - 873\text{ Hz} = 227\text{ Hz}$$

Choose the pair of frequencies that gives the more severe requirement—227 Hz.

Next, compute the notch filter's steepness factor, which is given by

$$A_S = \frac{\text{3-dB bandwidth}}{\text{stop-band bandwidth}} = \frac{400\ \text{Hz}}{227\ \text{Hz}} = 1.7$$

To come up with the final notch filter design, start out by treating the steepness factor as the steepness factor for a high-pass filter. Next, apply the same tricks you used earlier to construct a high-pass filter. Horizontally flip the high-pass response to get a low-pass response. Then normalize the low-pass response (setting the normalized stop frequency to 1.7 rad/s) and use Fig. 9.6 (n = 3 provides at least −20 dB at 1.7 rad/s). Next, use Table 9.1 and the π network to come up with a normalized low-pass design. Then apply the low-pass to high-pass transformational tricks to get a normalized high-pass filter:

$$L_{1(\text{transf})} = 1/C_{1(\text{table})} = 1/1 = 1\ \text{H}$$

$$L_{3(\text{transf})} = 1/C_{3(\text{table})} = 1/1 = 1\ \text{H}$$

$$C_{2(\text{transf})} = 1/L_{2(\text{table})} = 1/1 = 0.5\ \text{F}$$

The first three circuits in the figure show the low-pass to high-pass transformational process.

Next, impedance and frequency scale the normalized high-pass filter to require an impedance level of 600 Ω and a −3-dB frequency equal to the desired notch filter's bandwidth ($\Delta f_{BW} = f_2 - f_1$)—which in the example equals 400 Hz. Notice the frequency-scaling trick! The results follow:

$$L_{1(\text{actual})} = \frac{R_L L_{1(\text{transf})}}{2\pi(\Delta f_{BW})} = \frac{(600\ \Omega)(1\ \text{H})}{2\pi(400\ \text{Hz})} = 0.24\ \text{H}$$

$$L_{3(\text{actual})} = \frac{R_L L_{3(\text{transf})}}{2\pi(\Delta f_{BW})} = \frac{(600\ \Omega)(1\ \text{H})}{2\pi(400\ \text{Hz})} = 0.24\ \text{H}$$

$$C_{2(\text{actual})} = \frac{C_{1(\text{transf})}}{2\pi(\Delta f_{BW})R_L} = \frac{(0.5\ \text{F})}{2\pi(400\ \text{Hz})(600\ \Omega)} = 0.33\ \mu\text{F}$$

And finally, the important modification—resonate each branch to the notch filter's center frequency f_0 by adding a series capacitor to each inductor and a parallel inductor to each capacitor. The values for these additional components must be

$$C_{(\text{series with } L1)} = \frac{1}{(2\pi f_0)^2 L_{1(\text{actual})}} = \frac{1}{(2\pi \cdot 400\ \text{Hz})^2(0.24\ \text{H})} = 0.11\ \mu\text{F}$$

$$C_{(\text{series with } L3)} = \frac{1}{(2\pi f_0)^2 L_{3(\text{actual})}} = \frac{1}{(2\pi \cdot 400\ \text{Hz})^2(0.24\ \text{H})} = 0.11\ \mu\text{F}$$

$$L_{(\text{parallel with } L1)} = \frac{1}{(2\pi f_0)^2 C_{2(\text{actual})}} = \frac{1}{(2\pi \cdot 400\ \text{Hz})^2(0.33\ \mu\text{F})} = 80\ \text{mH}$$

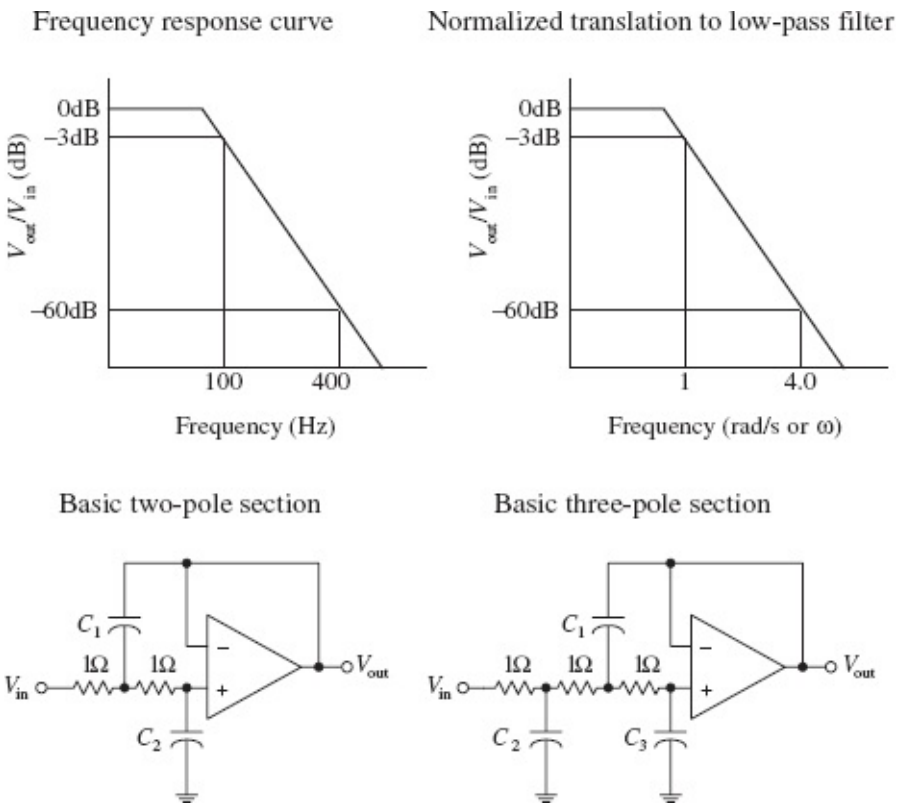
The final circuit is shown at the bottom of the figure.

FIGURE 9.13

9.8 Active Filter Design

This section covers some basic Butterworth active filter designs. I already discussed the pros and cons of active filter design earlier in this chapter. Here I will focus on the actual design techniques used to make unity-gain active filters. To begin, let's design a low-pass filter.

9.8.1 Active Low-Pass Filter Example



Suppose that you wish to design an active low-pass filter that has a 3-dB point at 100 Hz and at least 60 dB worth of attenuation at 400 Hz—which we'll call the *stop frequency* f_s.

The first step in designing the filter is to normalize low-pass requirements. The steepness factor is

$$A_s = \frac{f_s}{f_{3dB}} = \frac{400\text{ Hz}}{100\text{ Hz}} = 4$$

This means that the normalized position of f_s is set to 4 rad/s. See the graphs in Fig. 9.14. Next, use the Butterworth low-pass filter response curves in Fig. 9.6 to determine the order of filter you need. In this case, the $n = 5$ curve provides over −60 dB at 4 rad/s. In other words, you need a fifth-order filter.

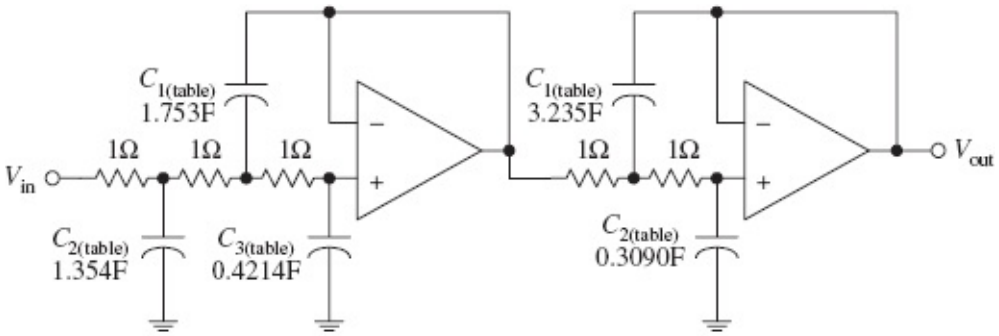
Now, unlike passive filters design, active filter design requires the use of a different set of basic normalized filter networks and a different table to provide the components of the networks. The active filter networks are shown in Fig. 9.14—there are two of them. The one to the left is called a *two-pole section*, while the one on the right is called a *three-pole section*. To design a Butterworth low-pass normalized filter of a given order, use Table 9.2. (Filter handbooks provide Chebyshev and Bessel tables as well.) In this example, a five-pole filter is needed, so according to the table, two sections are required—a three-pole and a two-pole section. These sections are cascaded together, and the component values listed in Table 9.2 are placed next to the corresponding components within the cascaded network. The resulting normalized low-pass filter is shown in Fig. 9.15.

FIGURE 9.14

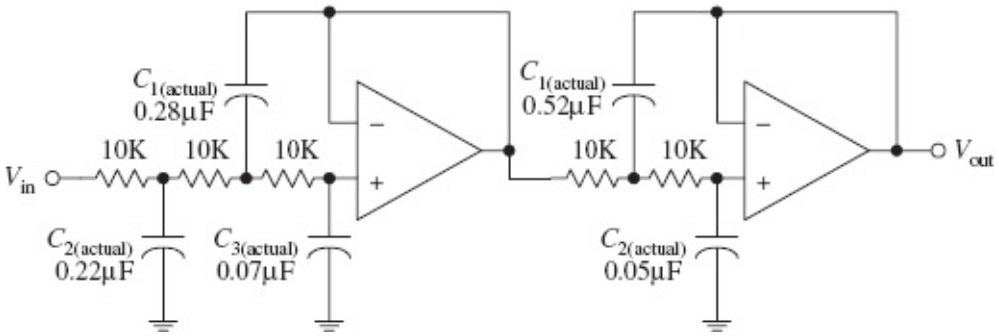
TABLE 9.2 Butterworth Normalized Active Low-Pass Filter Values

ORDER n	NUMBER OF SECTIONS	SECTIONS	C_1	C_2	C_3
2	1	2-pole	1.414	0.7071	
3	1	3-pole	3.546	1.392	0.2024
4	2	2-pole	1.082	0.9241	
		2-pole	2.613	0.3825	
5	2	3-pole	1.753	1.354	0.4214
		2-pole	3.235	0.3090	
6		2-pole	1.035	0.9660	
	3	2-pole	1.414	0.7071	
		2-pole	3.863	0.2588	
7		3-pole	1.531	1.336	0.4885
	3	2-pole	1.604	0.6235	
		2-pole	4.493	0.2225	
8		2-pole	1.020	0.9809	
	4	2-pole	1.202	0.8313	
		2-pole	2.000	0.5557	
		2-pole	5.758	0.1950	

Normalized low-pass filter



Final low-pass filter



The normalized filter will provide the correct response, but the component values are impractical—they're too big. In order to bring these values down to size, the circuit must be frequency and impedance scaled. To frequency scale, simply divide the capacitor values by $2\pi f_{3dB}$ (you need not frequency scale the resistors—they aren't reactive). In terms of impedance scaling, you do not have to deal with source/load impedance matching. Instead, simply multiply the normalized filter circuit's resistors by a factor of Z and divide the capacitors by the same factor. The value of Z is chosen to scale the normalized filter components to more practical values. A typical value for Z is 10,000 Ω. In summary, the final scaling rules are expressed as follows:

$$C_{(\text{actual})} = \frac{C_{(\text{table})}}{Z \cdot 2\pi f_{3\text{dB}}}$$

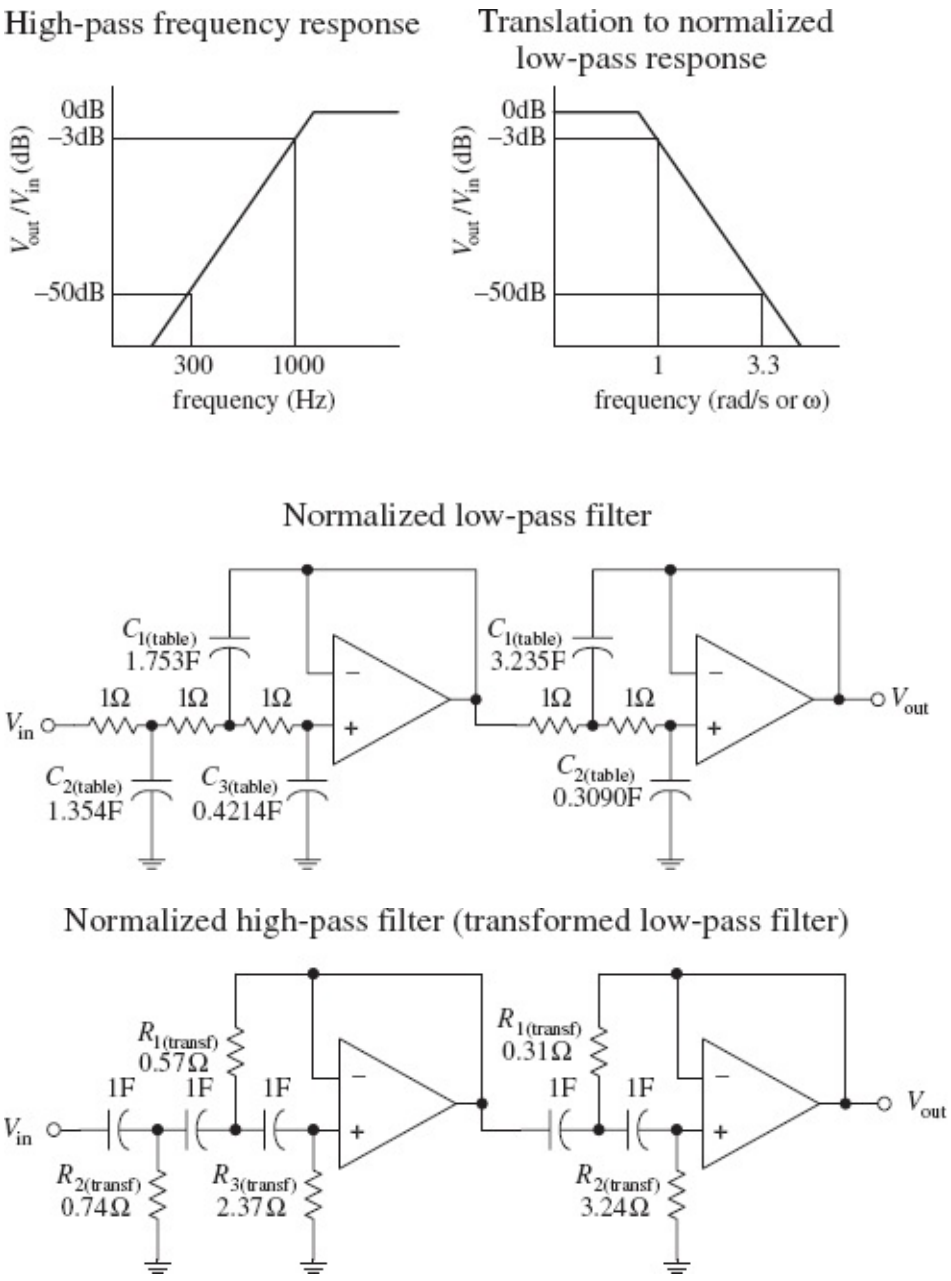
$$R_{(\text{actual})} = ZR_{(\text{table})}$$

Taking *Z* to be 10,000 Ω, you get the final low-pass filter circuit shown at the bottom of the figure.

FIGURE 9.15

9.8.2 Active High-Pass Filter Example

The approach used to design active high-pass filters is similar to the approach used to design passive high-pass filters. Take a normalized low-pass filter, transform it into a high-pass circuit, and then frequency and impedance scale it. For example, suppose that you want to design a high-pass filter with a −3-dB frequency of 1000 Hz and 50 dB worth of attenuation at 300 Hz. What do you do?



The first step is to convert the high-pass response into a normalized low-pass response, as shown in the figure. The steepness factor for the low-pass equivalent response is given by

$$A_s = \frac{f_{3\text{dB}}}{f_s} = \frac{1000 \text{ Hz}}{300 \text{ Hz}} = 3.3$$

This means that the stop frequency is set to 3.3 rad/s on the normalized graph. The Butterworth response curve shown in Fig. 9.6 tells you that a fifth-order (n = 5) filter will provide the needed attenuation response. Like the last example, a cascaded three-pole/two-pole normalized low-pass filter is required. This filter is shown in Fig. 9.16.

Next, the normalized low-pass filter must be converted into a normalized high-pass filter. To make the conversion, exchange resistors for capacitors that have values of 1/R F, and exchange capacitors with resistors that have values of 1/C Ω. The second circuit in Fig. 9.16 shows the transformation.

Like the last example problem, to construct the final circuit, the normalized high-pass filter's component values must be frequency and impedance scaled:

$$C_{(\text{actual})} \frac{C_{(\text{transf})}}{Z \cdot 2\pi f_{3\text{dB}}}$$

$$R_{(\text{actual})} = ZR_{(\text{transf})}$$

Again, let Z = 10,000 Ω. The final circuit is shown in Fig. 9.17.

FIGURE 9.16

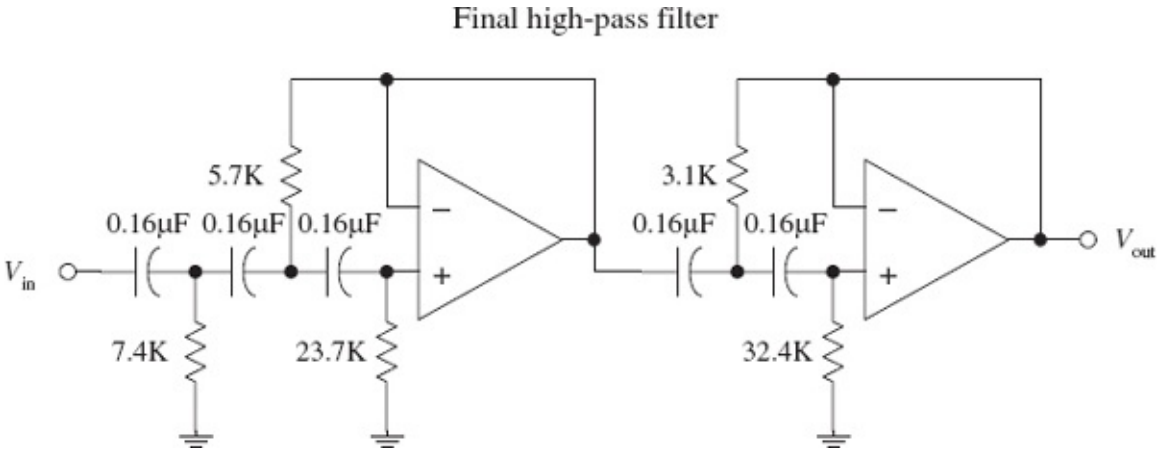


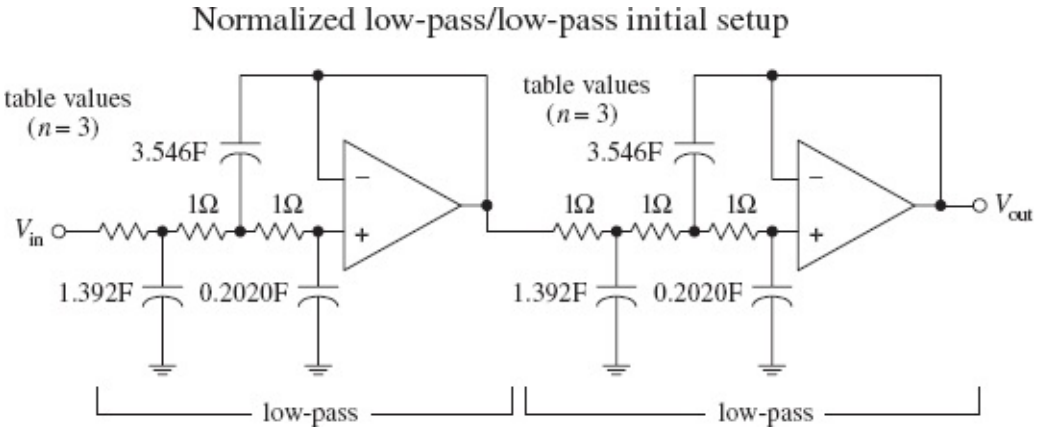
FIGURE 9.17

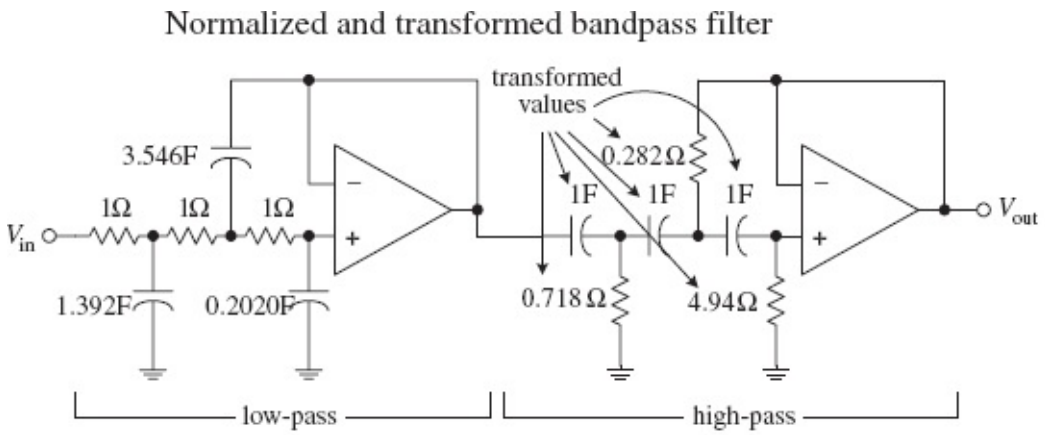
9.8.3 Active Bandpass Filters

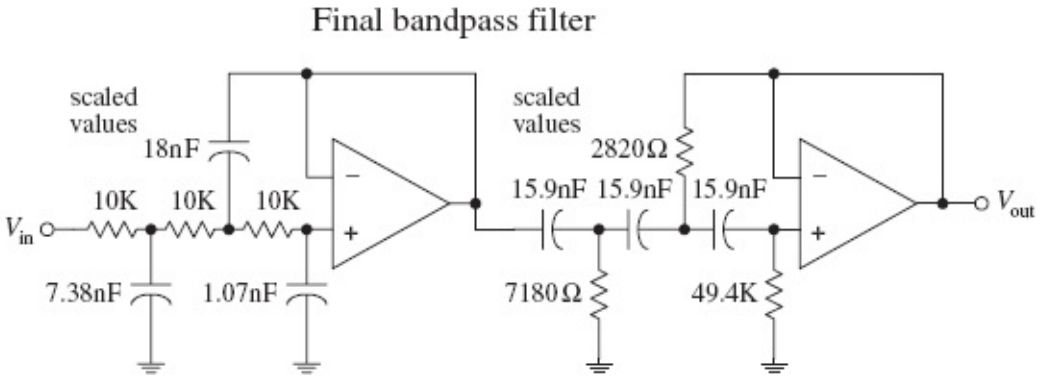
To design an active bandpass filter, it is necessary to determine if a wide-band or narrow-band type is needed. If the upper 3-dB frequency divided by the lower 3-dB frequency is greater than 1.5, the bandpass filter is a wide-band type; below 1.5, it is a narrow-band type. To design a wide-band bandpass filter, simply cascade a high-pass and low-pass active filter together. To design a narrow-band bandpass filter, you have to use some special tricks.

Wide-Band Example

Suppose that you want to design a bandpass filter that has Ω3-dB points at f_1 = 1000 Hz and f_2 = 3000 Hz and at least −30 dB at 300 and 10,000 Hz. What do you do?







First, confirm that this is a wide-band situation:

$$\frac{f_2}{f_1} = \frac{3000\ \text{Hz}}{1000\ \text{Hz}} = 3$$

Yes it is—it is greater than 1.5. This means that you simply have to cascade a low-pass and high-pass filter together. Next, the response requirements for the bandpass filter are broken down into low-pass and high-pass requirements:

Low-pass: –3 dB at 3000 Hz
–30 dB at 10,000 Hz
High-pass: –3 dB at 1000 Hz
–30 dB at 300 Hz

The steepness factor for the low-pass filter is

$$A_s = \frac{f_s}{f_{3\text{dB}}} = \frac{10{,}000\ \text{Hz}}{3000\ \text{Hz}} = 3.3$$

while the steepness factor for the high-pass filter is

$$A_s = \frac{f_{3\text{dB}}}{f_s} = \frac{1000\ \text{Hz}}{300\ \text{Hz}} = 3.3$$

This means that the normalized stop frequencies for both filters will be 3.3 rad/s. Next, use the response curves in Fig. 9.6 to determine the needed filter orders—*n* = 3 provides over −30 dB at 3.3 rad/s. To create the cascaded, normalized low-pass/high-pass filter, follow the steps in the last two examples. The upper two circuits in the figure show the steps involved in this process. To construct the final bandpass filter, the normalized bandpass filter must be frequency and impedance scaled.

Low-pass section:

$$C_{(\text{actual})} = \frac{C_{\text{table}}}{Z \cdot 2\pi f_{3\text{dB}}} = \frac{C_{\text{table}}}{Z \cdot 2\pi(3000\ \text{Hz})}$$

High-pass section:

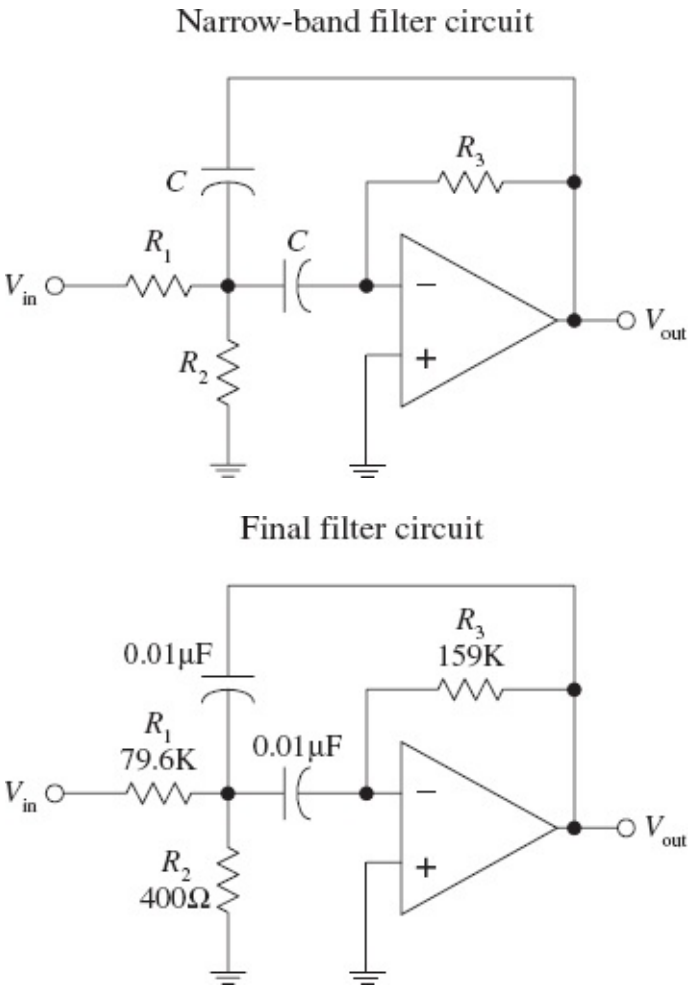
$$C_{(\text{actual})} = \frac{C_{\text{table}}}{Z \cdot 2\pi f_{2\text{dB}}} = \frac{C_{\text{table}}}{Z \cdot 2\pi(1000\ \text{Hz})}$$

Choose Z = 10,000 Ω. to provide convenient scaling of the components. In the normalized circuit, resistors are multiplied by a factor of Z. The final bandpass filter is shown at the bottom of the figure.

FIGURE 9.18

Narrow-Band Example

Suppose that you want to design a bandpass filter that has a center frequency f_0 = 2000 Hz and a −3-dB bandwidth $\Delta f_{BW} = f_2 - f_1$ = 40 Hz. How do you design the filter? Since f_2/f_1 = 2040 Hz/1960 Hz = 1.04, it is not possible to used the low-pass/high-pass cascading technique you used in the wide-band example. Instead, you must use a different approach. One simple approach is shown below.



In this example, simply use the circuit in Fig. 9.19 and some important equations that follow. No detailed discussion will ensue.
First, find the quality factor for the desired response:

$$Q = \frac{f_0}{f_2 - f_1} = \frac{2000\ \text{Hz}}{40\ \text{Hz}} = 50$$

Next, use the following design equations:

$$R_1 = \frac{Q}{2\pi f_0 C} \qquad R_2 = \frac{R_1}{2Q^2 - 1} \qquad R_3 = 2R_1$$

Picking a convenient value for C—which we'll set to 0.01 μF—the resistors' values become

$$R_1 = \frac{50}{2\pi(2000\ \text{Hz})(0.01\ \mu\text{F})} = 79.6\ \text{k}\Omega$$

$$R_2 = \frac{79.6\ \text{k}\Omega}{2(50)^2 - 1} = 400\ \Omega$$

$$R_3 = 2(79.6\ \text{k}\Omega) = 159\ \text{k}\Omega$$

The final circuit is shown at the bottom of the figure. R_2 can be replaced with a variable resistor to allow for tuning.

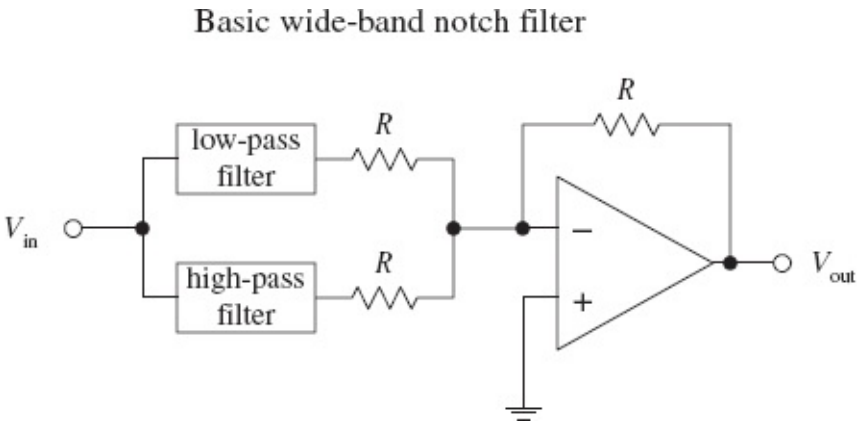
FIGURE 9.19

9.8.4 Active Notch Filters

Active notch filters come in narrow- and wide-band types. If the upper −3-dB frequency divided by the lower −3-dB frequency is greater than 1.5, the filter is called a *wide-band notch filter*—less than 1.5, the filter is called a *narrow-band notch filter*.

Wide-Band Notch Filter Example

To design a wide-band notch filter, simply combine a low-pass and high-pass filter together as shown in Fig. 9.20.



For example, if you need a notch filter to have −3-dB points at 500 and 5000 Hz and at least −15 dB at 1000 and 2500 Hz, simply cascade a low-pass filter with a response of

3 dB at 500 Hz
15 dB at 1000 Hz

with a high-pass filter with a response of

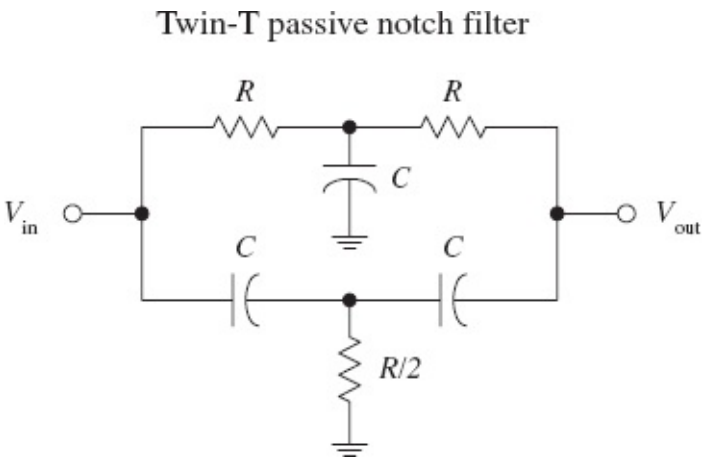
3 dB at 5000 Hz
15 dB at 2500 Hz

After that, go through the same low-pass and high-pass design procedures covered earlier. Once these filters are constructed, combine them as shown in the circuit in Fig. 9.20. In this circuit, R = 10 k is typically used.

FIGURE 9.20

Narrow-Band Notch Filters Example

To design a narrow-band notch filter ($f_2/f_1 < 1.5$), an *RC* network called the *twin-T* (see Fig. 9.21) is frequently used. A deep null can be obtained at a particular frequency with this circuit, but the circuit's *Q* is only ¼. (Recall that the *Q* for a notch filter is given as the center or null frequency divided by the −3-dB bandwidth.) To increase the *Q*, use the active notch filter shown in Fig. 9.22.

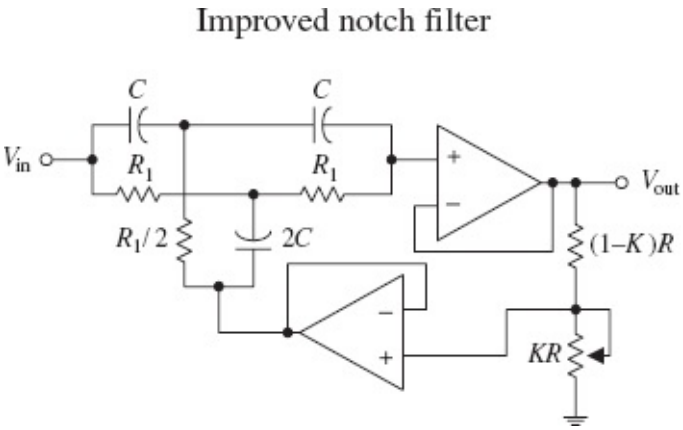


Like the narrow-bandpass example, let's simply go through the mechanics of how to pick the component values of the active notch filter. Here's an example.

Suppose that you want to make a "notch" at f_0 = 2000 Hz and desire a −3-dB bandwidth of Δf_{BW} = 100 Hz. To get this desired response, do the following. First determine the *Q:*

$$Q = \frac{\text{"notch" frequency}}{-3\text{-dB bandwidth}} = \frac{f_0}{\Delta f_{BW}} = \frac{2000\text{ Hz}}{100\text{ Hz}} = 20$$

FIGURE 9.21



The components of the active filter are found by using

$$R_1 = \frac{1}{2\pi f_0 C} \quad \text{and} \quad K = \frac{4Q-1}{4Q}$$

Now arbitrarily choose *R* and *C*; say, let *R* = 10 k and *C* = 0.01 µF. Next, solve for R_1 and *K:*

$$R_1 = \frac{1}{2\pi f_0 C} = \frac{1}{2\pi(2000\text{ Hz})(0.01\ \mu\text{F})} = 7961\ \Omega$$

$$K = \frac{4Q-1}{4Q} = \frac{4(20)-1}{4(20)} = 0.9875$$

Substitute these values into the circuit in Fig. 9.22. Notice the variable potentiometer—it is used to fine-tune the circuit.

FIGURE 9.22

9.9 Integrated Filter Circuits

A number of filter ICs are available on the market today. Two of the major categories of integrated filter circuits include the state-variable and switched-capacitor filter ICs. Both these filter ICs can be programmed to implement all the second-order functions described in the preceding sections. To design

higher-order filters, a number of these ICs can be cascaded together. Typically, all that's needed to program these filter ICs is a few resistors. Using IC filters allows for great versatility, somewhat simplified design, good precision, and limited design costs. Also, in most applications, frequency and selectivity factors can be adjusted independently.

An example of a state-variable filter IC is the AF100 made by National Semiconductor. This IC can provide low-pass, high-pass, bandpass, and notch filtering capabilities (see Fig. 9.23). Unlike the preceding filters covered in this chapter, the state-variable filter also can provide voltage gain.

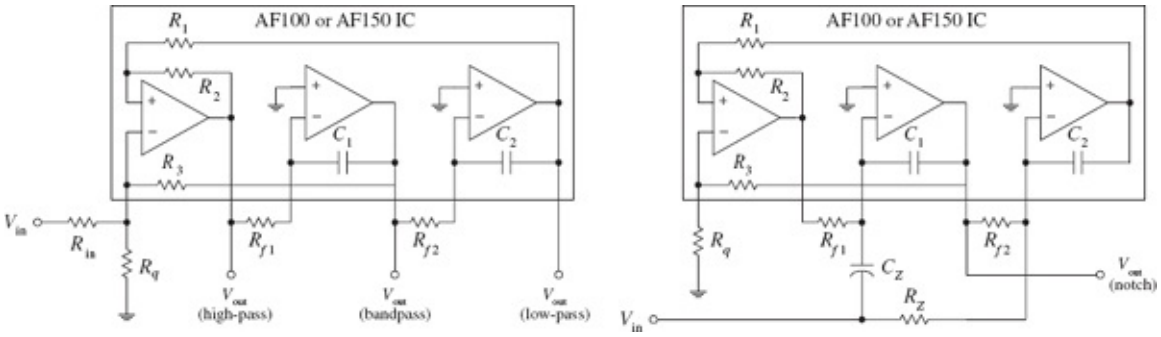


FIGURE 9.23

For the AF100, the low-pass gain is set using resistors R_1 and R_{in} (gain = $-R_1/R_{in}$). For the high-pass filter, the gain is set by resistors R_2 and R_{in} (gain = $-R_2/R_{in}$). (The negative sign indicates that the output is inverted relative to the input.) Setting the gain for the bandpass and notch functions is a bit more complex. Other parameters, such as Q, can be tweaked by using design formulas provided by the manufacturer. A good filter design handbook will discuss state-variable filters in detail and will provide the necessary design formulas. Also, check out the electronics catalogs to see what kinds of state-variable ICs exist besides the AF100.

Switched-capacitor filters are functionally similar to the other filters already discussed. However, instead of using external resistors to program the desired characteristics, switched-capacitor filters use a high-frequency capacitor-switching network technology. The capacitor-switching networks act like resistors whose values can be changed by changing the frequency of an externally applied clock voltage. The frequency of the clock signal determines which frequencies are passed and which frequencies get rejected. Typically, a digital clock signal is used to drive the filter—a useful feature if you are looking to design filters that can be altered by digital circuits. An example of a switched-capacitor IC is National Semiconductor's MF5 (see Fig. 9.24). By simply using a few external resistors, a power source, and an applied clock signal, you can program the filter for low-pass, high-pass, and bandpass functions. Again, like the state-variable ICs, manufacturers will provide you with necessary formulas needed for selecting the resistors and the frequency of the clock signal.

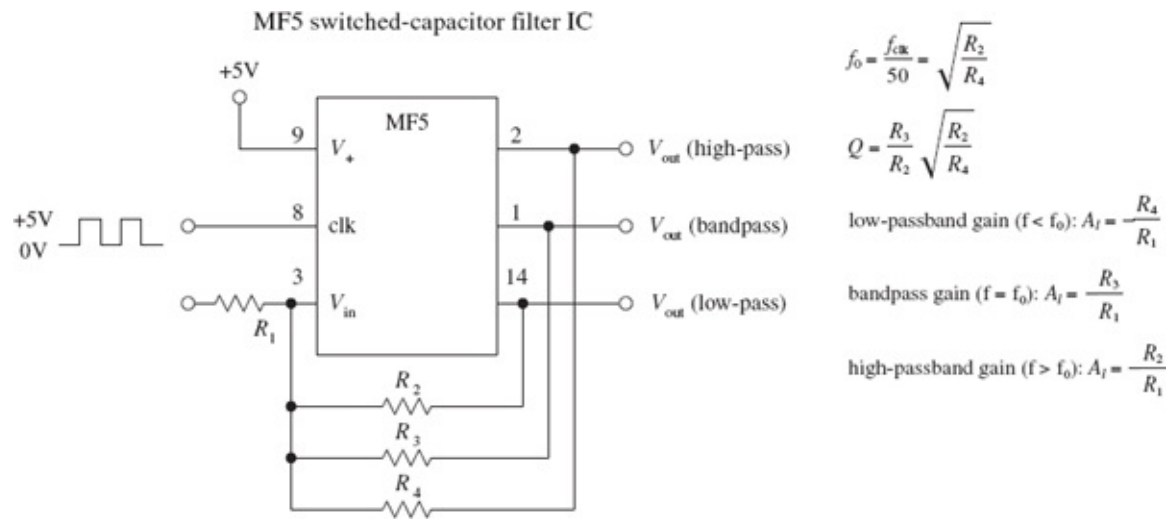


FIGURE 9.24

Switched-capacitor filters come in different filter orders. For example, the MF4 is a fourth-order Butterworth low-pass filter, and the MF6 is a sixth-order low-pass Butterworth filter; both are made by National Semiconductor. These two ICs have unity passband gain and require no external components, but they do require a clock input. There are a number of different kinds of switched-capacitor filters out there, made by a number of different manufacturers. Check the catalogs.

As an important note, the periodic clock signal applied to a switched-capacitor filter can generate a significant amount of noise (around 10 to 25 m V) present in the output signal. Typically, this is not of much concern because the frequency of the noise—which is the same as the frequency of the clock—is far removed from the signal band of interest. Usually a simple *RC* filter can be used to get rid of the problem.

CHAPTER 10

Oscillators and Timers

Within practically every electronic instrument there is an oscillator of some sort. The task of the oscillator is to generate a repetitive waveform of desired shape, frequency, and amplitude that can be used to drive other circuits. Depending on the application, the driven circuit(s) may require either a pulsed, sinusoidal, square, sawtooth, or triangular waveform.

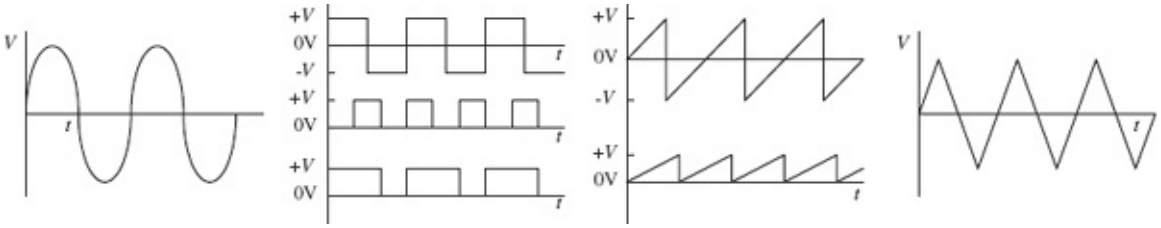


FIGURE 10.1

In digital electronics, squarewave oscillators, called *clocks*, are used to drive bits of information through logic gates and flip-flops at a rate of speed determined by the frequency of the clock. In radio circuits, high-frequency sinusoidal oscillators are used to generate carrier waves on which information can be encoded via modulation. The task of modulating the carrier also requires an oscillator. In oscilloscopes, a sawtooth generator is used to generate a horizontal electron sweep to establish the time base. Oscillators are also used in synthesizer circuits, counter and timer circuits, and LED/lamp flasher circuits. The list of applications is endless.

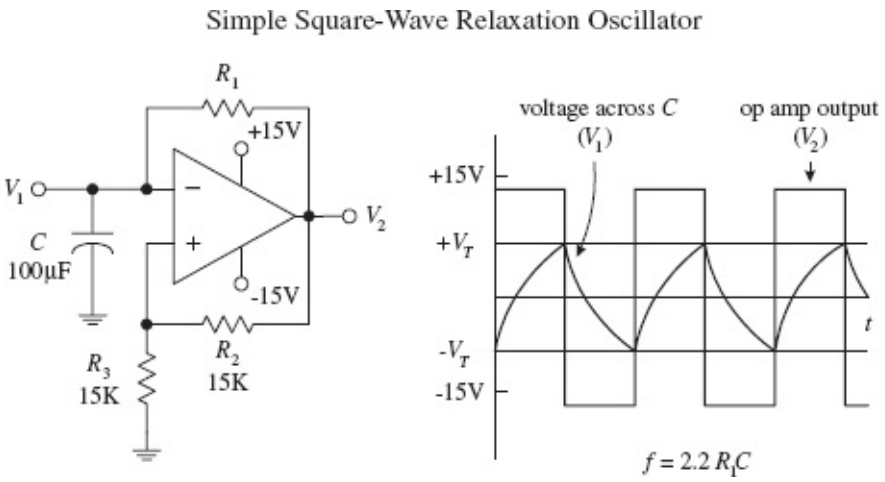
The art of designing good oscillator circuits can be fairly complex. There are a number of designs to choose from and a number of precision design techniques required. The various designs make use of different timing principles (e.g., *RC* charge/discharge cycle, *LC* resonant tank networks, crystals), and each is best suited for use within a specific application. Some designs are simple to construct but may have limited frequency stability. Other designs may have good stability within a certain frequency range, but poor stability outside that range. The shape of the generated waveform is obviously another factor that must be considered when designing an oscillator.

This chapter discusses the major kinds of oscillators, such as the *RC* relaxation oscillator, the Wien-bridge oscillator, the *LC* oscillator, and the crystal oscillator. The chapter also takes a look at popular oscillator ICs.

10.1 *RC* Relaxation Oscillators

Perhaps the easiest type of oscillator to design is the *RC* relaxation oscillator. Its oscillatory nature is explained by the following principle: Charge a capacitor through a resistor and then discharge it when the capacitor voltage reaches a certain threshold voltage. After that, the cycle is repeated, continuously. In order to control the charge/discharge cycle of the capacitor, an amplifier wired with positive feedback is used. The amplifier acts like a charge/discharge switch—triggered by the threshold voltage—and also

provides the oscillator with gain. Figure 10.2 shows a simple op amp relaxation oscillator.



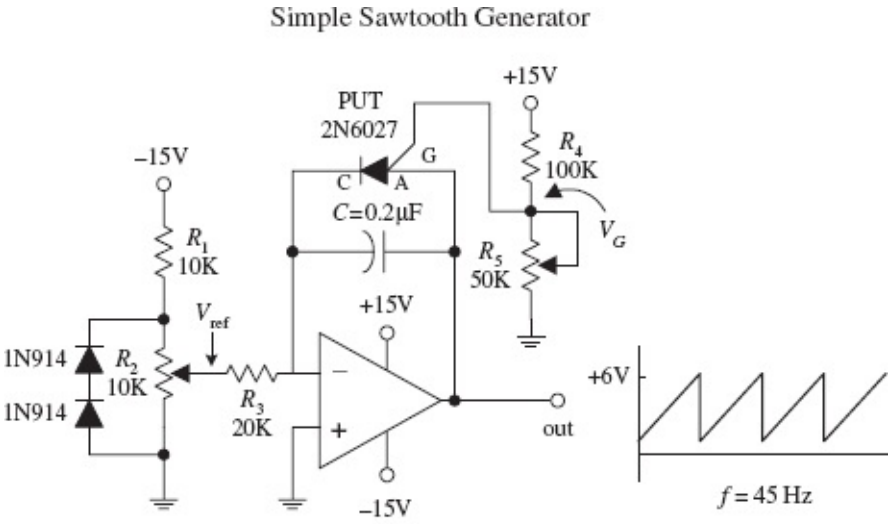
Assume that when power is first applied, the op amp's output goes toward positive saturation (it is equally likely that the output will go to negative saturation—see Chap. 8 for the details). The capacitor will begin to charge up toward the op amp's positive supply voltage (around +15 V) with a time constant of R_1C. When the voltage across the capacitor reaches the threshold voltage, the op amp's output suddenly switches to negative saturation (around −15 V). The threshold voltage is the voltage set at the inverting input, which is

$$V_T = \frac{R_3}{R_3 + R_2} = V_2 \frac{15\,\text{k}\Omega}{15\,\text{k}\Omega + 15\text{k}\Omega}(+15\,\text{V}) = +7.5\,\text{V}$$

The threshold voltage set by the voltage divider is now −7.5 V. The capacitor begins discharging toward negative saturation with the same R_1C time constant until it reaches −7.5 V, at which time the op amp's output switches back to the positive saturation voltage. The cycle repeats indefinitely, with a period equal to $2.2R_1C$.

FIGURE 10.2

Here's another relaxation oscillator that generates a sawtooth waveform (see Fig. 10.3). Unlike the preceding oscillator, this circuit resembles an op amp integrator network—with the exception of the PUT (programmable unijuction transistor) in the feedback loop. The PUT is the key ingredient that makes this circuit oscillate. Here's a rundown on how this circuit works.



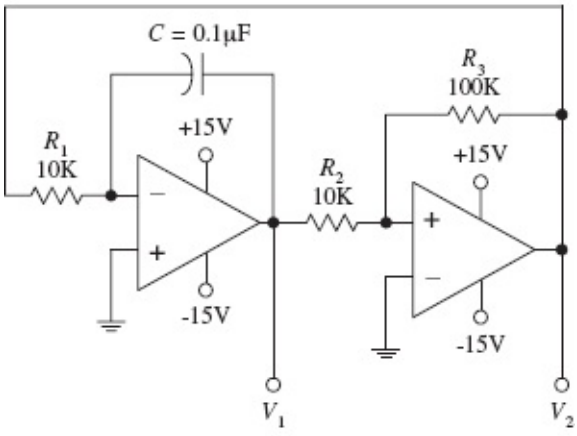
Let's initially pretend the circuit shown here does not contain the PUT. In this case, the circuit would resemble a simple integrator circuit; when a negative voltage is placed at the inverting input (−), the capacitor charges up at a linear rate toward the positive saturation voltage (+15 V). The output signal would simply provide a one-shot ramp voltage—it would not generate a repetitive triangular wave. In order to generate a repetitive waveform, we must now include the PUT. The PUT introduces oscillation into the circuit by acting as an active switch that turns on (anode-to-cathode conduction) when the anode-to-cathode

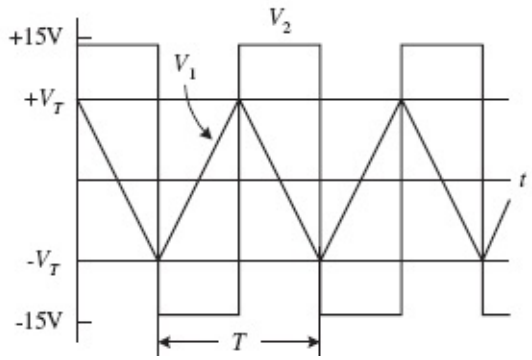
voltage is greater by one diode drop than its gate-to-cathode voltage. The PUT will remain on until the current through it falls below the minimum holding current rating. This switching action acts to rapidly discharge the capacitor before the output saturates. When the capacitor discharges, the PUT turns off, and the cycle repeats. The gate voltage of the PUT is set via voltage-divider resistors *R*4 and R_5. The R_1 and R_2 voltage-divider resistors set the reference voltage at the inverting input, while the diodes help stabilize the voltage across R_2 when it is adjusted to vary the frequency. The output-voltage amplitude is determined by *R*4, while the output frequency is approximated by the expression below the figure. (The 0.5 V represents a typical voltage drop across a PUT.)

FIGURE 10.3

Here's a simple dual op amp circuit that generates both triangular and square waveforms (see Fig. 10.4). This circuit combines a triangle-wave generator with a comparator.

Simple Triangle-Wave/Square-Wave Generator





The rightmost op amp in the circuit acts as a comparator—it is wired with positive feedback. If there is a slight difference in voltage between the inputs of this op amp, the V_2 output voltage will saturate in either the positive or negative direction. For sake of argument, let's say the op amp saturates in the positive direction. It will remain in that saturated state until the voltage at the noninverting input (+) drops below zero, at which time V_2 will be driven to negative saturation. The threshold voltage is given by

$$V_T = \frac{V_{sat}}{(R_3 \quad R_2)}$$

where V_{sat} is a volt or so lower than the op amp's supply voltage (see Chap. 8) Now this comparator is used with a ramp generator (leftmost op amp section). The output of the ramp generator is connected to the input of the comparator, while its output is fed back to the input of the ramp generator. Each time the ramp voltage reaches the threshold voltage, the comparator changes states. This gives rise to oscillation. The period of the output waveform is determined by the R_1C time constant, the saturation voltage, and the threshold voltage:

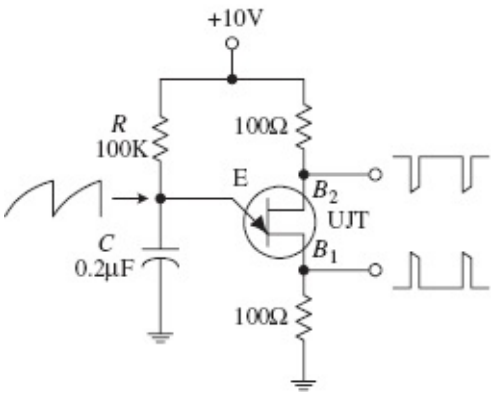
$$T = \frac{4V_T}{V_{sat}} R_1 C$$

The frequency is 1/*T*.

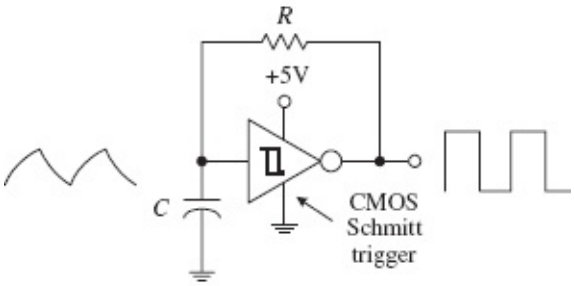
FIGURE 10.4

Now op amps are not the only active ingredient used to construct relaxation oscillators. Other components, such as transistors and digital logic gates, can take their place.

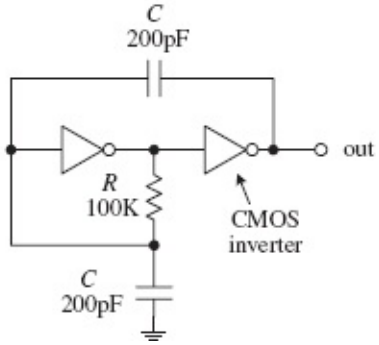
Unijunction oscillator



Digital oscillator
(using a Schmitt trigger inverter)



Digital oscillator (using inverters)



Here is a unijunction transistor (UJT), along with some resistors and a capacitor, that makes up a relaxation oscillator that is capable of generating three different output waveforms. During operation, at one instant in time, *C* charges through *R* until the voltage present on the emitter reaches the UJT's triggering voltage. Once the triggering voltage is exceeded, the *E*-to-B_1 conductivity increases sharply, which allows current to pass from the capacitor-emitter region through the emitter-base 1 region and then to ground. When this occurs, *C* suddenly loses its charge, and the emitter voltage suddenly falls below the triggering voltage. After that, the cycle repeats itself. The resulting waveforms generated during this process are shown in the figure. The frequency of oscillation is given by

$$f = \frac{1}{R_E C_E \ln[1/(1-\eta)]}$$

where η is the UJT's intrinsic standoff ratio, which is typically around 0.5. See Chap. 4 for more details.

Here a simple relaxation oscillator is built from a Schmitt trigger inverter IC and an *RC* network. (Schmitt triggers are used to transform slowly changing input waveforms into sharply defined, jitter-free output waveforms (see Chap. 12). When power is first applied to the circuit, the voltage across *C* is zero, and the output of the inverter is high (+5 V). The capacitor starts charging up toward the output voltage via *R*. When the capacitor voltage reaches the positive-going threshold of the inverter (e.g., 1.7 V), the output of the inverter goes low (~0 V). With the output low, *C* discharges toward 0 V. When the capacitor voltage drops below the negative-going threshold voltage of the inverter (e.g., 0.9 V), the output of the inverter goes high. The cycle repeats. The on/off times are determined by the positive- and negative-going threshold voltages and the *RC* time constant.

The third example is a pair of CMOS inverters that are used to construct a simple squarewave *RC* relaxation oscillator. The circuit can work with voltages ranging from 4 to 18 V. The frequency of oscillation is given by

$$f = \frac{1}{4RC \ln 2} \simeq \frac{1}{2.8RC}$$

***R* can be adjusted to vary the frequency. I will discuss CMOS inverters in Chap. 12.**

FIGURE 10.5

All the relaxation oscillators shown in this section are relatively simple to construct. Now, as it turns out, there is even an easier way to generate basic waveforms. The easy way is to use an IC especially designed for the task. An incredibly popular squarewave-generating chip that can be programmed with resistors and a capacitor is the 555 timer IC.

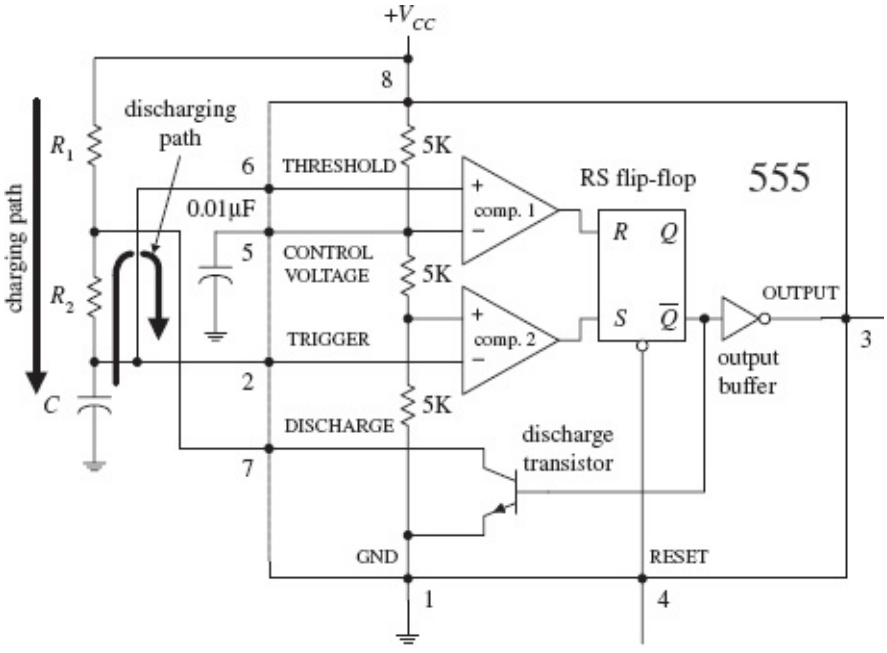
10.2 The 555 Timer IC

The 555 timer IC is an incredibly useful precision timer that can act as either a timer or an oscillator. In timer mode—better known as *monostable mode*—the 555 simply acts as a "one-shot" timer; when a trigger voltage is applied to its trigger lead, the chip's output goes from low to high for a duration set by an external *RC* circuit. In oscillator mode—better know as *astable mode*—the 555 acts as a rectangular-wave generator whose output waveform (low duration, high duration, frequency, etc.) can be adjusted by means of two external *RC* charge/discharge circuits.

The 555 timer IC is easy to use (requires few components and calculations) and inexpensive and can be used in an amazing number of applications. For example, with the aid of a 555, it is possible to create digital clock waveform generators, LED and lamp flasher circuits, tone-generator circuits (sirens, metronomes, etc.), one-shot timer circuits, bounce-free switches, triangular-waveform generators, frequency dividers, etc.

10.2.1 How a 555 Works (Astable Operation)

Figure 10.6 is a simplified block diagram showing what is inside a typical 555 timer IC. The overall circuit configuration shown here (with external components included) represents the astable 555 configuration.



The 555 gets its name from the three 5-kΩ resistors shown in the block diagram. These resistors act as a three-step voltage divider between the supply voltage (V_{CC}) and ground. The top of the lower 5-kΩ resistor (+ input to comparator 2) is set to $\frac{1}{3}V_{CC}$, while the top of the middle 5-kΩ resistor (− input to comparator 2) is set to $\frac{2}{3}V_{CC}$. The two comparators output either a high or low voltage based on the analog voltages being compared at their inputs. If one of the comparator's positive inputs is more positive than its negative input, its output logic level goes high; if the positive input voltage is less than the negative input voltage, the output logic level goes low. The outputs of the comparators are sent to the inputs of an *SR* (set/reset) flip-flop. The flip-flop looks at the *R* and *S* inputs and produces either a high or a low based on the voltage states at the inputs (see Chap. 12).

FIGURE 10.6

Pin 1 (ground). IC ground.

Pin 2 (trigger). Input to comparator 2, which is used to set the flip-flop. When the voltage at pin 2 crosses from above to below $\frac{1}{3}V_{CC}$, the comparator switches to high, setting the flip-flop.

Pin 3 (output). The output of the 555 is driven by an inverting buffer capable of sinking or sourcing around 200 mA. The output voltage levels depend on the output current but are approximately $V_{out(high)} = V_{CC} - 1.5$ V and $V_{out(low)} = 0.1$ V.

Pin 4 (reset). Active-low reset, which forces $\overline{Q}$ high and pin 3 (output) low.

Pin 5 (control). Used to override the $\frac{2}{3}V_{CC}$ level, if needed, but is usually grounded via a 0.01-μF bypass capacitor (the capacitor helps eliminate V_{CC} supply noise). An external voltage applied here will set a new trigger voltage level.

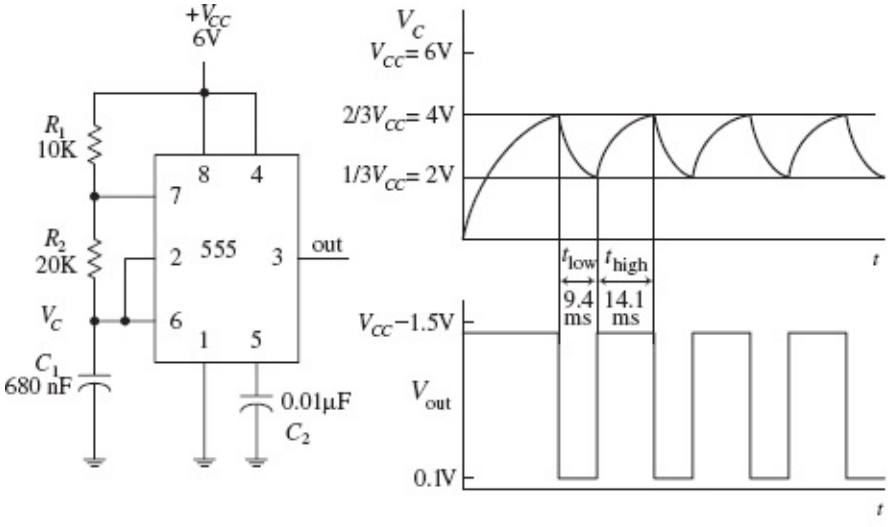
Pin 6 (threshold). Input to the upper comparator, which is used to reset the flip-flop. When the voltage at pin 6 crosses from below to above $\frac{2}{3}V_{CC}$, the comparator switches to a high, resetting the flip-flop.

Pin 7 (discharge). Connected to the open collector of the *npn* transistor. It is used to short pin 7 to ground when $\overline{Q}$ is high (pin 3 low). This causes the capacitor to discharge.

Pin 8 (Supply voltage V_{CC}). Typically between 4.5 and 16 V for general-purpose TTL 555 timers. (For CMOS versions, the supply voltage may be as low as 1 V.)

In the astable configuration, when power is first applied to the system, the capacitor is uncharged. This means that 0 V is placed on pin 2, forcing comparator 2 high. This in turn sets the flip-flop so that $\overline{Q}$ is high and the 555's output is low (a result of the inverting buffer). With $\overline{Q}$ high, the discharge transistor is turned on, which allows the capacitor to charge toward V_{CC} through R_1 and R_2. When the capacitor voltage exceeds $\frac{1}{3}V_{CC}$, comparator 2 goes low, which has no effect on the *SR* flip-flop. However, when the capacitor voltage exceeds $\frac{2}{3}V_{CC}$, comparator 1 goes high, resetting the flip-flop and forcing $\overline{Q}$ high and the output low. At this point, the discharge transistor turns on and shorts pin 7 to ground, discharging the capacitor through R_2. When the capacitor's voltage drops below $\frac{1}{3}V_{CC}$, comparator 2's output jumps back to a high level, setting the flip-flop and making $\overline{Q}$ low and the output high. With $\overline{Q}$ low, the transistor turns on, allowing the capacitor to start charging again. The cycle repeats over and over again. The net result is a squarewave output pattern whose voltage level is approximately $V_{CC} - 1.5$ V and whose on/off periods are determined by the C, R_1 and R_2.

10.2.2 Basic Astable Operation

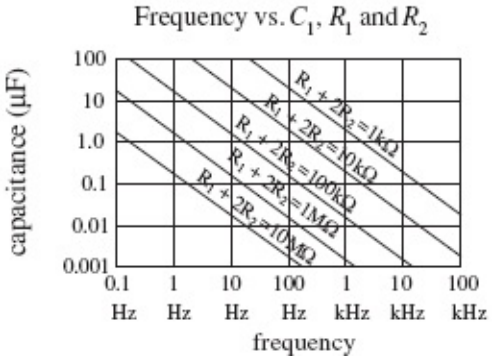


$t_{low} = 0.693(20K)(680nF) = 9.6ms$

$t_{high} = 0.693(10K + 20K)(680nF) = 14.1ms$

$$f = \frac{1}{9.4ms + 14.1ms} = 42Hz$$

$$\text{duty cycle} = \frac{14.1ms}{14.1ms + 9.4ms} = 0.6$$



When a 555 is set up in astable mode, it has no stable states; the output jumps back and forth. The time duration V_{out} remains low (around 0.1 V) is set by the R_1C_1 time constant and the $\frac{1}{3}V_{CC}$ and $\frac{2}{3}V_{CC}$ levels; the time duration V_{out} stays high (around V_{CC} −1.5 V) is determined by the (R_1 + R_2) C_1 time constant and the two voltage levels (see graphs). After doing some basic calculations, the following two practical expressions arise:

$t_{low} = 0.693R_2C_1$
$t_{high} = 0.693(R_1 + R_2)C_1$

The duty cycle (the fraction of the time the output is high) is given by

$$\text{Duty cycle} = \frac{t_{high}}{t_{high} + t_{low}}$$

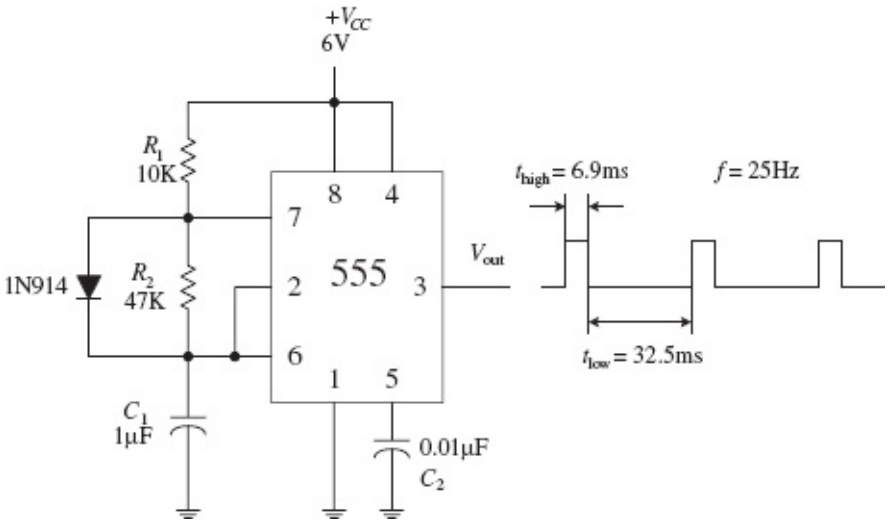
The frequency of the output waveform is

$$f = \frac{1}{t_{high} + t_{low}} = \frac{1.44}{(R_1 + 2R_2)C_1}$$

For reliable operation, the resistors should be between approximately 10 kΩ and 14 MΩ, and the timing capacitor should be from around 100 pF to 1000 µF. The graph will give you a general idea of how the frequency responds to the component values.

FIGURE 10.7

Low-Duty-Cycle Operation (Astable Mode)



Now there is a slight problem with the last circuit—you cannot get a duty cycle that is below 0.5 (or 50 percent). In other words, you cannot make t_{high} shorter than t_{low}. For this to occur, the R_1C_1 network (used to generate t_{low}) would have to be larger the $(R_1 + R_2)C_1$ network (used to generate t_{high}). Simple arithmetic tells us that this is impossible; $(R_1 + R_2)C_1$ is always greater than R_1C_1. How do you remedy this situation? You attach a diode across R_2, as shown in the figure. With the diode in place, as the capacitor is charging (generating t_{high}), the preceding time constant $(R_1 + R_2)C_1$ is reduced to R_1C_1 because the charging current is diverted around R_2 through the diode. With the diode in place, the high and low times become

$$t_{high} = 0.693(10K)(1\,\mu F) = 6.9\,ms$$
$$t_{low} = 0.693(47K)(1\,\mu F) = 32.5\,ms$$
$$f = \frac{1}{6.9\,ms + 32.5\,ms} = 25Hz$$
$$\text{duty cycle} = \frac{6.9\,ms}{6.9\,ms + 32.5\,ms} = 0.18$$

$$t_{high} = 0.693R_1C_1$$
$$t_{low} = 0.693R_2C_1$$

To generate a duty cycle of less than 0.5, or 50 percent, simply make R_1 less than R_2.

FIGURE 10.8

10.2.3 How a 555 Works (Monostable Operation)

Figure 10.9 shows a 555 hooked up in the monostable configuration (one-shot mode). Unlike the astable mode, the monostable mode has only one stable state. This means that for the output to switch states, an externally applied signal is needed.

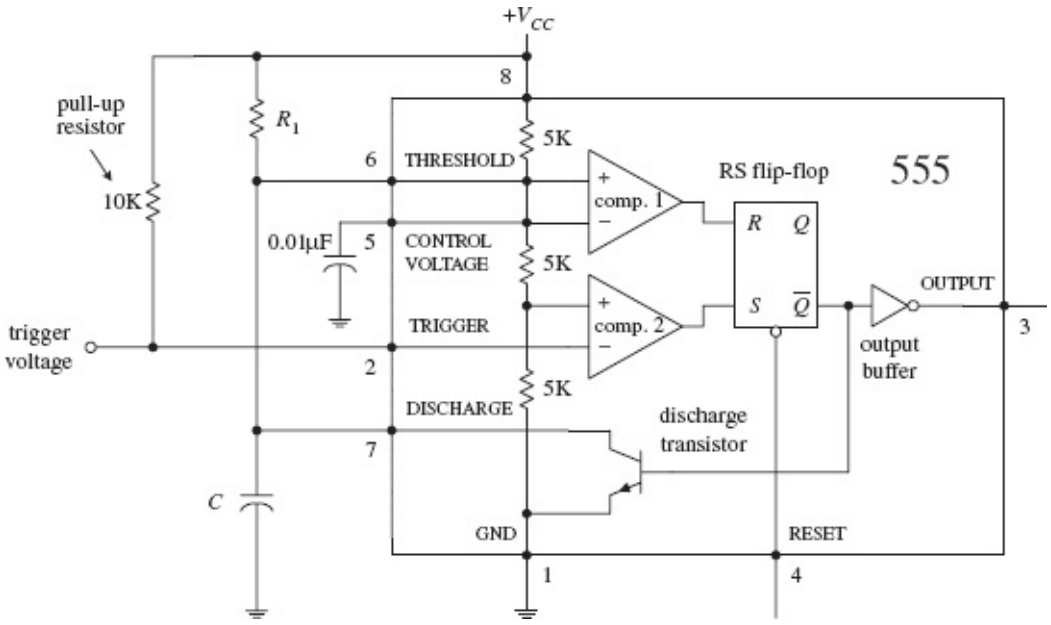
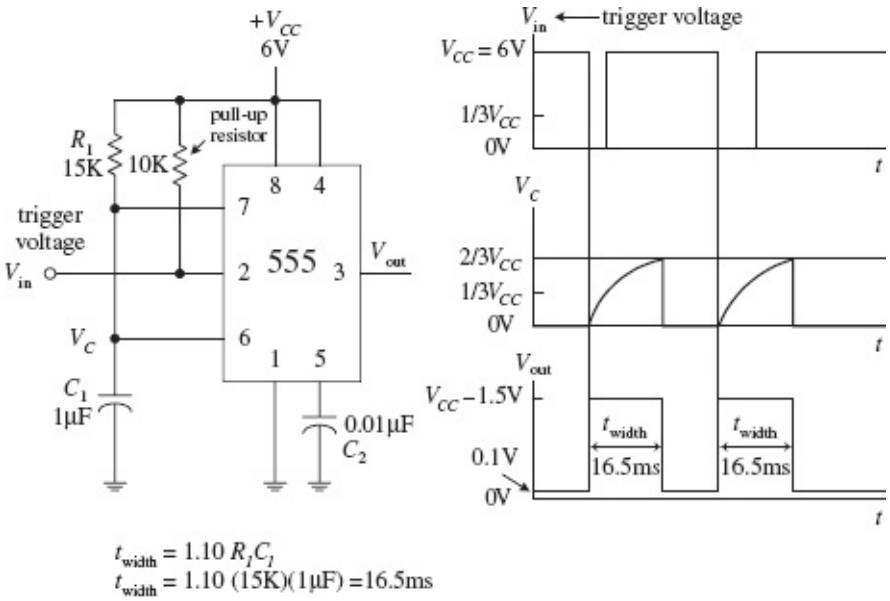


FIGURE 10.9

In the monostable configuration, initially (before a trigger pulse is applied) the 555's output is low, while the discharge transistor is on, shorting pin 7 to ground and keeping *C* discharged. Also, pin 2 is normally held high by the 10-k pull-up resistor. Now, when a negative-going trigger pulse (less than $\frac{1}{3}V_{CC}$) is applied to pin 2, comparator 2 is forced high, which sets the flip-flop's $\overline{Q}$ to low, making the output high (due to the inverting buffer), while turning off the discharge transistor. This allows *C* to charge up via R_1 from 0 V toward V_{CC}. However, when the voltage across the capacitor reaches $\frac{2}{3}V_{CC}$, comparator 1's output goes high, resetting the flip-flop and making the output low, while turning on the discharge transistor, allowing the capacitor to quickly discharge toward 0 V. The output will be held in this stable state (low) until another trigger is applied.

10.2.4 Basic Monostable Operation



$t_{width} = 1.10\,R_1C_1$
$t_{width} = 1.10\,(15\text{K})(1\mu\text{F}) = 16.5\text{ms}$

The monostable circuit only has one stable state. That is, the output rests at 0 V (in reality, more like 0.1 V) until a negative-going trigger pulse is applied to the trigger lead—pin 2. (The negative-going pulse can be implemented by momentarily grounding pin 2, say, by using a pushbutton switch attached from pin 2 to ground.) After the trigger pulse is applied, the output will go high (around V_{CC} − 1.5 V) for the duration set by the R_1C_1 network. Without going through the derivations, the width of the high output pulse is

$$t_{width} = 1.10R_1C_1$$

For reliable operation, the timing resistor R_1 should be between around 10 kΩ and 14 MΩ, and the timing capacitor should be from around 100 pF to 1000 μF.

FIGURE 10.10

10.2.5 Some Important Notes About 555 Timers

555 ICs are available in both bipolar and CMOS types. Bipolar 555s, like the ones you used in the preceding examples, use bipolar transistors inside, while CMOS 555s use MOSFET transistors instead. These two types of 555s also differ in terms of maximum output current, minimum supply voltage/current, minimum triggering current, and maximum switching speed. With the exception of maximum output current, the CMOS 555 surpasses the bipolar 555 in all regards. A CMOS 555 IC can be distinguished from a bipolar 555 by noting whether the part number contains a *C* somewhere within it (e.g., ICL7555, TLC555, LMC555, etc.). (Note that there are hybrid versions of the 555 that incorporate the best features of both the bipolar and CMOS technologies.) Table 10.1 shows specifications for a few 555 devices.

TABLE 10.1 Sample Specifications for Some 555 Devices

TYPE	SUPPLY VOLTAGE MIN. (V)	SUPPLY VOLTAGE MAX. (V)	SUPPLY CURRENT (V_{cc} = 5 V) TYP. (μA)	SUPPLY CURRENT (V_{cc} = 5 V) MAX. (μA)	TRIG. CURRENT (THRES. CURRENT) TYP. (nA)	TRIG. CURRENT (THRES. CURRENT) MAX. (nA)	TYPICAL FREQUENCY (MHZ)	$I_{out,max}$ (V_{cc} = 5 V) SOURCE (mA)	$I_{out,max}$ (V_{cc} = 5 V) SINK (mA)
SN555	4.5	18	3000	5000	100	500	0.5	200	200
ICL7555	2	18	60	300	—	10	1	4	25
TLC555	2	18	170	—	0.01	—	2.1	10	100
LMC555	1.5	15	100	250	0.01	—	3	—	—
NE555	4.5	15	—	6000	—	—	—	—	200

If you need more than one 555 timer per IC, check out the 556 (dual version) and 558 (quad version). The 556 contains two functionally independent 555 timers that share a common supply lead, while the 558 contains four slightly simplified 555 timers. In the 558, not all functions are brought out to the pins, and in fact, this device is intended to be used in monostable mode—although it can be tricked into astable mode with a few alterations (see manufacturer's literature for more information).

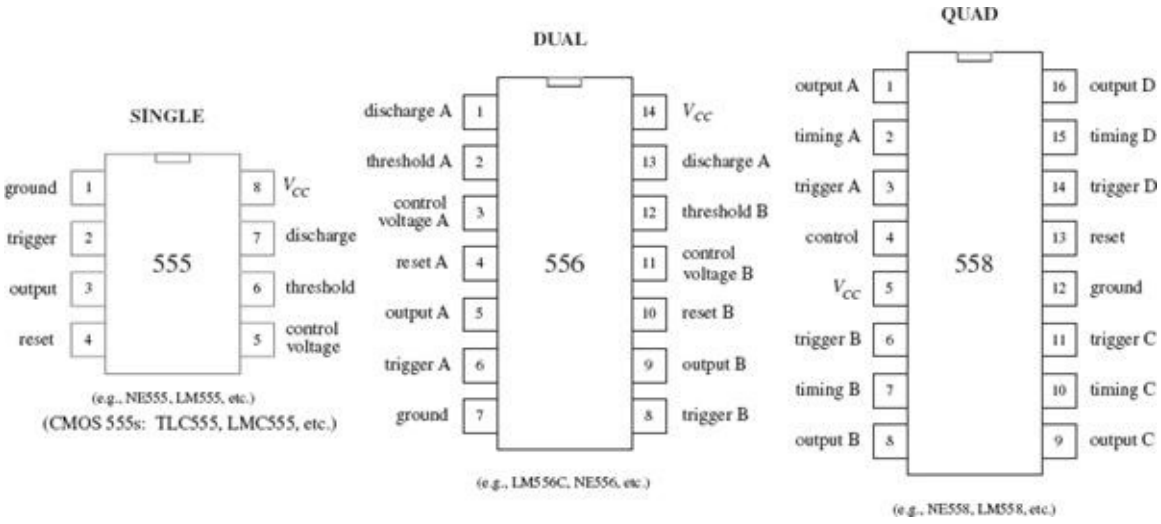


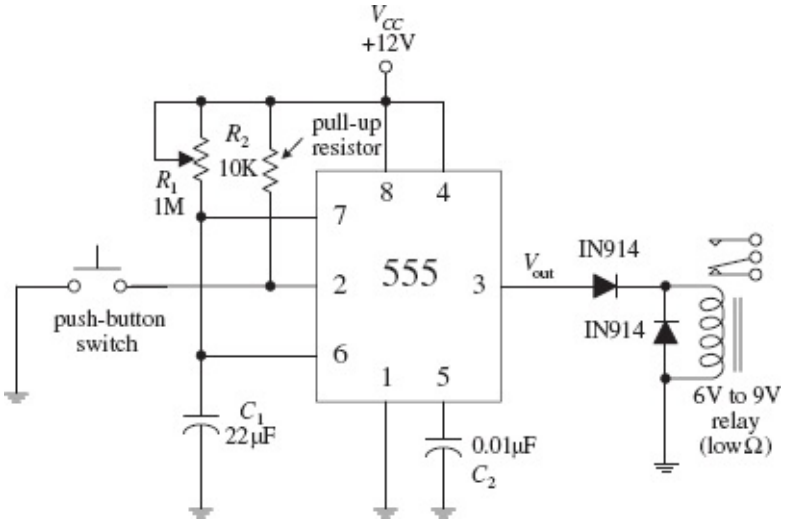
FIGURE 10.11

PRACTICAL TIP

To avoid problems associated with false triggering, connect the 555's pin 5 to ground through a 0.01-μF capacitor (we applied this trick already in this section). Also, if the power supply lead becomes long or the timer does not seem to function for some unknown reason, try attaching a 0.1-μF or larger capacitor between pins 8 and 1.

10.2.6 Simple 555 Applications

Relay Driver (Delay Timer)



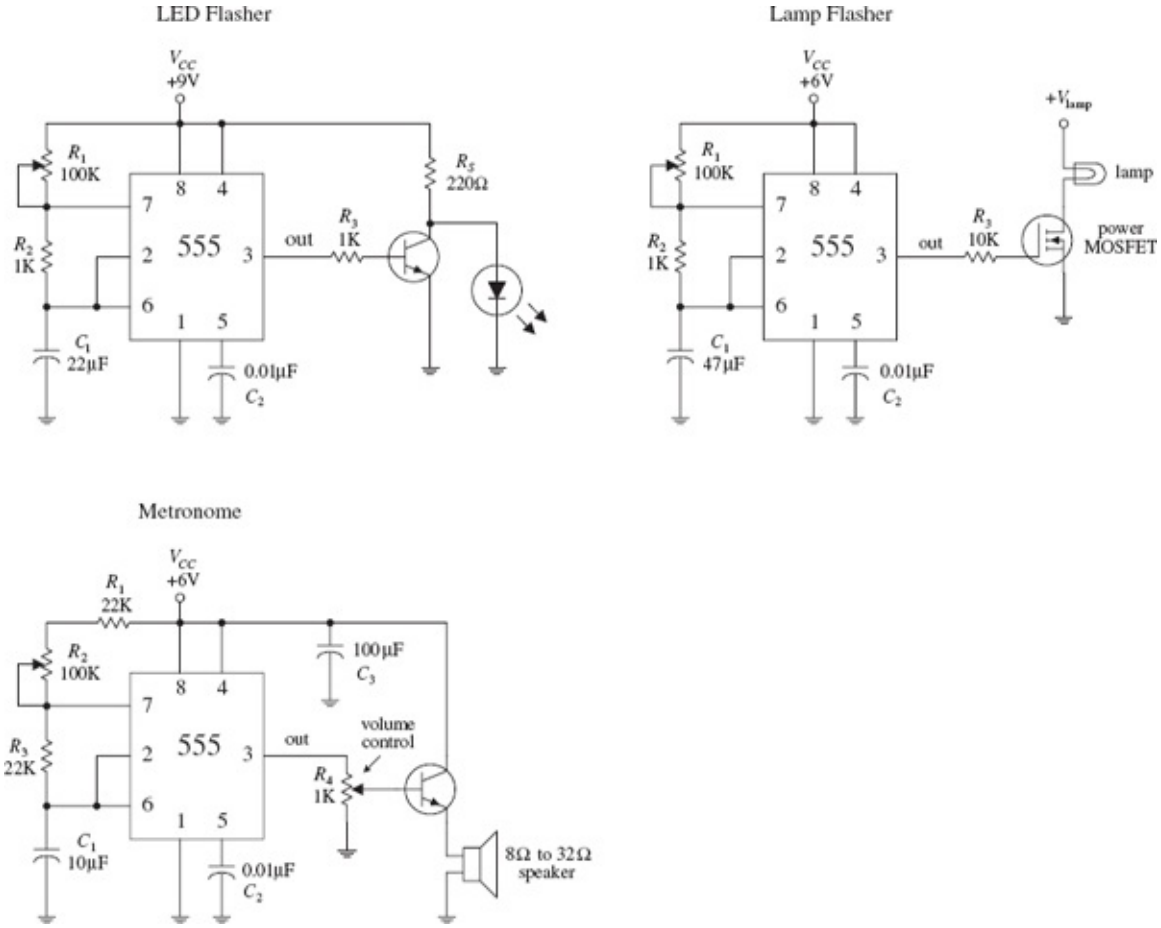
The monostable circuit shown here acts as a delay timer that is used to actuate a relay for a given duration. With the pushbutton switch open, the output is low (around 0.1 V), and the relay is at rest. However, when the switch is momentarily closed, the 555 begins its timing cycle; the output goes high (in this case ~10.5 V) for a duration equal to

$$t_{\text{delay}} = 1.10R_1C_1$$

The relay will be actuated for the same time duration. The diodes help prevent damaging current surges—generated when the relay switches states—from damaging the 555 IC, as well as the relay's switch contacts.

FIGURE 10.12

LED and Lamp Flasher and Metronome



All these circuits are oscillator circuits (astable multivibrators). In the LED flasher circuit, a transistor is used to amplify the 555's output in order to provide sufficient current to drive the LED, while R_S is used to prevent excessive current from damaging the LED. In the lamp-flasher circuit, a MOSFET amplifier is used to control current flow through the lamp. A power MOSFET may be needed if the lamp draws a considerable amount of current. The metronome circuit produces a series of "clicks" at a rate determined by R_2. To control the volume of the clicks, R_4 can be adjusted.

FIGURE 10.13

10.3 Voltage-Controlled Oscillators

Besides the 555 timer IC, there are a number of other voltage-controlled oscillators (VCOs) on the market—some of which provide more than just a squarewave output. For example, the NE566 function generator is a very stable, easy-to-use triangular-wave and squarewave generator. In the 566 circuit below, R_1 and C_1 set the center frequency, while a control voltage at pin 5 varies the frequency; the control voltage is applied by means of a voltage-divider network (R_2, R_3, R_4). The output frequency of the 566 can be determined by using the formula shown in Fig. 10.14.

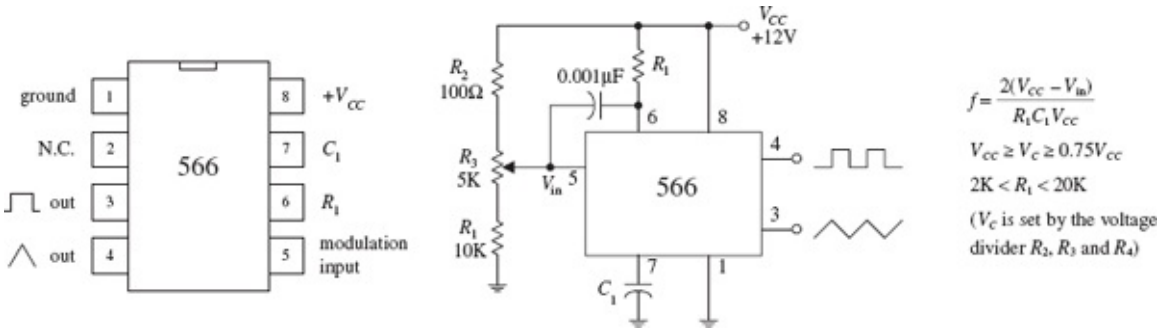
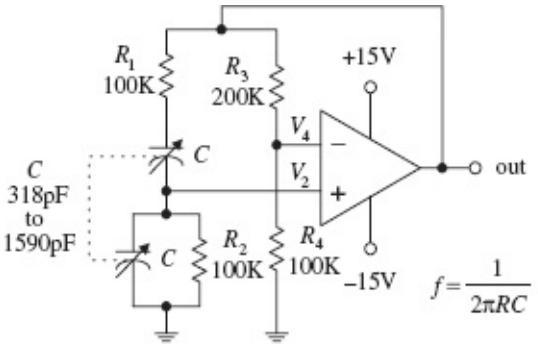


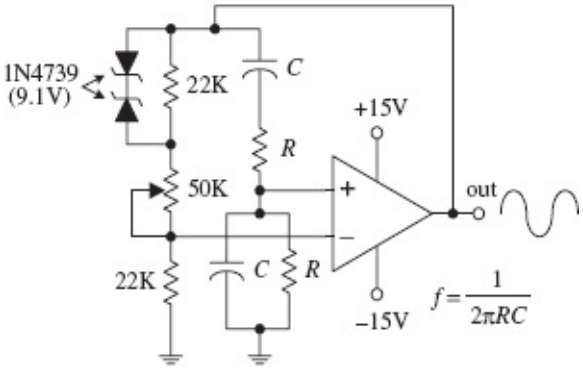
FIGURE 10.14

Other VCOs, such as the 8038 and the XR2206, can create a trio of output waveforms, including a sine wave (approximation of one, at any rate), a square wave, and triangular wave. Some VCOs are designed specifically for digital waveform generation and may use an external crystal in place of a capacitor for improved stability. To get a feel for what kinds of VCOs are out there, check the electronics catalogs.

10.4 Wien-Bridge and Twin-T Oscillators

A popular *RC*-type circuit used to generate low-distortion sinusoidal waves at low to moderate frequencies is the Wien-bridge oscillator. Unlike the oscillator circuits discussed already in this chapter, this oscillator uses a different kind of mechanism to provide oscillation, namely, a frequency-selective filter network.





The heart of the Wien-bridge oscillator is its frequency-selective feedback network. The op amp's output is fed back to the inputs in phase. Part of the feedback is positive (makes its way through the frequency-selective *RC* branch to the noninverting terminal), while the other part is negative (is sent through the resistor branch to the inverting input of the op amp). At a particular frequency $f_0 = 1/(2\pi RC)$, the inverting input voltage (V_4) and the noninverting input voltage (V_2) will be equal and in phase—the positive feedback will cancel the negative feedback, and the circuit will oscillate. At any other frequency, V_2 will be too small to cancel V_4, and the circuit will not oscillate. In this circuit, the gain must be set to +3. The resistors must satisfy the condition R3/R4 = 2 (which gives a noninverting gain of 3). Anything less than this value will cause oscillations to cease; anything more will cause the output to saturate. With the component values listed in the figure, this oscillator can cover a frequency range of 1 to 5 kHz. The frequency can be adjusted by means of a two-ganged variable-capacitor unit.

The second circuit shown in the figure is a slight variation of the first. Unlike the first circuit, the positive feedback must be greater than the negative feedback to sustain oscillations. The potentiometer is used to adjust the amount of negative feedback, while the *RC* branch controls the amount of positive feedback based on the operating frequency. Now, since the positive feedback is larger than the negative feedback, you have to contend with the "saturation problem," as encountered in the last example. To prevent saturation, two zener diodes placed face to face (or back to back) are connected across the upper 22-kΩ resistor. When the output voltage rises above the zener's breakdown voltage, one or the other zener diode conducts, depending on the polarity of the feedback. The conducting zener diode shunts the 22-kΩ resistor, causing the resistance of the negative feedback circuit to decrease. More negative feedback is applied to the op amp, and the output voltage is controlled to a certain degree.

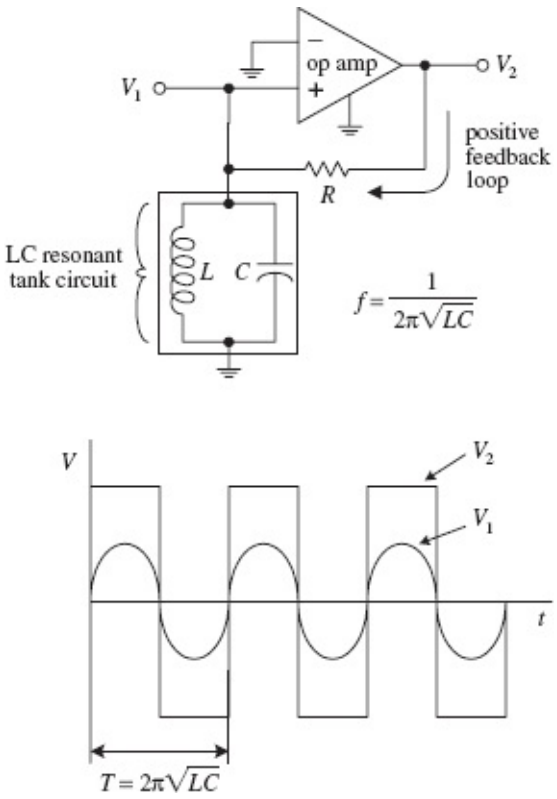
FIGURE 10.15

10.5 *LC* Oscillators (Sinusoidal Oscillators)

When it comes to generating high-frequency sinusoidal waves, commonly used in radiofrequency applications, the most common approach is to use an *LC* oscillator. The *RC* oscillators discussed so far have difficulty handling high frequencies, mainly because it is difficult to control the phase shifts of feedback signals sent to the amplifier input and because, at high frequencies, the capacitor and resistor values often become impractical to work with. *LC* oscillators, on the other hand, can use small inductances in conjunction with capacitance to create feedback oscillators that can reach frequencies up to around 500 MHz. However, it is important to note that at low frequencies (e.g., audio range), *LC* oscillators become highly unwieldy.

LC oscillators basically consist of an amplifier that incorporates positive feedback through a frequency-selective *LC* circuit (or tank). The *LC* tank acts to eliminate from the amplifier's input any

frequencies significantly different from its natural resonant frequency. The positive feedback, along with the tank's resonant behavior, acts to promote sustained oscillation within the overall circuit. If this is a bit confusing, envision shock exciting a parallel *LC* tank circuit. This action will set the tank circuit into sinusoidal oscillation at the *LC*'s resonant frequency—the capacitor and inductor will "toss" the charge back and forth. However, these oscillations will die out naturally due to internal resistance and loading. To sustain the oscillation, the amplifier is used. The amplifier acts to supply additional energy to the tank circuit at just the right moment to sustain oscillations. Here is a simple example to illustrate the point.



Here, an op amp incorporates positive feedback that is altered by an *LC* resonant filter or tank circuit. The tank eliminates from the noninverting input of the op amp any frequencies significantly different from the tank's natural resonant frequency:

$$f = \frac{1}{2\pi\sqrt{LC}}$$

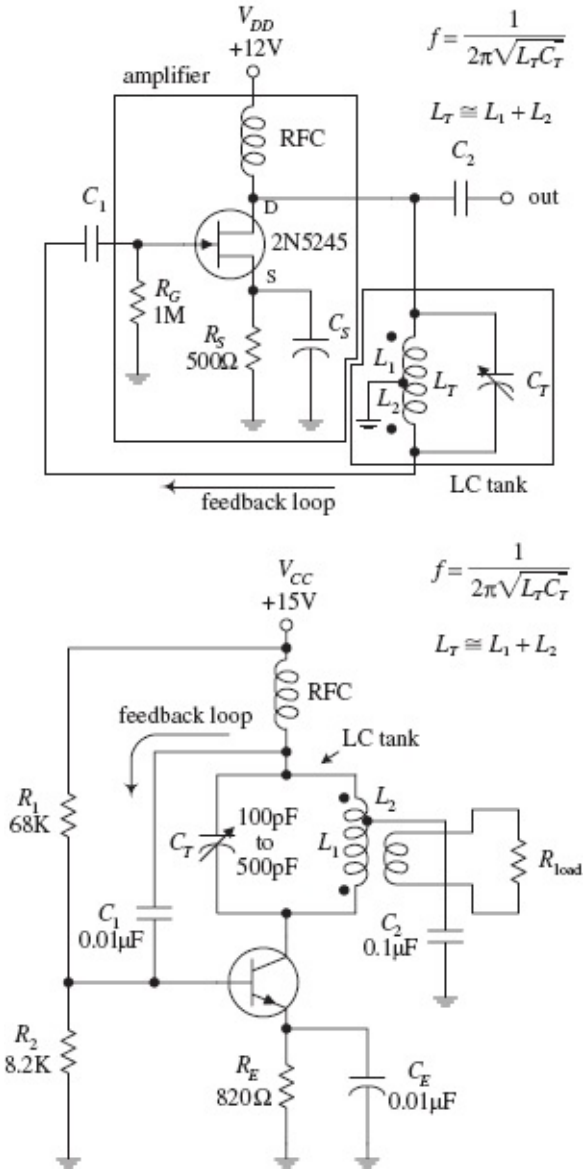
(Recall from Chap. 2 that a parallel *LC* resonant circuit's impedance becomes large at the resonant frequency but falls off on either side, allowing the feedback signal to be filtered out to ground.) If a sinusoidal voltage set at the resonant frequency is present at V_1, the amplifier is alternately driven to saturation in the positive and negative directions, resulting in a square wave at the output V_2. This square wave has a strong fundamental Fourier component at the resonant frequency, part of which is fed back to the noninverting input through the resistor to keep oscillations from dying out. If the initially applied sinusoidal voltage at V_1 is removed, oscillations will continue, and the voltage at V_1 will be sinusoidal. Now in practice (considering real-life components and not theoretical models), it is not necessary to apply a sine wave to V_1 to get things going (this is a fundamentally important point to note). Instead, due to imperfections in the amplifier, the oscillator will self-start. Why? With real amplifiers, there is always some inherent noise present at the output even when the inputs of the amplifier are grounded (see Chap. 8). This noise has a Fourier component at the resonant frequency, and because of the positive feedback, it rapidly grows in amplitude (perhaps in just a few cycles) until the output amplitude saturates.

FIGURE 10.16

Now, in practice, *LC* oscillators usually do not incorporate op amps into their designs. At very high frequencies (e.g., RF range), op amps tend to become unreliable due to slew-rate and bandwidth limitations. When frequencies above around 100 kHz are needed, it is essential to use another kind of amplifier arrangement. For high-frequency applications, what is typically used is a transistor amplifier (e.g., bipolar or FET type). The switching speeds for transistors can be incredibly high—a 2000-MHz ceiling is not uncommon for special RF transistors. However, when using a transistor amplifier within an

oscillator, there may be a slight problem to contend with—one that you did not have to deal with when you used the op amp. The problem stems from the fact that transistor-like amplifiers often take their outputs at a location where the output happens to be 180° out of phase with its input (see Chap. 4). However, for the feedback to sustain oscillations, the output must be in phase with the input. In certain *LC* oscillators this must be remedied by incorporating a special phase-shifting network between the output and input of the amplifier. Let's take a look at a few popular *LC* oscillator circuits.

Hartley *LC* Oscillator

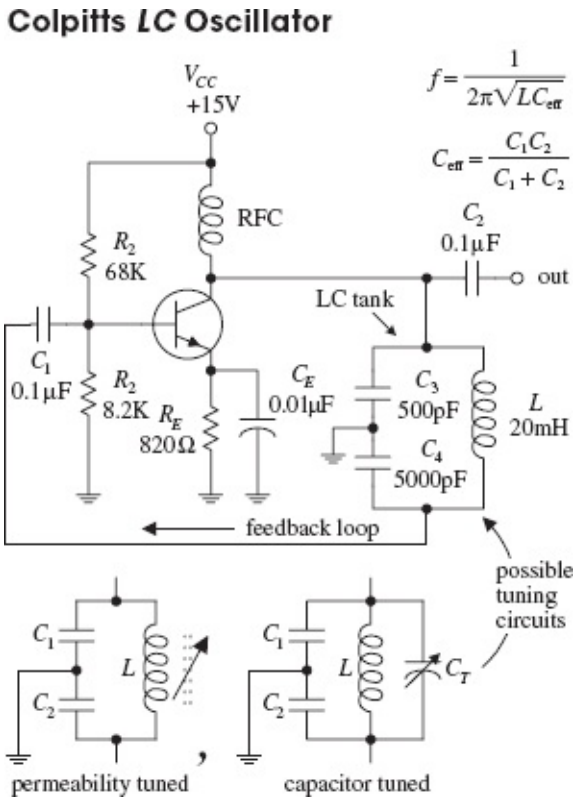


A Hartley oscillator uses an inductive voltage divider to determine the feedback ratio. The Hartley oscillator can take on a number of forms (FET, bipolar, etc.)—a JFET version is shown here. This oscillator achieves a 180° phase shift needed for positive feedback by means of a tapped inductor in the tank circuit. The phase voltage at the two ends of the inductor differ by 180° with respect to the ground tap. Feedback via L_2 is coupled through C_1 to the base of the transistor amplifier. (The tapped inductor is basically an autotransformer, where L_1 is the primary and L_2 is the secondary.) The frequency of the Hartley is determined by the tank's resonant frequency:

$$f = \frac{1}{2\pi\sqrt{L_T C_T}}$$

This frequency can be adjusted by varying C_T. R_G acts as a gate-biasing resistor to set the gate voltage. R_S is the source resistor. C_S is used to improve amplifier stability, while C_1 and C_2 act as a dc-blocking capacitor that provides low impedance at the oscillator's operating frequency while preventing the transistor's dc operating point from being disturbed. The radiofrequency choke (RFC) aids in providing the amplifier with a steady dc supply while eliminating unwanted ac disturbances.
The second circuit is another form of the Hartley oscillator that uses a bipolar transistor instead of a JFET as the amplifier element. The frequency of operation is again determined by the resonant frequency of the *LC* tank. Notice that in this circuit the load is coupled to the oscillator via a transformer's secondary.

FIGURE 10.17



The Colpitts oscillator is adaptable to a wide range of frequencies and can have better stability than the Hartley. Unlike the Hartley, feedback is obtained by means of a tap between two capacitors connected in series. The 180° phase shift required for sustained oscillation is achieved by using the fact that the two capacitors are in series; the ac circulating current in the *LC* circuit (see Chap. 2) produces voltage drops across each capacitor that are of opposite signs—relative to ground—at any instant in time. As the tank circuit oscillates, its two ends are at equal and opposite voltages, and this voltage is divided across the two capacitors. The signal voltage across C_4 is then connected to the transistor's base via coupling capacitor C_1, which is part of the signal from the collector. The collector signal is applied across C_3 as a feedback signal whose energy is coupled into the tank circuit to compensate for losses. The operating frequency of the oscillator is determined by the resonant frequency of the *LC* tank:

$$f = \frac{1}{2\pi\sqrt{LC_{eff}}}$$

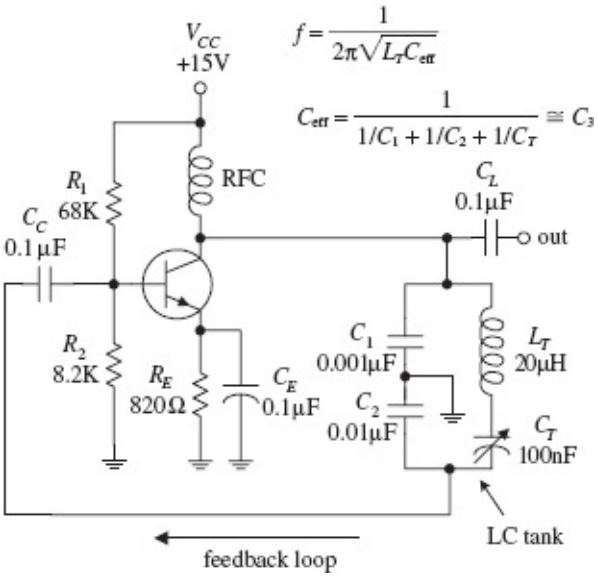
where C_{eff} is the series capacitance of C_3 and C_4:

$$\frac{1}{C_{eff}} = \frac{1}{C_1} + \frac{1}{C_2}$$

C_1 and C_2 are dc-blocking capacitors, while R_1 and R_2 act to set the bias level of the transistor. The RFC choke is used to supply steady dc to the amplifier. This circuit's tank can be exchanged for one of the two adjustable tank networks. One tank uses permeability tuning (variable inductor), while the other uses a tuning capacitor placed across the inductor to vary the resonant frequency of the tank.

FIGURE 10.18

Clapp Oscillator



The Clapp oscillator has exceptional frequency stability. It is a simple variation of the Colpitts oscillator. The total tank capacitance is the series combination of C_1 and C_2. The effective inductance L of the tank is varied by changing the net reactance by adding and subtracting capacitive reactance via C_T from inductive reactance of L_T. Usually C_1 and C_2 are much larger than C_T, while L_T and C_T are series resonant at the desired frequency of operation. C_1 and C_2 determine the feedback ratio, and they are so large compared with C_T that adjusting C_T has almost no effect on feedback. The Clapp oscillator achieves its reputation for stability since stray capacitances are swamped out by C_1 and C_2, meaning that the frequency is almost entirely determined by L_T and C_T. The frequency of operation is determined by

$$f = \frac{1}{2\pi\sqrt{L_T C_{eff}}}$$

where C_{eff} is

$$C_{eff} = \frac{1}{1/C_1 + 1/C_2 + 1/C_T} = C_3$$

FIGURE 10.19

10.6 Crystal Oscillators

When stability and accuracy become critical in oscillator design—which is often the case in high-quality radio and microprocessor applications—one of the best approaches is to use a crystal oscillator. The stability of a crystal oscillator (from around 0.01 to 0.001 percent) is much greater than that of an *RC* oscillator (around 0.1 percent) or an *LC* oscillator (around 0.01 percent at best).

When a quartz crystal is cut in a specific manner and placed between two conductive plates that act as leads, the resulting two-lead device resembles an *RLC* tuned resonant tank. When the crystal is shock-excited by either a physical compression or an applied voltage, it will be set into mechanical vibration at a specific frequency and will continue to vibrate for some time, while at the same time generating an ac voltage between its plates. This behavior, better know as the *piezoelectric effect*, is similar to the damped electron oscillation of a shock-excited *LC* circuit. However, unlike an *LC* circuit, the oscillation of the crystal after the initial shock excitation will last longer—a result of the crystal's naturally high *Q* value. For a high-quality crystal, a *Q* of 100,000 is not uncommon. *LC* circuits typically have a *Q* of around a few hundred.

The *RLC* circuit shown in Fig. 10.20 is used as an equivalent circuit for a crystal. The lower branch of the equivalent circuit, consisting of R_1, C_1, and L_1 in series, is called the *motional arm*. The motional arm represents the series mechanical resonance of the crystal. The upper branch containing C_0 accounts for the

stray capacitance in the crystal holder and leads. The *motional inductance* L_1 is usually many henries in size, while the motional capacitance C_1 is very small (<<1 pF). The ratio of L_1 to C_1 for a crystal is much higher than could be achieved with real inductors and capacitors. Both the internal resistance of the crystal R_1 and the value of C_0 are both fairly small. (For a 1-MHz crystal, the typical components values within the equivalent circuit would be L_1 = 3.5 H, C_1 = 0.007 pF, R_1 = 340 Ω, C_0 = 3 pF. For a 10-MHz fundamental crystal, the typical values would be L_1 = 9.8 mH, C_1 = 0.026 pF, R_1 = 7 Ω, C_0 = 6.3 pF.)

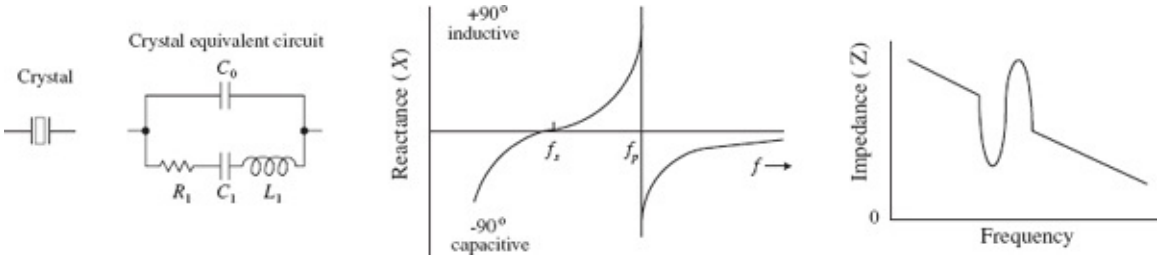


FIGURE 10.20

In terms of operation, a crystal can be driven at *series resonance* or *parallel resonance*. In series resonance, when the crystal is driven at a particular frequency, called the *series resonant frequency* f_s, the crystal resembles a series-tuned resonance *LC* circuit; the impedance across it goes to a minimum—only R_1 remains. In parallel resonance, when the crystal is driven at what is called the *parallel resonant frequency* f_p, the crystal resembles a parallel-tuned *LC* tank; the impedance across it peaks to a high value (see the graphs in Fig. 10.20).

Quartz crystals come in series-mode and parallel-mode forms and may either be specified as a fundamental-type or an overtone-type crystal. Fundamental-type crystals are designed for operation at the crystal's fundamental frequency, while overtone-type crystals are designed for operation at one of the crystal's overtone frequencies. (The fundamental frequency of a crystal is accompanied by harmonics or overtone modes, which are odd multiples of the fundamental frequency. For example, a crystal with a 15-MHz fundamental also will have a 45-MHz third overtone, a 75-MHz fifth overtone, a 135-MHz ninth overtone, etc. Figure 10.21 below shows an equivalent *RLC* circuit for a crystal, along with a response curve, both of which take into account the overtones.) Fundamental-type crystals are available from around 10 kHz to 30 MHz, while overtone-type crystals are available up to a few hundred megahertz. Common frequencies available are 100 kHz and 1.0, 2.0, 4, 5, 8, and 10 MHz.

Designing crystal oscillator circuits is similar to designing *LC* oscillator circuits, except that now you replace the *LC* tank with a crystal. The crystal will supply positive feedback and gain at its series or parallel resonant frequency, hence leading to sustained oscillations. Here are a few basic crystal oscillator circuits to get you started.

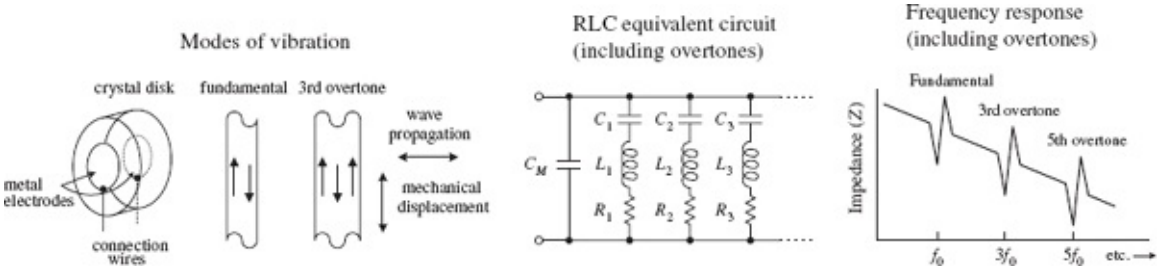
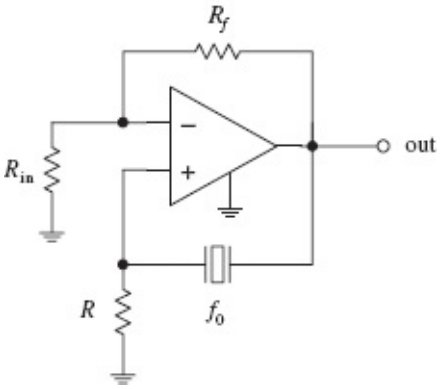
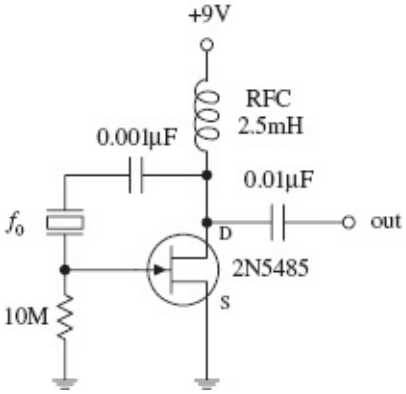


FIGURE 10.21

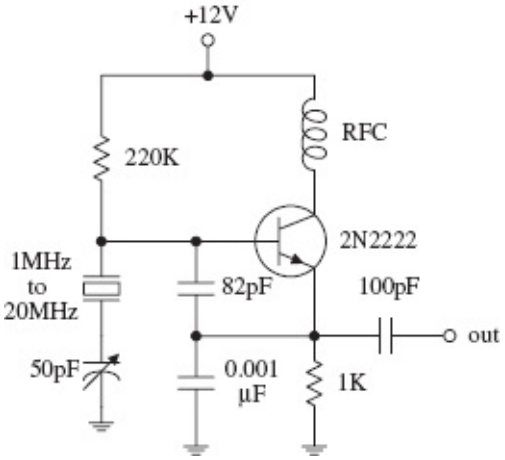
Basic Crystal Oscillator



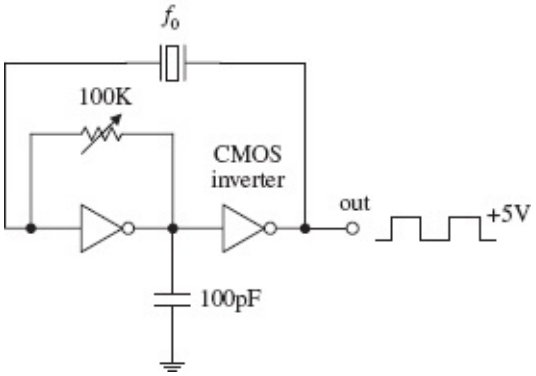
Pierce Oscillator



Colpitts Oscillator



CMOS Inverter Oscillator



The simple op amp circuit shown here resembles the *LC* oscillator circuit in Fig. 10.16, except that it uses the series resonance of the crystal instead of the parallel resonance of an *LC* circuit to provide positive feedback at the desired frequency. Other crystal oscillators, such as the Pierce oscillator, Colpitts oscillator, and a CMOS inverter oscillator, shown below, also incorporate a crystal as a frequency-determining component. The Pierce oscillator, which uses a JFET amplifier stage, employs a crystal as a series-resonant feedback element; maximum positive feedback from drain to gate occurs only at the crystal's

series-resonant frequency. The Colpitts circuit, unlike the Pierce circuit, uses a crystal in the parallel feedback arrangement; maximum base-emitter voltage signal occurs at the crystal's parallel-resonant frequency. The CMOS circuit uses a pair of CMOS inverters along with a crystal that acts as a series-resonant feedback element; maximum positive feedback occurs at the crystal's series resonant frequency.

FIGURE 10.22

There are a number of ICs available that can make designing crystal oscillators a breeze. Some of these ICs, such as the 74S124 TTL VCO (squarewave generator), can be programmed by an external crystal to output a waveform whose frequency is determined by the crystal's resonant frequency. The MC12060 VCO, unlike the 74S124, outputs a pair of sine waves. Check the catalogs to see what other types of oscillator ICs are available.

Now there are also crystal oscillator modules that contain everything (crystal and all) in one single package. These modules resemble a metal-like DIP package, and they are available in many of the standard frequencies (e.g., 1, 2, 4, 5, 6, 10, 16, 24, 25, 50, and 64 MHz, etc.). Again, check out the electronics catalogs to see what is available.

10.7 Microcontroller Oscillators

In Chap. 13, we will also see how a microcontroller can be used to generate a waveform using a digital-to-analog convertor. The basic technique is to store the waveform in memory and then play it through the digital-to-analog converter.

In the case where just a squarewave is required, a simple 8-pin microcontroller with a built-in clock can be an effective alternative to a 555 timer, requiring fewer external components.

CHAPTER 11

Voltage Regulators and Power Supplies

Circuits usually require a dc power supply that can maintain a fixed voltage while supplying enough current to drive a load. Batteries make good dc supplies, but their relatively small current capacities make them impractical for driving high-current, frequently used circuits. An alternative solution is to take a 120-V ac, 60-Hz line voltage and convert it into a usable dc voltage.

There are two approaches to this. The more traditional approach is to use a step-down transformer. The other approach is to use a "switch-mode" power supply. This latter method has all but taken over from step-down transformers in recent years, and is the reason your "wall-wart" black plastic power adapters have become smaller and lighter. They also have the advantage that they will often work without modification or switch flipping on the higher line voltages found in other parts of the world. We'll discuss both methods in this chapter, beginning with the step-down transformer approach.

The trick to converting the ac line voltage into a usable (typically lower-level) dc voltage is to first use a transformer to step down the ac voltage. After that, the transformed voltage is applied through a rectifier network to get rid of the negative swings (or positive swings if you are designing a negative voltage supply). Once the negative swings are eliminated, a filter network is used to flatten out the rectified signal into a nearly flat (rippled) dc voltage pattern. Figure 11.1 shows the process in action.

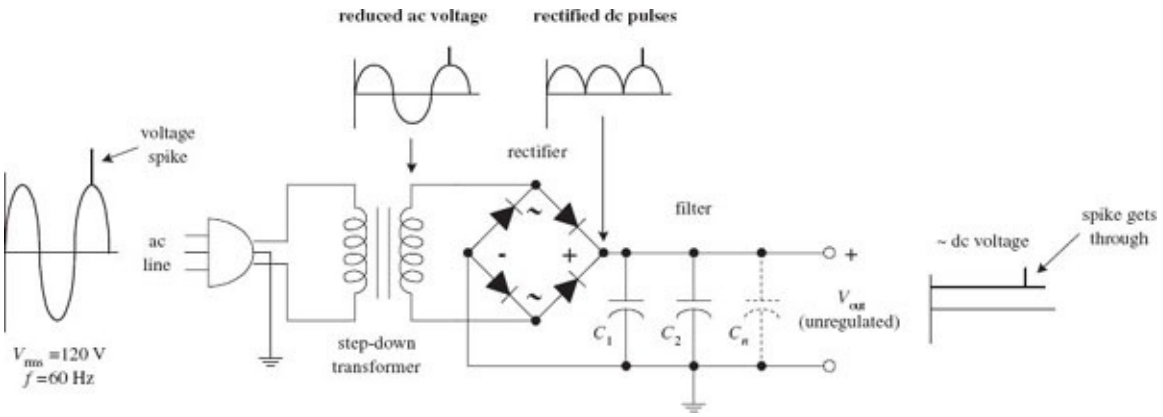


FIGURE 11.1

Now there is one problem with this supply—it is *unregulated*. This means that if there are any sudden surges within the ac input voltage (spikes, dips, etc.), these variations will be expressed at the supply's output (notice the spike that gets through in Fig. 11.1). Using an unregulated supply to run sensitive circuits (e.g., digital IC circuits) is a bad idea. The current spikes can lead to improper operating characteristics (e.g., false triggering, etc.) and may destroy the ICs in the process. An unregulated supply also has a problem maintaining a constant output voltage as the load resistance changes. If a highly resistive (low-current) load is replaced with a lower-resistance (high-current) load, the unregulated output voltage will drop (Ohm's law).

Fortunately, there is a special circuit that can be placed across the output of an unregulated supply to convert it into a regulated supply—a supply that eliminates the spikes and maintains a constant output voltage with load variations (see Fig. 11.2). This special circuit is called a *voltage regulator*.

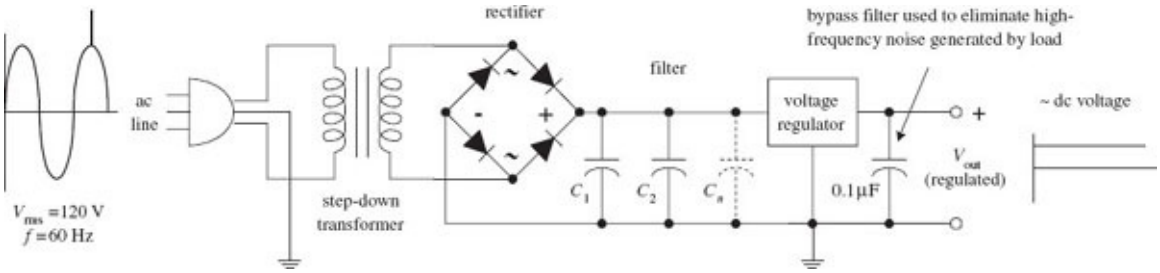


FIGURE 11.2

A voltage regulator is designed to automatically adjust the amount of current flowing through a load—so as to maintain a constant output voltage—by comparing the supply's dc output with a fixed or programmed internal reference voltage. A simple regulator consists of a sampling circuit, an error amplifier, a conduction element, and a voltage reference element (see Fig. 11.3).

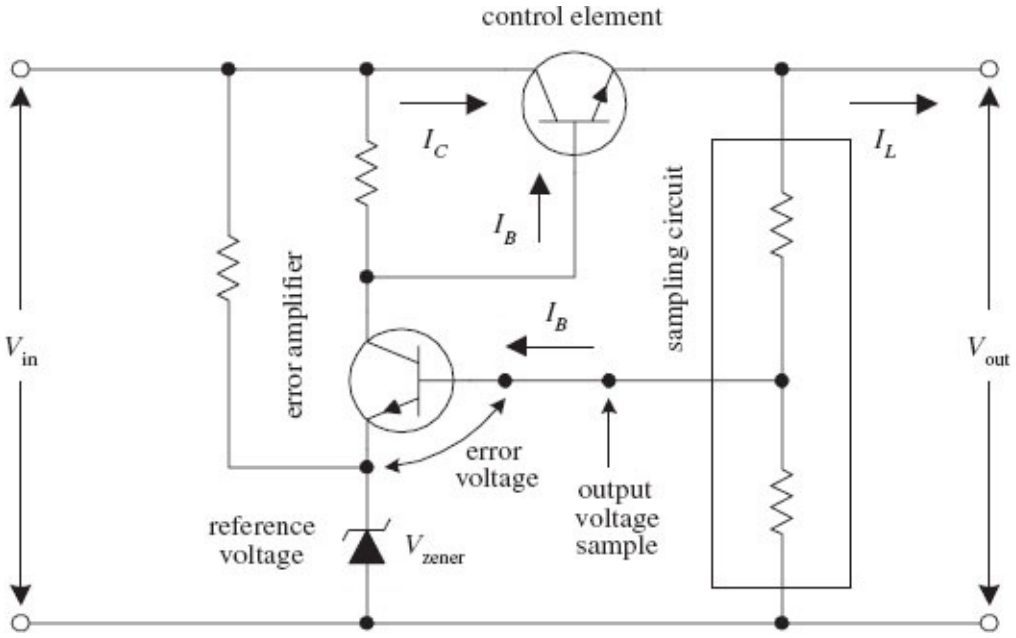


FIGURE 11.3

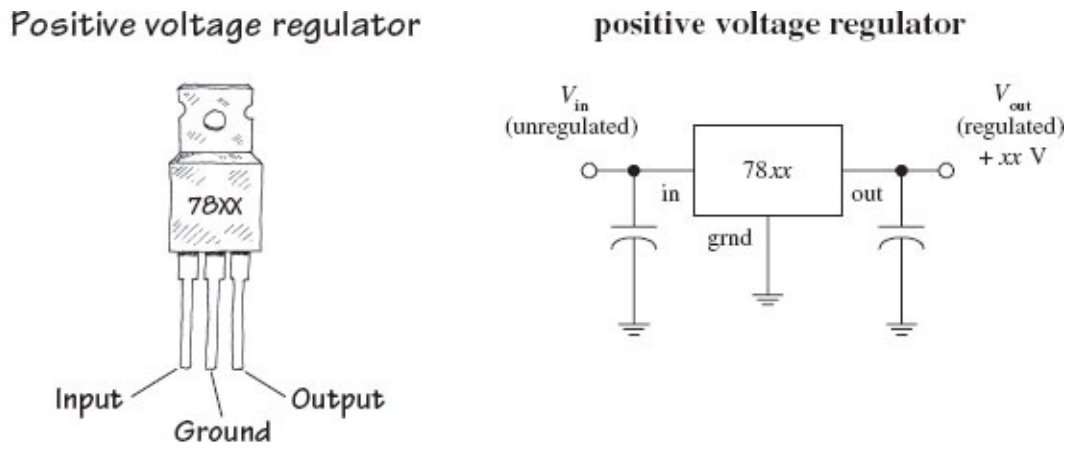
The regulator's sampling circuit (voltage divider) monitors the output voltage by feeding a *sample voltage* back to the error amplifier. The reference voltage element (zener diode) acts to maintain a constant *reference voltage* that is used by the error amplifier. The error amplifier compares the output sample voltage with the reference voltage and then generates an *error voltage* if there is any difference between the two. The error amplifier's output is then fed to the current-control element (transistor), which is used to control the load current.

In practice, you do not have to worry about designing voltage-regulator circuits from scratch. Instead, what you do is spend 50 cents for a voltage-regulator IC. Let's take a closer look at these integrated devices.

11.1 Voltage-Regulator ICs

There are a number of different kinds of voltage-regulator ICs on the market today. Some of these devices are designed to output a fixed positive voltage, some are designed to output a fixed negative voltage, and others are designed to be adjustable.

11.1.1 Fixed Regulator ICs



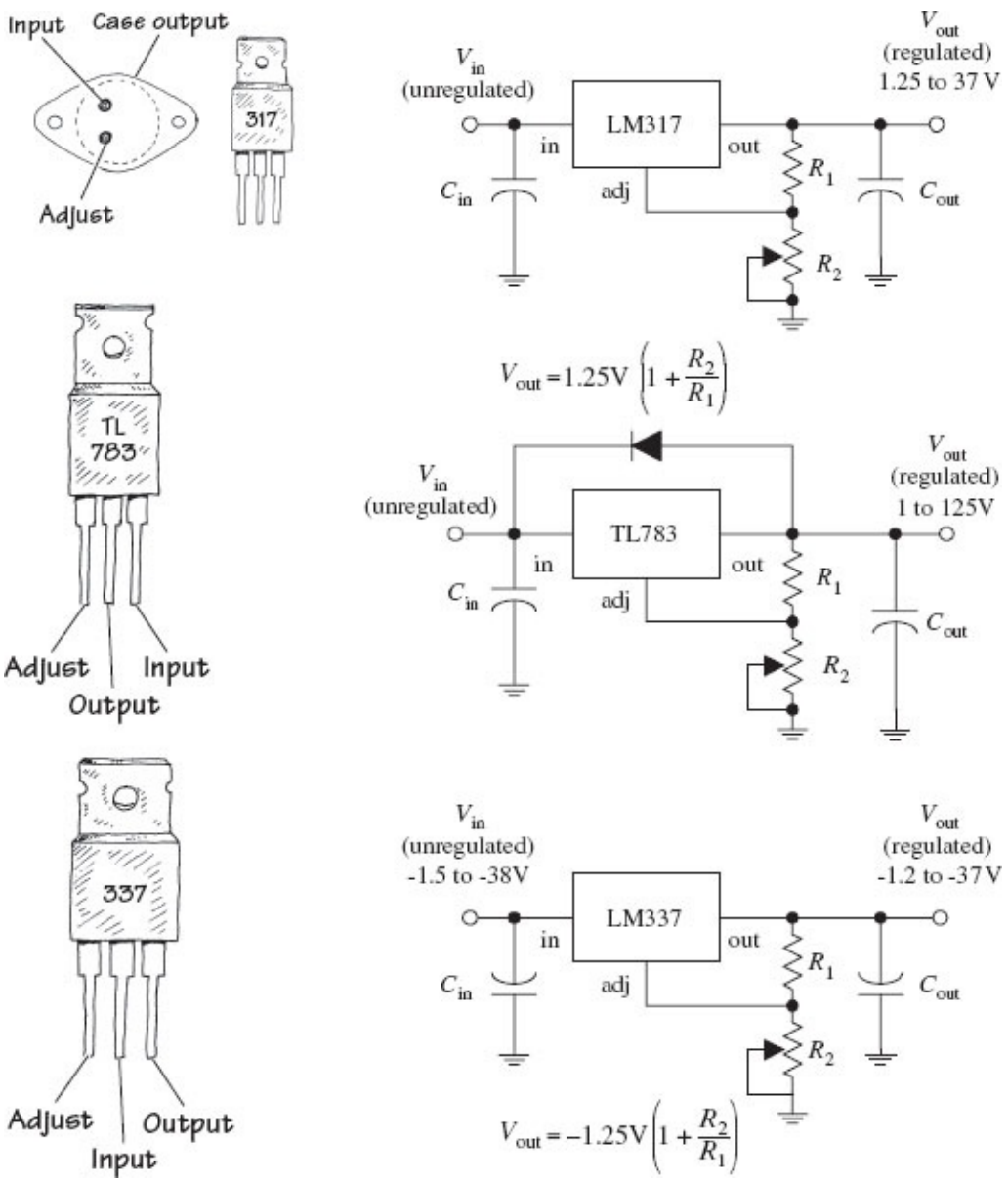
One popular line of regulators includes the three-terminal LM78xx series shown here. The "xx" digits represent the output voltage, e.g., 7805 (5 V), 7806 (6 V), 7808 (8 V), 7810 (10 V), 7812 (12 V), 7815 (15 V), 7818 (18 V), and 7824 (24 V). These devices can handle a maximum output current of 1.5 A if properly heat-sunk. To remove unwanted input or output spikes/noise, capacitors can be attached to the regulator's input and output terminals, as shown in the figure. A popular series of negative voltage-regulator IC is the LM79xx regulators, where "xx" represent the negative output voltage. These devices can handle a maximum output current of 1.5 A. A number of different manufacturers make their own kinds of voltage regulators. Some of the regulators can handle more current than others.

These devices are also available as SMD parts, typically in a SOT-89 package. Be sure to check the data sheets though, as these parts often have a lower maximum output current than their through-hole siblings.

Check out the catalogs to see what is available.

FIGURE 11.4

11.1.2 Adjustable Regulator ICs



The LM317 regulator shown here is a popular 3-terminal adjustable positive voltage regulator. Unlike the 7800 fixed voltage-regulator series, the LM317 is a floating regulator—it sees only the input-to-output differential voltage—and it can be programmed via two external resistors to set the output voltage. In operation, the LM317 develops a nominal 1.25 V reference voltage between the output and adjust terminals. This reference voltage is impressed across program resistor R_1, and since this voltage is constant, a constant current I_1 flows through the output set resistor R_2, giving an output voltage given by the equation shown in the figure. Increasing R_2 forces the regulator's output to a higher level. The LM317 is designed to accept an unregulated input voltage of up to 37 V and can output a maximum current of 1.5 A. The TL783 is another positive adjustable regulator that can output a regulated voltage of from 1 to 125 V, with a maximum output current of 700 mA. The LM337T, unlike the previous two regulators, is an adjustable negative voltage regulator. It can output a regulated voltage of from –1.2 V to –37 V, with a maximum output current of 1.5 A. Again, check the electronics catalogs to see what other kinds of adjustable regulators are available. (C_{in} should be included if the regulator is far from the power source; it should be around 0.1 µF or so. C_{out} is used to eliminate voltage spikes at the output; it should be around 0.1 µF or larger.)

FIGURE 11.5

11.1.3 Regulator Specifications

The specifications tables for regulators typically will provide you with the following information: output voltage, accuracy (percent), maximum output current, power dissipation, maximum and minimum input voltage, 120-Hz ripple rejection (decibels), temperature stability ($\Delta V_{out}/\Delta T$), and output impedance (at specific frequencies). A regulator's ripple rejection feature can greatly reduce voltage variations in a power supply's output, as you will discover later in this chapter.

11.2 A Quick Look at a Few Regulator Applications

Before we take a look at how voltage regulators are used in power supplies, it is worthwhile seeing how they are used in other types of applications. Here are a few examples.

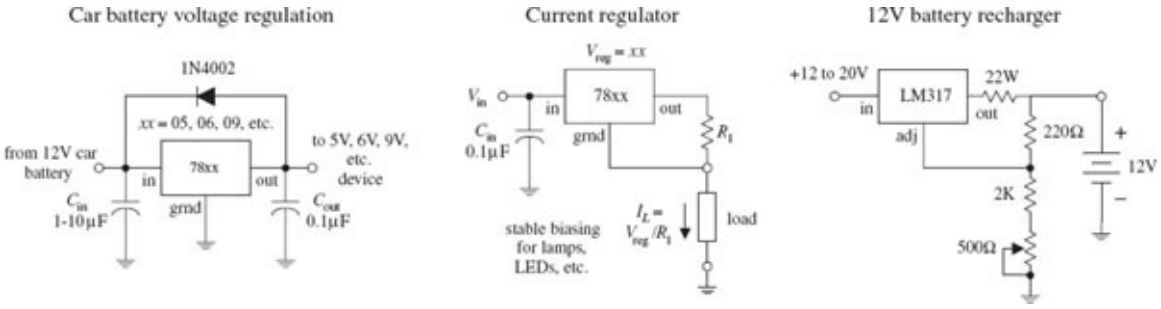


FIGURE 11.6

The constant current regulator is often used as a power supply for LEDs, especially the higher-power devices.

11.3 The Transformer

It is important that you choose the right transformer for your power supply. The transformer's secondary voltage should not be much larger than the output voltage of the regulator; otherwise, energy will be wasted because the regulator will be forced to dissipate heat. However, at the same time, the secondary voltage must not drop below the required minimum input voltage of the regulator (typically 2 to 3 V above its output voltage).

11.4 Rectifier Packages

Three basic rectifier networks used in power supply designs include the half-wave, full-wave, and bridge rectifiers, shown in Fig. 11.7. To understand how these rectifiers work, see Chap. 4.

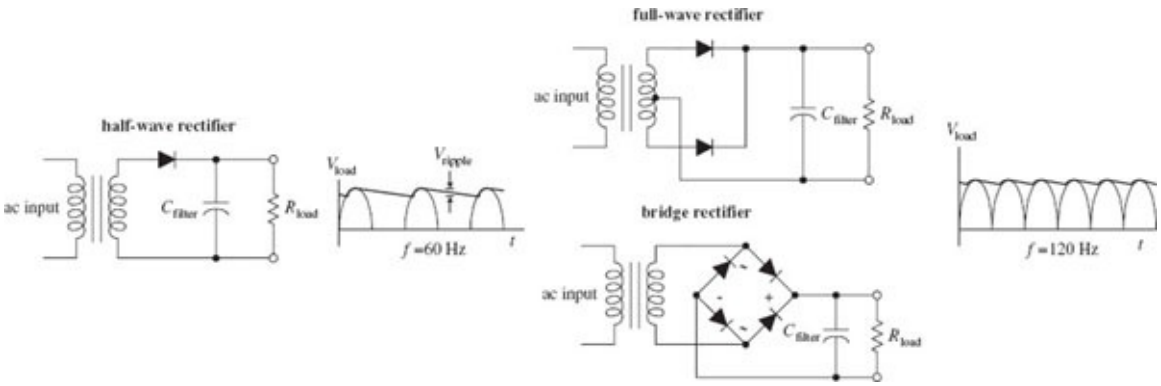


FIGURE 11.7

Half-wave, full-wave, and bridge rectifiers can be constructed entirely from individual diodes. However, both full-wave and bridge rectifiers also come in preassembled packages (see Fig. 11.8).

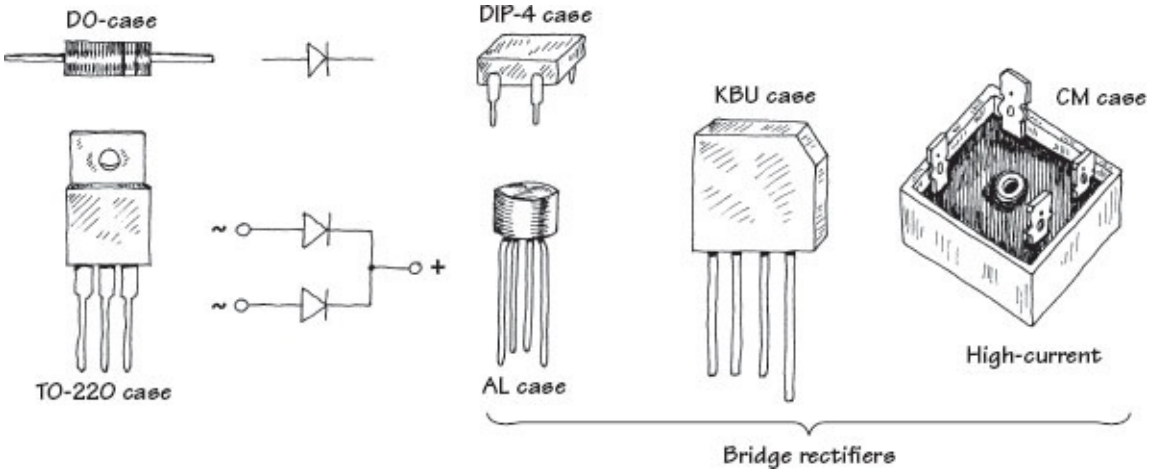
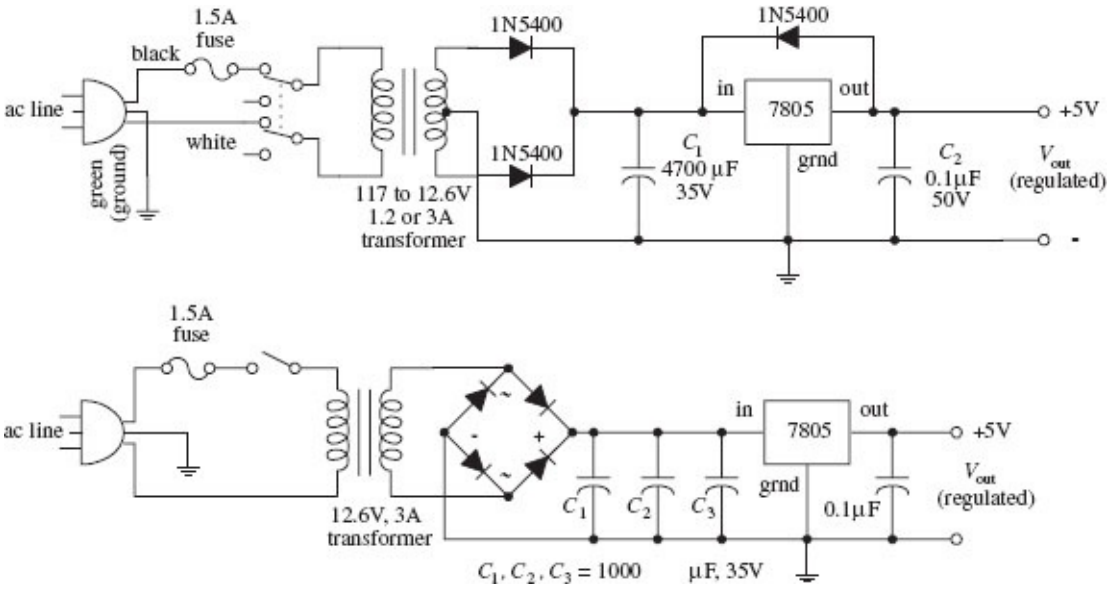


FIGURE 11.8

Make sure that the power supply's rectifier diodes have the proper current and peak-inverse-voltage (PIV) ratings. Typical rectifier diodes have current ratings from 1 to 25 A, PIV ratings from 50 to 1000 V, and surge-current ratings from around 30 to 400 A. Popular general-purpose rectifier diodes include the 1N4001 to 1N4007 series (rated at 1 A, 0.9-V forward voltage drop), the 1N5059 to 1N5062 series (rated at 2 A, 1.0-V forward voltage drop), the 1N5624 to 1N5627 series (rated at 5 A, 1.0-V forward voltage drop), and the 1N1183A-90A (rated at 40 A, 0.9-V forward voltage drop). For low-voltage applications, Schottky barrier rectifiers can be used; the voltage drop across these rectifiers is smaller than a typical rectifier (typically less than 0.4 V); however, their breakdown voltages are significantly smaller. Popular full-wave bridge rectifiers include the 3N246 to 3N252 series (rated at 1 A, 0.9-V forward voltage drop) and the 3N253 to 3N259 series (rated at 2 A, 0.85-V forward voltage drop).

11.5 A Few Simple Power Supplies

Regulated +5-V Supplies



The first supply uses a center-tapped transformer rated at 12.6 V at 1.2 to 3 A. The voltage after rectification resides at an 8.9-V peak pulse. The filter capacitor (C_1) smoothes the pulses, and the 7805 outputs a regulated +5 V. C_2 is placed across the output of the regulator to bypass high-frequency noise that might be generated by the load. The diode placed across the 7805 helps protect the regulator from damaging reverse-current surges generated by the load. Such surges may result when the power supply is turned off. For example, the capacitance across the output may discharge more slowly than the capacitance across the input. This would reverse-bias the regulator and could damage it in the process. The diode diverts the unwanted current away from the regulator. The second power supply is similar to the first but uses a bridge rectifier.

FIGURE 11.9

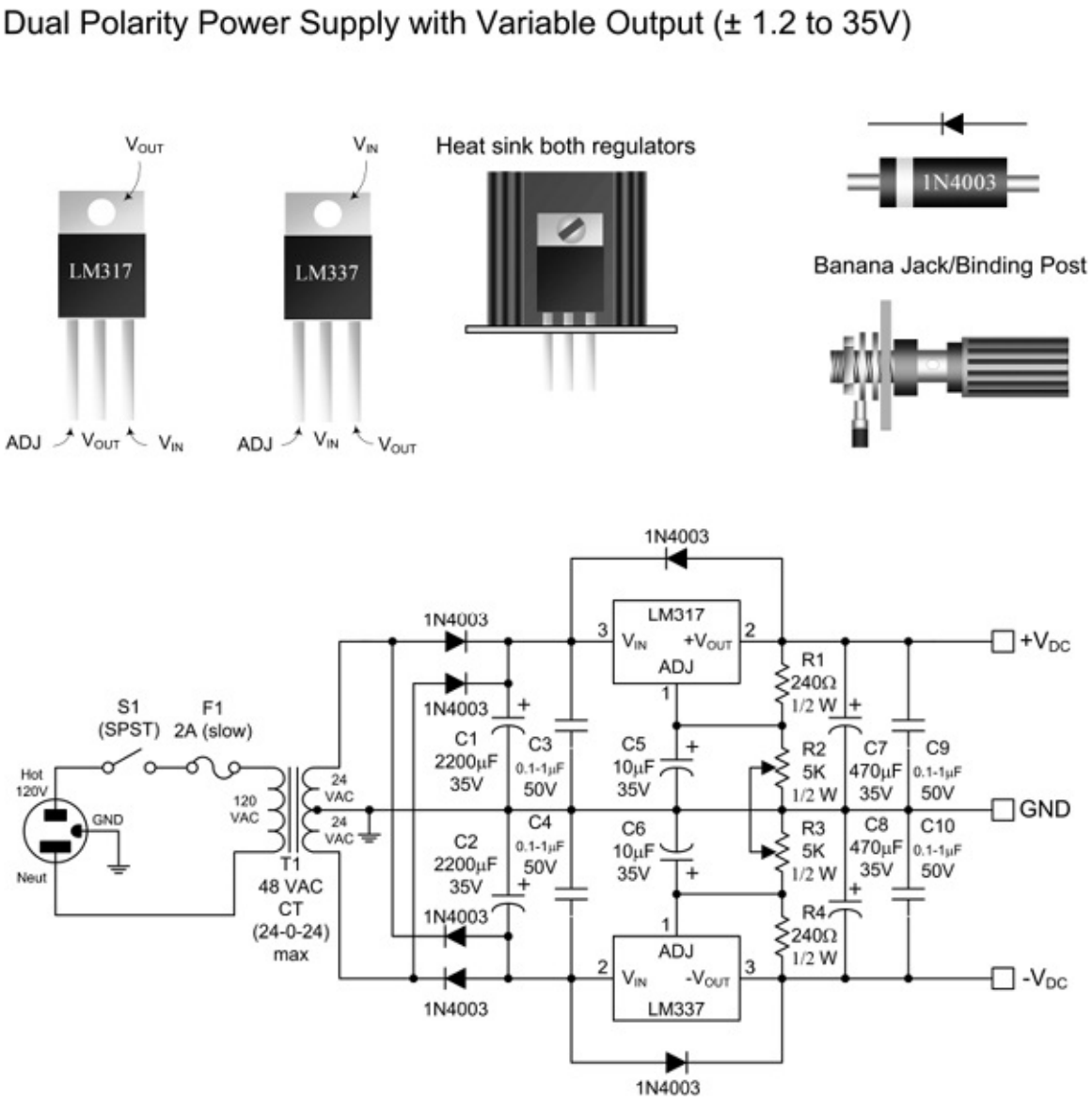
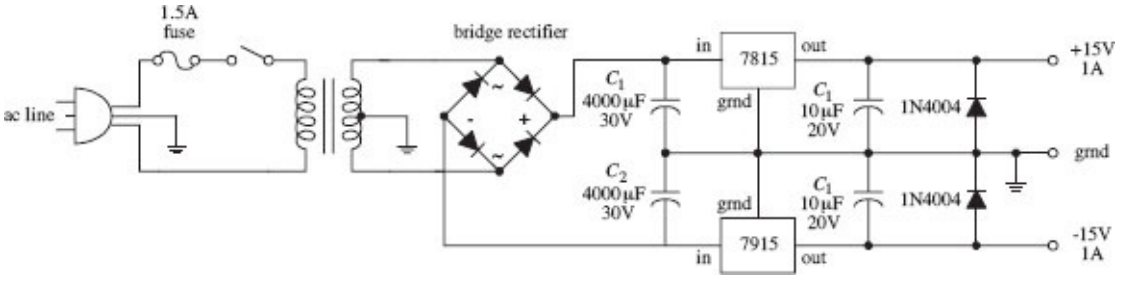


FIGURE 11.10a This dual-polarity linear power supply will provide any positive or negative voltage between 1.2 and 35 V. The complementary regulators—LM317 (+) and LM337 (−)—can deliver up to 1.5 A (output voltage dependent) if provided adequate heat sinking. This is sufficient for testing or powering a wide variety of everyday circuits. The key is the center-tap transformer with secondary tapped to ground that supplies both positive and negative voltages relative to ground. It converts the line voltage from 120 VAC to 48 VAC at the secondary, which is center-tapped and divided into 24-VAC portions. By using the center tap as a ground or common connection, it is possible to get both positive and negative outputs relative to ground. The diodes rectify the ac from the transformer output into a pulsing dc waveform. C_1 and C_2 electrolytic capacitors perform bulk filtering of the pulsing dc waveform, resulting in a raw dc voltage. C_3 and C_4 are small film capacitors bypassing the electrolytic capacitors to improve transient response and filter high-frequency line noise. The LM317 (+) and LM337 (−) are complementary adjustable voltage regulators, whose output can be programmed with two external resistors. The output voltage is given by:

$$V_{out} = 1.25\left(1 + \frac{R_2}{R_1}\right)$$

C_5 and C_6 improve the regulators' ripple rejection from 65 dB to 80 dB by preventing ripple voltage from being amplified at the output of the regulator. C_7 and C_8 electrolytic capacitors stiffen the output voltage and reduce output impedance. C_9 and C_{10} are small bypass capacitors to filter any high-frequency noise present at the output. All bypass capacitors should have low impedance (e.g., polyester, polypropylene, polystyrene, or Mylar film capacitors are okay). A 48-VAC CT (24-0-24) transformer is about the practical upper limit of commonly available models. This limit is set by the maximum input voltage of the regulators, and also by the bulk filter capacitors that are rated at 35 VDC.



Here, a 7815 positive voltage regulator and a 7915 negative voltage regulator are used to construct a ±15-V supply.

FIGURE 11.10b

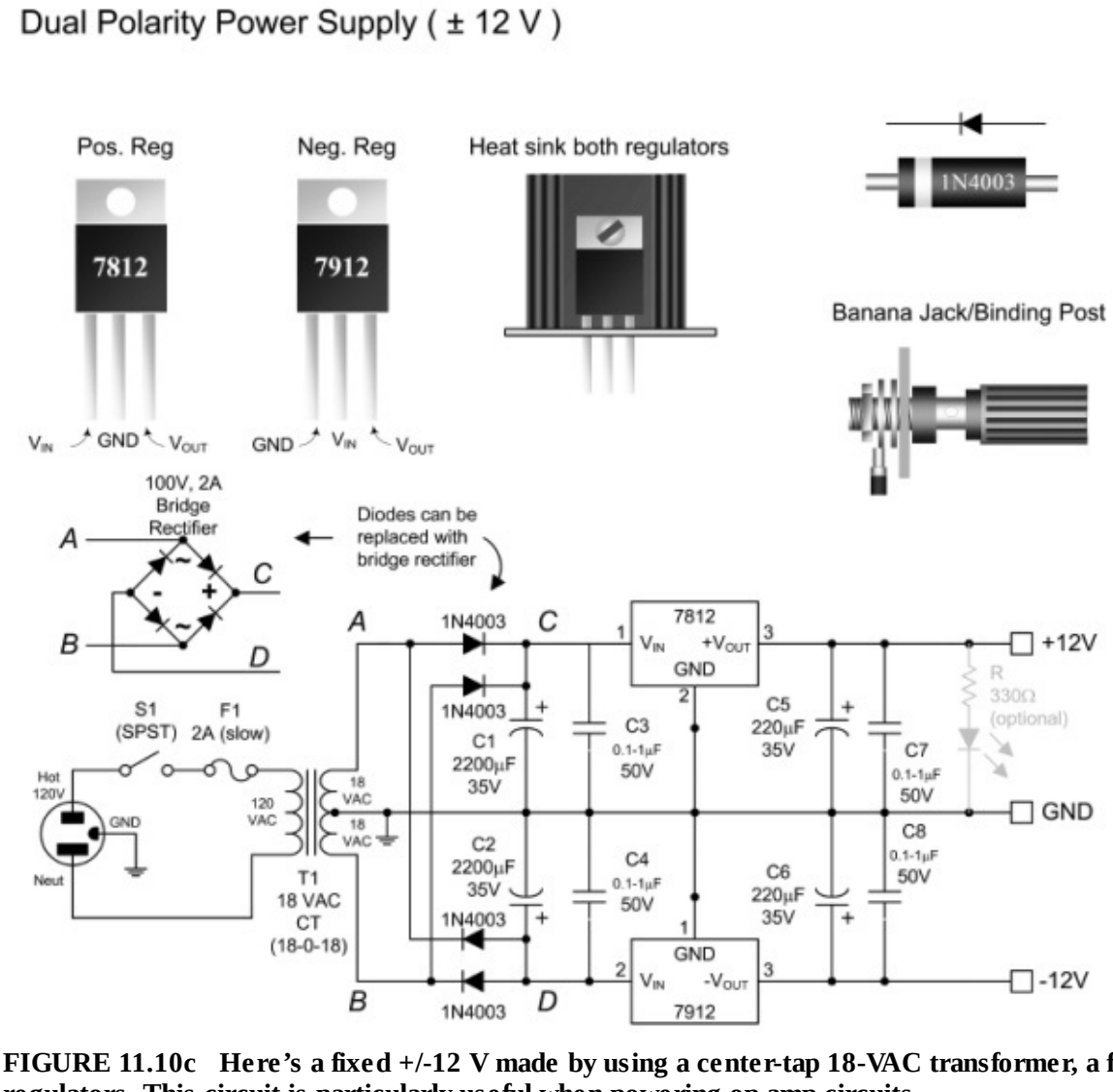
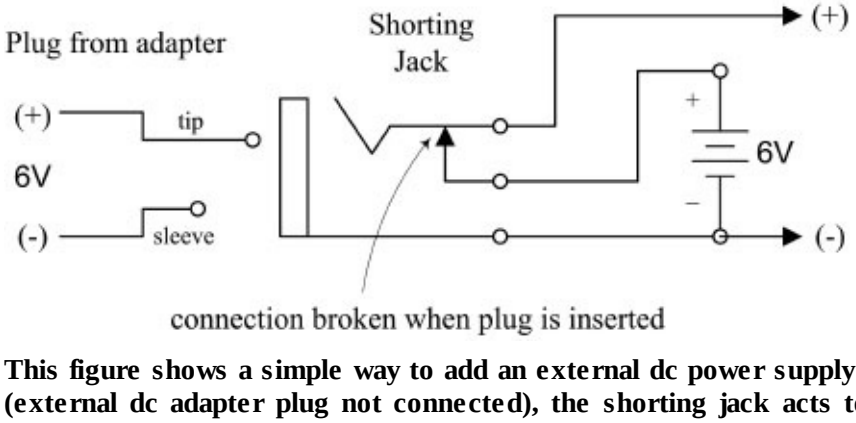


FIGURE 11.10c Here's a fixed +/-12 V made by using a center-tap 18-VAC transformer, a full-wave rectifier, and 7812 and 7912 regulators. This circuit is particularly useful when powering op amp circuits.



This figure shows a simple way to add an external dc power supply input to a battery-powered device. With no external power (external dc adapter plug not connected), the shorting jack acts to switch power to the battery. However, when external dc

power is applied via a dc adapter plug, the shorting jack switches power from the battery to the dc adapter connects.

FIGURE 11.11

11.6 Technical Points About Ripple Reduction

When using a supply to power sensitive circuits, it is essential to keep the variation in output voltage as small as possible. For example, when driving digital circuits from a 5-V supply, the variation in output voltage should be no more than 5 percent, or 0.25 V, if not lower. In fact, digital logic circuits usually have a minimum 200-mV noise margin around critical logic levels. Small analog signal circuits can be especially finicky when it comes to output variations. For example, they may require a variation less than 1 percent to operate properly. How do you keep the output variation error low? You use filter capacitors and voltage regulators.

Filter capacitors act to reduce the fluctuations in output by storing charge during the positive-going rectifier cycle and then releasing charge through the load—at a slow enough rate to maintain a level output voltage—during the negative-going rectifier cycle. If the filter capacitor is too small, it will not be able to store enough charge to maintain the load current and output voltage during the negative-going cycle.

As it turns out, the amount of current drawn by a load influences the rate of capacitor discharge. If a low-resistance, high-current load is placed across the supply's output, the capacitor will discharge relatively quickly, causing the voltage across the capacitor and therefore the voltage across the load to drop relatively quickly. On the other hand, a highly resistive, low-current load will cause the capacitor to discharge more slowly, which means the dip in output voltage will not be as significant. To calculate the drop in voltage of the capacitor during its discharge cycle, use

$$I = C\frac{dV}{dt} \approx C\frac{\Delta V}{\Delta t}$$

where I is the load current, Δt is the discharge time, and ΔV is the fluctuation in output above and below the average dc level. ΔV is also referred to as the *peak-to-peak ripple voltage* V_{ripple}(pp). (Here, I cheated by substituting straight lines for exponential decays to describe the discharge cycle; see Fig. 11.12.) Δt can be approximated by dividing 1 by the rectified output voltage frequency. For a full-wave rectifier, the period is 1/120 Hz, or 8.3×10^{-3} s. In reality, the actual amount of time the capacitor spends discharging during a peak-to-peak variation is about 5 ms. The other 3.3 ms is used up during the charge. To make life easier on you, the following equation simplifies matters:

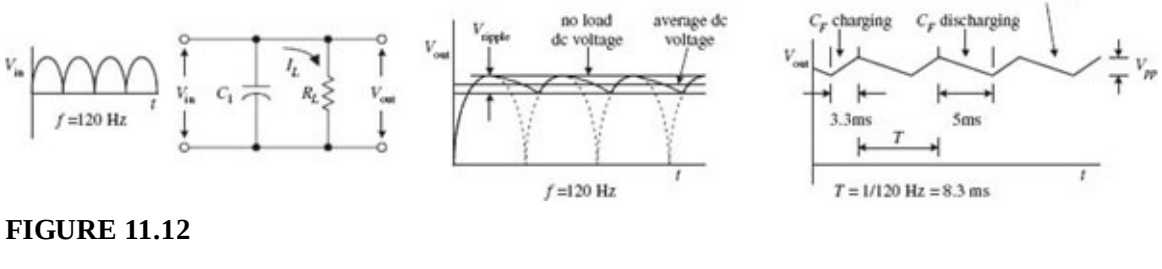


FIGURE 11.12

$$V_{ripple(rms)} = (0.0024\ \text{s})\frac{I_L}{C_f}$$

Note that the ripple voltage is not given in the peak-to-peak form but is given in the rms form (recall that $V_{pp} = \sqrt{2}V_{rms}$). To test the equation out, let's find the ripple voltage for a 5-V supply that has a 4700-μF

filter capacitor and a maximum load current of 1.0 A. (Here I am pretending that no voltage regulator is present at this point.) After plugging the numbers into the equation, you get $V_{ripple(rms)}$ = 510 mV. But now, recall that a minute ago I said that the amount of variation in the output had to be ±0.25 V% to run digital ICs—510 mV is too big. Now you could keep fiddling around with the capacitance value in the equation and come up with an even better answer, say, letting *C* equal infinity. In theory, this is fine, but in reality, it is not fine. It is not fine for three basic reasons. The first reason has to do with the simple fact that you cannot find a capacitor at Radio Shack that has an infinite capacitance. If an infinite-capacitance capacitor existed, the universe would not be the same, and you and I probably would not be around to talk about it. The second reason has to do with capacitor tolerances. Unfortunately, the high-capacity electrolytic capacitors used in power supplies have some of the worst tolerances among the capacitor families. It is not uncommon to see a 5 to 20 percent or even larger percentage tolerance for these devices. The mere fact that the tolerances are so bad makes being "nitpicky" about the equation a questionable thing to do. The third and perhaps most important reason for avoiding fiddling around with the equation too much has to do with the inherent ripple-rejection characteristics of voltage regulators. As you will see, the voltage regulator can save us.

Voltage regulators often come with a ripple-rejection parameter given in decibels. For example, the 7805 has a ripple-rejection characteristic of about 60 dB. Using the attenuation expression, you can find the extent of the ripple reduction:

$$-60 \text{ dB} = 20 \log_{10} \frac{V_{out}}{V_{in}}$$

$$-3 = \log_{10} \frac{V_{out}}{V_{in}}$$

$$10^{-3} = \frac{V_{out}}{V_{in}}$$

The last expression says that the output ripple is reduced by a factor of 1000. This means that if you use the regulator with the initial setup, you will only have an output ripple of 0.51 mV—a value well within the safety limits. At this point, it is important to note that the 7805 requires a minimum voltage difference between its input and output of 3 V to function properly. This means that to obtain a 5-V output, the input to the regulator must be at least 8 V. At the same time, it is important to note the voltage drop across the rectifier (typically around 1 to 2 V). The secondary voltage from the transformer therefore should be even larger than 8 V. A transformer with a secondary voltage of 12 V or so would be a suitable choice for the 5-V supply.

Now let's see how well the LM319 adjustable regulator rejects the ripple. Let's say that an LM317 is used in a power supply that has a transformer secondary rms voltage of 12.6 V. The capacitor's peak voltage during a cycle will be 17.8 V (the peak-to-peak voltage of the secondary). An LM317 has a ripple-rejection characteristic of around 65 dB, but this value can be raised to approximately 80 dB by bypassing the LM317's voltage divider with a 10-µF capacitor (see Fig. 11.13).

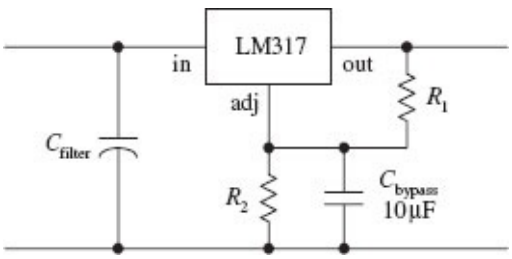


FIGURE 11.13

If you let C = 4700 μF and assume a 1.5-A maximum load current, you get a ripple voltage of

$$V_{r(rms)} = 0.0024\ s(1.5\ A/4700\ \mu F) = 760\ mV$$

Again, the ripple voltage is too much for sensitive ICs to handle. However, if you consider the LM319's ripple-rejection characteristic (assuming that you use the bypassing capacitor), you get a reduction of

$$-80\ db = 20 \log_{10} \frac{V_{out}}{V_{in}}$$

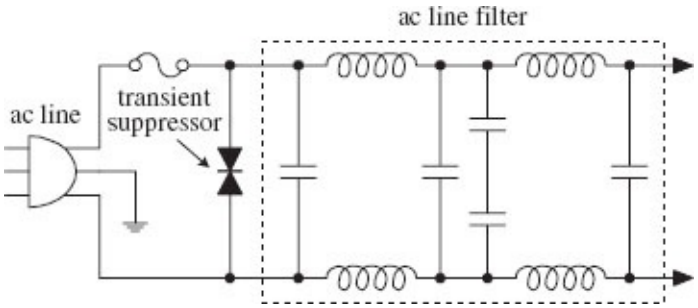
$$-4 = \log_{10} \frac{V_{out}}{V_{in}}$$

$$10^{-4} = \frac{V_{out}}{V_{in}}$$

In other words, the output ripple is reduced by a factor of 10,000, so the final output ripple voltage is only 0.076 mV.

11.7 Loose Ends

Line Filter and Transients Suppressors

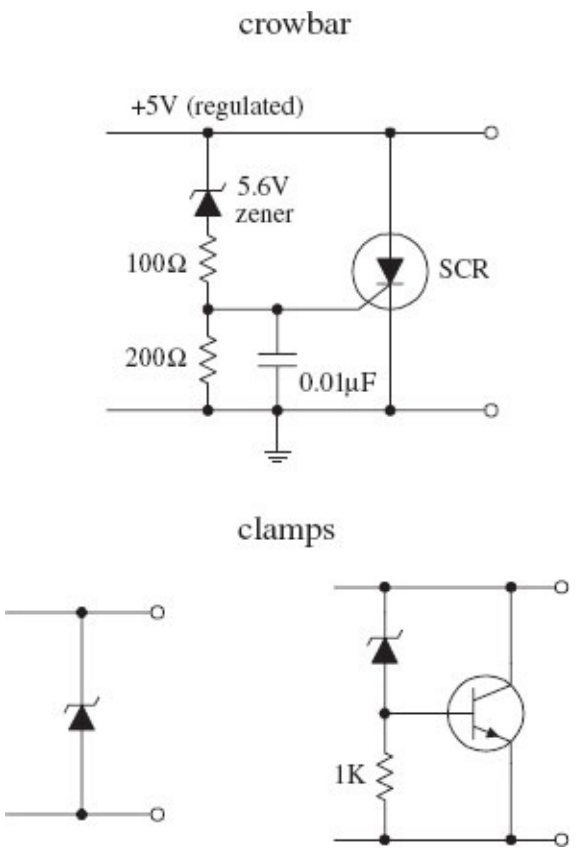


A line filter is an *LC* filter circuit that is inserted into a supply to filter out unwanted high-frequency interference present in the input line supply. Line filters also can help reduce voltage spikes, as well as help eliminate the emission of radiofrequency interference by the power supply. Line filters are placed before the transformer, as shown in the figure. AC line filters can be purchased in preassembled packages. See the electronics catalogs for more info.

A transient suppressor is a device that acts to short out when the terminal voltage exceeds safe limits (e.g., spikes). These devices act like bidirectional high-power zener diodes. They are inexpensive, come in diode-like packages, and come with low-voltage and peak-pulse-voltage ratings.

FIGURE 11.14

Overvoltage Protection



The crowbar and clamp circuits shown here can be placed across the output of a regulated supply to protect a load against an unregulated voltage that would be present at the output if the voltage regulator failed (shorted internally).

CROWBAR

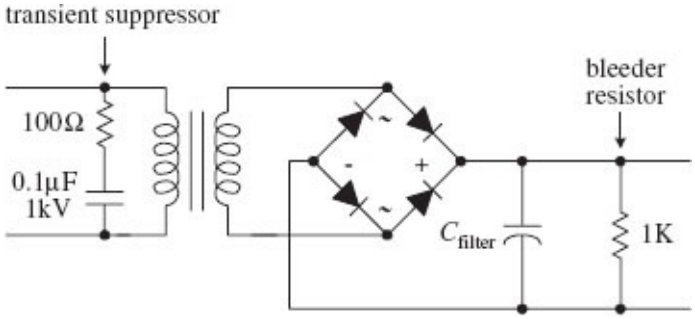
For the crowbar circuit, when the supply voltage exceeds the zener diode's breakdown voltage by 0.6 V, the zener diode conducts, triggering the SCR into conduction. The SCR then diverts potentially harmful current to ground. The crowbar's SCR will not turn off until the power supply is turned off or the SCR's anode-to-cathode current is interrupted, say, by means of a switch.

CLAMP

A zener diode placed across the output of a supply also can be used for overvoltage protection. However, it may "fry" if the unregulated current is too large. To avoid frying the zener diode, use a high-power transistor to help divert the current. When the zener diode's breakdown voltage is exceeded, some of the current that flows through it will enter the transistor's base, allowing the excessive current to flow toward ground. Using clamps can eliminate false triggering caused by voltage spikes; the crowbar, on the other hand, would need to be reset in such a case.

FIGURE 11.15

Bleeder Resistors and Transient Suppressors



When a resistor is placed across the output of the unregulated supply, it will act to discharge the high-voltage (potentially lethal) filter capacitor when the supply is turned off and the load removed. Such a resistor is referred to as a *bleeder resistor*. A 1-k, 1/2-W resistor is suitable for most applications.

An *RC* network placed across the primary coil of the transformer can prevent large, potentially damaging inductive transients from forming when the supply is turned off. The capacitor must have a high voltage rating. A typical *RC* network consists of a 100-Ω resistor and a 0.1-μF, 1-kV capacitor. Special z-lead transient suppressor devices can also be used, as was mentioned earlier.

FIGURE 11.16

11.8 Switching Regulator Supplies (Switchers)

A switching power supply, or *switcher*, is a unique kind of power supply that can achieve power conversion efficiencies far exceeding those of the linear supplies covered earlier in this chapter. With linear regulated supplies, the regulator converts the input voltage that is higher than needed into a desired lower output voltage. To lower the voltage, the extra energy is dissipated as heat from the regulator's control element. The power-conversion efficiency (P_{out}/P_{in}) for these supplies is typically lower than 50 percent. This means that more than half the power is dissipated as heat.

Switchers, on the other hand, can achieve power-conversion efficiencies exceeding 85 percent, meaning that they are much more energy efficient than linearly regulated supplies. Switchers also have a wide current and voltage operating range and can be configured in either step-down (output voltage smaller than input voltage), step-up (output voltage larger than input voltage), or inverting (output is the opposite polarity of the input) configurations. Also, switchers can be designed to run directly off ac line power, without the need for a power transformer. By eliminating the hefty power transformer, the switcher can be made light and small. This makes switchers good supplies for computers and other small devices.

A switching supply resembles a linear supply in many ways. However, two unique features include an energy-storage inductor and a nonlinear regulator network. Unlike a linear supply, which provides regulation by varying the resistance of the regulator's control element, a switcher incorporates a regulation system in which the control element is switched on and off very rapidly. The on/off pulses are controlled by an oscillator/error amplifier/pulse-width modulator network (see Fig. 11.17).

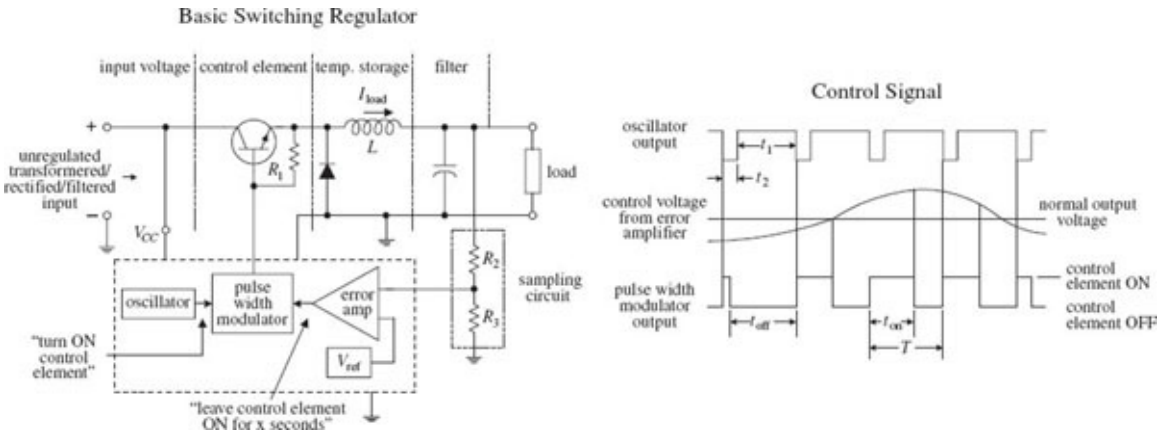


FIGURE 11.17

During the on cycle, energy is pumped into the inductor (energy is stored in the magnetic fields about the inductor's coil). When the control element is turned off, the stored energy in the inductor is directed by the diode into the filter and into the load. The sampling circuit (R_2 and R_3) takes a sample of the output voltage and feeds the sample to one of the inputs of the error amplifier. The error amplifier then compares the sample voltage with a reference voltage applied to its other input. If the sample voltage is below the reference voltage, the error amplifier increases its output control voltage. This control voltage is then sent to the pulse-width modulator. (If the sample voltage is above the reference voltage, the error amplifier will decrease the output voltage it sends to the modulator.) While this is going on, the oscillator is supplying a steady series of triggering voltage pulses to the pulse-width modulator. The modulator uses both the oscillator's pulses and the error amplifier's output to produce a modified on/off signal that is sent to the control element's base. The modified signal represents a squarewave whose on time is determined by the input error voltage. If the error voltage is low (meaning the sample voltage is higher than it should

be), the modulator sends a short-duration on pulse to the control element. However, if the error voltage is high (meaning the sample voltage is lower than it should be), the pulse-width modulator sends a long-duration on pulse to the control element. (The graph in Fig. 11.17 shows how the oscillator, error amplifier, and pulse-width modulator outputs are related.) Using a series of on/off pulses that can be varied in frequency and duration gives the switching regulator its exceptional efficiency; releasing a series of short pulses of energy over time is more efficient than taking excessive supply energy and radiating it off as heat (linear supply).

Figure 11.18 shows a typical switching regulator arrangement. The 556 dual-timer IC houses both the oscillator and pulse-width modulator, while the UA723 voltage-regulator IC acts as the error amplifier. R_2 and R_3 comprise the sampling network, R_6 and R_7 set the reference voltage, and R_4 and R_5 set the final control voltage that is sent to the pulse-width modulator.

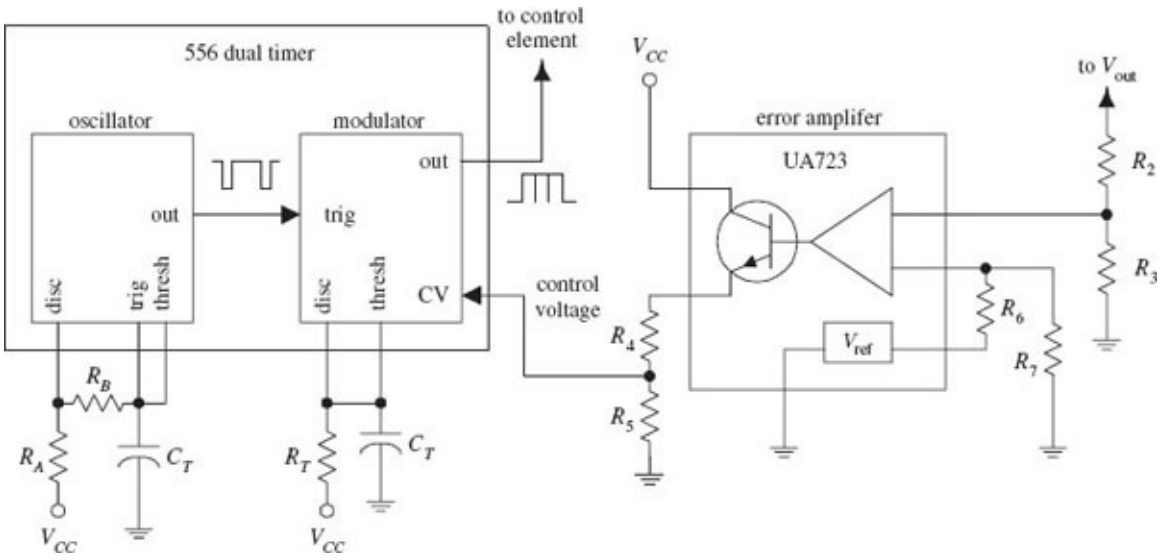
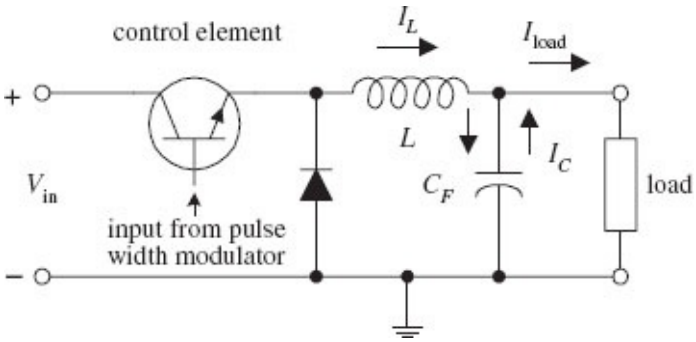


FIGURE 11.18

Step-Up, Step-Down, and Inverting Configurations

The switching regulator in Fig. 11.17 is referred to as a *step-down regulator*. It is used when the regulated output voltage is to be lower than the regulator input voltage. Now, switching regulators also come in step-up and inverting configurations. The step-up version is used when the output is to be higher than the input, whereas the inverting version is used when the output voltage is to be the opposite polarity of the input voltage. Here's an overview of the three configurations.

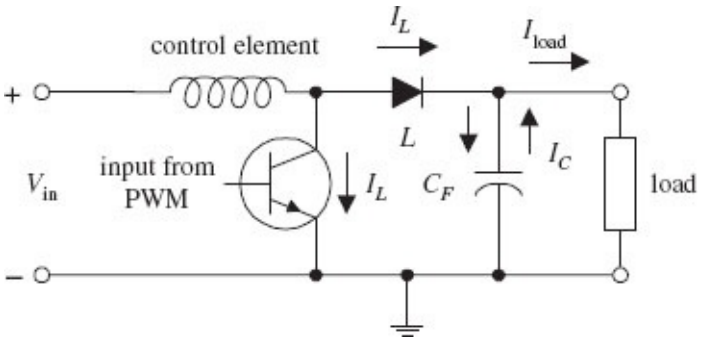
STEP-DOWN REGULATOR



This is used when output voltage is to be lower than input voltage. When the control element is on, *L* stores energy, helps supply load current, and supplies current to the filter capacitor. When the control element is off, the energy stored in *L* helps supply load current but again restores charge on C_F—the charge on C_F is used to supply the load when the control element turns off and *L* has discharged its energy.

FIGURE 11.19

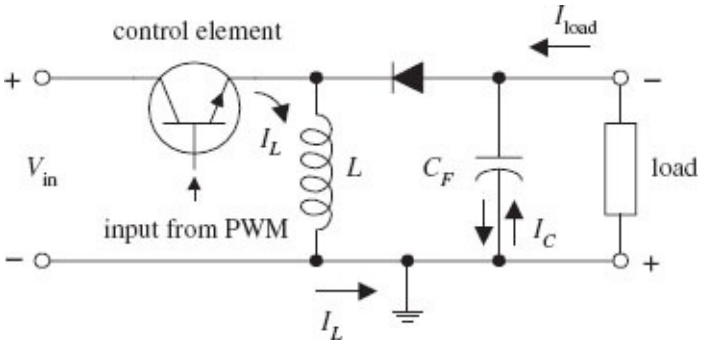
STEP-UP SWITCHER REGULATOR



This is used when output is to be higher than input voltage. When the control element is on, energy is stored in the inductor. The load, isolated by the diode, is supplied by the charge stored in C_F. When the control element is off, the energy stored in L is added to the input voltage. At the same time, L supplies load current, as well as restoring the charge on C_F—the charge on C_F is used to supply the load current when the control element is off and when the energy in L is discharged.

FIGURE 11.20

INVERTING SWITCHER REGULATOR



This is used when output voltage is to be the opposite polarity of the input voltage. When the control element is on, energy is stored in L, while the diode isolates L from load. The load current is supplied by the charge on C_F. When the control element is turned off, the energy stored in L charges C_F to a polarity such that V_{out} is negative. I_L supplies load current and restores the charge on C_F while it is discharging its energy. C_F supplies load current when the control element is off and the inductor is discharged. An inverting switcher regulator can be designed to step up or step down the inverted output.

FIGURE 11.21

11.9 Switch-Mode Power Supplies

The logical extension of a switching regulator is to ditch the transformer entirely and work directly from rectified line voltage.

Transformer-based designs have become a rare occurrence in consumer electronics. Transformers are relatively expensive and heavy components. Watt for watt, a switch-mode power supply is generally cheaper to produce.

By using the unique switching action of the switcher, it is possible to design a supply that does not require the hefty 60-Hz power transformer at the input stage. In other words, you can design a switching power supply to run directly off a 120-V ac line—you still must rectify and filter the line voltage before feeding it to the regulator. However, if you remove the power transformer, you remove the protective isolation that is present between the 120-V ac line and the dc input to the supply. Without the isolation, the dc input voltage will be around 160 V. To avoid this potentially “shocking” situation, the switching regulator must be modified. One method for providing isolation involves replacing the energy-storage inductor with the secondary coil of a high-frequency transformer while using another high-frequency transformer or optoisolator to link the feedback from the error amplifier to the modulating element (see

Fig. 11.22).

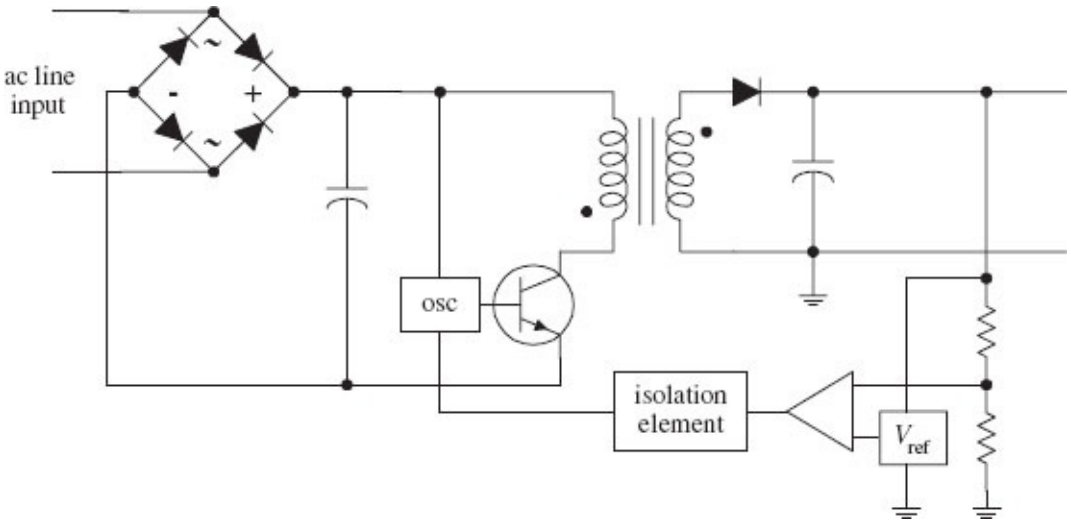


FIGURE 11.22

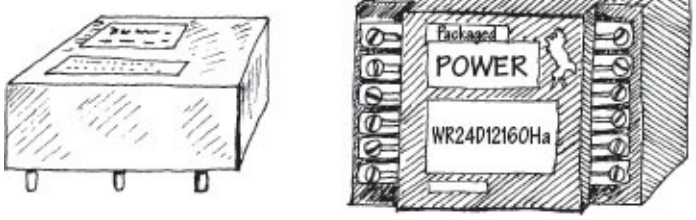
Now, you may be wondering how removing one transformer and adding another transformer (if not two) makes things smaller and lighter. Well, according to the laws of physics, as the frequency of an alternating signal increases, the need for a large iron core within the transformer decreases. You can use the high-frequency transformer(s) because the switcher's oscillator is beating so fast (e.g., 65 kHz). The difference in size and weight between a switching supply that uses high-frequency transformers and a supply that uses a 60-Hz power transformer is significant. For example, a 500-W switching supply takes up around 640 in^3 as compared with 1520 in^3 for a linear supply rated at the same power. Also, switching supplies run cooler than linear supplies. In terms of watts per cubic inch, a switching power supply can achieve 0.9 W/in^3, while a linear supply usually provides 0.4 W/in^3.

There is a slight problem with switchers that should be noted. As a result of the on/off pulsing action of the switching regulator, a switching supply's output will contain a small switching ripple voltage (typically in the tens of millivolts). Usually the ripple voltage does not pose too many problems (e.g., 200-mV noise margins for most digital ICs are not exceeded). However, if a circuit is not responding well to the ripple, an external high-current, low-pass filter can be added.

11.10 Kinds of Commercial Power Supply Packages

To make life easier, you can forget about designing your own supplies and buy one that has been made by the pros. These supplies come in either linear or switcher form and come in a variety of different packages. Here are some of the packages that are available.

Small Modular Units

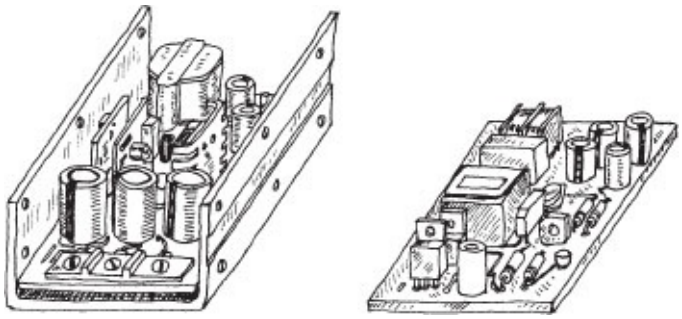


These are used in low-power applications (e.g., ±5, ±10, ±15 V). Supplies are housed in small modules, usually around 2.5 × 3.5 × 1 in. They often come with pinlike leads that can be mounted directly into circuit boards or come with terminal-strip screw connections along their sides. These supplies may come with single (e.g., +5 V), dual (e.g., ±15 V), or triple (e.g., +5 V, ±15 V) output terminals. Linear units have power ratings from around 1 to 10 W, while switching units have power ratings from around

10 to 25 W. You must supply the fuses, switches, and filters.

FIGURE 11.23

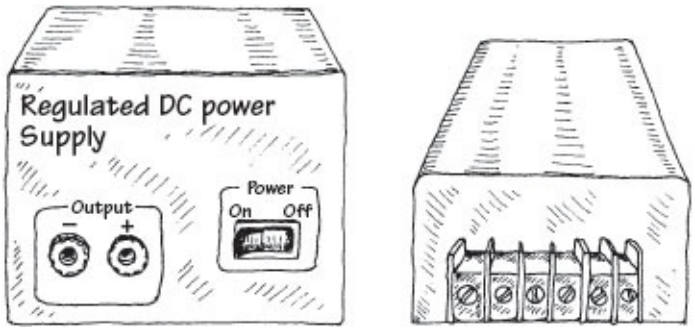
Open Frame



This supply's circuit board, transformer, etc. are mounted on a metal platform (if it is a low-voltage supply, it may simply be mounted on a circuit board) that is inserted into an instrument. These supplies come in linear and switching types and come with a wide range of voltage, current, and power ratings (around 10 to 200 W for linear supplies, 20 to 400 W for switching supplies). You will probably have to supply the fuses, switches, and filters.

FIGURE 11.24

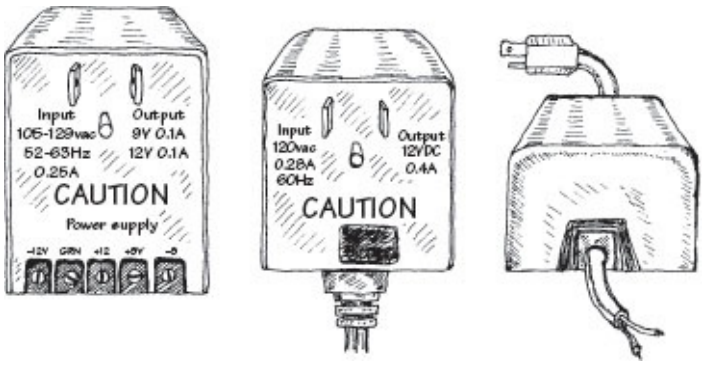
Enclosed



These supplies are enclosed in a metal box that is especially designed to efficiently radiate off excessive heat. They come in both linear and switching forms. Power rating ranges from around 10 to 800 W for linear supplies and 20 to 1500 W for switching supplies.

FIGURE 11.25

Wall Plug-In



Wall plug-in power supplies (often called "wall-warts") get inserted directly into an ac wall socket. Some of these devices only provide ac transformation, others supply an unregulated dc voltage, and others supply a regulated dc output. Typical output voltages include +3, +5, +6, +7.5, +9, +12, and +15 V. They also come in dual-polarity form.

These devices are mostly switched mode, and are manufactured in millions, as they are not specific to a product. They are small and light, and it is possible to get a unit that can supply 2 A at 12 V at a cost of less than $10.

This approach is helped by the largely standard 6.3 mm plug for the dc voltage, which, in most consumer appliances, has a positive inner connection and a negative outer sheath connection. Note that this is often the other way around in musical equipment, such as guitar effects pedals.

FIGURE 11.26

11.11 Power Supply Construction

When building a power supply, the following suggestions should help:

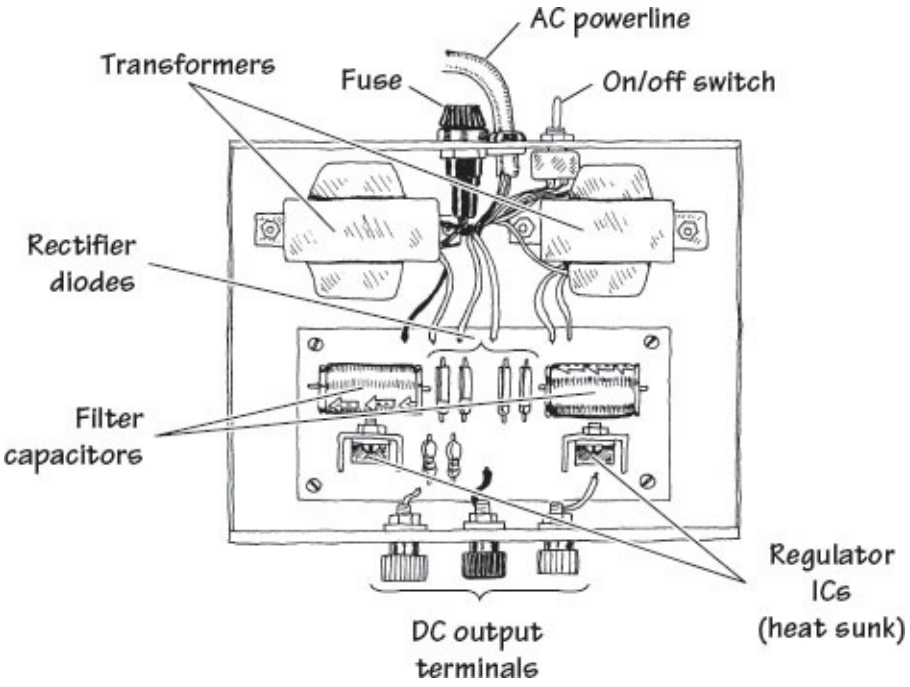


FIGURE 11.27

- Mount the transformer directly to the metal enclosure box, toward the rear.
- Install fuses, power switch, and binding posts at the rear of the box.
- Mount circuit boards on standoffs within the box.
- Place diode or rectifier modules, along with the capacitors and voltage regulators, on the circuit board.
- Make sure to heat-sink voltage regulators.
- Place supply output jacks on the front of the box.
- Drill holes in box to allow cooling.
- Ground the box.
- Place the power-line core through a hole in the rear. Use a rubber grommet for strain relief.
- To avoid shocks, make sure to insulate all exposed 120-V power connections inside the box with heat-shrink tubing.

CHAPTER 12

Digital Electronics

Before we begin, I'll warn you that there is a lot of information in this chapter, and it may be difficult to absorb all this at once. Some information is present largely for historical interest and to provide a better understanding of how complex digital systems such as microcontrollers work. My advice is to skim to your heart's content, and pull out whatever information you find practical. The basic principles are still the same, but if you find that your design uses more than three ICs, you probably could be using a microcontroller (the subject of Chapter 13).

12.1 The Basics of Digital Electronics

Until now, we have mainly covered the analog realm of electronics—circuits that accept and respond to voltages that vary continuously over a given range. Such analog circuits included rectifiers, filters, amplifiers, simple RC timers, oscillators, simple transistor switches, and so on. Although each of these analog circuits is fundamentally important in its own right, these circuits lack an important feature: they cannot store and process bits of information needed to make complex logical decisions. To incorporate logical decision-making processes into a circuit, you need to use digital electronics.

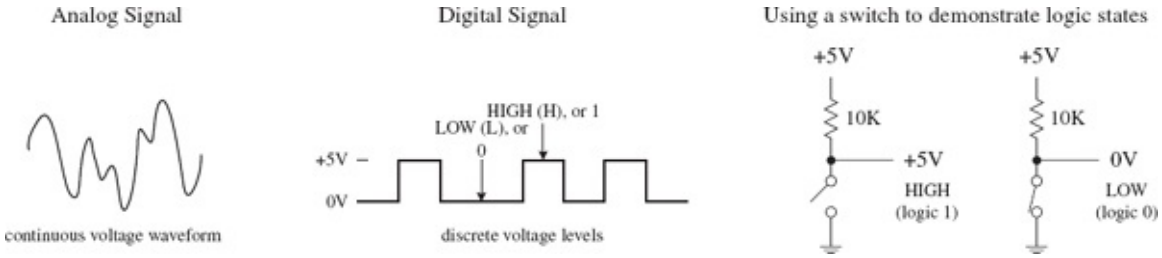


FIGURE 12.1

12.1.1 Digital Logic States

In digital electronics, there are only two voltage states present at any point within a circuit. These voltage states are either *high* or *low*. The voltage being high or low at a particular location within a circuit can signify a number of things. For example, it may represent the on or off state of a switch or saturated transistor, one bit of a number, whether an event has occurred, or whether some action should be taken.

The high and low states can be represented as true and false statements, which are used in Boolean logic. In most cases, high equals true and low equals false. However, this does not need to be the case—you could make high equal to false and low equal to true. The decision to use one convention over the other is a matter left ultimately to the designer. In digital lingo, to avoid people getting confused over which convention is in use, the term *positive true logic* is used when high equals true, while the term *negative true logic* is used when high equals false.

In Boolean logic, the symbols 1 and 0 are used to represent true and false, respectively. Now, unfortunately, 1 and 0 are also used in electronics to represent high and low voltage states, where high equals 1 and low equals 0. As you can see, things can get a bit confusing, especially if you are not sure

which type of logic convention is being used: positive true or negative true logic. In Section 12.3, you will see some examples that deal with this confusing issue.

The exact voltages assigned to high or low voltage states depend on the specific logic IC that is used (as it turns out, digital components are IC-based). As a general rule of thumb, +5 *V* is considered high, while 0 *V* (ground) is considered low. However, as you will see in Section 12.4, this does not need to be the case. For example, some logic ICs may interpret a voltage from +2.4 to +5 *V* as high and a voltage from +0.8 to 0 *V* as low. Other ICs may use an entirely different range.

12.1.2 Number Codes Used in Digital Electronics

Binary

Because digital circuits work with only two voltage states, it is logical to use the binary number system to keep track of information. A binary number is composed of two binary digits, 0 and 1, which are also called *bits* (for example, 0 = low voltage and 1 = high voltage). By contrast, a decimal number such as 736 is represented by successive powers of 10:

$$736_{10} = 7 \times 10^2 + 3 \times 10^1 + 6 \times 10^0$$

Similarly, a binary number such as 11100 (28_{10}) can be expressed as successive powers of 2:

$$11100_2 = 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0$$

The subscript tells which number system is in use (X_{10} = decimal number and X_2 = binary number). The highest-order bit (leftmost bit) is called the *most significant bit* (MSB), while the lowest-order bit (rightmost bit) is called the *least significant bit* (LSB). Methods used to convert from decimal to binary and vice versa are shown in Fig. 12.2.

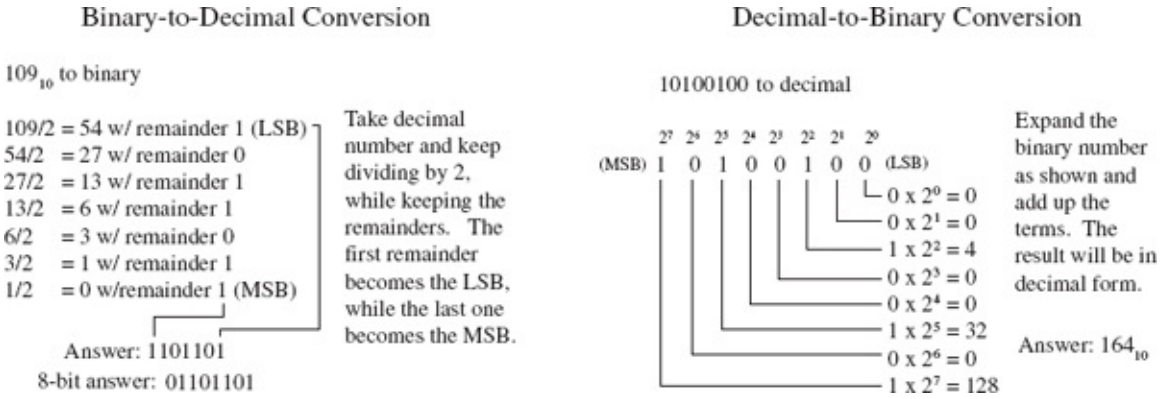


FIGURE 12.2

It should be noted that most digital systems deal with 4, 8, 16, or 32 bits at a time. The decimal-to-binary conversion example given here has a 7-bit answer. In an 8-bit system, you would need to put an additional 0 in front of the MSB (for example, 01101101). In a 16-bit system, nine additional 0s would need to be added (for example, 0000000001101101).

As a practical note, the easiest way to convert a number from one base to another is to use a calculator. For example, to convert a decimal number into a binary number, type in the decimal number (in base 10 mode) and then change to binary mode (which usually entails a second function key). The number will now be in binary (1s and 0s). To convert a binary number to a decimal number, start out in binary mode, type in the number, and then switch to decimal mode.

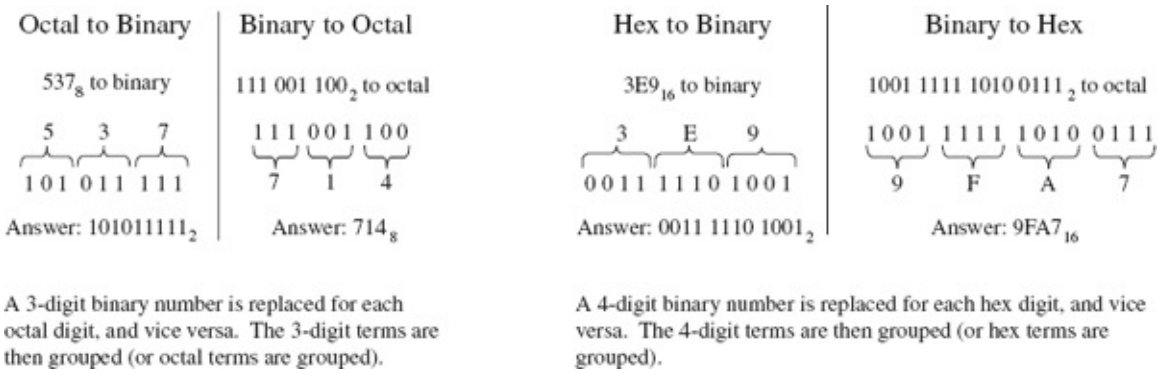
Octal and Hexadecimal

Two other number systems used in digital electronics include the octal and hexadecimal systems. In the octal system (base 8), there are 8 allowable digits: 0, 1, 2, 3, 4, 5, 6, and 7. In the hexadecimal system (base 16), there are 16 allowable digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, and F. Here are examples of octal and hexadecimal numbers with decimal equivalents:

$$247_8 \text{ (octal)} = 2 \times 8_2 + 4 \times 8^1 + 7 \times 8^0 = 167_{10} \text{ (decimal)}$$
$$2D5_{16} \text{ (hex)} = 2 \times 16^2 + D\ (=13_{10}) \times 16^1 + 9 \times 16^0 = 725_{10} \text{ (decimal)}$$

Of course, binary numbers are the natural choice for digital systems, but since these binary numbers can become long and difficult to interpret by our decimal-based brains (a result of our ten fingers), it is common to write them out in hexadecimal or octal form.

Unlike decimal numbers, octal and hexadecimal numbers can be translated easily to and from binary. This is because a binary number, no matter how long, can be broken up into 3-bit groupings (for octal) or 4-bit groupings (for hexadecimal). You simply add zero to the beginning of the binary number if the total numbers of bits is not divisible by 3 or 4. Figure 12.3 should paint the picture better than words.



A 3-digit binary number is replaced for each octal digit, and vice versa. The 3-digit terms are then grouped (or octal terms are grouped).

A 4-digit binary number is replaced for each hex digit, and vice versa. The 4-digit terms are then grouped (or hex terms are grouped).

FIGURE 12.3

Today, the hexadecimal system has essentially replaced the octal system. The octal system was popular at one time, when microprocessor systems used 12-bit and 36-bit words, along with a 6-bit alphanumeric code, which are all divisible by 3-bit units (1 octal digit). Today, microprocessor systems mainly work with 8-bit, 16-bit, 20-bit, 32-bit, or 64-bit words, which are all divisible by 4-bit units (1 hex digit). In other words, an 8-bit word can be broken down into 2 hex digits, a 16-bit word into 4 hex digits, a 20-bit word into 5 hex digits, and so on.

Hexadecimal representation of binary numbers pops up in many memory and microprocessor applications that use programming codes (for example, within assembly language) to address memory locations and initiate other specialized tasks that would otherwise require typing in long binary numbers. For example, a 20-bit address code used to identify one of a million memory locations can be replaced with a hexadecimal code (in the assembly program) that reduces the count to five hex digits. Note that a compiler program later converts the hex numbers within the assembly language program into binary numbers (machine code), which the microprocessor can use. Table 12.1 shows a conversion table.

TABLE 12.1 Decimal, Binary, Octal, Hex, BCD Conversion Table

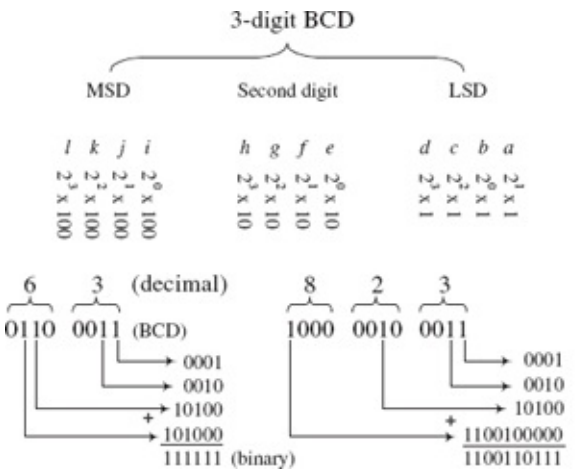
DECIMAL	BINARY	OCTAL	HEXADECIMAL	BCD
00	0000 0000	00	00	0000 0000
01	0000 0001	01	01	0000 0001
02	0000 0010	02	02	0000 0010
03	0000 0011	03	03	0000 0011
04	0000 0100	04	04	0000 0100
05	0000 0101	05	05	0000 0101
06	0000 0110	06	06	0000 0110
07	0000 0111	07	07	0000 0111
08	0000 1000	10	08	0000 1000
09	0000 1001	11	09	0000 1001
10	0000 1010	12	0A	0001 0000
11	0000 1011	13	0B	0001 0001
12	0000 1100	14	0C	0001 0010
13	0000 1101	15	0D	0001 0011
14	0000 1110	16	0E	0001 0100
15	0000 1111	17	0F	0001 0101
16	0001 0000	20	10	0001 0110
17	0001 0001	21	11	0001 0111
18	0001 0010	22	12	0001 1000
19	0001 0011	23	13	0001 1001
20	0001 0100	24	14	0010 0000

Binary-Coded Decimal

Binary-coded decimal (BCD) is used to represent each digit of a decimal number as a 4-bit binary number. For example, the number 150_{10} in BCD is expressed as follows:

```
   1    5    0          150₁₀ = 0001 0101 000(BCD)
  /     |     \
0001   0101   000
```

To convert from BCD to binary is vastly more difficult, as shown in Fig. 12.4. Of course, you could cheat by converting the BCD into decimal first and then convert to binary, but that does not show you the mechanics of how machines do things with 1s and 0s. You will rarely need to do BCD-to-binary conversion, so I will not dwell on this topic.



bit position	Weighting factor: decimal	Weighting factor: binary
a	1	1
b	2	10
c	4	100
d	8	1000
e	10	1010
f	20	10100
g	40	101000
h	80	1010000
i	100	1100100
j	200	11001000
k	400	110010000
l	800	1100100000

FIGURE 12.4

BCD is commonly used when outputting to decimal (0–9) displays, such as those found in digital clocks and multimeters. BCD will be discussed in Section 12.3.

Sign-Magnitude and 2's Complement Numbers

Up to now, we have not considered negative binary numbers. How do you represent them? A simple method is to use *sign-magnitude representation*. In this method, you simply reserve a bit, usually the MSB, to act as a sign bit. If the sign bit is 0, the number is positive; if the sign bit is 1, the number is negative (see Fig. 12.5).

Decimal to 2's complement

$+4_{10}$ to 2's complement

true binary = 0010 1001
2's comp = 0010 1001

If the decimal number is positive, the 2's complement number is equal to the true binary equivalent of the decimal number.

If the decimal number is negative, the 2's complement number is found by:
1) Complementing each bit of the true binary equivalent of the decimal (making 1's into 0's and vice versa). This is is called taking the 1's complement.
2) Adding 1 to the 1's complement number to get the magnitude bits. The sign bit will always end up being 1.

-41_{10} to 2's complement

true binary = 0010 1001
1's comp = 1101 0110
Add 1 = +1
2's comp = 1101 0111

2's complement to decimal

1100 1101 (2's comp) to decimal

2's comp = 1100 1101
Complement = 0011 0010
Add 1 = +1
True binary = 0011 0011
Decimal eq. = -51_{10}

If the 2's complement number is positive (sign bit = 0), perform a regular binary-to-decimal conversion.

If the 2's complement number is negative (sign bit = 1), the decimal sign will be negative. The decimal is found by:
1) Complementing each bit of the 2's complement number.
2) Adding 1 to get the true binary equivalent.
3) Performing a true binary-to-decimal conversion, and including negative sign.

FIGURE 12.5

Although the sign-magnitude representation is simple, it is seldom used, because adding requires a different procedure than subtracting (as you will see in the next section). Occasionally, you will see sign-magnitude numbers used in display and analog-to-digital applications, but you will hardly ever see them in circuits that perform arithmetic.

A more popular choice when dealing with negative numbers is to use *2's complement representation*. In 2's complement, the positive numbers are exactly the same as unsigned binary numbers. A negative number, however, is represented by a binary number, which when added to its corresponding positive equivalent results in zero. In this way, you can avoid two separate procedures for doing addition and subtraction. You will see how this works in the next section. A simple procedure outlining how to convert a decimal number into a binary number and then into a 2's complement number, and vice versa, is outlined in Fig. 12.5.

Decimal, Sign-Magnitude, 2's Complement Conversion Table

DECIMAL	SIGN-MAGNITUDE	2'S COMPLEMENT
+7	0000 0111	0000 0111
+6	0000 0110	0000 0110
+5	0000 0101	0000 0101
+4	0000 0100	0000 0100
+3	0000 0011	0000 0011
+2	0000 0010	0000 0010
+1	0000 0001	0000 0001
0	0000 0000	0000 0000
–1	1000 0001	1111 1111
–2	1000 0010	1111 1110
–3	1000 0011	1111 1101
–4	1000 0100	1111 1100
–5	1000 0101	1111 1011
–6	1000 0110	1111 1010
–7	1000 0111	1111 1001
–8	1000 1000	1111 1000

Arithmetic with Binary Numbers

Adding, subtracting, multiplying, and dividing binary numbers, hexadecimal numbers, and other representations can be done with a calculator set to that particular base mode. But that's cheating, and it doesn't help you understand the mechanics of how it is done. The mechanics become important when designing the actual arithmetical circuits. Here are the basic techniques used to add and subtract binary numbers.

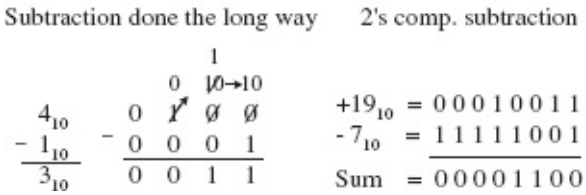
ADDING

$$+\ \begin{array}{l} 5_{10} = 0\ 1\ 0\ 1 \\ 3_{10} = 0\ 0\ 1\ 1 \\ \hline 1\ 0\ 0\ 0 \end{array} \qquad + \begin{array}{l} 20_{10} = 0\,0\,0\,1\,0\,1\,0\,0 \\ 87_{10} = 0\,1\,0\,1\,0\,1\,1\,1 \\ \hline 0\,1\,1\,0\,1\,0\,1\,1 \end{array}$$

Adding binary numbers is just like adding decimal numbers. Whenever the result of adding one column of numbers is greater than one digit, a 1 is carried over to the next column to be added.

FIGURE 12.6

SUBTRACTING



Subtracting binary numbers is not as easy as it looks. It is similar to decimal subtraction but can be confusing. For example, you might think that if you were to subtract a 1 from a 0, you would borrow a 1 from the column to the left. No! You must borrow a 10 (2_{10}). It becomes a headache if you try to do this by hand. The trick to subtracting binary numbers is to use the 2's complement representation that provides the sign bit, and then just add the positive number with the negative number to get the sum. This method is often used by digital circuits because it allows both addition and subtraction, without the headache of needing to subtract the smaller number from the larger number.

FIGURE 12.7

ASCII

American Standard Code for Information Interchange (ASCII) is an alphanumeric code used to transmit letters, symbols, numbers, and special nonprinting characters between computers and computer peripherals (such as printers and keyboards). ASCII consists of 128 different 7-bit codes.

Codes from 000 0000 (or hex 00) to 001 1111 (or hex 1F) are reserved for nonprinting characters or special machine commands like ESC (escape), DEL (delete), CR (carriage return), and LF (line feed). Codes from 010 0000 (or hex 20) to 111 1111 (or hex 7F) are reserved for printing characters like a, A, #, &, {, @, and 3. Tables 12.2 and 12.3 show the ASCII nonprinting and printing characters.

In practice, when ASCII code is sent, an additional bit is added to make it compatible with 8-bit systems. This bit may be set to 0 and ignored, it may be used as a parity bit for error detection (Section 12.3.8 covers parity bits), or it may act as a special function bit used to implement an additional set of specialized characters.

TABLE 12.2 ASCII Nonprinting Characters

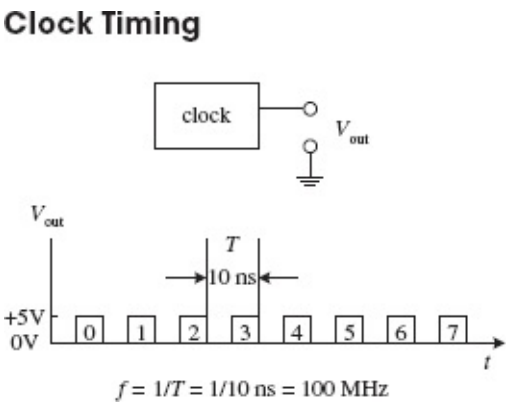
DEC	HEX	7-BIT CODE	CONTROL CHAR	CHAR	MEANING	DEC	HEX	7-BIT	CONTROL CHAR	CHAR	MEANING
00	00	000 0000	CTRL-@	NUL	Null	16	10	001 0000	CTRL-P	DLE	Data line escape
01	01	000 0001	CTRL-A	SOH	Start of heading	17	11	001 0001	CTRL-Q	DC1	Device control 1
02	02	000 0010	CTRL-B	STX	Start of text	18	12	001 0010	CTRL-R	DC2	Device control 2
03	03	000 0011	CTRL-C	ETX	End of text	19	13	001 0011	CTRL-S	DC3	Device control 3
04	04	000 0100	CTRL-D	EOT	End of transmit	20	14	001 0100	CTRL-T	DC4	Device control 4
05	05	000 0101	CTRL-E	ENQ	Enquiry	21	15	001 0101	CTRL-U	NAK	Neg acknowledge
06	06	000 0110	CTRL-F	ACK	Acknowledge	22	16	001 0110	CTRL-V	SYN	Synchronous idle
07	07	000 0111	CTRL-G	BEL	Bell	23	17	001 0111	CTRL-W	ETB	End of transmit block
08	08	000 1000	CTRL-H	BS	Backspace	24	18	001 1000	CTRL-X	CAN	Cancel
09	09	000 1001	CTRL-I	HT	Horizontal tab	25	19	001 1001	CTRL-Y	EM	End of medium
10	0A	000 1010	CTRL-J	LF	Line feed	26	1A	001 1010	CTRL-Z	SUB	Substitute
11	0B	000 1011	CTRL-K	VT	Vertical tab	27	1B	001 1011	CTRL-[	ESC	Escape
12	0C	000 1100	CTRL-L	FF	Form feed	28	1C	001 1100	CTRL-\	FS	File separator
13	0D	000 1101	CTRL-M	CR	Carriage return	29	1D	001 1101	CTRL-]	GS	Group separator
14	0E	000 1110	CTRL-N	S0	Shift out	30	1E	001 1110	CTRL-^	RS	Record separator
15	0F	000 1111	CTRL-O	SI	Shift in	31	1F	001 1111	CTRL-_	US	Unit separator

TABLE 12.3 ASCII Printing Characters

DEC	HEX	7-BIT CODE	CHAR	DEC	HEX	7-BIT	CHAR	DEC	HEX	7-BIT CODE	CHAR
32	20	010 0000	SP	64	40	100 0000	@	96	60	110 0000	'
33	21	010 0001	!	65	41	100 0001	A	97	61	110 0001	a
34	22	010 0010	"	66	42	100 0010	B	98	62	110 0010	b
35	23	010 0011	#	67	43	100 0011	C	99	63	110 0011	c
36	24	010 0100	$	68	44	100 0100	D	100	64	110 0100	d
37	25	010 0101	%	69	45	100 0101	E	101	65	110 0101	e
38	26	010 0110	&	70	46	100 0110	F	102	66	110 0110	f
39	27	010 0111	`	71	47	100 0111	G	103	67	110 0111	g
40	28	010 1000	(	72	48	100 1000	H	104	68	110 1000	h
41	29	010 1001	)	73	49	100 1001	I	105	69	110 1001	i
42	2A	010 1010	*	74	4A	100 1010	J	106	6A	110 1010	j
43	2B	010 1011	+	75	4B	100 1011	K	107	6B	110 1011	k
44	2C	010 1100	,	76	4C	100 1100	L	108	6C	110 1100	l
45	2D	010 1101	-	77	4D	100 1101	M	109	6D	110 1101	m
46	2E	010 1110	.	78	4E	100 1110	N	110	6E	110 1110	n
47	2F	010 1111	/	79	4F	100 1111	O	111	6F	110 1111	o
48	30	011 0000	0	80	50	101 0000	P	112	70	111 0000	p
49	31	011 0001	1	81	51	101 0001	Q	113	71	111 0001	q
50	32	011 0010	2	82	52	101 0010	R	114	72	111 0010	r
51	33	011 0011	3	83	53	101 0011	S	115	73	111 0011	s
52	34	011 0100	4	84	54	101 0100	T	116	74	111 0100	t
53	35	011 0101	5	85	55	101 0101	U	117	75	111 0101	u
54	36	011 0110	6	86	56	101 0110	V	118	76	111 0110	v
55	37	011 0111	7	87	57	101 0111	W	119	77	111 0111	w
56	38	011 1000	8	88	58	101 1000	X	120	78	111 1000	x
57	39	011 1001	9	89	59	101 1001	Y	121	79	111 1001	y
58	3A	011 1010	:	90	5A	101 1010	Z	122	7A	111 1010	z
59	3B	011 1011	;	91	5B	101 1011	[	123	7B	111 1011	{
60	3C	011 1100	<	92	5C	101 1100	\	124	7C	111 1100	|
61	3D	011 1101	=	93	5D	101 1101	]	125	7D	111 1101	}
62	3E	011 1110	>	94	5E	101 1110	^	126	7E	111 1110	~
63	3F	011 1111	?	95	5F	101 1111	_	127	7F	111 1111	DEL

12.1.3 Clock Timing and Parallel Versus Serial Transmission

Before moving on to the next section, let’s take a brief look at three important items: clock timing, parallel transmission, and serial transmission.



Most digital circuits require precise timing to function properly. Usually, a clock circuit that generates a series of high and low pulses at a fixed frequency is used as a reference on which to base all critical actions executed within a system. The clock is also used to push bits of data through the digital circuitry. The period of a clock pulse is related to its frequency by $T = 1/f$. So, if T = 10 ns, then f = 1/(10 ns) = 100 MHz.

FIGURE 12.8

Serial Versus Parallel Representation

Binary information can be transmitted from one location to another in either a serial or parallel manner. The serial format uses a single electrical conductor (and a common ground) for data transfer. Each bit from the binary number occupies a separate clock period, with the change from one bit to another occurring at each falling or leading clock edge; the type of edge depends on the circuitry used.

Figure 12.9 shows an 8-bit (10110010) word that is transmitted from circuit *A* to circuit *B* in 8 clock pulses (0–7). In computer systems, serial communications are used to transfer data between keyboard and computer, as well as to transfer data between two computers via a telephone line.

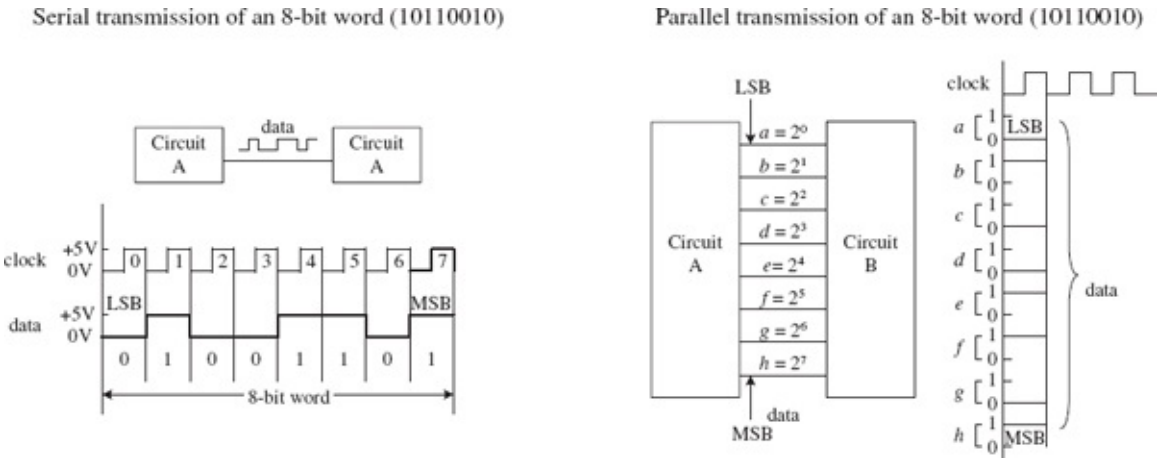


FIGURE 12.9

Parallel transmission uses separate electrical conductors for each bit (and a common ground). In Fig. 12.9, an 8-bit string (01110110) is sent from circuit A to circuit B. As you can see, unlike serial transmission, the entire word is transmitted in only one clock cycle, not eight clock cycles. In other words, it is eight times faster. Parallel communications are most frequently found within microprocessor systems that use multiline data and control buses to transmit data and control instructions from the microprocessor to other microprocessor-based devices (such as memory and output registers).

12.2 Logic Gates

Logic gates are the building blocks of digital electronics. The fundamental logic gates include the INVERT (NOT), AND, NAND, OR, NOR, exclusive OR (XOR), and exclusive NOR (XNOR) gates.

Each of these gates performs a different logical operation. Figure 12.10 provides a description of what each logic gate does and gives a switch and transistor analogy for each gate.

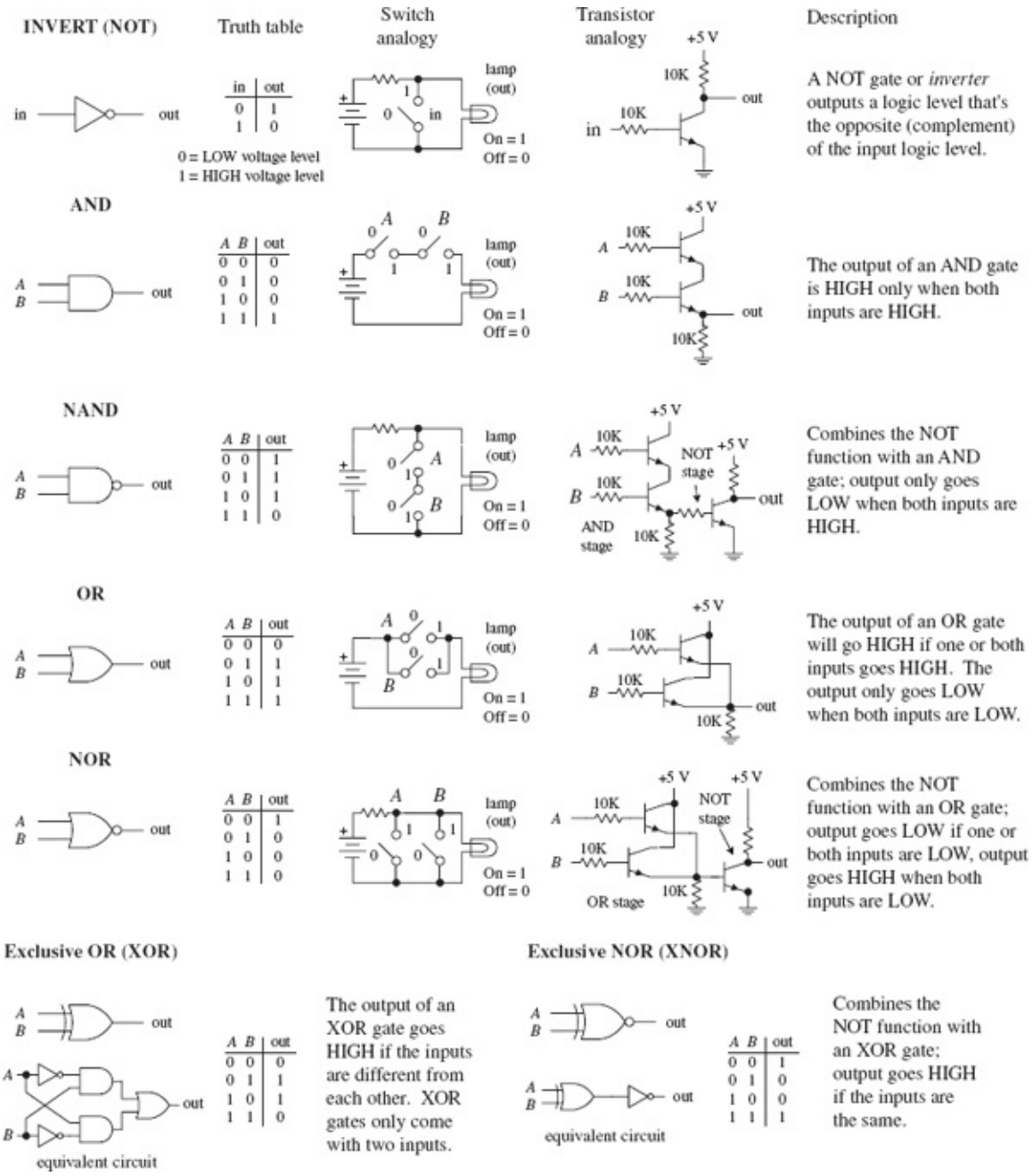


FIGURE 12.10

12.2.1 Multiple-Input Logic Gates

AND, NAND, OR, and NOR gates often come with more than two inputs (this is not the case with XOR and XNOR gates, which require two inputs only). Figure 12.11 shows a four-input AND, an eight-input AND, a three-input OR, and an eight-input OR gate. With the eight-input AND gate, all inputs must be high for the output to be high. With the eight-input OR gate, at least one of the inputs must be high for the output to go high.

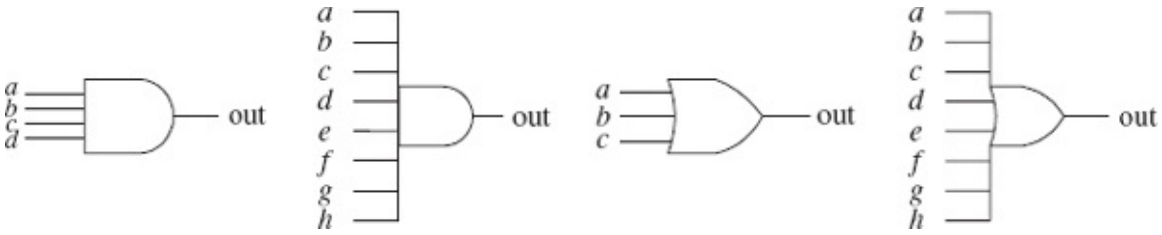


FIGURE 12.11

12.2.2 Digital Logic Gate ICs

The construction of digital gates is best left to the IC manufacturers. In fact, making gates from discrete components is highly impractical in regard to both overall performance (power consumption, speed, drive capacity, and so on) and overall cost and size.

As we mentioned in the introduction to this chapter, the use of individual logic ICs has almost completely been superseded by the use of microcontrollers. However, one or two logic ICs are still often used together in simple applications.

There are a number of technologies used in the fabrication of digital logic. The two most popular technologies include transistor-transistor logic (TTL) and complementary MOSFET (CMOS) logic. TTL incorporates bipolar transistors into its design, while CMOS incorporates MOSFET transistors. Both technologies perform the same basic functions, but certain characteristics (such as power consumption, speed, and output drive capacity) differ. There are many subfamilies within both TTL and CMOS. These subfamilies, as well as the various characteristics associated with each subfamily, will be discussed in greater detail in Section 12.4.

A logic IC, be it TTL or CMOS, typically houses more than one logic gate (for example, a quad two-input NAND, hex inverter, and so on). Each of the gates within the IC shares a common supply voltage that is implemented via two supply pins: a positive supply pin ($+V_{CC}$ or $+V_{DD}$) and a ground pin (GND). The vast majority of TTL and CMOS ICs are designed to run off a +5-V supply. (This does not apply for all the logic families, but I will get to that in Section 12.4.)

Generally speaking, input and output voltage levels are assumed to be 0 *V* (low) and +5 *V* (high). However, the actual input voltage required and the actual output voltage provided by the gate are not set in stone. For example, the 74*xx* TTL series will recognize a high input from 2.0 to 5 *V* and a low from 0 to 0.8 *V*, and will guarantee a high output from 2.4 to 5 *V* and a low output from 0 to 0.4 *V*. However, for the CMOS 4000B series (V_{CC} = +5 *V*), recognizable input voltages range from 3.3 to 5 *V* for high and 0 to 1.7 *V* for low. Guaranteed high and low output levels range from 4.9 to 5 *V* and 0 to 0.1 *V*, respectively. Again, I will discuss specifics later in Section 12.4. For now, let's just get acquainted with what some of these ICs look like, as shown in Figs. 12.12 and 12.13. The CMOS devices listed in the figures include 74HC*xx* and 4000(B). The TTL devices shown include the 74*xx*, 74F*xx*, and 74LS.

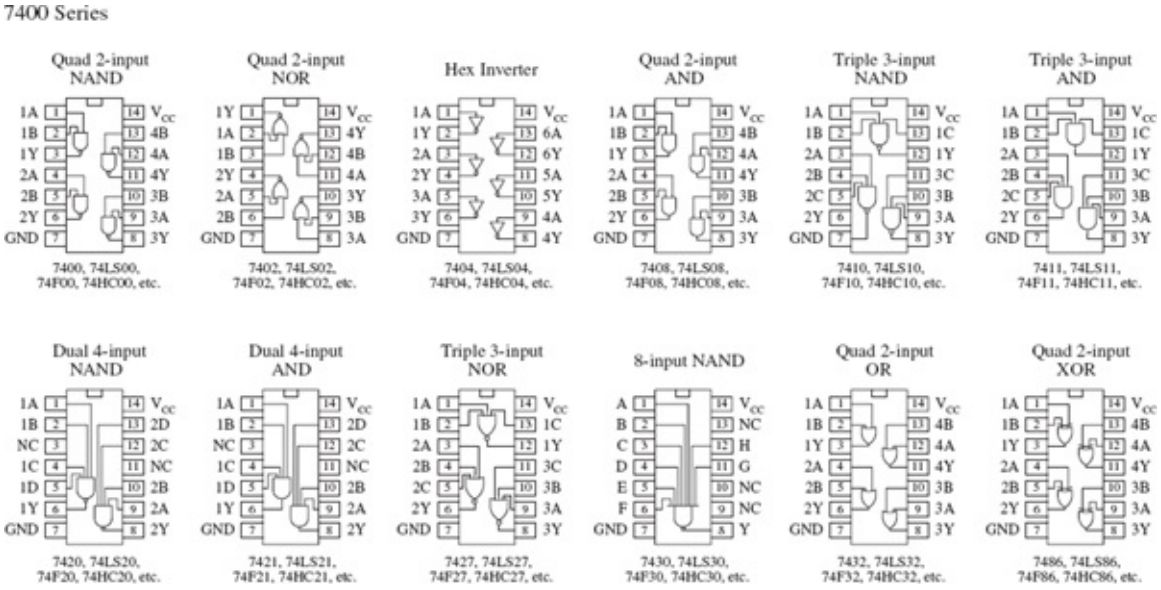


FIGURE 12.12

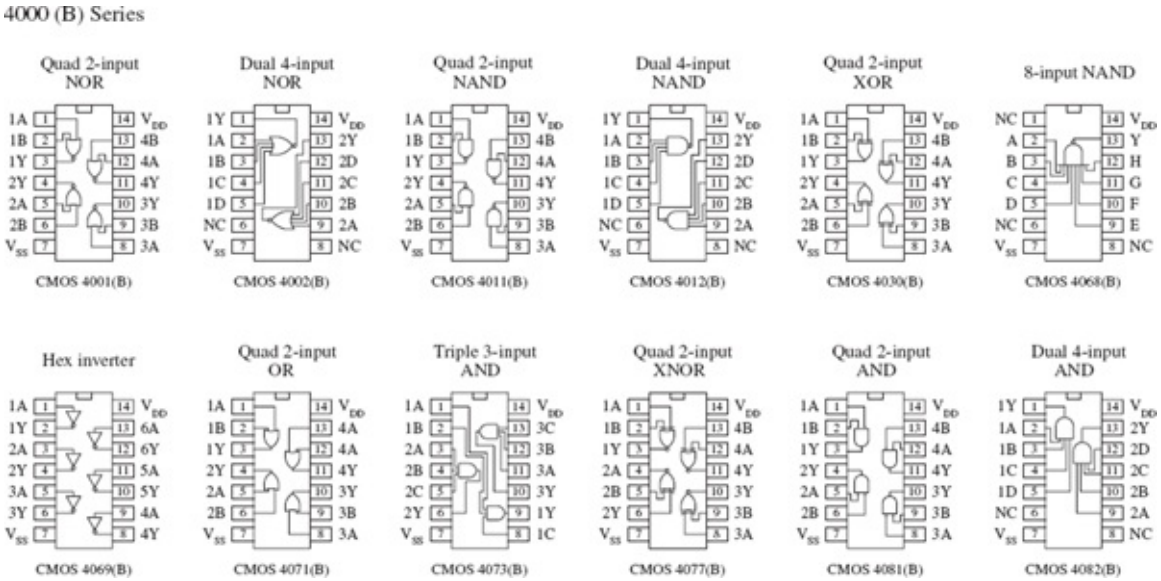


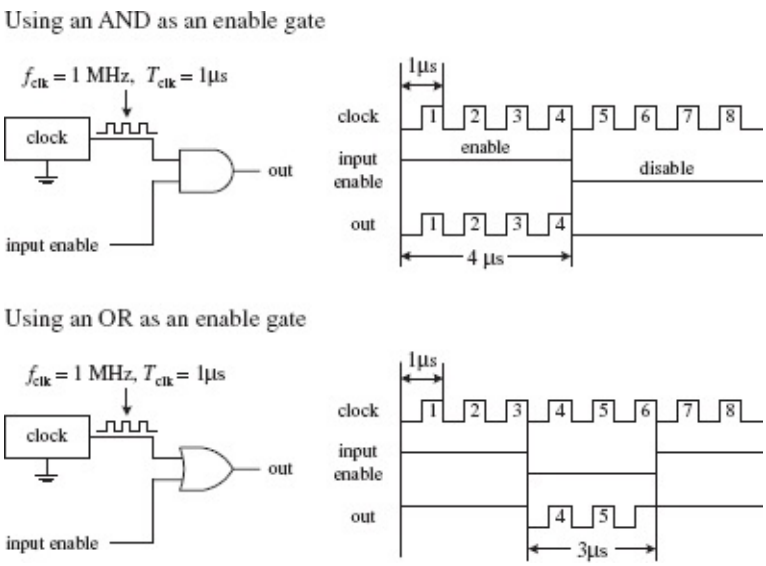
FIGURE 12.13

12.2.3 Applications for a Single Logic Gate

Before we jump into the heart of logic gate applications that involve combining logic gates to form complex decision-making circuits, let's take a look at a few simple applications that require the use of a single logic gate.

Enable/Disable Control

An enable/disable gate is a logic gate that acts to control the passage of a given waveform. The waveform —say, a clock signal—is applied to one of the gate's inputs, while the other input acts as the enable/disable control lead. Enable/disable gates are used frequently in digital systems to enable and disable control information from reaching various devices. Figure 12.14 shows two enable/disable circuits: the first uses an AND gate, and the second uses an OR gate. NAND and NOR gates are also frequently used as enable gates.



In the upper part of the figure, an AND gate acts as the enable gate. When the input enable lead is made high, the clock signal will pass to the output. In this example, the input enable is held high for 4 μs, allowing 4 clock pulses (where T_{clk} = 1 μs) to pass. When the input enable lead is low, the gate is disabled, and no clock pulses make it through to the output.

Below, an OR gate is used as the enable gate. The output is held high when the input enable lead is high, even as the clock signal is varying. However, when the enable input is low, the clock pulses are passed to the output.

FIGURE 12.14

Waveform Generation

By using the basic enable/disable function of a logic gate, as illustrated in the previous example, it is possible, with the help of a repetitive waveform generator circuit, to create specialized waveforms that can be used for the digital control of sequencing circuits.

An example waveform generator circuit is the Johnson counter. The Johnson counter will be discussed in Section 12.8. For now, let's simply focus on the outputs. In Fig. 12.15, a Johnson counter uses clock pulses to generate different output waveforms, as shown in the timing diagram. Outputs *A*, *B*, *C*, and *D* go high for 4 μs (four clock periods) and are offset from each other by 1 μs. Outputs $\overline{A}$, $\overline{B}$, $\overline{C}$, and $\overline{D}$ produce waveforms that are complements of outputs *A*, *B*, *C*, and *D*, respectively.

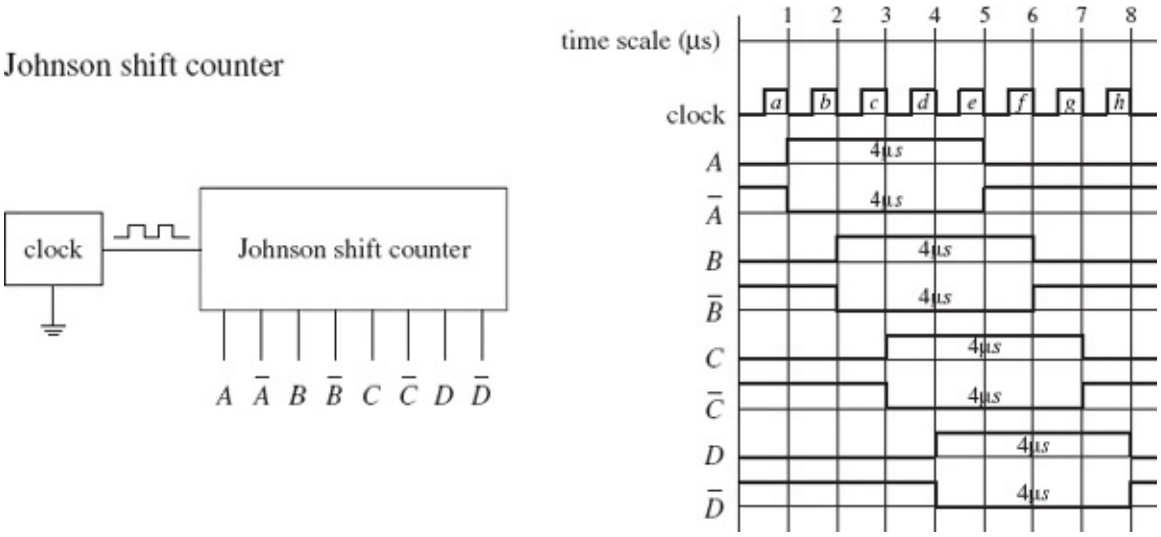


FIGURE 12.15

Now, there may be certain applications that require 4-μs high/low pulses applied at a given time, as the counter provides. However, what would you do if the application requires a 3-μs high waveform that begins at 2 μs and ends at 5 μs (relative to the time scale indicated in Fig. 12.15)? This is where the logic gates come in handy. For example, if you attach an AND gate's inputs to the counter's *A* and *B* outputs, you will get the desired 2- to 5-μs high waveform at the AND gate's output: from 1 to 2 μs the AND gate

outputs a low ($A = 1$, $B = 0$), from 2 to 5 μs the AND gate outputs a high ($A = 1$, $B = 1$), and from 5 to 6 μs the AND gate outputs a low ($A = 0$, $B = 1$). See the leftmost area of Fig. 12.16.

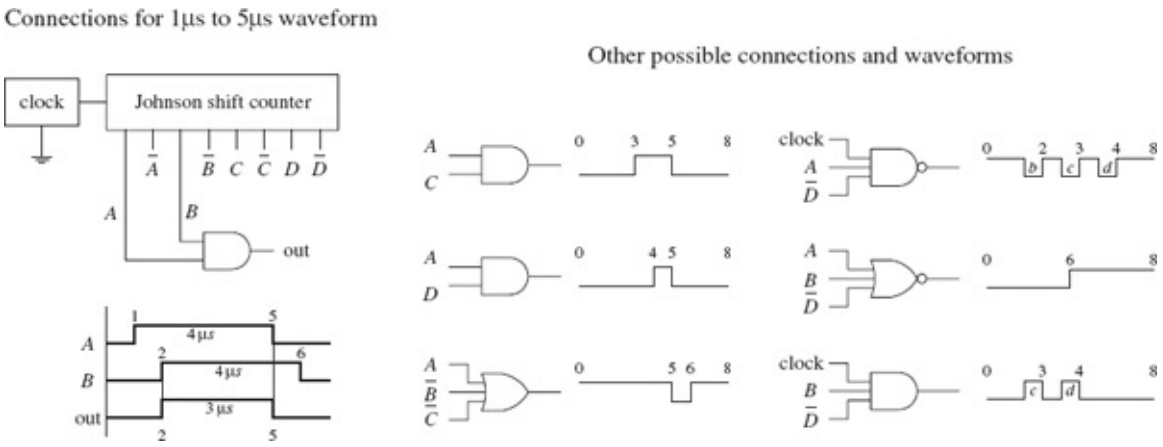


FIGURE 12.16

Various other specialized waveforms can be generated by using different logic gates and tapping different outputs of the Johnson shift counter. In Fig. 12.16, six other possibilities are shown.

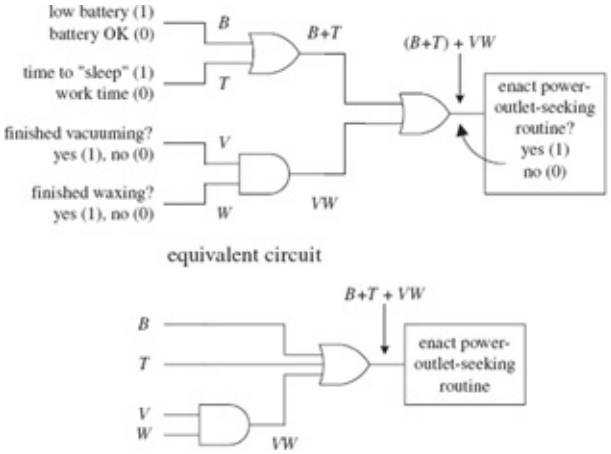
12.2.4 Combinational Logic

Combinational logic involves combining logic gates together to form circuits capable of enacting more useful, complex functions. For example, let's design the logic used to instruct a janitor-type robot to recharge itself (seek out a power outlet) only when a specific set of conditions is met. The "recharge itself" condition is specified as follows:

- When its battery is low (indicated by a high output signal from a battery-monitor circuit)
- When the workday is over (indicated by a high output signal from a timer circuit)
- When vacuuming is complete (indicated by a high voltage output from a vacuum-completion monitor circuit)
- When waxing is complete (indicated by a high output signal from a wax-completion monitor circuit).

Let's also assume that the power-outlet-seeking routine circuit is activated when a high is applied to its input.

Two simple combinational circuits that perform the desired logic function for the robot are shown in Fig. 12.17. The two circuits use a different number of gates but perform the same function. Now, the question remains, how did we come up with these circuits? In either circuit, it is not hard to predict which gates are needed. You simply exchange the word *and* present within the conditional statement with an AND gate within the logic circuit, and exchange the word *or* present within the conditional statement with an OR gate within the logic circuit. Common sense takes care of the rest.



B	T	V	W	B+T	VW	(B+T)+VW
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	0	0	0	0
0	0	1	1	0	1	1
0	1	0	0	1	0	1
0	1	0	1	1	0	1
0	1	1	0	1	0	1
0	1	1	1	1	1	1
1	0	0	0	1	0	1
1	0	0	1	1	0	1
1	0	1	0	1	0	1
1	0	1	1	1	1	1
1	1	0	0	1	0	1
1	1	0	1	1	0	1
1	1	1	1	1	1	1

FIGURE 12.17

However, when you begin designing more complex circuits, using intuition to figure out what kind of logic gates to use and how to join them together becomes exceedingly difficult. To make designing combinational circuits easier, a special symbolic language called *Boolean algebra* is used, which uses only true and false variables. A Boolean expression for the robot circuit would appear as follows:

$$E = (B + T) + VW$$

This expression amounts to saying that if B (battery-check circuit's output) *or T* (timer circuit's output) is true, or *V and W* (vacuum and waxing circuit outputs) are true, then *E* (enact power-outlet circuit input) is true.

Note that the word or is replaced by the symbol +, and the word and is simply expressed in a way similar to multiplying two variables together (placing them side by side or using a dot between variables). Also note that the term true in Boolean algebra is expressed as a 1, and false is expressed as a 0. Here, we are assuming positive logic, where true equals high voltage. Using the Boolean expression for the robot circuit, we can come up with some of the following results (the truth table in Fig. 12.17 provides all possible results):

$E = (B + T) + VW$

$E = (1 + 1) + (1 \cdot 1) = 1 + 1 = 1$	(battery is low, time to sleep, finished with chores = go recharge)
$E = (1 + 0) + (0 \cdot 0) = 1 + 0 = 1$	(battery is low = go recharge)
$E = (0 + 0) + (1 \cdot 0) = 0 + 0 = 0$	(hasn't finished waxing = don't recharge yet)
$E = (0 + 0) + (1 \cdot 1) = 0 + 1 = 1$	(has finished all chores = go recharge)
$E = (0 + 0) + (0 \cdot 0) = 0 + 0 = 0$	(hasn't finished vacuuming and waxing = don't recharge yet)

The robot example showed you how to express AND and OR functions in Boolean algebraic terms. But what about the negation operations (NOT, NAND, and NOR) and the exclusive operations (XOR and XNOR)? How do you express these in Boolean terms?

- For a NOT condition, place a line over the NOT'ed variable or variables.
- For a NAND expression, place a line over an AND expression.
- For a NOR expression, place a line over an OR expression.
- For exclusive operations, use the symbol ⊕.

Figure 12.18 shows a rundown of all the possible Boolean expressions for the various logic gates.

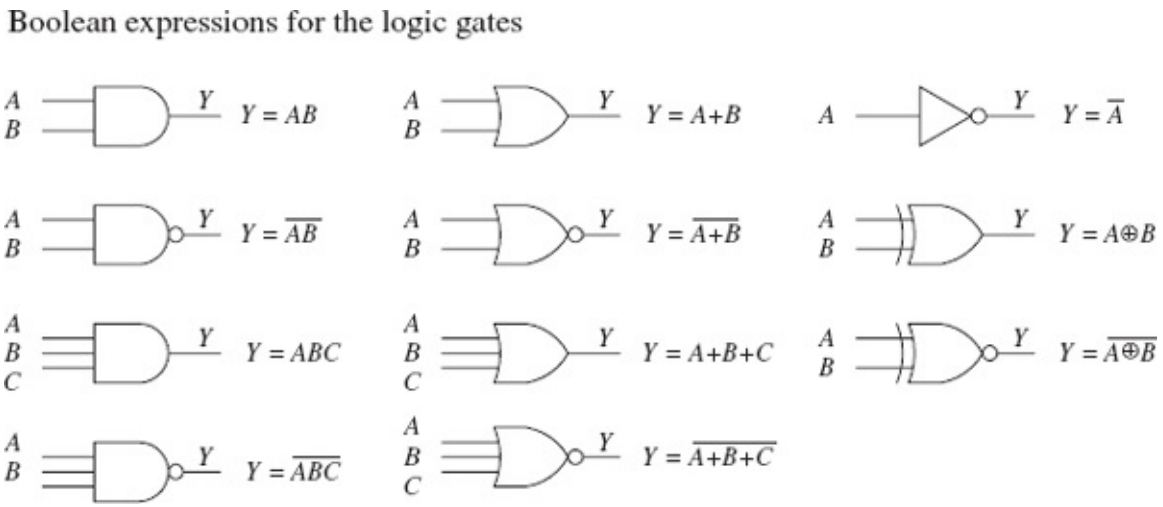


FIGURE 12.18

Like conventional algebra, Boolean algebra has a set of logic identities that can be used to simplify the Boolean expressions and thus make circuits more compact. These identities go by names such as the *commutative law of addition*, *associate law of addition*, and *distributive law*. Instead of worrying about what the various identities are called, simply make reference to the list of identities provided on the next page. Most of these identities are self-explanatory, although a few are not so obvious, as you will see in a minute. The various circuits in Fig. 12.19 show some of the identities in action.

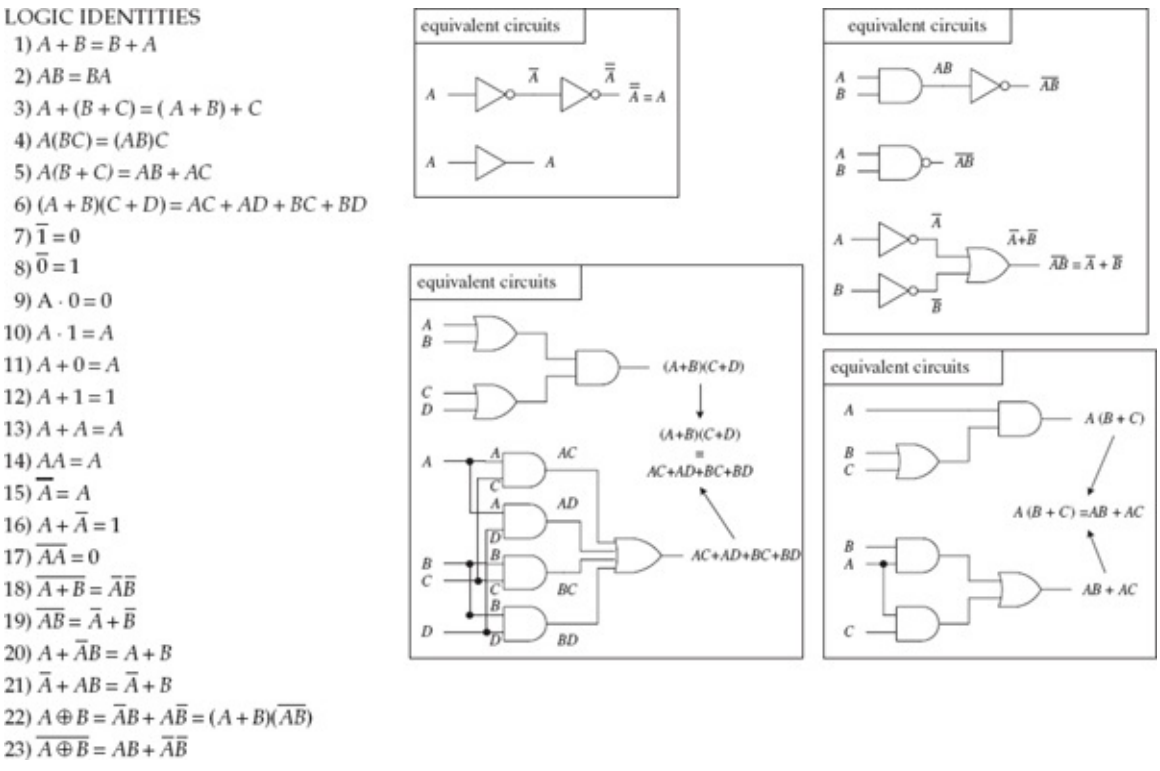
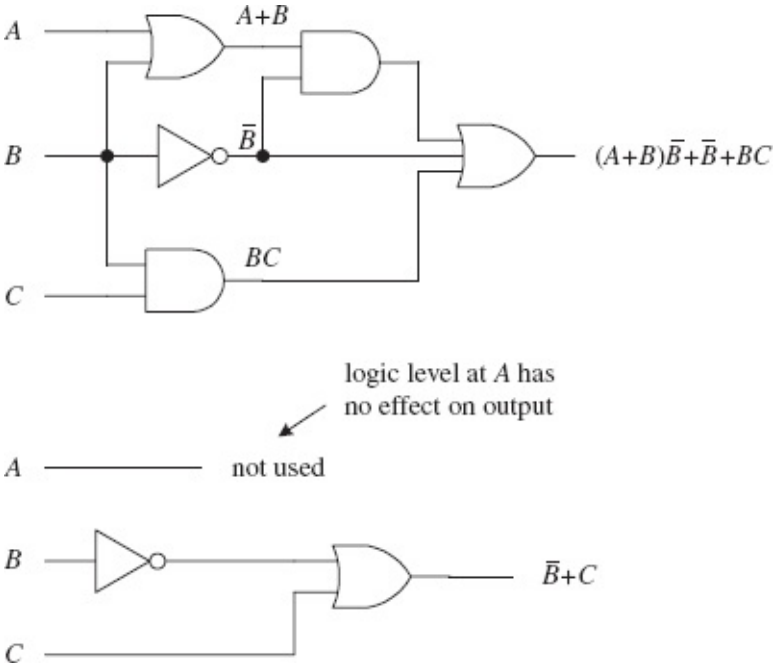


FIGURE 12.19

Example

Let's find the initial Boolean expression for the circuit in Fig. 12.20, and then use the logic identities to come up with a circuit that requires fewer gates.



The circuit shown here is expressed by the following Boolean expression:

out* = (*A* + *B*)$\overline{B}$ + $\overline{B}$ + *BC

This expression can be simplified by using Identity 5:

(*A* + *B*)$\overline{B}$ = *A*$\overline{B}$ + *B*$\overline{B}$

This makes:

out* = *A*$\overline{B}$ + *B*$\overline{B}$ + $\overline{B}$ = *BC

Using Identities 17 (*B*$\overline{B}$ = 0) and ($\overline{B}$ + 0 = $\overline{B}$), you get:

***out* = *A*$\overline{B}$ + 0 + $\overline{B}$ + *BC* = *A*$\overline{B}$ + *BC* + $\overline{B}$**

Factoring a $\overline{B}$ from the preceding term gives:

out* = $\overline{B}$(*A* + 1) + *BC

Using Identity 10, you get:

out = $\overline{B}$(1) + *BC* = $\overline{B}$ + *BC*

Finally, using Identity 21, you get the simplified expression:

out* = $\overline{B}$ + *C

Notice that *A* is now missing. This means that the logic input at *A* has no effect on the output and therefore can be omitted. From the reduction, you get the simplified circuit in the bottom part of the figure.

FIGURE 12.20

Dealing with Exclusive Gates (Identities 22 and 23)

Now let's take a look at a couple of not-so-obvious logic identities: those that involve the XOR (Identity 22) and XNOR (Identity 23) gates. The leftmost section in Fig. 12.21 shows equivalent circuits for the XOR gate. In the lower two equivalent circuits, Identity 22 is proved by Boolean reduction. Equivalent circuits for the XNOR gate are shown in the rightmost section of the figure. To prove Identity 23, you can simply invert Identity 22.

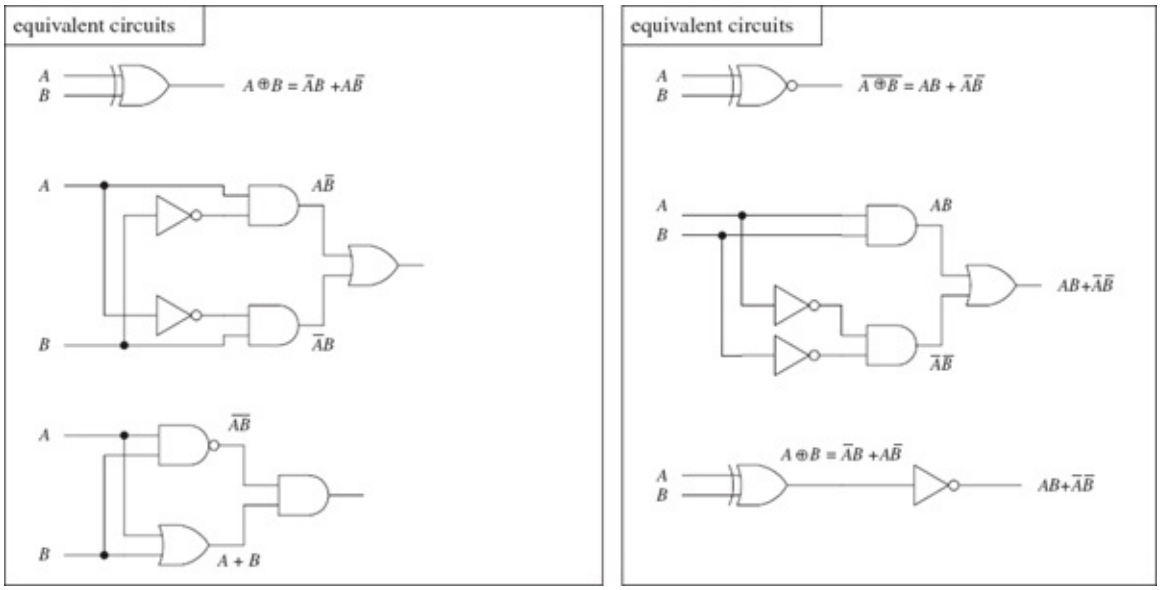


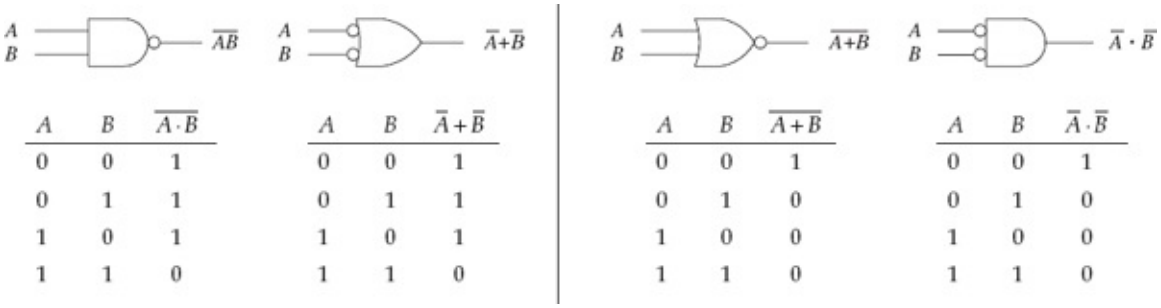
FIGURE 12.21

De Morgan's Theorem (Identities 18 and 19)

To simplify circuits containing NANDs and NORs, you can use an incredibly useful theorem known as *De Morgan's theorem*. This theorem allows you to convert an expression having an inversion bar over two or more variables into an expression having inversion bars over single variables only. De Morgan's theorem (Identities 18 and 19) is as follows:

$$\overline{A \cdot B} = \bar{A} + \bar{B} \quad \text{(2 variables)} \qquad \overline{A \cdot B \cdot C} = \bar{A} + \bar{B} + \bar{C} \quad \text{(3 or more variables)}$$
$$\overline{A + B} = \bar{A} \cdot \bar{B} \qquad \overline{A + B + C} = \bar{A} \cdot \bar{B} \cdot \bar{C}$$

The easiest way to prove that these identities are correct is to use Fig. 12.22, noting that the truth tables for the equivalent circuits are the same. Note the inversion bubbles present on the inputs of the corresponding leftmost gates. The inversion bubbles mean that before inputs *A* and *B* are applied to the base gate, they are inverted (negated). In other words, the bubbles are simplified expressions for NOT gates.



A	B	$\overline{A \cdot B}$
0	0	1
0	1	1
1	0	1
1	1	0

A	B	$\bar{A}+\bar{B}$
0	0	1
0	1	1
1	0	1
1	1	0

A	B	$\overline{A+B}$
0	0	1
0	1	0
1	0	0
1	1	0

A	B	$\bar{A} \cdot \bar{B}$
0	0	1
0	1	0
1	0	0
1	1	0

FIGURE 12.22

Why do you use the inverted-input OR gate symbol instead of a NAND gate symbol? Or why would you use the inverted-input AND gate symbol instead of a NOR gate symbol? This is left up to the designer to choose whatever symbol seems most logical to use. For example, when designing a circuit, it may be easier to think about ORing or ANDing inverted inputs than to think about NANDing or NORing inputs. Similarly, it may be easier to create truth tables or work with Boolean expressions using the inverted-input gate. It is typically easier to create truth tables and Boolean expressions that do not have variables joined under a common inversion bar. Of course, when it comes time to construct the actual working circuit, you probably will want to convert to the NAND and NOR gates because they do not require additional NOT gates at their inputs.

Bubble Pushing

A shortcut method for forming equivalent logic circuits, based on De Morgan's theorem, is to use what's called *bubble pushing*.

Bubble pushing involves the following tricks:

- Change an AND gate to an OR gate or change an OR gate to an AND gate.
- Add inversion bubbles to the inputs and outputs where there were none, while removing the original bubbles.

That's it. You can prove to yourself that this works by examining the corresponding truth tables for the original gate and the bubble-pushed gate, or you can work out the Boolean expressions using De Morgan's theorem. Figure 12.23 shows examples of bubble pushing.

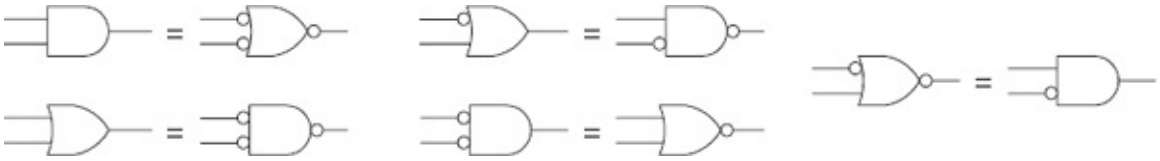


FIGURE 12.23

Universal Capability of NAND and NOR Gates

NAND and NOR gates are referred to as *universal gates* because each alone can be combined together with itself to form all other possible logic gates. The ability to create any logic gate from NAND or NOR gates is obviously a handy feature. For example, if you do not have an XOR IC handy, you can use a single multigate NAND gate (such as 74HC00) instead. Figure 12.24 shows how to wire NAND or NOR gates together to create equivalent circuits of the various logic gates.

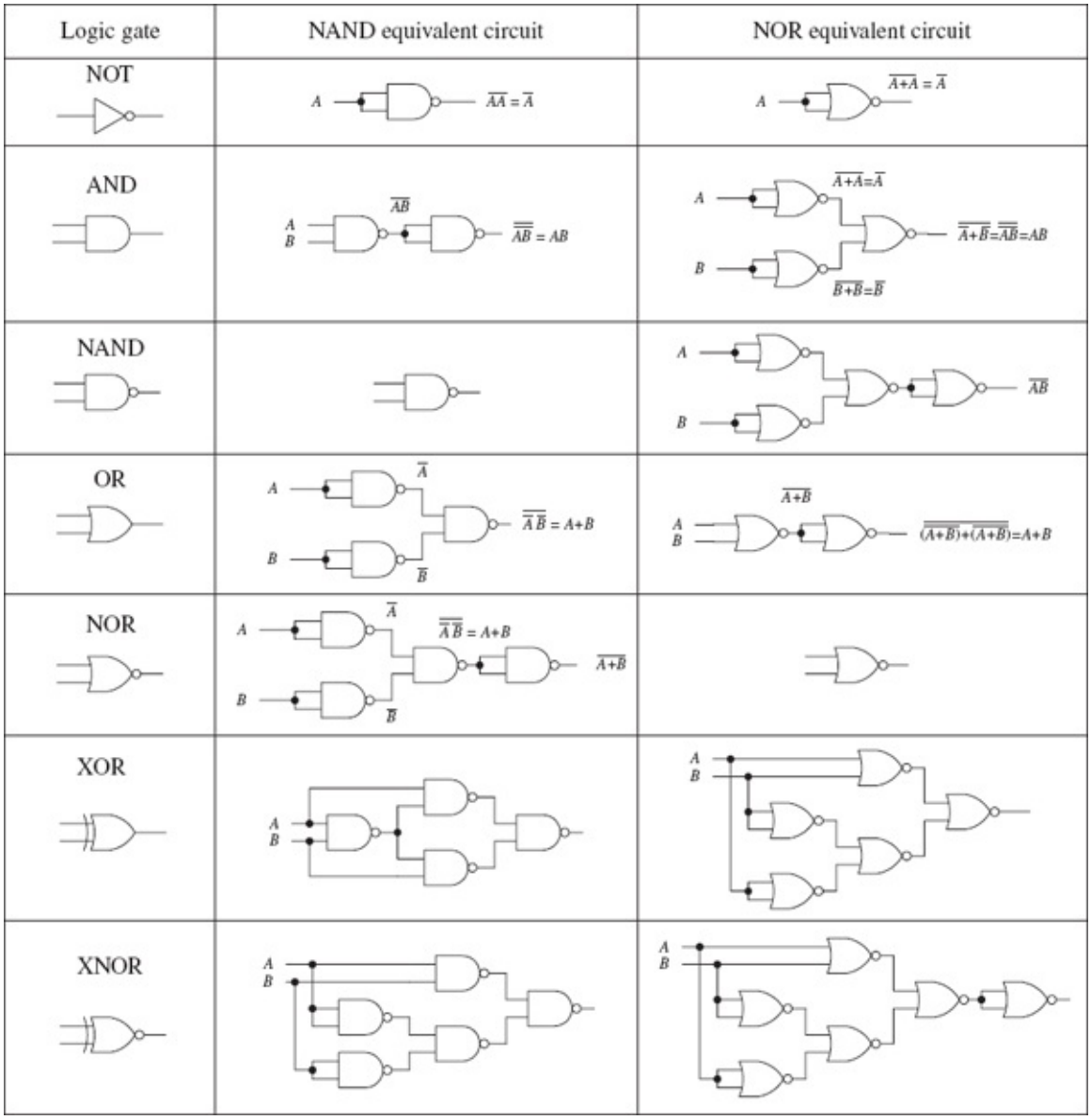


FIGURE 12.24

AND-OR-INVERTER Gates

When a Boolean expression is reduced, the equation that is left over typically will be of one of the following two forms: *product of sums* (POS) or *sum of products* (SOP). A POS expression appears as two or more OR'ed variables AND'ed together with two or more additional OR'ed variables. An SOP expression appears as two or more AND'ed variables OR'ed together with additional AND'ed variables. Figure 12.25 shows two circuits that provide the same logic function (they are equivalent), but the circuit to the left is designed to yield a POS expression, while the circuit to the right is designed to yield a SOP expression.

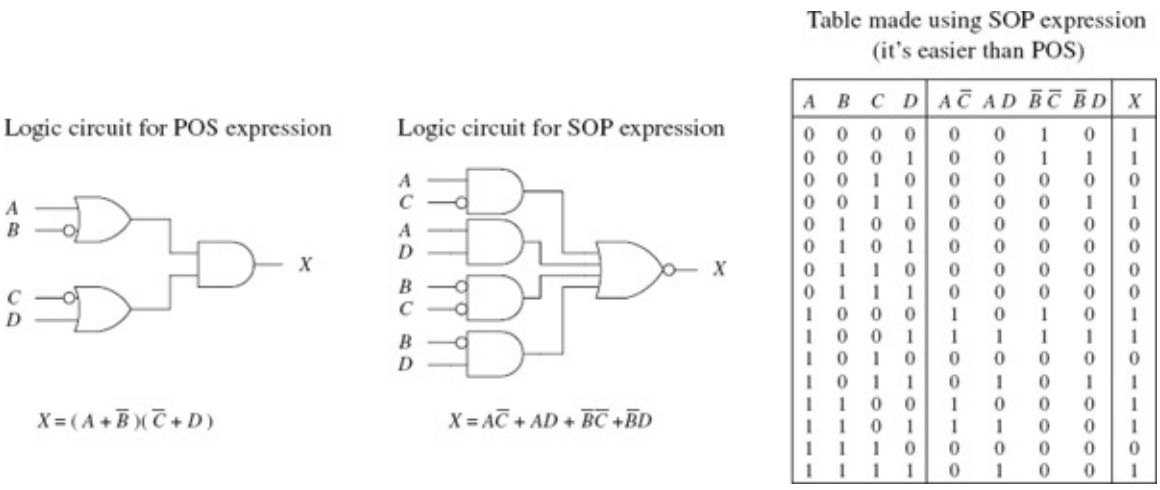


FIGURE 12.25

Which circuit is best for design: the one that implements the POS expression or the one that implements the SOP expression? The POS design shown here would appear to be the better choice because it requires fewer gates. However, the SOP design is nice because it is easy to work with the Boolean expression. For example, which Boolean expression in Fig. 12.25 (POS or SOP) would you rather use to create a truth table? The SOP expression seems the obvious choice.

A more down-to-earth reason for using an SOP design has to do with the fact that special ICs called AND-OR-INVERTER (AOI) gates are designed to handle SOP expressions. For example, the 74LS54 AOI IC shown in Fig. 12.26 creates an inverted SOP expression at its output, via two two-input AND gates and two three-input AND gates NOR'ed together. A NOT gate can be attached to the output to get rid of the inversion bar, if desired. If specific inputs are not used, they should be held high, as shown in the example circuit in Fig. 12.26. AOI ICs come in many different configurations—check out the catalogs to see what's available.

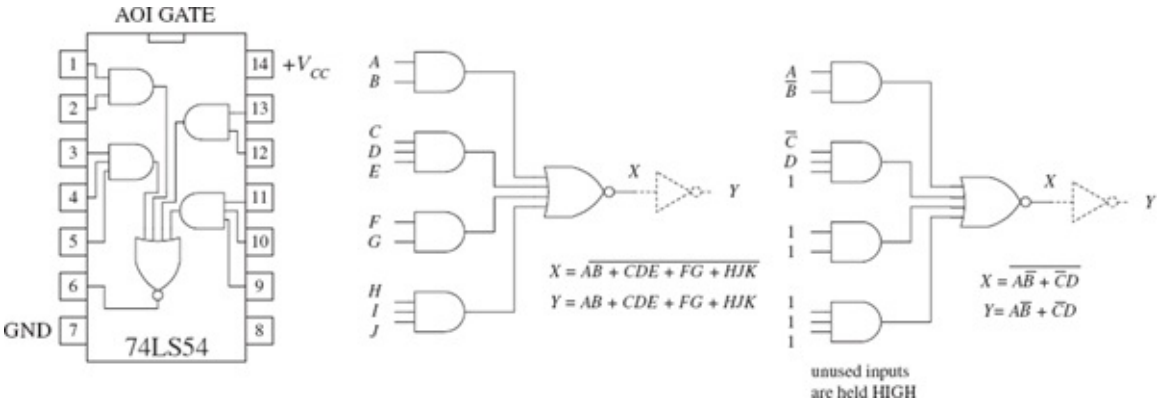


FIGURE 12.26

12.2.5 Keeping Circuits Simple (Karnaugh Maps)

We have just covered how using the logic identities can simplify a Boolean expression. This is important because it reduces the number of gates needed to construct the logic circuit. However, as I am sure you will agree, having to work out Boolean problems in longhand is not easy. It takes time and ingenuity. A simple way to avoid the unpleasant task of using your ingenuity is to get a computer program that accepts a truth table or Boolean expression, and then provides you with the simplest expression, and perhaps even the circuit schematic.

However, let's assume that you do not have such a program to help you out. Are you stuck with the Boolean longhand approach? No. You can use a technique referred to as *Karnaugh mapping*. With this

technique, you take a given truth table (or Boolean expression that can be converted into a truth table), convert it into a Karnaugh map, apply some simple graphic rules, and come up with the simplest (most of the time) possible Boolean expression for your final circuit. Karnaugh mapping works best for circuits with three to four inputs—below this, things usually do not require much thought anyway; beyond four inputs, things get quite tricky.

Here's a basic outline showing how to apply Karnaugh mapping to a three-input system:

1. Select a desired truth table. Let's choose the one shown in Fig. 12.27. (If you have only a Boolean expression, transform it into an SOP expression and use the SOP expression to create the truth table; refer to Fig. 12.26 to figure out how this is done.)

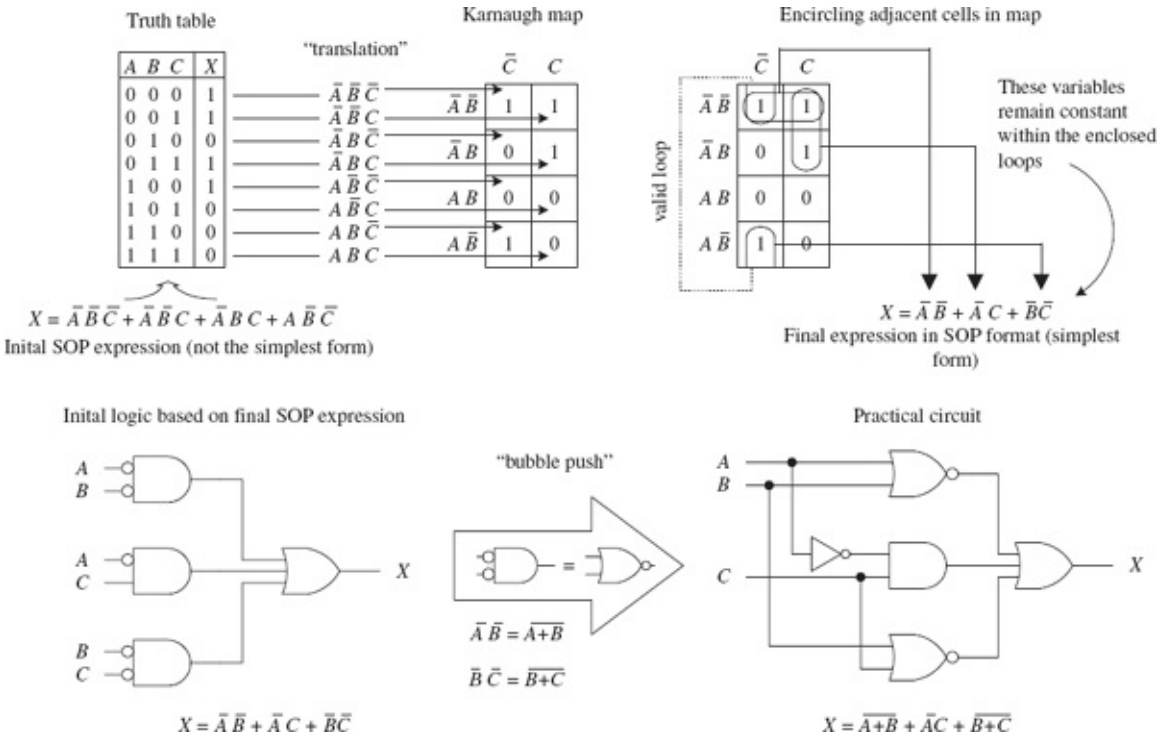


FIGURE 12.27

2. Translate the truth table into a Karnaugh map. A Karnaugh map is similar to a truth table but has its variables represented along two axes. Translating the truth table into a Karnaugh map reduces the number of 1s and 0s needed to present the information. Figure 12.27 shows how the translation is carried out.
3. After you create the Karnaugh map, proceed to encircle adjacent cells of 1s into groups of 2, 4, or 8. The more groups you can encircle, the simpler the final equation will be. In other words, take all possible loops.
4. Identify the variables that remain constant within each loop, and write out an SOP equation by OR'ing these variables together. Here, *constant* means that a variable and its inverse are not present together within the loop. For example, the top horizontal loop in Fig. 12.27 yields $\overline{A}B$ (the first term in the SOP expression), since $\overline{A}$'s and $\overline{B}$'s inverses (A and B) are not present. However, the C variable is omitted from this term because C and $\overline{C}$ are both present.
5. The SOP expression you end up with is the simplest possible expression. With it, you can create your logic circuit. You may need to apply some bubble pushing to make the final circuit practical, as shown in Fig. 12.28.

inputs				out
A	B	C	D	Y
0	0	0	0	0
0	0	0	1	1
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	1
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

Unsimplified SOP expression:

$$\bar{A}\cdot\bar{B}\cdot\bar{C}\cdot D+\bar{A}\cdot\bar{B}\cdot C\cdot D+\bar{A}\cdot B\cdot\bar{C}\cdot D+\bar{A}\cdot B\cdot C\cdot\bar{D}$$
$$+\bar{A}\cdot B\cdot C\cdot D+A\cdot\bar{B}\cdot\bar{C}\cdot D+A\cdot\bar{B}\cdot C\cdot D+A\cdot B\cdot\bar{C}\cdot D+A\cdot B\cdot C\cdot D=Y$$

Simplified SOP expression and circuit

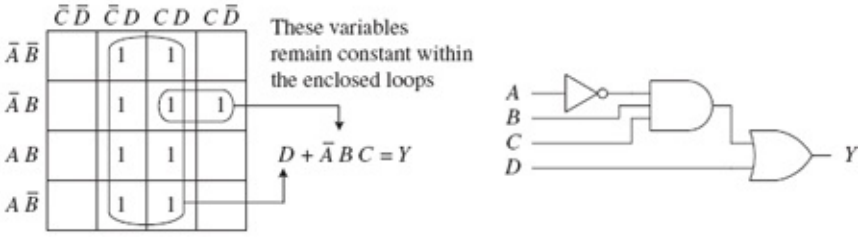


FIGURE 12.28

To apply Karnaugh mapping to four-input circuits, you apply the same basic steps used in the three-input scheme. However, now you must use a 4 × 4 Karnaugh map to hold all the necessary information. Figure 12.28 shows an example of how a four-input truth table (or unsimplified four-variable SOP expression) can be mapped and converted into a simplified SOP expression that can be used to create the final logic circuit.

Figure 12.29 shows an example that uses an AOI IC to implement the final SOP expression after mapping. I've thrown in variables other than the traditional A, B, C, and *D* just to let you know you are not limited to them alone. The choice of variables is up to you and usually depends on the application.

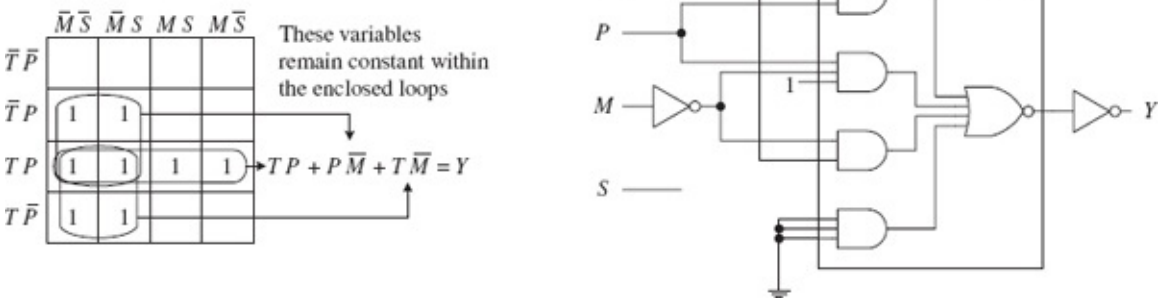


FIGURE 12.29

Other Looping Configurations

Figure 12.30 shows examples of other looping arrangements used with 4 × 4 Karnaugh maps.

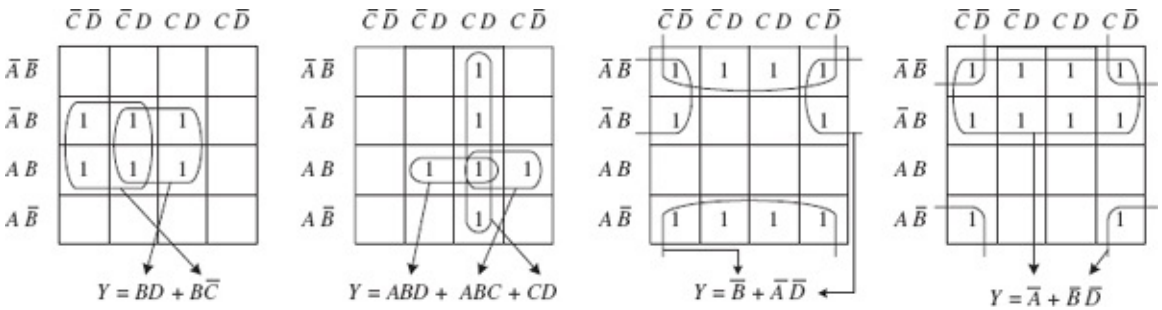


FIGURE 12.30

There are also useful online resources for exploring truth tables and minimizing logical expressions, such as these:

http://turner.faculty.swau.edu/mathematics/materialslibrary/truth/
http://courseware.ee.calpoly.edu/~rsandige/KarnaughExplorer.html

12.3 Combinational Devices

Now that you know a little something about how to use logic gates to enact functions represented within truth tables and Boolean expressions, it is time to take a look at some common functions that are used in the real world of digital electronics. As you will see, these functions are usually carried out by an IC that contains all the necessary logic.

As with almost everything discussed in this chapter, before using these ideas, you need to ask yourself if using a microcontroller would be more appropriate. However, many of the devices described here can be used with a microcontroller, especially when it comes to decoders. They can be a useful and low-cost solution for tasks such as driving more LEDs than there are pins on the microcontroller that you are using.

A word on IC part numbers before we begin. As with the logic gate ICs, the combinational ICs that follow will be of either the 4000 or 7400 series. It is important to note that an original TTL IC, like the 74138, is essentially the same device (usually with the same pinouts and function, but not always) as its newer counterparts, such as the 74F138, 74HC128 (CMOS), and 74LS138. The practical difference resides in the overall performance of the device (speed, power dissipation, voltage level rating, and so on). I will get into these gory details in a bit.

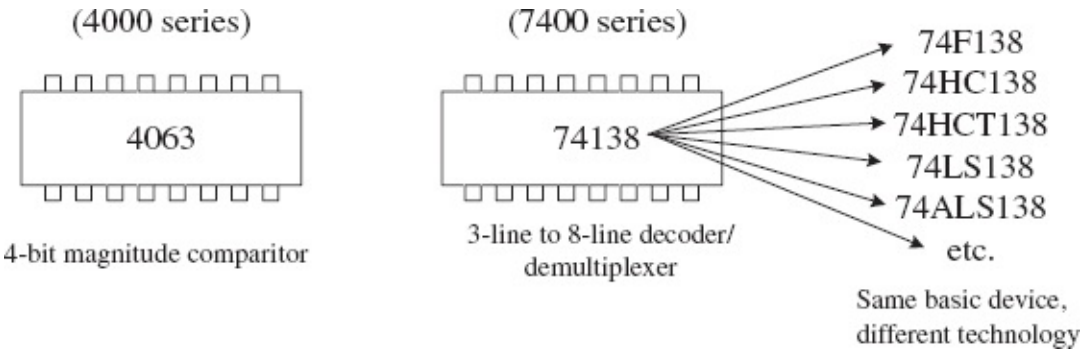


FIGURE 12.31

12.3.1 Multiplexers (Data Selectors) and Bilateral Switches

Multiplexers or data selectors act as digitally controlled switches. The term *data selector* appears to be the accepted term when the device is designed to act like an SPDT switch, while the term *multiplexer* is used when the throw count of the switch exceeds two, such as an SP8T. I will stick with this convention (although others may not).

A simple 1-of-2 data selector built from logic gates is shown in Fig. 12.32. The data select input of this circuit acts to control which input (*A* or *B*) gets passed to the output: When data select is low, input *A* passes while *B* is blocked. When data select is high, input *B* is passed while *A* is blocked. To understand how this circuit works, think of the AND gates as enable gates.

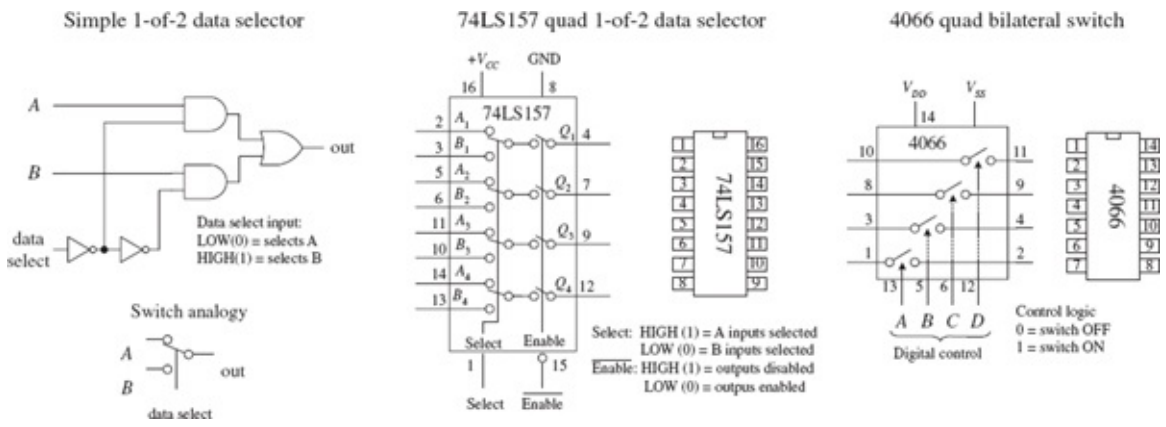


FIGURE 12.32

There are a number of different types of data selectors that come in IC form. For example, the 74LS157 quad 1-of-2 data selector IC, shown in Fig. 12.32, acts like an electrically controlled quad SPDT switch (or if you like, a 4PDT switch). When its select input is set high (1), inputs A_1, A_2, A_3, and A_4 are allowed to pass to outputs Q_1, Q_2, Q_3, and Q_4. When its select input is low (0), inputs B_1, B_2, B_3, and B_4 are allowed to pass to outputs Q_1, Q_2, Q_3, and Q_4. Either of these two conditions, however, ultimately depends on the state of the enable input.

When the enable input is low, all data-input signals are allowed to pass to the output; however, if the enable is high, the signals are not allowed to pass. This type of enable control is referred to as *active-low* enable, since the active function (passing the data to the output) occurs only with a low-level input voltage. The active-low input is denoted with a bubble (inversion bubble), and the outer label of the active-low input is represented with a line over it. Sometimes people omit the bubble and place a bar over the inner label. Both conventions are used commonly.

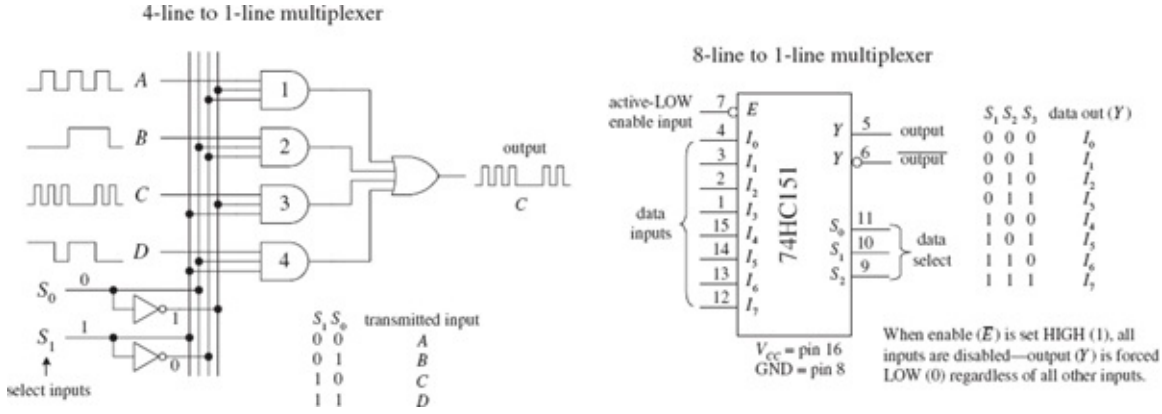


FIGURE 12.33

Figure 12.33 shows a 4-line-to-1-line multiplexer built with logic gates. This circuit resembles the 2-of-1 data selector shown in Fig. 12.32 but requires an additional select input to provide four address combinations.

In terms of ICs, there are multiplexers of various input line capacities. For example, the 74151 8-line-to-1-line multiplexer uses three select inputs (S_0, S_1, S_2) to choose among one of eight possible data inputs (I_0 to I_7) to be funneled to the output. Note that this device actually has two outputs: one true (pin 5) and one inverted (pin 6). The active-low enable forces the true output low when set high, regardless of the inputs.

To create a larger multiplexer, you combine two smaller multiplexers together. For example, Fig. 12.34 shows two 8-line-to-1-line 74HC151s combined to create a 16-line-to-1-line multiplexer. Another alternative is to use a 16-line-to-1-line multiplexer IC like the 74HC150 shown in the figure. Check the

catalogs to see what other kinds of multiplexers are available.

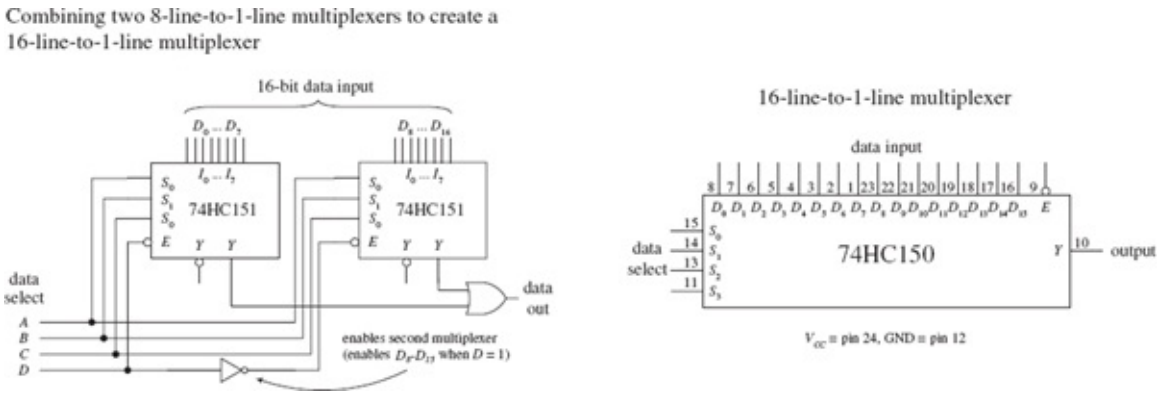


FIGURE 12.34

Finally, let's take a look at a very useful device called a *bilateral switch*. An example bilateral switch IC is the 4066, shown to the far right in Fig. 12.32. Unlike the multiplexer, this device merely acts as a digitally controlled quad SPST switch or quad transmission gate. Using a digital control input, you select which switches are on and which switches are off. To turn on a given switch, apply a high level to the corresponding switch select input; otherwise, keep the select input low.

In Section 12.10, we will look at analog switches and multiplexers. These devices use digital select inputs to control analog signals. Analog switches and multiplexers become important when you start linking the digital world to the analog world.

12.3.2 Demultiplexers (Data Distributors) and Decoders

A demultiplexer (or data distributor) is the opposite of a multiplexer. It takes a single data input and routes it to one of several possible outputs. A simple four-line demultiplexer built from logic gates is shown on the left side of Fig. 12.35. To select the output (*A*, *B*, *C*, or *D*) to which you want to send the input signal (applied at *E*), you apply logic levels to the data select inputs (S_0, S_1), as shown in the truth table.

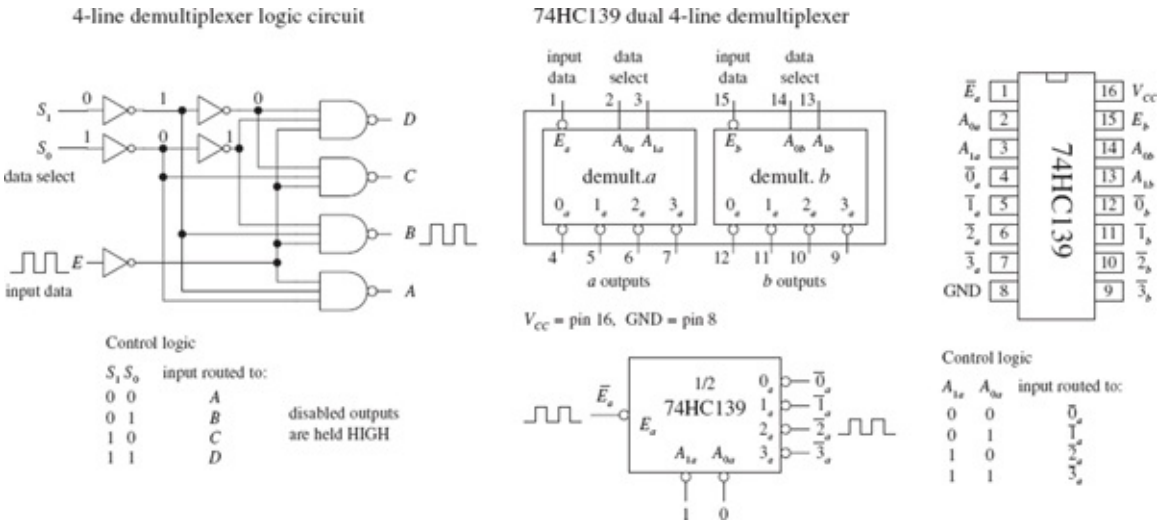


FIGURE 12.35

Notice that the unselected outputs assume a high level, while the selected output varies with the input signal. An IC that contains two functionally separate four-line demultiplexers is the 74HC139, shown on the right side of Fig. 12.35. If you need more outputs, check out the 75*xx*154 16-line demultiplexer. This IC uses four data select inputs to choose from 1 of 16 possible outputs. Check out the catalogs to see what

other demultiplexers exist.

A decoder is somewhat like a demultiplexer, but it does not route input data to a specific output via data select inputs. Instead, it simply uses the data select inputs to choose which output (or outputs) among many are to be made high or low. The number of address inputs, the number of outputs, and the active state of the selected output vary from decoder to decoder. The variance is based on what the decoder is designed to do. For example, the 74LS138 1-of-8 decoder shown in Fig. 12.36 uses a 3-bit address input to select which of eight outputs will be made low; all other outputs are held high. Like the demultiplexer in Fig. 12.35, this decoder has active-low outputs.

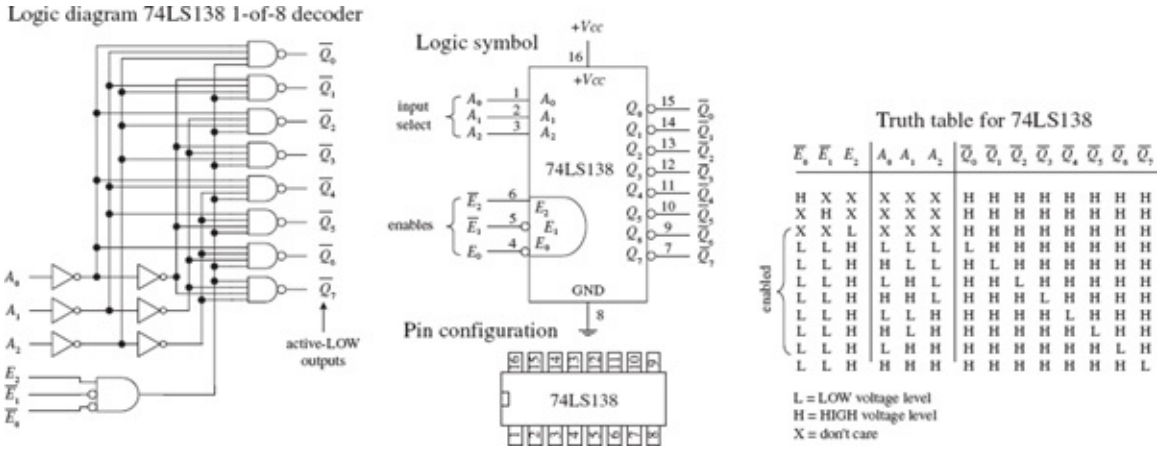


FIGURE 12.36

Now what exactly does it mean to say an output is an active-low output? It simply means that when an active-low output is selected, it is forced to a low logic state; otherwise, it is held high. Active-high outputs behave in the opposite manner. An active-low output is usually indicated with a bubble, although sometimes it is indicated with a barred variable within the IC logic symbol—no bubble included. Active-high outputs have no bubbles. Both active-low and active-high outputs are equally common among ICs.

By placing a load (for example, a warning LED) between $+V_{CC}$ and an active-low output, you can sink current through the load and into the active-low output when the output is selected. By placing a load between an active-high output and ground, you can source current from the active-high output and sink it through the load when the output is selected. The limits to how much current an IC can source or sink will be discussed in Section 12.4, and various schemes used to drive analog loads will be presented in Section 12.10.

Now let's get back to the 74LS138 decoder and discuss the remaining enable inputs ($\overline{E}_0$,$\overline{E}_1$,E_2). For the 74LS138 to "decode," you must make the active-low inputs $\overline{E}_0$ and $\overline{E}_1$ low, while making the active-high input E_2 high. If any other set of enable inputs is applied, the decoder is disabled, making all active-low outputs high regardless of the selected inputs.

Other common decoders include the 7442 BCD-to-DEC (decimal) decoder, the 74154 1-of-16 (hex) decoder, and the 7447 BCD-to-seven-segment decoder shown in Figure 12.37. Like the preceding decoder, these devices also have active-low outputs. The 7442 uses a binary-coded decimal input to select 1 of 10 (0 through 9) possible outputs. The 74154 uses a 4-bit binary input to address 1 of 16 (or 0 of 15) outputs, making that output low (all others high), provided the enables are both set low.

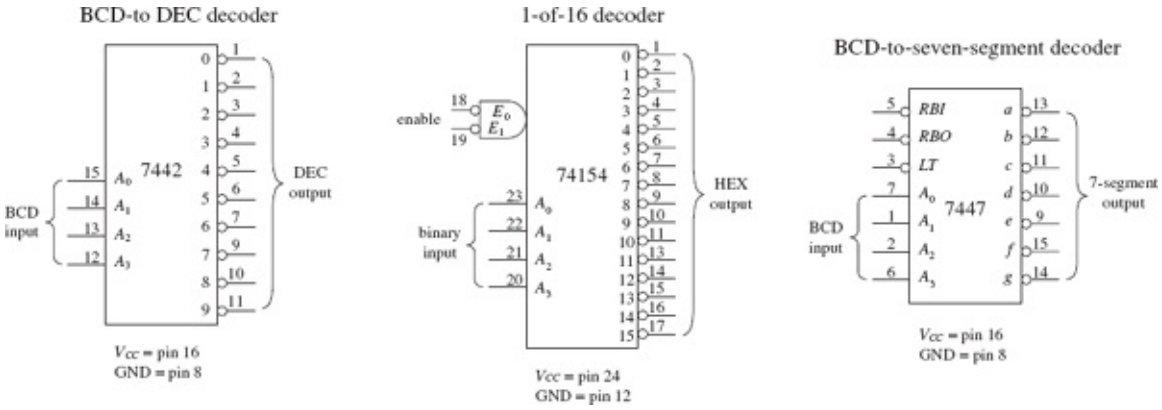


FIGURE 12.37

Now the 7447 is a bit different from the other decoders. With this device, more than one output can be driven low at a time. This is important because it allows the 7447 to drive a seven-segment LED display; to create different numbers requires driving more than one LED segment at a time. For example, in Fig. 12.38, when the BCD number for 5 (0101) is applied to the 7447's inputs, all outputs except $\overline{b}$ and $\overline{e}$ go low. This causes LED segments *a*, *c*, *d*, *f*, and *g* to light up—the 7447 sinks current through these LED segments, as indicated by the internal wiring of the display and the truth table.

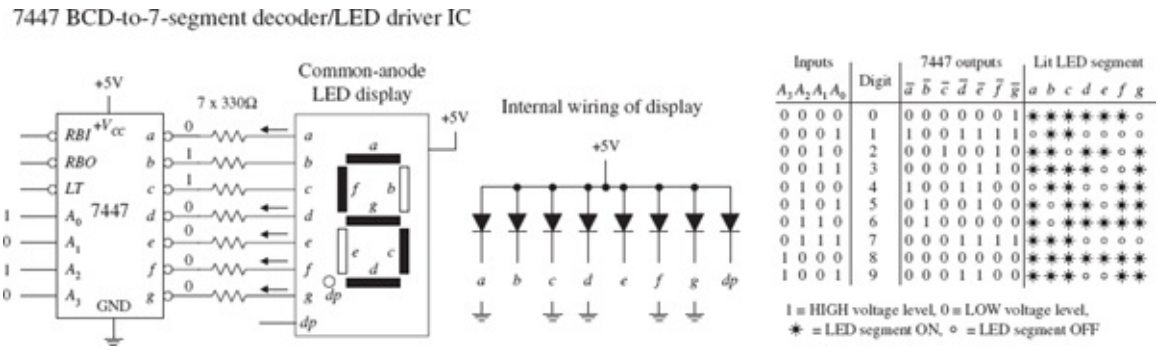
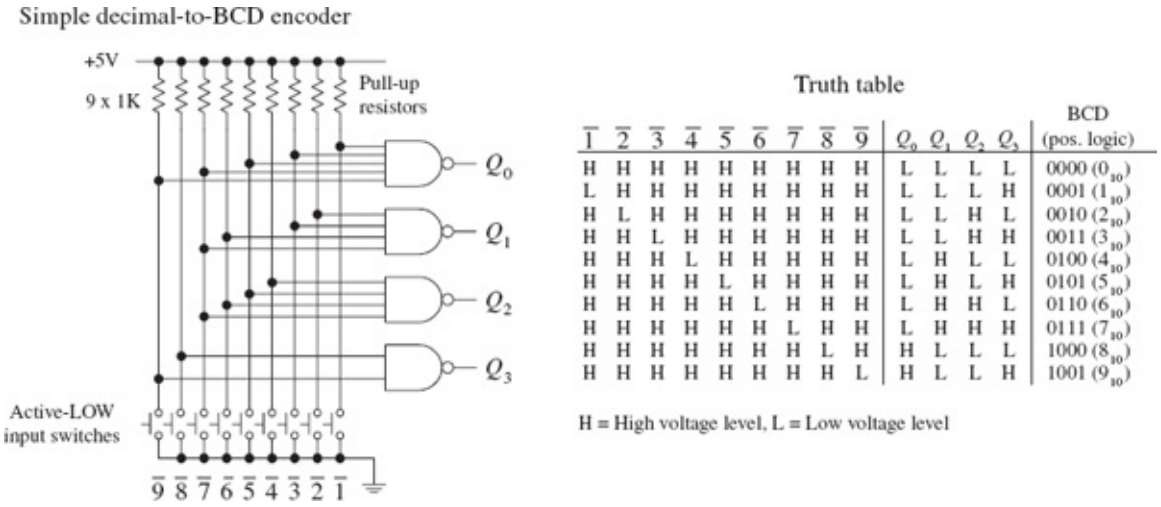


FIGURE 12.38

The 7447 also comes with a lamp test active-low input ($\overline{LT}$) that can be used to drive all LED segments at once to see if any of the segments are faulty. The ripple blanking input ($\overline{RBI}$) and ripple blanking output ($\overline{RBO}$) can be used in multistage display applications to suppress a leading-edge and/or trailing-edge zero in a multidigit decimal. For example, using the ripple blanking inputs and outputs, it is possible to take an eight-digit expression like 0056.020 and display 56.02, suppressing the two leading zeros and the one trailing zero. Leading-edge zero suppression is obtained by connecting the ripple blanking output of a decoder to the ripple blanking input of the next lower-stage device. The most significant decoder stage should have its ripple blanking input grounded. A similar procedure is used to provide automatic suppression of trailing zeros in the fractional part of the decimal.

12.3.3 Encoders and Code Converters

Encoders are the opposite of decoders. They are used to generate a coded output from a single active numeric input. To illustrate this in a simple manner, let's take a look at the simple decimal-to-BCD encoder circuit shown in Fig. 12.39.



Truth table

$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	Q_0	Q_1	Q_2	Q_3	BCD (pos. logic)
H	H	H	H	H	H	H	H	H	L	L	L	L	0000 (0_{10})
L	H	H	H	H	H	H	H	H	L	L	L	H	0001 (1_{10})
H	L	H	H	H	H	H	H	H	L	L	H	L	0010 (2_{10})
H	H	L	H	H	H	H	H	H	L	L	H	H	0011 (3_{10})
H	H	H	L	H	H	H	H	H	L	H	L	L	0100 (4_{10})
H	H	H	H	L	H	H	H	H	L	H	L	H	0101 (5_{10})
H	H	H	H	H	L	H	H	H	L	H	H	L	0110 (6_{10})
H	H	H	H	H	H	L	H	H	L	H	H	H	0111 (7_{10})
H	H	H	H	H	H	H	L	H	H	L	L	L	1000 (8_{10})
H	H	H	H	H	H	H	H	L	H	L	L	H	1001 (9_{10})

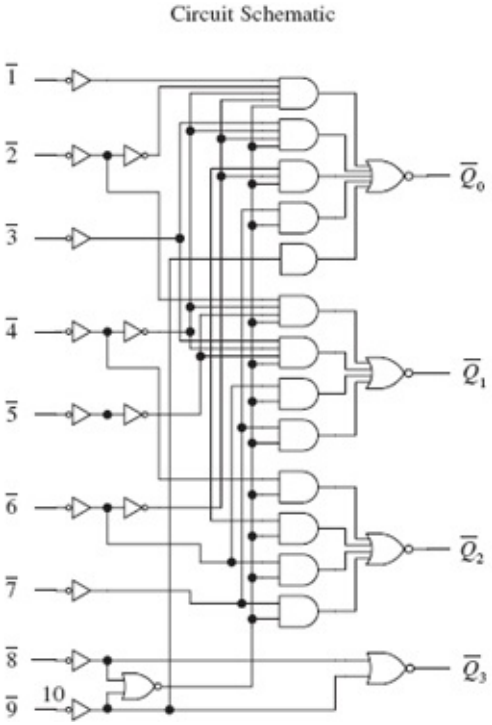
H = High voltage level, L = Low voltage level

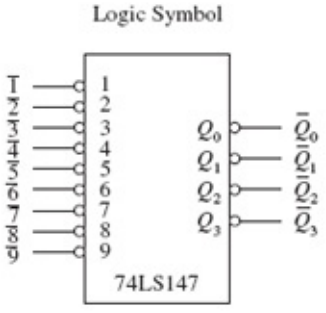
FIGURE 12.39

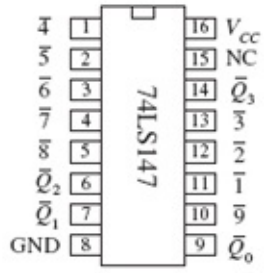
In this circuit, normally all lines are held high by the pullup resistors connected to +5 V. To generate a BCD output that is equivalent to a single selected decimal input, the switch corresponding to that decimal is closed. (The switch acts as an active-low input.) The truth table in Fig. 12.39 explains the rest.

Figure 12.40 shows a 74LS147 decimal-to-BCD (ten-line-to-four-line) priority encoder IC. The 74LS147 provides the same basic function as the circuit shown in Fig. 12.39, but it has active-low outputs. This means that instead of getting an LLHH output when 3 is selected, as in the previous encoder, you get HHLL. The two outputs represent the same thing (3); one is expressed in positive true logic, and the other (the 74LS147) is expressed in negative true logic. If you do not like negative true logic, you can slap inverters on the outputs of the 74LS147 to get positive true logic. The choice to use positive or negative true logic really depends on what you are planning to drive. For example, negative true logic is useful when the device that you wish to drive uses active-low inputs.

74LS147 decimal-to-4-bit BCD Priority Encoder IC







Truth table for 74LS147

$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	$\overline{Q}_0$	$\overline{Q}_1$	$\overline{Q}_2$	$\overline{Q}_3$	BCD (neg. logic)
H	H	H	H	H	H	H	H	H	H	H	H	H	1111 (0_{10})
X	X	X	X	X	X	X	X	L	L	H	H	L	0110 (9_{10})
X	X	X	X	X	X	X	L	H	L	H	H	H	0111 (8_{10})
X	X	X	X	X	X	L	H	H	H	L	L	L	1000 (7_{10})
X	X	X	X	X	L	H	H	H	H	L	L	H	1001 (6_{10})
X	X	X	X	L	H	H	H	H	H	L	H	L	1010 (5_{10})
X	X	X	L	H	H	H	H	H	H	L	H	H	1011 (4_{10})
X	X	L	H	H	H	H	H	H	H	H	L	L	1100 (3_{10})
X	L	H	H	H	H	H	H	H	H	H	L	H	1101 (2_{10})
L	H	H	H	L	H	H	H	H	H	H	H	L	1110 (1_{10})

H = High voltage level, L = Low voltage level, X = don't care

FIGURE 12.40

Another important difference between the two encoders is the priority that is used with the 74LS147 and not used with the encoder in Fig. 12.39. The term *priority* is applied to the 74LS147 because this

encoder is designed so that if two or more inputs are selected at the same time, it will select only the larger-order digit. For example, if 3, 5, and 8 are selected at the same time, only the 8 (negative true BCD LHHH or 0111) will be output. The truth table in Fig. 12.40 demonstrates this; look at the "don't care" or "X" entries. With the nonpriority encoder, if two or more inputs are applied at the same time, the output will be unpredictable.

The circuit shown in Fig. 12.41 provides a simple illustration of how an encoder and a decoder can be used together to drive an LED display via a 0 through 9 keypad. The 74LS147 encodes a keypad's input into BCD (negative logic). A set of inverters then converts the negative true BCD into positive true BCD. The transformed BCD is then fed into a 7447 seven-segment LED display decoder/driver IC.

A decimal-to-BCD encoder being used to convert keypad instructions into BCD instructions used to drive a LED display circuit (7447 decoder plus common-anode display)

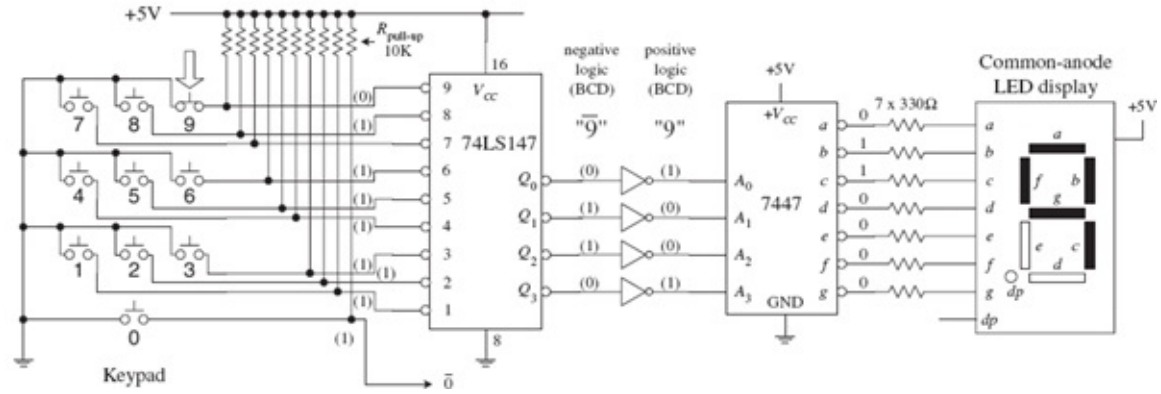
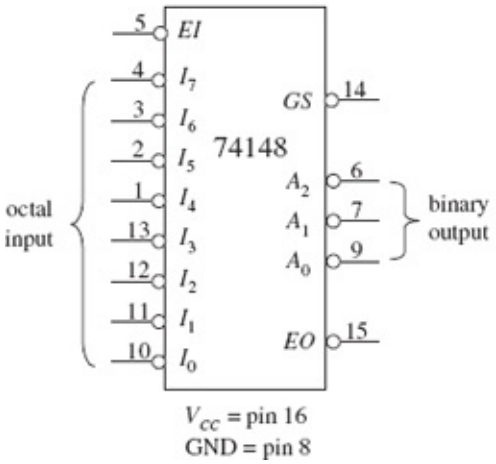


FIGURE 12.41

Figure 12.42 shows a 74148 octal-to-binary priory encoder IC. It is used to transform a specified single octal input into a binary 3-bit output code. As with the 74LS147, the 74148 comes with a priority feature, so if two or more inputs are selected at the same time, only the higher-order number is selected.

74148 octal-to-binary priority encoder



74148 truth table

$\overline{EI}$	$\overline{I_0}$	$\overline{I_1}$	$\overline{I_2}$	$\overline{I_3}$	$\overline{I_4}$	$\overline{I_5}$	$\overline{I_6}$	$\overline{I_7}$	$\overline{GS}$	$\overline{A_0}$	$\overline{A_1}$	$\overline{A_2}$	$\overline{EO}$
H	X	X	X	X	X	X	X	X	H	H	H	H	H
L	H	H	H	H	H	H	H	H	H	H	H	H	L
L	X	X	X	X	X	X	X	L	L	L	L	L	H
L	X	X	X	X	X	X	L	H	L	H	L	L	H
L	X	X	X	X	X	L	H	H	L	L	H	L	H
L	X	X	X	X	L	H	H	H	L	H	H	L	H
L	X	X	X	L	H	H	H	H	L	L	L	H	H
L	X	X	L	H	H	H	H	H	L	H	L	H	H
L	X	L	H	H	H	H	H	H	L	L	H	H	H
L	L	H	H	H	H	H	H	H	L	H	H	H	H

FIGURE 12.42

A high applied to the input enable ($\overline{EI}$) forces all outputs to their inactive (high) state and allows new data to settle without producing erroneous information at the outputs. A group signal output ($\overline{GS}$) and an enable output ($\overline{EO}$) are also provided to allow for system expansion. The $\overline{GS}$ output is active level low when any input is low (active). The $\overline{EO}$ output is low (active) when all inputs are high. Using the output enable along with the input enable allows priority coding of N input signals. Both $\overline{EO}$ and $\overline{GS}$ are active high when the input enable is high (device disabled).

Figure 12.43 shows a 74184 BCD-to-binary converter (encoder) IC. This device has eight active-high

outputs ($Y_1 - Y_8$). Outputs Y_1 to Y_5 are outputs for regular BCD-to-binary conversion, while outputs Y_6 to Y_8 are used for a special BDC code called *nine's complement* and *ten's complement*. The active-high BCD code is applied to inputs *A* through *E*. The $\overline{G}$ input is an active-low enable input.

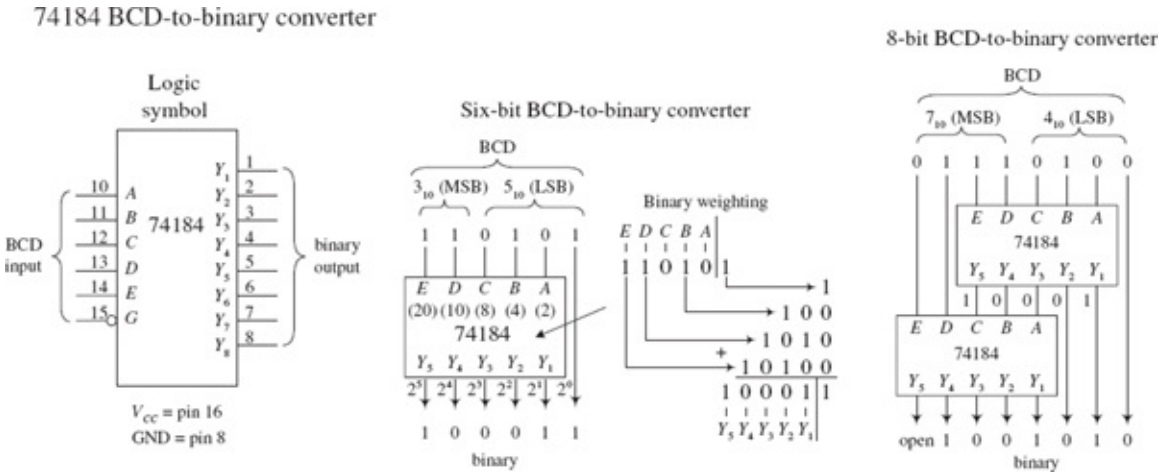


FIGURE 12.43

A sample 6-bit BCD-to-binary converter and a sample 8-bit BCD-to-binary converter that use the 74184 are shown to the right in Fig. 12.43. In the 6-bit circuit, since the LSB of the BCD input is always equal to the LSB of the binary output, the connection is made straight from input to output. The other BCD bits are applied directly to inputs *A* through *E*. The binary weighing factors for each input are $A = 2$, $B = 4$, $C = 8$, $D = 10$, and $E = 20$. Because only 2 bits are available for the MSD BCD input, the largest BCD digit in that position is 3 (binary 11). To get a complete 8-bit BCD converter, you connect two 74184s together, as shown to the far right in Fig. 12.43.

Figure 12.44 shows a 74185 binary-to-BCD converter (encoder). It is essentially the same as the 74184, but in reverse. The figure shows 6-bit and 8-bit binary-to-BCD converter arrangements.

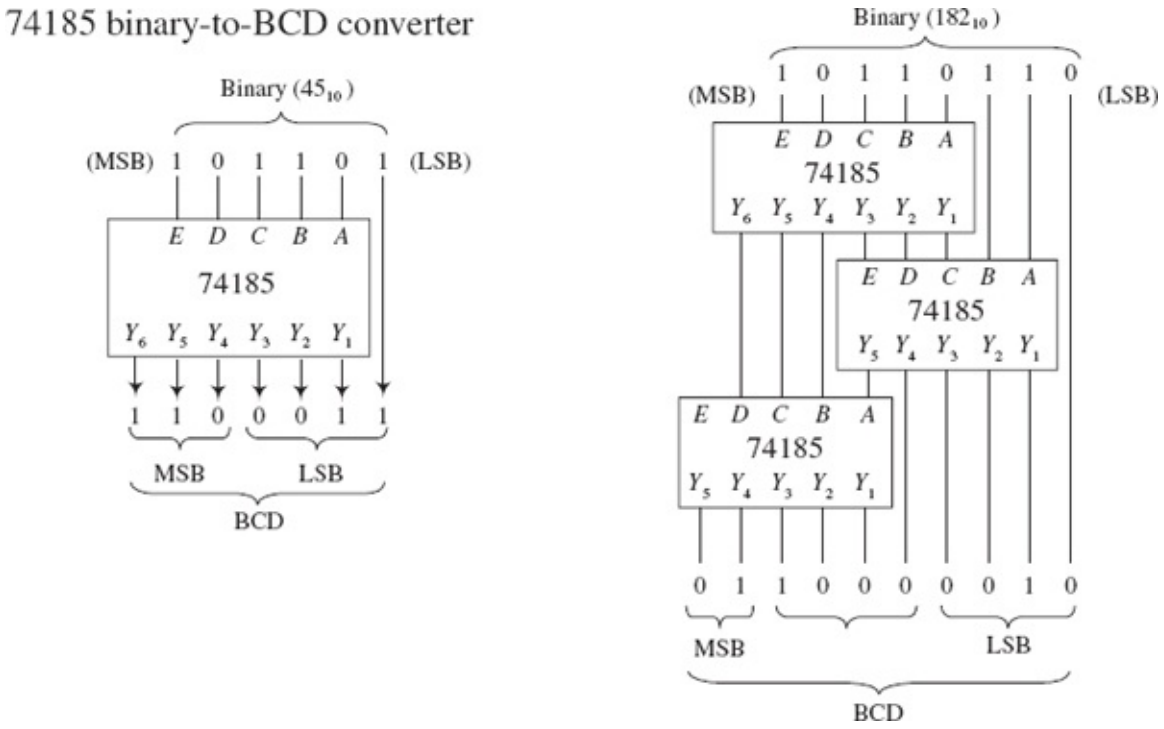


FIGURE 12.44

12.3.4 Binary Adders

If you find yourself needing to do arithmetic in logic, then that is a pretty sure sign that you need to use a microcontroller. However, that microcontroller will contain exactly the sort of logic that we describe here in its arithmetic logic unit (ALU), so it is instructive to see how it works under the hood.

With a few logic gates, you can create a circuit that adds binary numbers. The mechanics of adding binary numbers is basically the same as that of adding decimal numbers. When the first digit of a two-digit number is added, a 1 is carried and added to the next row whenever the count exceeds binary 2 (for example., 1 + 1 = 10, or = 0 carry a 1). For numbers with more digits, you have multiple carry bits.

To demonstrate how you can use logic gates to perform basic addition, start out by considering the half-adder circuits in Fig. 12.45. Both half-adders shown are equivalent; one simply uses XOR/AND logic, while the other uses NOR/AND logic. The half-adder adds two single-bit numbers A and B and produces a 2-bit number. The LSB is represented as Σ_0, and the MSB, or carry bit, is represented as C_{out}.

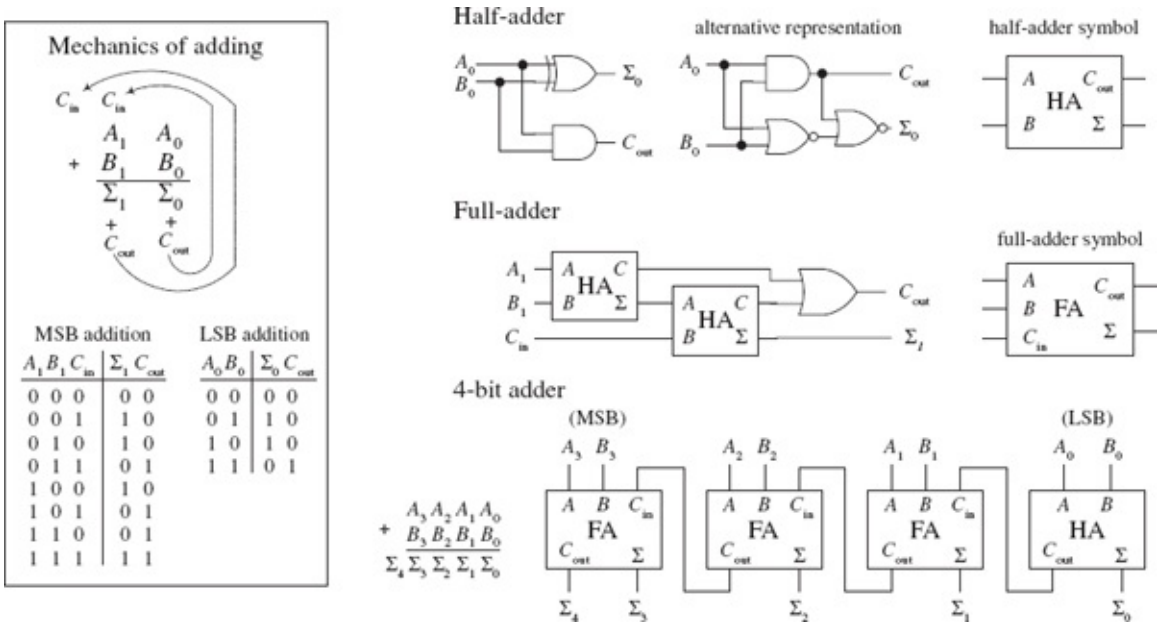


FIGURE 12.45

The most complicated operation the half-adder can do is 1 + 1. To perform addition on a two-digit number, you must attach a full-adder circuit (shown in Fig. 12.45) to the output of the half-adder. The full-adder has three inputs: two to input the second digits of the two binary numbers (A_1, B_1), and another that accepts the carry bit from the half-adder (the circuit that added the first digits, A_0 and B_0, of the two numbers). The two outputs of the full-adder will provide the 2d-place digit sum Σ_1 and another carry bit that acts as the third-place digit of the final sum. Now, you can keep adding more full-adders to the half-adder/full-adder combination to add larger numbers, linking the carry bit output of the first full-adder to the next full-adder, and so forth. To illustrate this point, a 4-bit adder is shown in Fig. 12.45.

A number of 4-bit full-adder ICs are available, such as the 74LS283 and 4008. These devices will add two 4-bit binary numbers and provide an additional input carry bit, as well as an output carry bit, so you can stack them together to get adders that are 8-bit, 12-bit, 16-bit, and so on. For example, Fig. 12.46 shows an 8-bit adder made by cascading two 74LS283 4-bit adders.

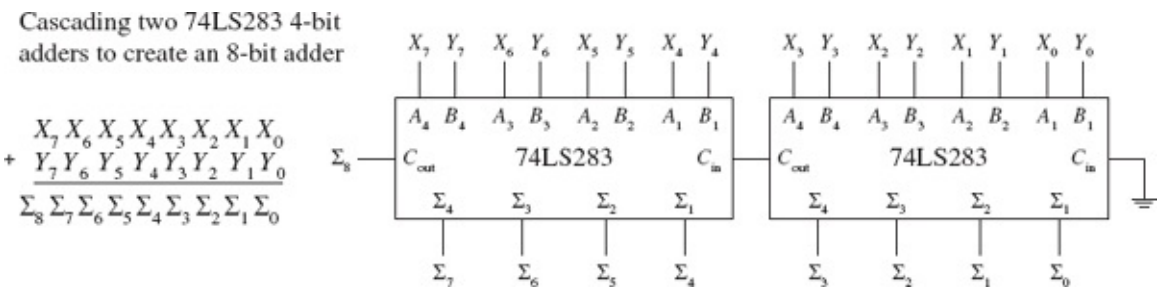


FIGURE 12.46

12.3.5 Binary Adder/Subtractor

Figure 12.47 shows how two 74LS283 4-bit adders can be combined with an XOR array to yield an 8-bit 2's complement adder/subtractor. The first number X is applied to the X_0 through X_7 inputs, while the second number Y is applied to the Y_0 through Y_7 inputs.

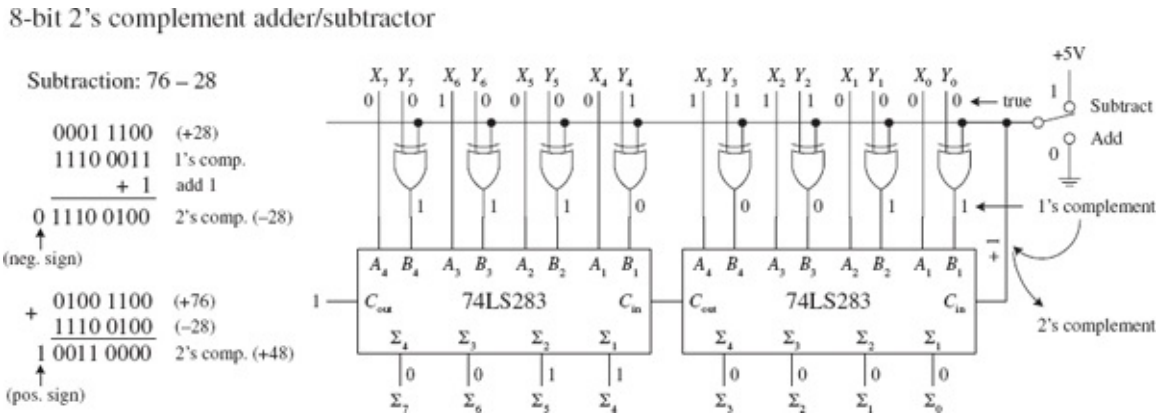


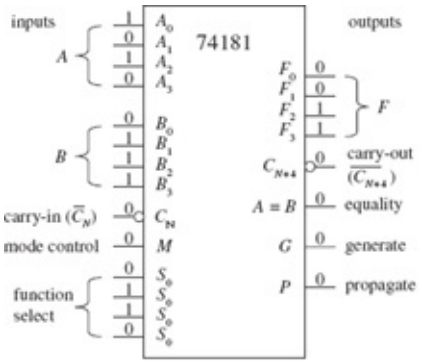
FIGURE 12.47

To add X and Y, the add/subtract switch is thrown to the add position, making one input of all XOR gates low. This has the effect of making the XOR gates appear transparent, allowing Y values to pass to the 74LS283s' B inputs (X values are passed to the A inputs). The 8-bit adder then adds the numbers and presents the result to the Σ outputs.

To subtract Y from X, you must first convert Y into 1's complement form; then you must add 1 to get Y into 2's complement form. After that you simply add X to the 2's complemented form of Y to get $X-Y$. When the add/subtract switch is thrown to the subtract position, one input to each XOR gate is set high. This causes the Y bits that are applied to the other XOR inputs to become inverted at the XOR outputs—you have just taken the 1's complement of Y. The 1's complement bits of Y are then presented to the inputs of the 8-bit adder. At the same time, C_{in} of the left 74LS283 is set high via the wire (see Fig. 12.47) so that a 1 is added to the 1's complement number to yield a 2's complement number. The 8-bit adder then adds X and the 2's complement of Y together. The final result is presented at the Σ outputs. In the figure, 76 is subtracted from 28.

12.3.6 Arithmetic Logic Units

An ALU is a multipurpose integrated circuit capable of performing various arithmetic and logic operations. To choose a specific operation to be performed, a binary code is applied to the IC's mode select inputs. The 74181, shown in Fig. 12.48, is a 4-bit ALU that provides 16 arithmetic and 16 logic operations.



Mode select				Logic functions	Arithmetic operations
S_3	S_2	S_1	S_0	$(M = 1)$	$(M = 0, \overline{C}_n = 1)$
0	0	0	0	$F = \overline{A}$	$F = A$
0	0	0	1	$F = \overline{A + B}$	$F = A + B$
0	0	1	0	$F = \overline{A}B$	$F = A + \overline{B}$
0	0	1	1	$F = 0$	$F =$ minus 1 (2's comp.)
0	1	0	0	$F = \overline{AB}$	$F = A$ plus $A\overline{B}$
0	1	0	1	$F = \overline{B}$	$F = (A + B)$ plus $A\overline{B}$
0	1	1	0	$F = A \oplus B$	$F = A$ minus B minus 1
0	1	1	1	$F = A\overline{B}$	$F = A\overline{B}$ minus 1
1	0	0	0	$F = \overline{A} + B$	$F = A$ plus AB
1	0	0	1	$F = \overline{A \oplus B}$	$F = A$ plus B
1	0	1	0	$F = B$	$F = (A + \overline{B})$ plus AB
1	0	1	1	$F = AB$	$F = AB$ minus 1
1	1	0	0	$F = 1$	$F = A$ plus A
1	1	0	1	$F = A + B$	$F = (A + B)$ plus A
1	1	1	0	$F = A + \overline{B}$	$F = (A + \overline{B})$ plus A
1	1	1	1	$F = A$	$F = A$ minus 1

FIGURE 12.48

To select an arithmetic operation, the 74181's mode control input (M) is set low. To select a logic operation, the mode control input is set high. Once you have decided whether you want to perform a logic or arithmetic operation, you apply a 4-bit code to the mode select inputs (S_0, S_1, S_2, and S_3) to specify which specific operation, as indicated within the truth table, is to be performed. For example, if you select $S_3 = 1$, $S_2 = 1$, $S_1 = 1$, $S_0 = 0$, while $M = 1$, then you get $F_0 = A_0 + B_0$, $F_1 = A_1 + B_1$, $F_2 = A_2 + B_2$, $F_3 = A_3 + B_3$. Note that the + shown in the truth table does not represent addition; it is used to represent the OR function. For addition, the tables use "plus." Carry-in ($\overline{C}_N$) and carry-out (C_{N+4}) leads are provided for use in arithmetic operations. All arithmetic results generated by this device are in 2's complement notation.

12.3.7 Comparators and Magnitude Comparator ICs

A digital comparator is a circuit that accepts two binary numbers and determines whether the two numbers are equal. For example, Fig. 12.49 shows a 1-bit and a 4-bit comparator. The 1-bit comparator outputs a high (1) only when the two 1-bit numbers A and B are equal. If A is not equal to B, then the output goes low (0). The 4-bit is basically four 1-bit comparators in one. When all individual digits of each number are equal, all XOR gates output a high, which in turn enables the AND gate, making the output high. If any two corresponding digits of the two numbers are not equal, the output goes low.

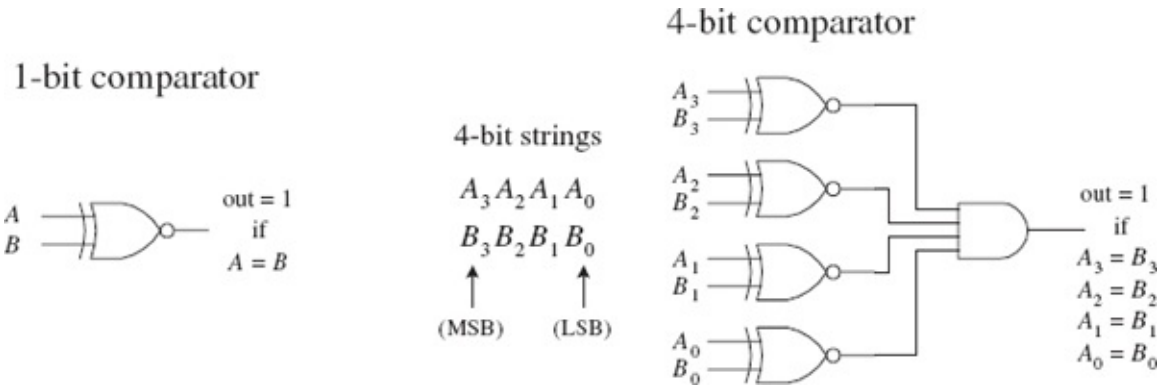


FIGURE 12.49

Now, say you want to know which number, A or B, is larger. The circuits in Fig. 12.49 will not do the trick. What you need instead is a *magnitude comparator* like the 74HC85 shown in Fig. 12.50. This device not only tells you if two numbers are equal, but also which number is larger. For example, if you

apply a 1001 (9_{10}) to the $A_3A_2A_1A_0$ inputs and a second number 1100 (12_{10}) to the $B_3B_2B_1B_0$ inputs, the $A < B$ output will go high (the other two outputs, $A > B$ and $A = B$, will remain low). If A and B were equal, the $A = B$ output would have gone high, and so on. If you wanted to compare larger numbers—say, two 8-bit numbers—you simply cascade two 74HC85 comparators together, as shown on the right side of Fig. 12.50. The leftmost 74HC85 compares the lower-order bits, while the rightmost 74HC85 compares the higher-order bits. To link the two devices together, you connect the output of the lower-order device to the expansion inputs of the higher-order device, as shown. The lower-order device's expansion inputs are always set low ($I_A < B$), high ($I_A = B$), and low ($I_A > B$).

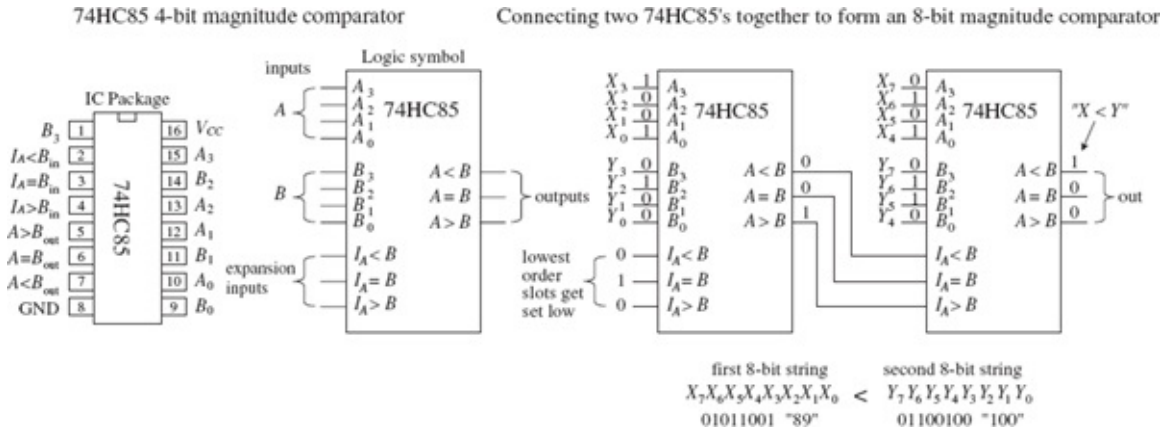


FIGURE 12.50

12.3.8 Parity Generator/Checker

Often, external noise can corrupt binary information (cause a bit to flip from one logic state to the other) as it travels along a conductor from one device to the next. For example, in the 4-bit system shown in Fig. 12.51, a BCD 4 (0100) picks up noise and becomes 0101 (or 5) before reaching its destination. Depending on the application, this type of error could lead to some serious problems.

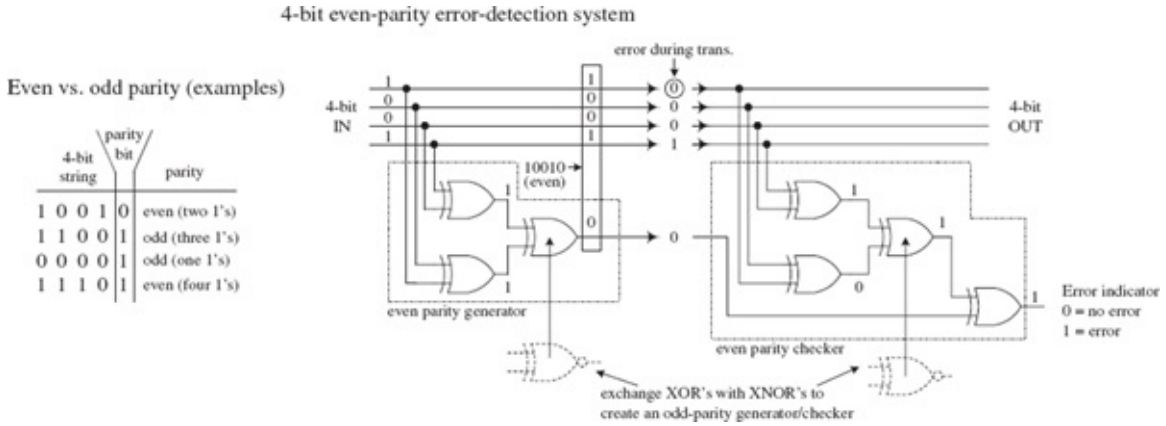


FIGURE 12.51

To avoid problems caused by unwanted data corruption, a parity generator/checker system, like the one shown in Fig. 12.51, can be used. The basic idea is to add an extra bit, called a parity bit, to the digital information being transmitted. If the *parity bit* makes the sum of all transmitted bits (including the parity bit) odd, the transmitted information is of odd parity. If the parity bit makes the sum even, the transmitted information is of even parity. A parity generator circuit creates the parity bit, while the parity checker on the receiving end determines if the information sent is of the proper parity. The type of parity (odd or even) is agreed to beforehand, so the parity checker knows what to look for. The parity bit can be placed next to the MSB or the LSB, provided the device on the receiving end knows which bit is the parity bit

and which bits are the data. The arrangement shown in Fig. 12.51 is designed with an even-parity error-detection system.

If you want to avoid building parity generators and checkers from scratch, use a parity generator/checker IC like the 74F280 9-bit odd-even parity generator/checker shown in Fig. 12.52. To make a complete error-detection system, two 74F280s are used: one acts as the parity generator, and the other acts as the parity checker. The generator's inputs *A* through *H* are connected to the eight data lines of the transmitting portion of the circuit. The ninth input (I) is grounded when the device is used as a generator. If you want to create an odd-parity generator, you tap the Σ_{odd} output; for even parity, you tap Σ_{even}. The 74F280 checker taps the main line at the receiving end and also accepts the parity bit line at input *I*. Figure 12.52 shows an odd-parity error-detection system used with an 8-bit system. If an error occurs, a high (1) is generated at the Σ_{odd} output.

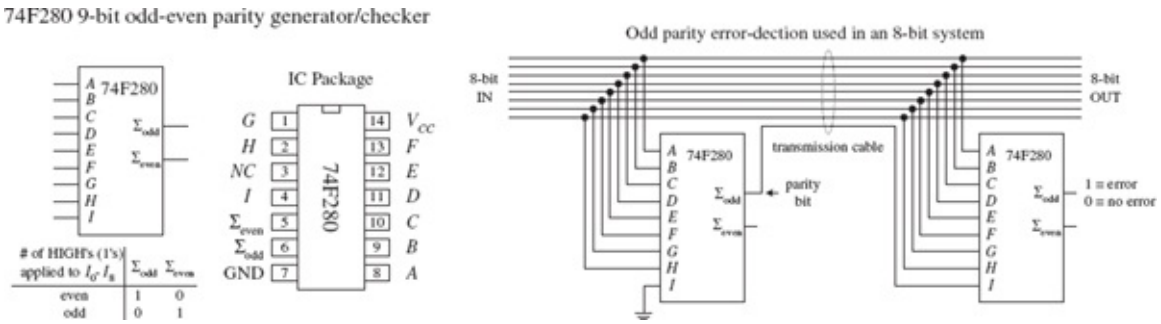


FIGURE 12.52

12.3.9 A Note on Obsolescence and the Trend Toward Microcontroller Control

We have just covered most of the combinational devices you will find discussed in textbooks and listed within electronic catalogs. Many of these devices are still used. However, some devices, such as the binary adders and code converters, are obsolete.

Today, the trend is to use software-controlled devices such as microprocessors and microcontrollers to carry out arithmetic operations and code conversions. Before you attempt to design any logic circuit, I suggest jumping to Chapter 13, which covers microcontrollers.

Microcontrollers can be used to collect data, store data, and perform logical operations using the input data. They also can generate output signals that can be used to control displays, audio devices, stepper motors, servos, and so on. The specific functions a microcontroller is designed to perform depend on the program you store in its internal ROM-type memory.

Programming the microcontroller typically involves simply using a special programming unit provided by the manufacturer. The programming unit usually consists of a prototyping platform that is linked to a host computer (via a USB port) that is running a development environment. In the development environment, you typically write out a program in a high-level language such as C, or some other specialized language designed for a certain microcontroller, and then, with the press of a key, the program is converted into machine language (1s and 0s) and downloaded into the microcontroller's memory.

In many applications, a single microcontroller can replace entire logic circuits composed of numerous discrete components. For this reason, it is tempting to skip the rest of this chapter and go directly to the chapter on microcontrollers. However, there are three basic problems with this approach:

- If you are a beginner, you will miss out on many important principles behind digital control that are most easily understood by learning how the discrete components work.
- Many digital circuits that you can build simply do not require the amount of sophistication a microcontroller provides.

- You may feel intimidated by the electronics catalogs that list every conceivable component available (even those that are obsolete). Knowing what's out there and knowing what to avoid are also important parts of the learning process.

12.4 Logic Families

Before moving on to sequential logic, let's touch on a few practical matters regarding the various logic families available and what kind of operating characteristics these families have. In this section, you will also encounter unique logic gates that have open-collector output stages and logic gates that have Schmitt-triggered inputs.

The key ingredient within any integrated logic device—a logic gate, a multiplexer, or a microprocessor—is the transistor. The kinds of transistors used within the IC, to a large extent, specify the type of logic family. The two most popular transistors used in ICs are bipolar and MOSFET transistors.

In general, ICs made from MOSFET transistors use less space due to their simpler construction, have very high noise immunity, and consume less power than equivalent bipolar transistor ICs. However, the high-input impedance and input capacitance of the MOSFET transistors (due to their insulated gate leads) result in longer time constants for transistor on/off switching speeds when compared with bipolar gates, and therefore typically result in a slower device. Over years of development, however, the performance gap between these two technologies has narrowed considerably.

Both the bipolar and MOSFET logic families can be divided into a number of subclasses. The major subclasses of the bipolar family include transistor-transistor logic (TTL), emitter-coupled logic (ECL), and integrated-injection logic (IIL or I^2L). The major subclasses of the MOSFET logic include P-channel MOSFET (PMOS), N-channel MOSFET (NMOS), and complementary MOSFET (CMOS). CMOS uses both NMOS and PMOS technologies. The two most popular technologies are TTL and CMOS. The other technologies are typically used in large-scale integration devices, such as microprocessors and memories. There are new technologies popping up all the time, which yield faster, more energy-efficient devices. Some examples include BiCMOS, GaAS, SOS, and Josephson junction technologies.

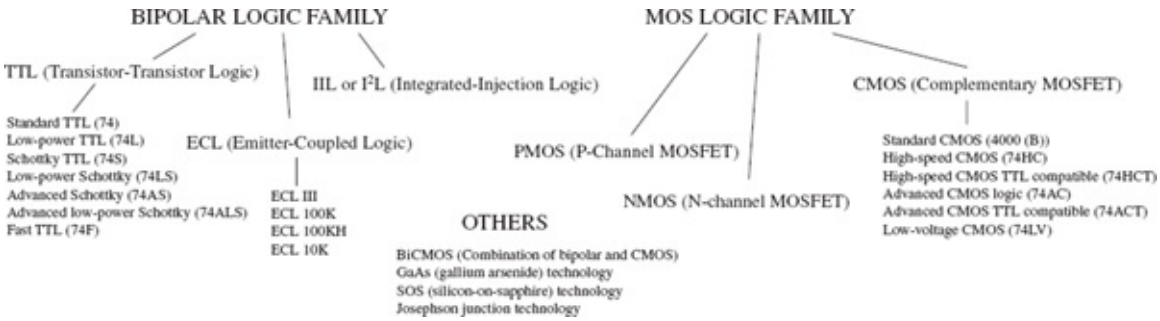
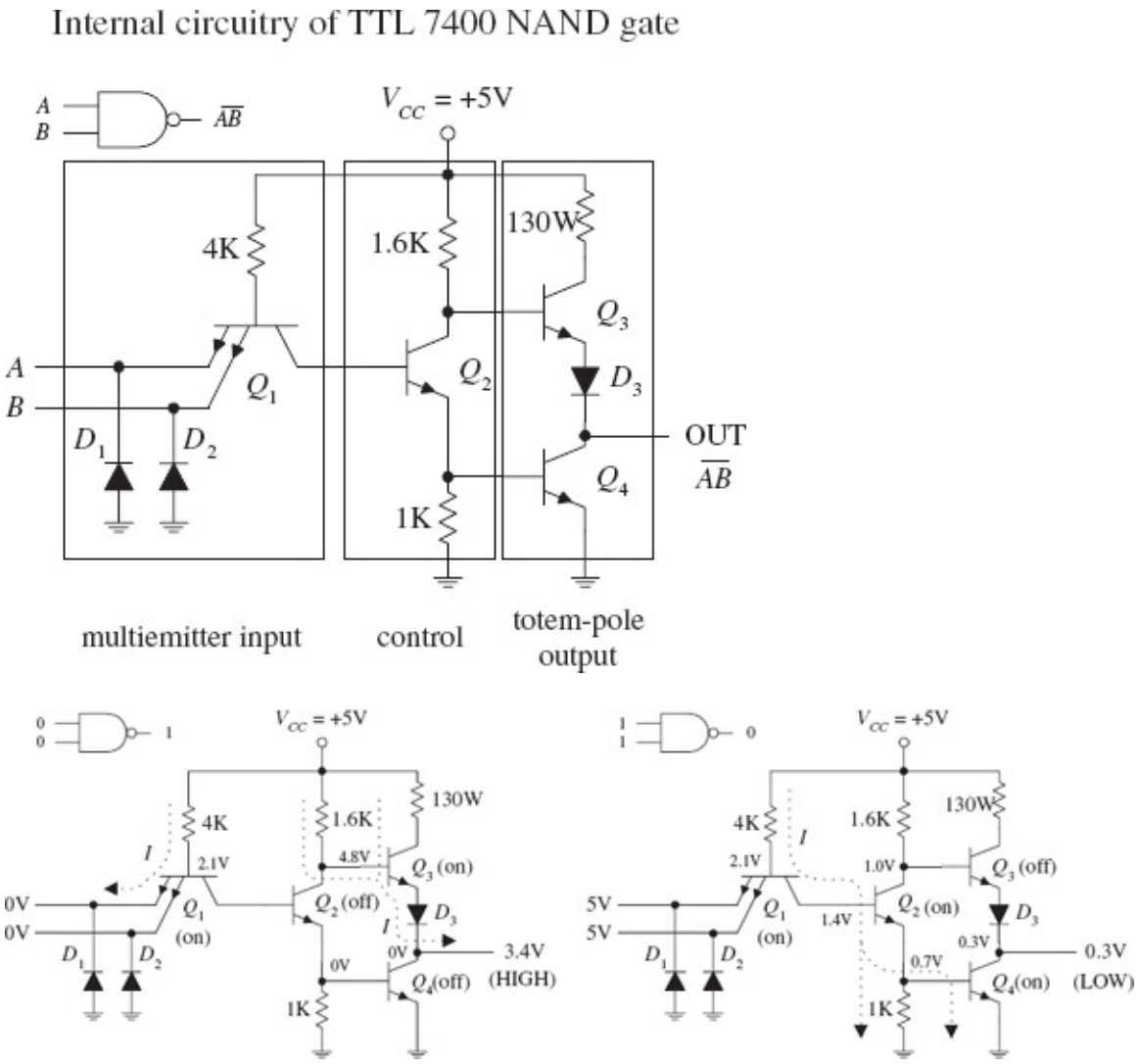


FIGURE 12.53

As you have already learned, TTL and CMOS devices are grouped into functional categories that get placed into either the 7400 series (74F, 74LS, 74HC (CMOS), and so on) or 4000 CMOS series (or the improved 4000B series). Another series you will run into is the 5400 series. This series is essentially equivalent to the 7400 series (with the same pinouts and same basic logic function), but it is a more expensive chip because it is designed for military applications that require increased supply voltage tolerances and temperature tolerances. For example, a 7400 IC typically has a supply voltage range from 4.75 to 5.25 *V* with a temperature range from 0 to 70°C. A 5400 IC typically has a voltage range between 4.5 and 5.5 *V* and a temperature range from -55 to 125°C.

12.4.1 TTL Family of ICs

The original TTL series, referred to as the *standard TTL* series (74xx), was developed early in the 1960s. This series is still in use, even though its overall performance is inferior to the newer line of TTL devices, such as the 74LSxx, 74ALSxx, and 74Fxx. The internal circuitry of a standard TTL 7400 NAND gate, along with a description of how it works, is provided next.



The TTL NAND gate is divided into three basic sections: multi-emitter input, control section, and totem-pole output stage. In the multi-emitter input section, a multi-emitter bipolar transistor Q_1 acts like a two-input AND gate, while diodes D_1 and D_2 act as negative clamping diodes used to protect the inputs from any short-term negative input voltages that could damage the transistor. Q_2 provides control and current boosting to the totem-pole output stage; when the output is high (1), Q_4 is off (open) and Q_3 is on (short). When the output is low (0), Q_4 is on and Q_3 is off. Because one or the other transistor is always off, the current flow from V_{CC} to ground in that section of the circuit is minimized. The lower figures show both a high and low output state, along with the approximate voltages present at various locations. Notice that the actual output voltages are not exactly 0 or +5 V—a result of internal voltage drops across resistor, transistor, and diode. Instead, the outputs are around 3.4 *V* for high and 0.3 *V* for low. Note that to create, say, an eight-input NAND gate, the multi-emitter input transistor would have eight emitters instead of just two as shown.

FIGURE 12.54

A simple modification to the standard TTL series was made early on by reducing all the internal resistor values in order to reduce the *RC* time constants and thus increase the speed (reduce propagation delays). This improvement to the original TTL series marked the 74H series. Although the 74H series offered improved speed (about twice as fast) over the 74 series, it had more than double the power consumption. Later, the 74L series emerged. Unlike the 74H, the 74L took the 74 and increased all internal resistances. The net effect led to a reduction in power but increased propagation delay.

A significant impr
series (Schottky TTL
collector junctions
charge buildup in th
diodes were the bes
effect was an increas

Continually over t
resistors, more powe
third the power dissi
emerged, which had
series, or FAST logi
ALS series) that led

Today you will fi
ultimately depends o

12.4.2 CMOS F

While the TTL series
The original CMOS
consumption than th
characteristics of its
to 18 V), with minin
more energy efficien
to electrostatic disch
gates. To figure out h
which transistor gate

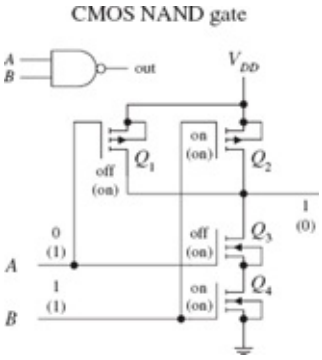


FIGURE 12.55

A further improv
40H00 series. Altho
TTL series. The 74C
compatible with the

Another significar
74HCT series. Both
74HC (high-speed C
power consumption.
interchangeable with
popular today.

12.10.2 Using Logic to Drive External Loads

Driving simple loads such as LEDs, relays, buzzers, or any device that assumes either an on or off state is relatively simple. When driving such loads, it is important to first check the driving logic's current specifications—how much current, say, a gate can sink or source. After that, you determine how much current the device to be driven will require. If the device draws more current than the logic can source or sink, a high-power transistor typically can be used as an output switch. Figure 12.137 shows some sample circuits used to drive various loads.

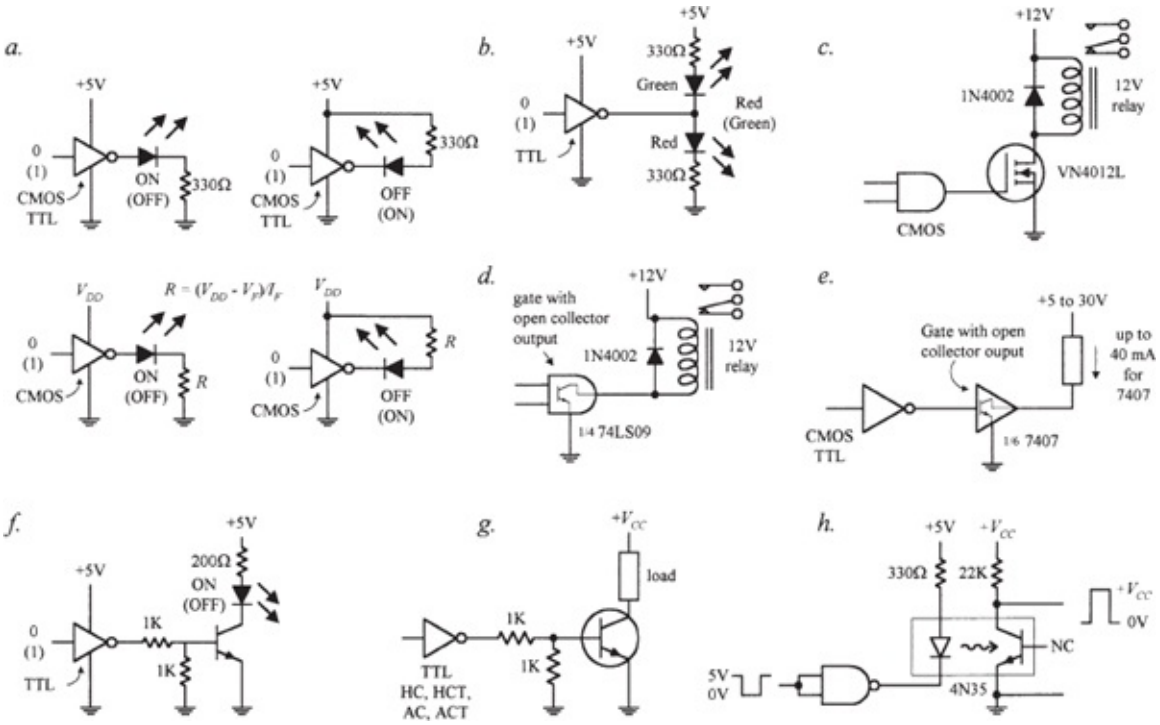


FIGURE 12.137

In Fig. 12.137a, LEDs can be driven directly by logic through a current-limiting resistor. Current can be sunk or sourced as shown, although it is better to sink when using TTL. Because an LED's forward voltage drop (V_F) and forward current (I_F) can vary from 1.0 to 3.6 *V* and 1 to 20 mA, respectively, it is best to select current-limiting resistors that obey the equation shown in the figure. If an LED requires more current than the logic can supply or sink, a transistor output stage like the one shown in Fig. 12.137f can be used.

Figure 12.137b shows a simple way to get dual-lighting action from a pair of LEDs. When the gate's output goes low, the upper green LED turns on, while the lower red LED turns off. The LEDs switch states when the output goes high.

Relays will draw considerable current. To avoid damaging the logic device, in Fig. 12.137c, a power MOSFET transistor is attached to the logic output. The diode is used to protect the circuit from current spikes generated by relay as it switches states.

A handy method for interfacing standard logic with loads is to use a gate with an open-collector output as a go-between. Recall that open-collector gates cannot source current; they can only sink current. However, they typically can sink ten times the current of a standard logic gate. In Fig. 12.137d, an open-collector gate is used to drive a relay. Check the current ratings of specific open-collector devices before using them to be sure they can handle the load current.

Figure 12.137e shows another open-collector application. In Fig. 12.137f, a bipolar transistor is used to increase the output drive current used to drive a high-current LED. Make sure the transistor is of the proper current rating.

Figure 12.137g is basically the same as the previous example, but the load can be something other than an LED.

In Fig. 12.137h, an optocoupler is used to drive a load that requires electrical isolation from the logic driving it. Electrical isolation is often used in situations where external loads use a separate ground system. The voltage level at the load side of the optical interface can be set via V_{CC}. There are many different types of optocouplers available (see Chapter 5).

12.10.3 Analog Switches

Analog switches are ICs designed to switch analog signals via digital control. The internal structure of these devices typically consists of a number of logic control gates interfaced with transistor stages used to control the flow of analog signals.

Figure 12.138 shows various types of analog switches. The CMOS 4066B quad bilateral switch uses a single-supply voltage from 3 to 15 V. It can switch analog or digital signals within ±7.5 V and has a maximum power dissipation of around 700 mW. Individual switches are controlled by digital inputs *A* through *D*. The TTL-compatible AH0014D DPDT analog switch can switch analog signals of ±10 V via the *A* and *B* logic control inputs. Note that this device has separate analog and digital supplies: *V*+ and *V*− are analog; V_{CC} and GND are digital. The DG302A dual-channel CMOS DPST analog switch can switch analog signals within the ±10-V range at switching speeds up to 15 ns.

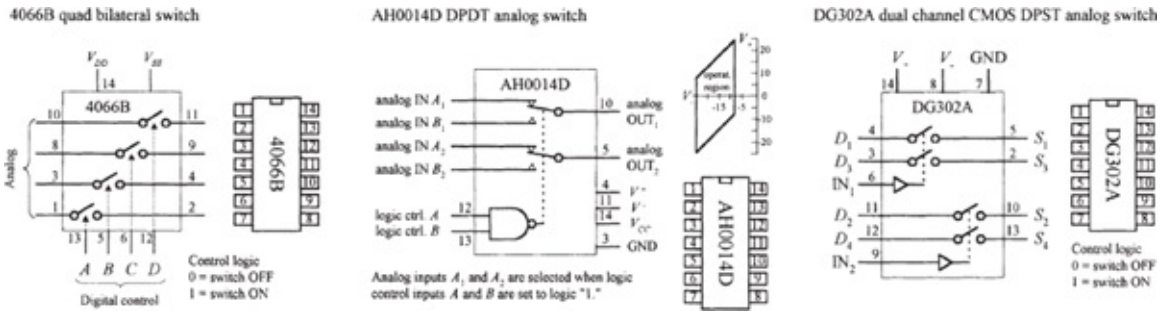


FIGURE 12.138

A number of circuits use analog switches. They are found in modulator/demodulator circuits, digitally controlled frequency circuits, analog signal-gain circuits, and analog-to-digital conversion circuits, where they often act as sample-hold switches. They can, of course, be used simply to turn a given analog device on or off.

12.10.4 Analog Multiplexer/Demultiplexer

Recall from Section 12.3 that a digital multiplexer acts like a data selector, while a digital demultiplexer acts like a data distributor. Analog multiplexers and demultiplexers act the same way but are capable of selecting or distributing analog signals. (They still use digital select inputs to select which pathways are open and which are closed to signal transmission.)

A popular analog multiplexer/demultiplexer IC is the 4051B, shown in Fig. 12.139. This device functions as either a multiplexer or demultiplexer, since its inputs and outputs are bidirectional (signals can flow in either direction). When used as a multiplexer, analog signals enter through I/O lines 0 through 7, while the digital code that selects which input is passed to the analog O/I line (pin 3) is applied to digital inputs *A*, *B*, and *C*. See the truth table in the figure. When used as a demultiplexer, the connections are reversed: The analog input comes in through the analog O/I line (pin 3) and passes out through one of the seven analog I/O lines. The specific output is again selected by the digital inputs *A*, *B*, and *C*. Note that

when the inhibit line (INH) is high, none of the addresses are selected.

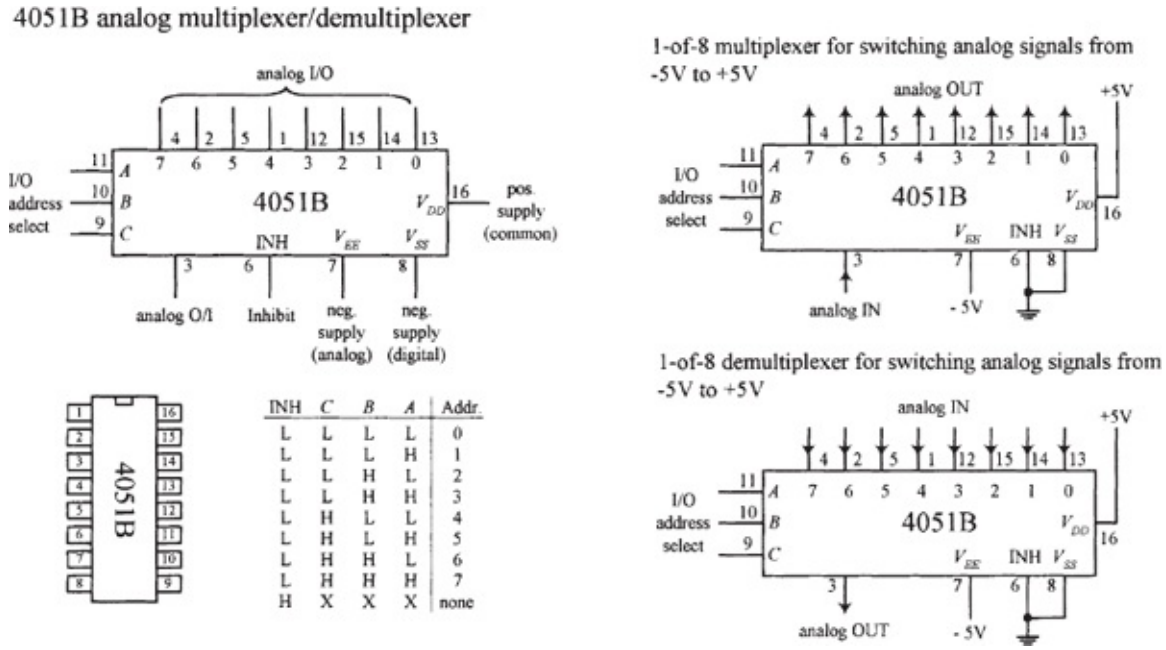


FIGURE 12.139

The I/O analog voltage levels for the 4051B are limited to a region between the positive supply voltage V_{DD} and the analog negative supply voltage V_{EE}. Note that the V_{SS} supply is grounded. If the analog signals you are planning to use are all positive, V_{EE} and V_{SS} can both be connected to a common ground. However, if you plan to use analog voltages that range from, say, −5 to +5 V, V_{EE} should be set to −5 V, while V_{DD} should be set to +5 V. The 4051B accepts digital signals from 3 to 15 V, while allowing for analog signals from −15 to +15 V.

12.10.5 Analog-to-Digital and Digital-to-Analog Conversion

In order for analog devices (speakers, temperature sensors, strain gauges, position sensors, light meters, and so on) to communicate with digital circuits in a manner that goes beyond simple threshold triggering, we use an analog-to-digital converter (ADC). An ADC converts an analog signal into a series of binary numbers, each number proportional to the analog level measured at a given moment. Typically, the digital words generated by the ADC are fed into a microprocessor or microcontroller, where they can be processed, stored, interpreted, and manipulated. Analog-to-digital conversion is used in data-acquisition systems, digital sound recording, and within simple digital display test instruments (such as light meters and thermometers).

In order for a digital circuit to communicate with the analog world, we use a digital-to-analog converter (DAC). A DAC takes a binary number and converts it to an analog voltage that is proportional to the binary number. By supplying different binary numbers, one after the other, a complete analog waveform is created. DACs are commonly used to control the gain of an op amp, which in turn can be used to create digitally controlled amplifiers and filters. They are also used in waveform generator and modulator circuits and as trimmer replacements, and are found in a number of process-control and autocalibration circuits.

Many digital consumer products such as MP3 players, DVDs, and CD players use digital signal processing ADCs and DACs.

ADC and DAC Basics

Figure 12.140 shows the basic idea behind analog-to-digital and digital-to-analog conversion. In the analog-to-digital figure, the ADC receives an analog input signal along with a series of digital sampling pulses. Each time a sampling pulse is received, the ADC measures the analog input voltage and outputs a 4-bit binary number that is proportional to the analog voltage measured during the specific sample. With 4 bits, we get 16 binary codes (0000 to 1111) that correspond to 16 possible analog levels (for example, 0 to 15 V).

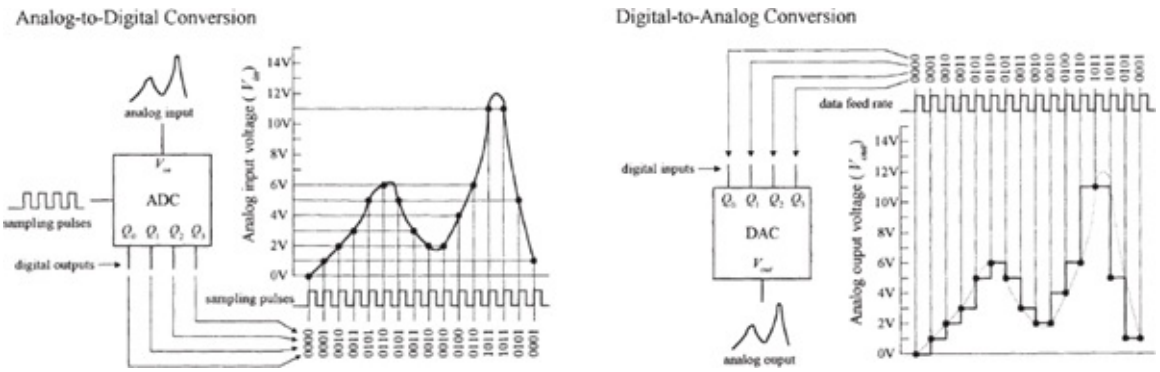


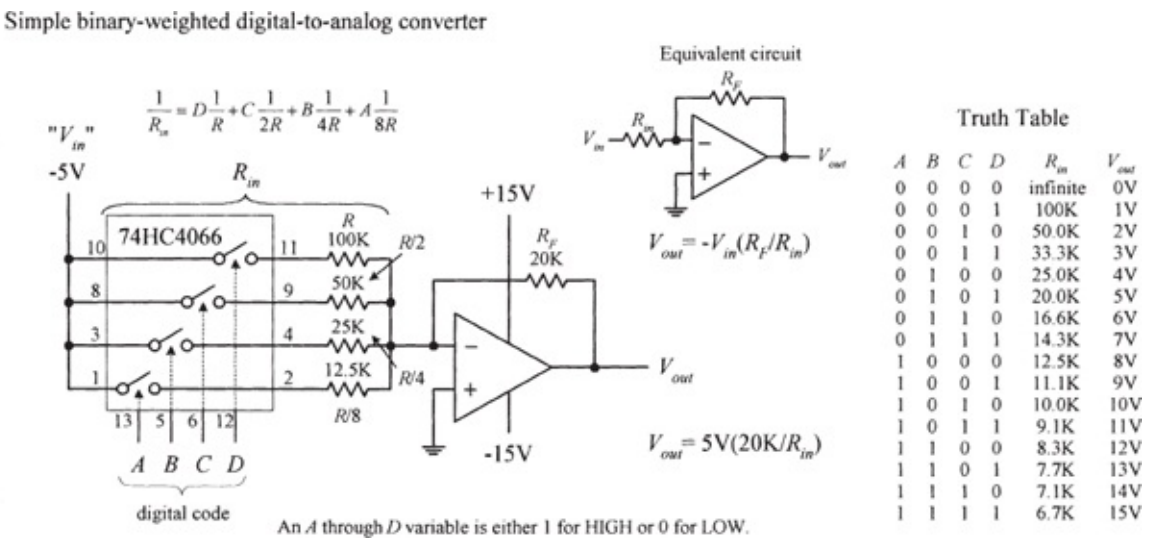
FIGURE 12.140

In the digital-to-analog conversion figure, the DAC receives a series of 4-bit binary numbers. The rate at which new binary numbers are fed into the DAC is determined by the logic that generates them. With each new binary number, a new analog voltage is generated. As with the ADC example, we have a total of 16 binary numbers to work with and 16 possible output voltages.

As you can see from the graphs, both these 4-bit converters lack the resolution needed to make the analog signal appear continuous (without steps). To make things appear more continuous, a converter with higher resolution is used. This means that instead of using 4-bit binary numbers, we use larger-bit numbers, such as 6-bit, 8-bit, 10-bit, 12-bit, 16-bit, or even 18-bit or higher numbers. If our converter has a resolution of 8 bits, we have $2^8 = 256$ binary numbers to work with, along with 256 analog steps. Now, if this 8-bit converter is set up to generate 0 V at binary 00000000 and 15 V at binary 11111111 (full scale), then each analog step is only 0.058 V high ($\frac{1}{256} \times 15$ V). With an 18-bit converter, the steps get incredibly tiny because we have $2^{18} = 262{,}144$ binary numbers and steps. With 0 V corresponding to binary 000000000000000000 and 15 V corresponding to 111111111111111111, the 18-bit converter yields steps that are only 0.000058 V high! As you can see in the 18-bit case, the conversion process between digital and analog appears practically continuous.

Simple Binary-Weighted DAC

Figure 12.141 shows a simple 4-bit DAC that is constructed from a digitally controlled switch (74HC4066), a set of binary-weighted resistors, and an operational amplifier. The basic idea is to create an inverting amplifier circuit whose gain is controlled by changing the input resistance R_{in}. The 74HC4066 and the resistors together act as a digitally controlled R_{in} that can take on one of 16 possible values. You can think of the 74HC4066 and resistor combination as a digitally controlled current source. Each new binary code applied to the inputs of the 74HC4066 generates a new discrete current level that is summed by R_F to provide a new discrete output voltage level.



A	B	C	D	R_{in}	V_{out}
0	0	0	0	infinite	0V
0	0	0	1	100K	1V
0	0	1	0	50.0K	2V
0	0	1	1	33.3K	3V
0	1	0	0	25.0K	4V
0	1	0	1	20.0K	5V
0	1	1	0	16.6K	6V
0	1	1	1	14.3K	7V
1	0	0	0	12.5K	8V
1	0	0	1	11.1K	9V
1	0	1	0	10.0K	10V
1	0	1	1	9.1K	11V
1	1	0	0	8.3K	12V
1	1	0	1	7.7K	13V
1	1	1	0	7.1K	14V
1	1	1	1	6.7K	15V

FIGURE 12.141

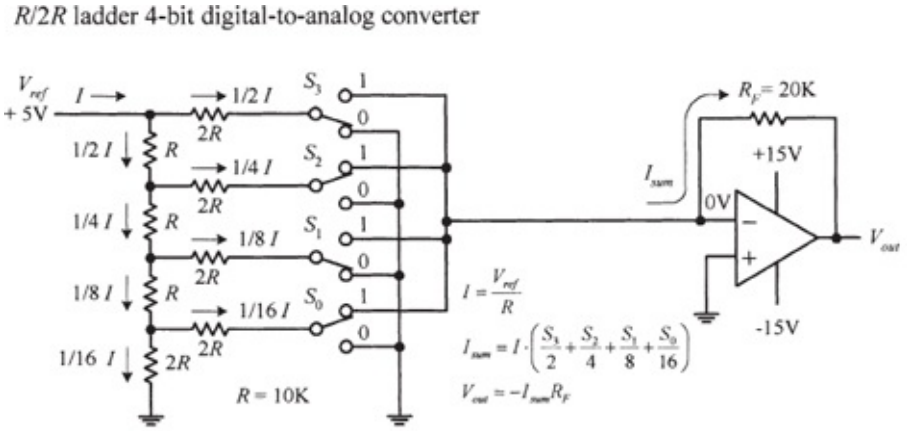
We choose scaled resistor values of R, $R/2$, $R/4$, and $R/8$ to give R_{in} discrete values that are equally spaced. To find all possible values of Rin, we use the formula provided in Fig. 12.141. This formula looks like the old resistors-in-parallel formula, but we must exclude those resistors that are not selected by the digital input code—that's what the coefficients A through D are for (a coefficient is either 1 or 0, depending on the digital input).

To find the analog output voltage, we simply use $V_{out} = -V_{ref}(R_F/R_{in})$—the expression used for the inverting amplifier (see Chapter 8). Figure 12.141 shows what we get when we set $V_{ref} = -5$ V, $R = 100$ kΩ, and $R_F = 20$ kΩ, and take all possible input codes.

The binary-weighted DAC shown in Fig. 12.141 is limited in resolution (4-bit, 16 analog levels). To double the resolution (make an 8-bit DAC), you might consider adding another 74HC4066 and $R/16$, $R/32$, $R/64$, and $R/128$ resistors. In theory, this works; in reality, it doesn't. The problem with this approach is that when we reach the $R/128$ resistor, we must find a 0.78125-kΩ resistor, assuming $R = 100$ kΩ. Assuming we can find or construct an equivalent resistor network for $R/128$, we're still in trouble because the tolerances of these resistors will cause problems. This scaled-resistor approach becomes impractical when we deal with resolutions of more than a few bits. To increase the resolution, we scrap the scaled-resistor network and replace it with an $R/2R$ ladder network. The manufacturers of DAC ICs do this as well.

R/2R Ladder DAC

An $R/2R$ DAC uses an $R/2R$ resistor ladder network instead of a scaled-resistor network, as was the case in the previous DAC. The benefit of using the $R/2R$ ladder is that we need only two resistor values: R and $2R$. Figure 12.142 shows a simple 4-bit $R/2R$ DAC. For now, assume that the switches are digitally controlled (in real DACs, they are replaced with transistors).



S_3	S_2	S_1	S_0	V_{out}
0	0	0	0	0.000V
0	0	0	1	-0.625V
0	0	1	0	-1.250V
0	0	1	1	-1.875V
0	1	0	0	-2.500V
0	1	0	1	-3.125V
0	1	1	0	-3.750V
0	1	1	1	-4.375V
1	0	0	0	-5.000V
1	0	0	1	-5.625V
1	0	1	0	-6.250V
1	0	1	1	-6.875V
1	1	0	0	-7.500V
1	1	0	1	-8.125V
1	1	1	0	-8.750V
1	1	1	1	-9.375V

FIGURE 12.142

The trick to understanding how the *R*/2*R* ladder works is realizing that the current drawn through any one switch is always the same, no matter if it is thrown up or down. If a switch is thrown down, current will flow through the switch into ground (0 V). If a switch is thrown up, current will flow toward virtual ground—located at the op amp's inverting input (recall that if the noninverting input of an op amp is set to 0 V, the op amp will make the inverting input 0 V, via negative feedback). Once you realize that the current through any given switch is always constant, you can figure that the total current (*I*) supplied by V_{ref} will be constant as well. Once you have that, you figure out what fractions of the total current pass through each of the branches within the *R*/2*R* network using simple circuit analysis. Figure 12.142 shows that ½*I* passes through S_3 (MSB switch), ¼*I* through S_2, ⅛ *I* through S_1 and $\frac{1}{16}I$ through S_0 (LSB switch). If you're interested in how that was figured out, the circuit reduction shown in Fig. 12.143 should help.

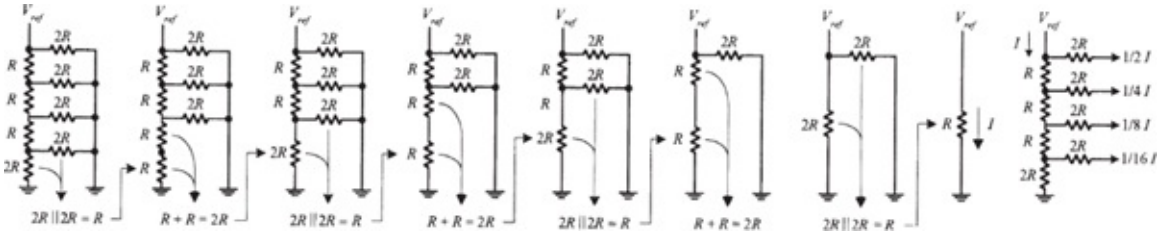


FIGURE 12.143

Now that we have a means of consistently generating fractions of ½*I*, ¼*I*, ⅛ *I*, and $\frac{1}{16}I$, we can choose, via the digital input switches, which fractions are summed together by the amplifier. For example, if switches S_3, S_2, S_1, and S_0 are thrown to 0101 (5), ¼*I* + $\frac{1}{16}I$ combine to form I_{sum}. But what is *I*? Using Ohm's law, it's just $I = V_{ref} / R$ = +5 V / 10 kΩ = 500 µA. This means that I_{sum} = ¼(500 µA) + $\frac{1}{16}$(500 µA) = 156.25 µA. The final output voltage is determined by $V_{out} = -I_{sum}R_F$ = − (156.25 µA)(20 *k*Ω) = -3.125 V. The formulas and the table in Fig. 12.142 show the other possible binary/analog combinations.

To create an *R*/2*R* DAC with higher resolution, we simply add more runs and switches to the ladder.

Integrated DACs

Often, making DACs from scratch isn't worth the effort. The cost as well as the likelihood for conversion errors is great. The best thing to do is to simply buy a DAC IC. You can buy these devices from a number of different manufacturers (such as National Semiconductor, Analog Devices, and Texas Instruments). The typical resolutions for these ICs are 6, 8, 10, 12, 16, and 18 bits. DAC ICs also may come with a serial digital input, as opposed to the parallel input scheme shown in Figs 12.141 and 12.142. Before a serial-input DAC can make a conversion, the entire digital word must be clocked into an internal shift register.

Most often, DAC ICs come with an external reference input that is used to set the analog output range.

There are some DACs that have fixed references, but these are becoming rare.

Often, you'll see a manufacturer list one of its DACs as being a multiplying DAC. A multiplying DAC can produce an output signal that is proportional to the product of a varying input reference level (voltage or current) times a digital code. As it turns out, most DACs, even those that are specifically designated as multiplying DAC on the data sheets, can be used for multiplying purposes simply by using the reference input as the analog input. However, many such ICs do not provide the same quality multiplying characteristics, such as a wide analog input range and fast conversion times, as those that are called *multiplying* DACs.

Multiplying is most commonly applied in systems that use ratiometeric transducers (for example, position potentiometers, strain gauges, and pressure transducers). These transducers require an external analog voltage to act as a reference level on which to base analog output responses. If this reference level is altered, say, by an unwanted supply surge, the transducer's output will change in response, and this results in conversion errors at the DAC end. However, if we use a multiplying DAC, we eliminate these errors by feeding the transducer's reference voltage to the DAC's analog input. If any supply voltage/current errors occur, the DAC will alter its output in proportion to the analog error.

DACs are capable of producing unipolar (single-polarity output) or bipolar (positive and negative) output signals. In most cases, when a DAC is used in unipolar mode, the digital code is expressed in standard binary. When used in bipolar mode, the most common code is either offset binary or 2's complement. Offset binary and 2's complement codes make it possible to express both positive and negative values. Figure 12.144 shows all three codes and their corresponding analog output levels (referenced from an external voltage source).

Common Digital Codes Used by DACs

	Unipolar Operation		Bipolar Operation					
	Binary	Analog Output		Offset Binary	Analog Output		2's Comp.	Analog Output
FS	1111 1111	$V_{ref}\left(\frac{255}{256}\right)$	FS	1111 1111	$+V_{ref}\left(\frac{127}{128}\right)$	FS	0111 1111	$+V_{ref}\left(\frac{127}{128}\right)$
FS-1	1111 1110	$V_{ref}\left(\frac{254}{256}\right)$	FS-1	1111 1110	$+V_{ref}\left(\frac{126}{128}\right)$	FS-1	0111 1110	$+V_{ref}\left(\frac{126}{128}\right)$
	↓			↓			↓	
$\frac{FS}{2}$	1000 0000	$V_{ref}\left(\frac{128}{256}\right)=\frac{V_{ref}}{2}$	0 + 1LSB	1000 0001	$+V_{ref}\left(\frac{1}{128}\right)$	0 + 1LSB	0000 0001	$+V_{ref}\left(\frac{1}{128}\right)$
			0	1000 0000	$V_{ref}\left(\frac{0}{128}\right)=0$	0	0000 0000	$V_{ref}\left(\frac{0}{128}\right)=0$
LSB	0000 0001	$V_{ref}\left(\frac{1}{256}\right)$	0– 1LSB	0111 1111	$-V_{ref}\left(\frac{1}{128}\right)$	0– 1LSB	1111 1111	$-V_{ref}\left(\frac{1}{128}\right)$
LSB– 1	0000 0000	$V_{ref}\left(\frac{0}{256}\right)=0$		↓			↓	
			–FS+1	0000 0001	$-V_{ref}\left(\frac{127}{128}\right)$	–FS+1	1000 0001	$-V_{ref}\left(\frac{127}{128}\right)$
	(FS = full scale)		–FS	0000 0000	$-V_{ref}\left(\frac{128}{128}\right)$	–FS	1000 0000	$-V_{ref}\left(\frac{128}{128}\right)$

FIGURE 12.144

Note that in the figure, FS stands for full scale, which is the maximum analog level that can be reached when applying the highest binary code. It is important to realize that at full scale, the analog output for an n-bit converter is actually $(2^n - 1) / 2^n \times V_{ref}$, not $2^n/2^n \times V_{ref}$. For example, for an 8-bit converter, the number of binary numbers is $2^8 = 256$, while the maximum analog output level is 255/256 V_{ref}, not 256/256 V_{ref}, since the highest binary number is 255 (1111 1111). The "missing count" is used up by the LSB-1 condition (0 state).

Example DAC ICs

DAC0808 8-BIT DAC

The DAC0808 (National Semiconductor) is a popular 8-bit DAC that requires an input reference current and supplies 1 of 256 analog output current levels. Figure 12.145 shows a block diagram of the DAC0808, along with its IC pin configuration and a sample application circuit.

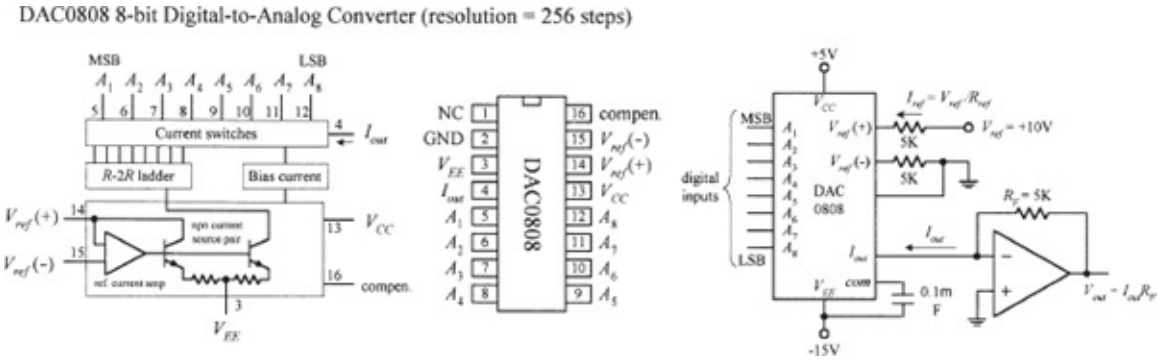


FIGURE 12.145

In the application circuit, the analog output range is set by applying a reference current (I_{ref}) to pin 14 ($+V_{ref}$). In this example, I_{ref} is set to 2 mA via an external +10 V/5 kΩ resistor combination. Note that another 5-kΩ resistor is required between pin 15 ($-V_{ref}$) and ground.

To determine the DAC's analog output current (I_{out}) for all possible binary inputs, we use the following formula:

$$I_{out} = I_{ref}\left(\frac{A_1}{2} + \frac{A_2}{4} + \ldots + \frac{A_8}{256}\right) = \frac{\text{decimal equivalent of input binary number}}{256}$$

At full scale (all A's high or binary 255), $I_{out} = I_{ref}$ (255/256) = (2 mA)(0.996) = 1.99 mA. Considering that the DAC has 256 analog output levels, we can figure that each corresponding level is spaced 1.99 mA/256 = 0.0078 mA apart.

To convert the analog output currents into analog output voltages, we attach the op amp. Using the op amp rules from Chapter 8, we find that the output voltage is $V_{out} = I_{out} \times R_f$. At full scale, V_{out} = (1.99 mA) (5 kΩ) = 9.95 V. Each analog output level is spaced 9.95 V/256 = 0.0389 V apart.

The DAC0808 can be configured as a multiplying DAC by applying the analog input signal to the reference input. In this case, however, the analog input current should be limited to a range from 16 μA to 4 mA to retain reasonable accuracy. See the National Semiconductor's data sheets for more details.

DAC8043A SERIAL 12-BIT INPUT MULTIPLYING DAC

The DAC8083A (Analog Devices) is a high-precision 12-bit CMOS multiplying DAC that comes with a serial digital input. Figure 12.146 shows a block diagram, pin configuration, and write cycle timing diagram for this device.

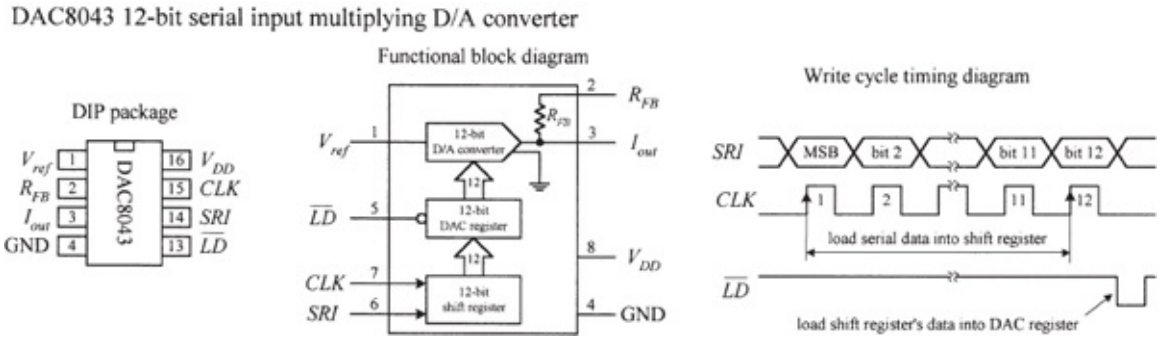
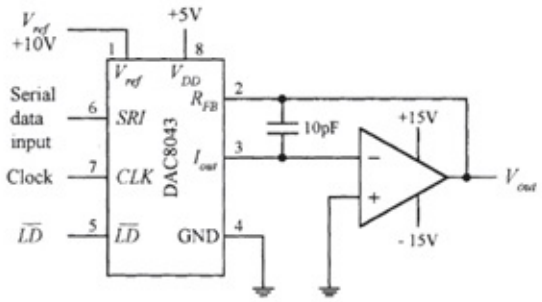


FIGURE 12.146

Before the DAC8043 can make a conversion, serial data must be clocked into the input register by supplying an external clock signal (each positive edge of the clock load one bit). Once loaded, the input register's contents are dumped off to the DAC register by applying a low pulse to the $\overline{LD}$ line. Data in the DAC register is then converted to an output current through the Iout terminal.

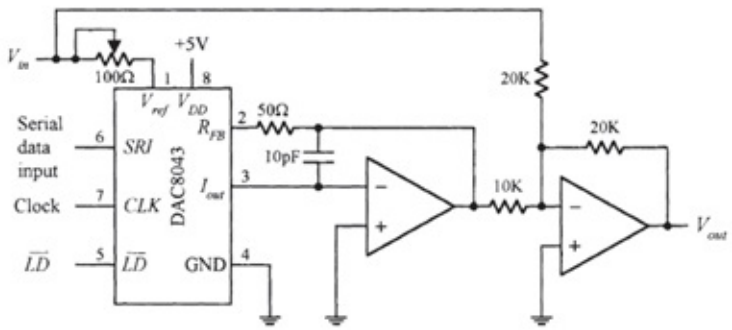
In most applications, this current is then transformed into a voltage by an op amp stage, as is the case within the two circuits shown in Fig. 12.147. In the unipolar (two-quadrant) circuit, a standard binary code is used to select from 4096 possible analog output levels. In the bipolar (four-quadrant) circuit, an offset binary code is used again to select from 4096 analog output levels, but now the range is broken up to accommodate both positive and negative polarities.

Unipolar Operation (2-Quadrant)



Digital Input (Binary)	Nominal Analog Output
1111 1111 1111	$-V_{ref}\left(\frac{4095}{4096}\right)$
1000 0000 0001	$-V_{ref}\left(\frac{2049}{4096}\right)$
1000 0000 0000	$-V_{ref}\left(\frac{2048}{4096}\right)=-\frac{V_{ref}}{2}$
0111 1111 1111	$-V_{ref}\left(\frac{2047}{4096}\right)$
0000 0000 0001	$-V_{ref}\left(\frac{1}{4096}\right)$
0000 0000 0000	$-V_{ref}\left(\frac{0}{4096}\right)=0$

Bipolar Operation (4-Quadrant)



Digital Input (Offset Binary)	Nominal Analog Output
1111 1111 1111	$+V_{ref}\left(\frac{2047}{2048}\right)$
1000 0000 0001	$+V_{ref}\left(\frac{1}{2048}\right)$
1000 0000 0000	
0111 1111 1111	$-V_{ref}\left(\frac{1}{2048}\right)$
0000 0000 0001	$-V_{ref}\left(\frac{2047}{2048}\right)$
0000 0000 0000	$-V_{ref}\left(\frac{2048}{2048}\right)$

FIGURE 12.147

If you're interested in learning more about the DAC8043, go to Analog Device's website and check out the data sheet.

Another very similar device worth considering is the MAX522.

12.10.6 Analog-to-Digital Converters

There are a number of techniques used to convert analog signals into digital signals. The most popular techniques include successive approximation conversion and parallel-encoded conversion (or flash conversion). Other techniques include half-flash conversion, delta-sigma processing, and pulse-code modulation (PCM). In this section, we'll focus on the successive approximation and parallel-encoded conversion techniques.

Successive Approximation

Successive approximation analog-to-digital conversion is the most common approach used in integrated ADCs. In this conversion technique, each bit of the binary output is found, one bit at a time—MSB first. This technique yields fairly fast conversion times (from around 10 to 300 μs) with a limited amount of circuitry. Figure 12.148 shows a simple 8-bit successive approximation ADC, along with an example analog-to-digital conversion sequence.

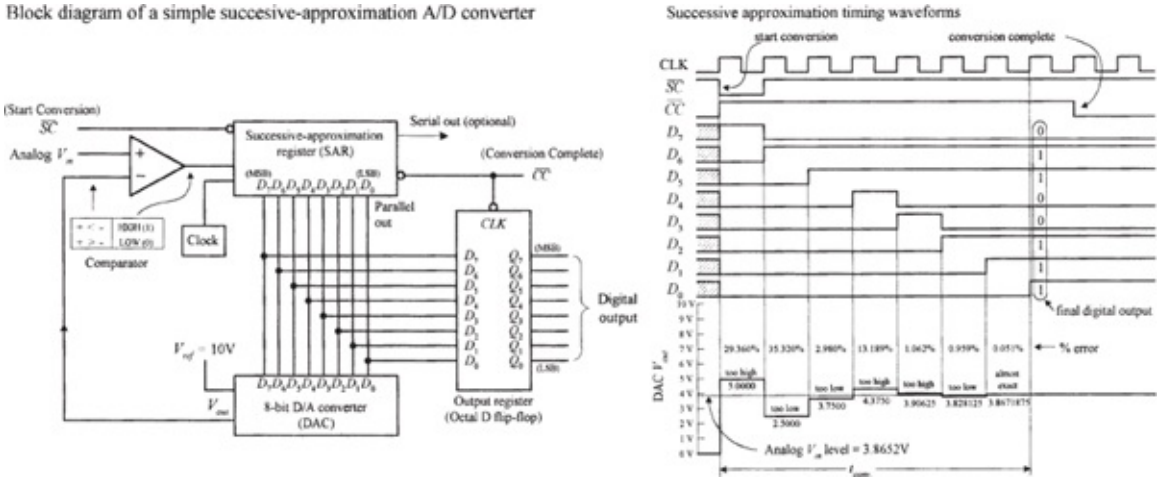


FIGURE 12.148

To begin a conversion, the $\overline{SC}$ (start conversion) input is pulsed low. This causes the successive approximation register (SAR) to first apply a high on the MSB (D_7) line of the DAC. With only D_7 high, the DAC's output is driven to one-half its full-scale level, which in this case is +5 *V* because the full-scale output is +10 V. The +5-V output level from the DAC is then compared with the analog input level, via the comparator. If the analog input level is greater than +5 V, the SAR keeps the D_7 line high; otherwise, the SAR returns the D_7 line low. At the next clock pulse, the next bit (D_6) is tried. Again, if the analog input level is larger than the DAC's output level, D_6 is left high; otherwise, it is returned low.

During the next six clock pulses, the rest of the bits are tried. After the last bit (LSB) is tried, the *CC* (conversion complete) output of the SAR goes low, indicating that a valid 8-bit conversion is complete, and the binary data is ready to be clocked into the octal flip-flop, where it can be presented to the Q_0–Q_7 outputs.

The timing diagram shows a 3.8652-V analog level being converted into an approximate digital equivalent. Note that after the first approximation (the D_7 try), the percentage error between the actual analog level and corresponding digital equivalent is 29.360 percent. However, after the final approximation, the percentage error is reduced to only 0.051 percent.

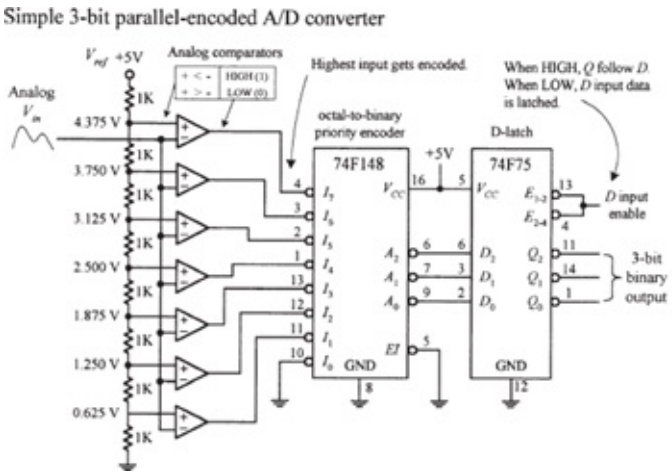
Until now, we've assumed that the analog input to our ADC was constant during the conversion. But what happens when the analog input changes during conversion time? Errors result. The more rapidly the analog input changes during the conversion time, the more pronounced the errors will become. To prevent such errors, a sample-and-hold circuit is often attached to the analog input. With an external control

signal, this circuit can be made to sample the analog input voltage and hold the sample while the ADC makes the conversion.

With the exception of very high-speed ADCs, separate ADC ICs are now largely redundant and have been replaced with microcontrollers containing 12-bit or higher ADC channels.

Parallel-Encoded Analog-to-Digital Conversion (Flash Conversion)

Parallel-encoded analog-to-digital conversion, or flash conversion, is perhaps the easiest conversion process to understand. To illustrate the basics behind parallel encoding (also referred to as *simultaneous multiple comparator or flash converting*), let's take a look at the simple 3-bit converter in Fig. 12.149.



Truth Table

Analog V_{in} (V)	I_7	I_6	I_5	I_4	I_3	I_2	I_1	I_0	$\overline{A}_2$	$\overline{A}_1$	$\overline{A}_0$	$\overline{Q}_2$	$\overline{Q}_1$	$\overline{Q}_0$
0.000 - 0.625	0	0	0	0	0	0	0	0	1	1	1	0	0	0
0.625 - 1.250	0	0	0	0	0	0	1	0	1	1	0	0	0	1
1.250 - 1.875	0	0	0	0	0	1	1	0	1	0	1	0	1	0
1.875 - 2.500	0	0	0	0	1	1	1	0	1	0	0	0	1	1
2.500 - 3.125	0	0	0	1	1	1	1	0	0	1	1	1	0	0
3.125 - 3.750	0	0	1	1	1	1	1	0	0	1	0	1	0	1
3.750 - 4.375	0	1	1	1	1	1	1	0	0	0	1	1	1	0
4.375 or greater	1	1	1	1	1	1	1	0	0	0	0	1	1	1

Table assumes that the D-latch's enable input is set HIGH, making the latch appear transparent (*D* inputs follow *Q* outputs).

FIGURE 12.149

The set of comparators is the key feature to note in this circuit. Each comparator is supplied with a different reference voltage from the 1 kΩ voltage divider network. Since we've set up a +5V reference voltage, the voltage drop across each resistor within the voltage divider network is 0.625 V. From this, you can determine the specific reference voltages given to each comparator (see Fig. 12.149).

To convert an analog signal into a digital number, the analog signal is applied to all the comparators at the same time, via the common line attached to the inverting inputs of all the comparators. If the analog voltage is between, say, 2.5000 and 3.125 V, only those comparators with reference voltage below 3.125 *V* will output a high. To create a 3-bit binary output, the eight comparator outputs are fed into an octal-to-binary priority encoder. A *D* latch also can be incorporated into the circuit to provide enable control of the binary output. The truth table should fill in the rest.

12.11 Displays

A number of displays can be interfaced with control logic to display numbers, letters, special characters, and graphics. Two popular displays that we'll consider here include the light-emitting diode (LED) display and the liquid-crystal display (LCD).

12.11.1 LED Displays

LED displays come in three basic configurations: numeric (numbers), alphanumeric (numbers and letters), and dot-matrix forms (see Fig. 12.150). Numeric displays consist of seven LED segments. Each LED segment is given a letter designation, as shown in the figure. Seven-segment LED displays are most frequently used to generate numbers (0–9), but they also can be used to display hexadecimal (0–9, A, B, C, D, E, F). The 14-segment, 16-segment, and special 4 × 7 dot matrix displays are alphanumeric. The 5 ×

7 dot matrix display is both alphanumeric and graphic—you can display unique characters and simple graphics. See Chapter 5 for information about other types of LED displays.

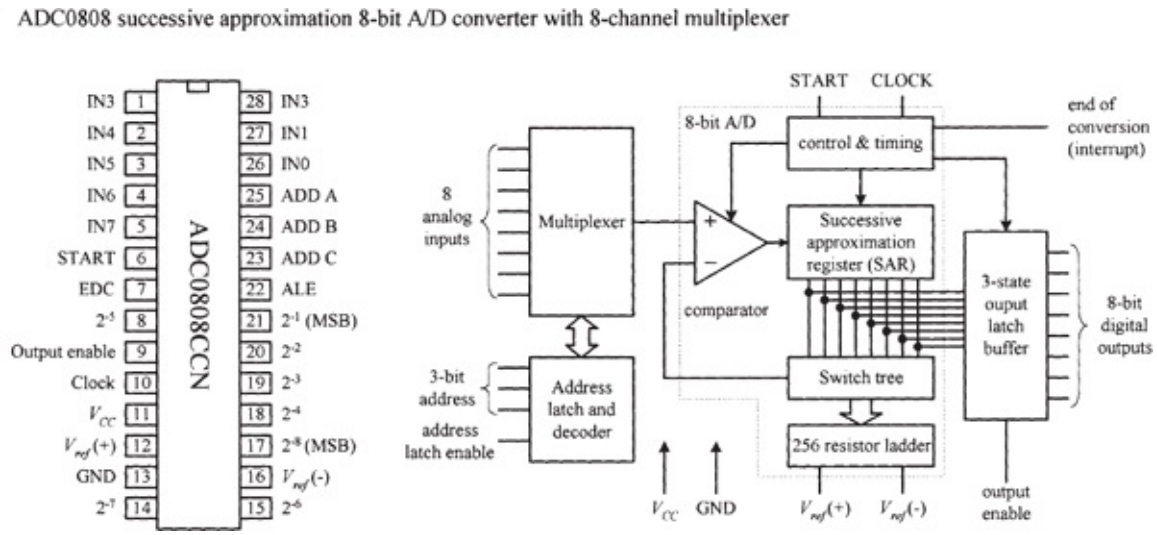


FIGURE 12.150

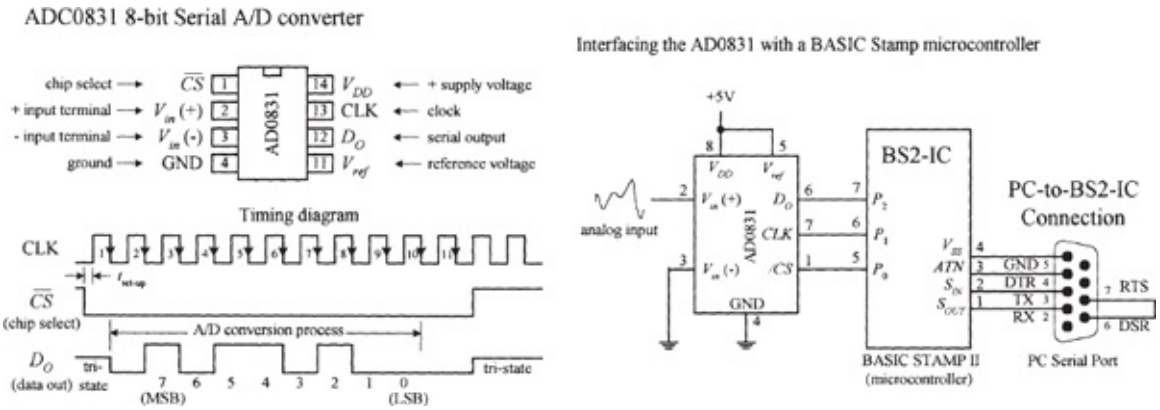


FIGURE 12.151

To drive a given segment of a common anode display, current must be sunk out through the corresponding segment's terminal. With the common cathode display, current must be sourced into the corresponding segment's terminal. A simple way to drive these displays is to use BCD to seven-segment display decoder/drivers, like the ones show in the figure. Applying a BCD input character results in a decimal digit being displayed (e.g., 0101 applied to A_0–A_3, or A–D displays a "5"). The 74LS47 active-low open-collector outputs are suited for a common anode display, while the 74HC4511's active-high outputs are suited for a common cathode display. Both ICs also come with extra terminals used for lamp testing and ripple blanking, as well as leading zero suppression (controlling the decimal point). We discussed how to use these terminals in Section 12.6.

Direct Drive of Numeric LED Displays

Seven-segment LED displays come in two varieties: common anode and common cathode. Figure 12.152 shows single digital eight-segment (seven digit segments + decimal point) displays of both varieties.

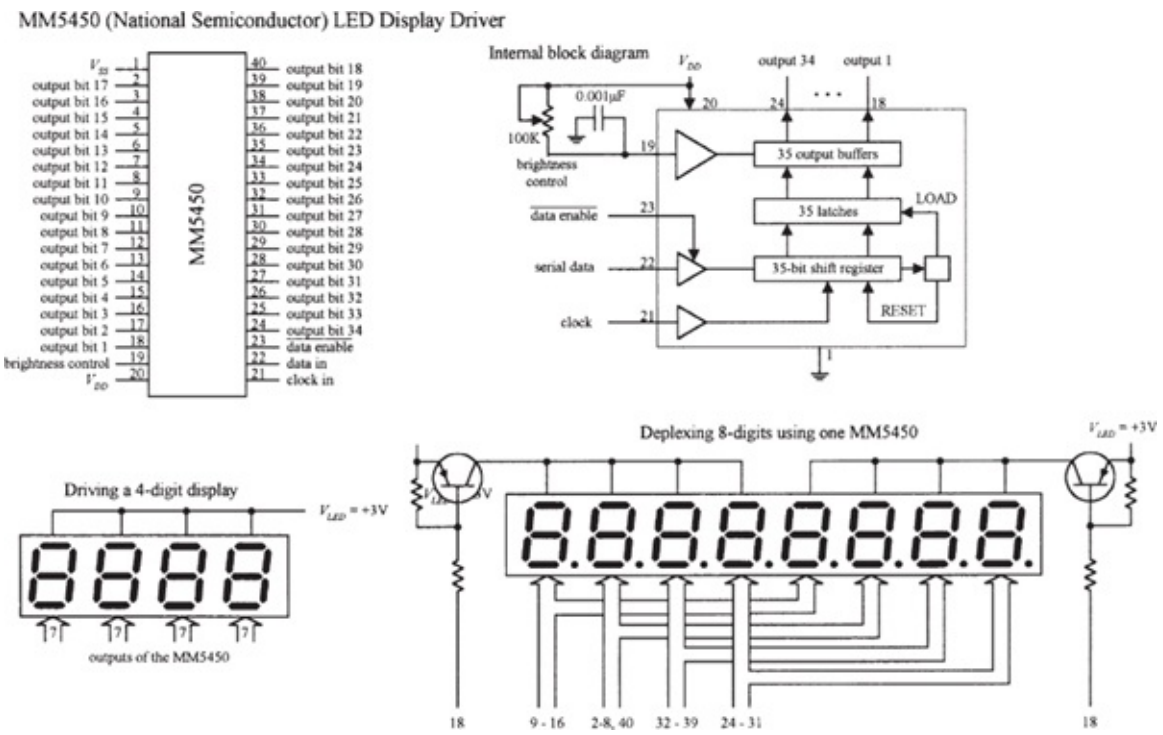


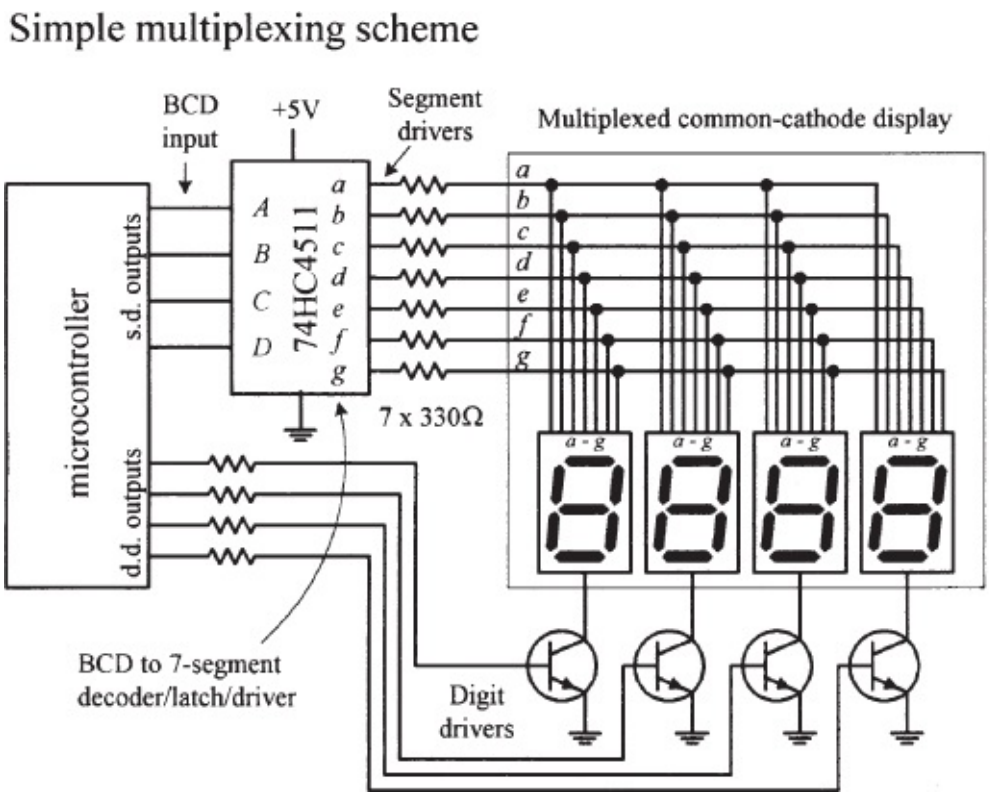
FIGURE 12.152

When driving a multidigit display, say, one with eight digits, the previous technique becomes awkward. It requires eight discrete decoder/driver ICs. One way to avoid this problem is to use a special direct-drive LED display driver IC.

For example, National Semiconductor's MM5450, shown in Fig. 12.157, is designed to drive 4- or 5-digit alphanumeric common anode LED displays. It comes with 34 TTL-compatible outputs that are used to drive desired LED segments within a display. Each of these outputs can sink up to 15 mA. In order to specify which output lines are driven high or low, serial input data are clocked into the driver's serial input. The serial data chain that is entered is 36 bits long. The first bit is a start bit (set to 1), and the remaining 35 bits are data bits. Each data bit corresponds to a given output data line that is used to drive a given LED segment within the display. At the thirty-sixth positive clock signal, a LOAD signal is generated that loads the 35 data bits into the latches (see the block diagram in Fig. 12.157). At the low state of the clock, a $\overline{\text{RESET}}$ signal is generated that clears the shift register for the next set of data. You can learn more about the MM5450 at http://www.micrel.com/_PDF/mm5450.pdf.

Multiplexed LED Displays

Another technique used to drive multidigit LED displays involves multiplexing. Multiplexing can drastically reduce the number of connections needed between display and control logic. In a multiplexed display, digits share common segment lines. Also, only one digit within the display is lighted at a time. To make it appear that a complete readout is displayed, all the digits must be flashed very rapidly in sequence, over and over again. The simple example in Fig. 12.153 shows multiplexing in action.



Here, we have a multiplexed common-cathode display—all digits share common segment lines (*a*–*g*). To supply a full one-digit readout, digits must be flashed rapidly, one at a time. To enable a given digit, the digit's common line is grounded via one of the digital drivers (transistors)—all other digits' common lines are left floating. In this example, the drivers are controlled by a microcontroller. To light the segments of a given digit, the microcontroller supplies the appropriate 4-bit BCD code to the seven-segment decoder/driver (74HC4511). As an example, if we wanted to display 1234, we would need to program the microcontroller (using software) to turn off all digits except the MSD (leftmost digit) and then supply the decoder/driver with the BCD code for 1. Then the next significant digit (2) would be driven, and then the next significant digit (3), and then the LSD (4). After that, the process would recycle for as long as we wanted our program to display 1234.

FIGURE 12.153

Simple Alphanumeric Display

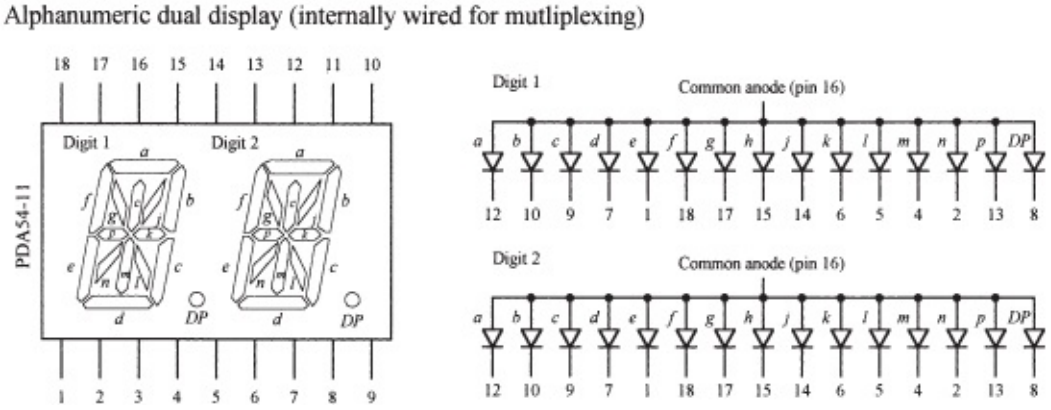
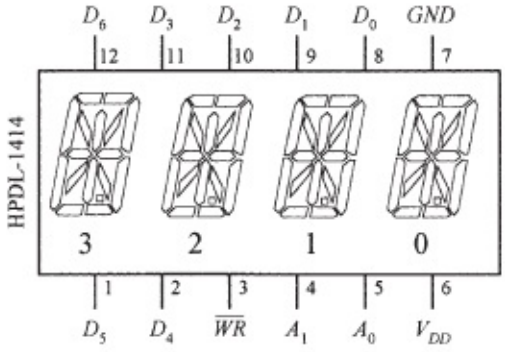


Figure 12.154 shows a common anode, 2-character, 14-segment (+ decimal) alphanumeric display. Notice that the segments of the two characters are internally wired together. This means that the display is designed for multiplexing. Though it is possible to use a microcontroller along with transistor drivers to control this display, the number of lines required is fairly large. Another option is to use a special driver IC, like Intersil's ICM7243B 14-segment 6-bit ASCII driver. Another alternative is simply to avoid using this kind of display and use a "smart" alphanumeric display that contains all the necessary control logic (drivers, code converters, and so on).

FIGURE 12.154

"Smart" Alphanumeric Display

HPDL-1414 (Hewlett Packard) 4-character smart alphanumeric display



Write Truth Table

$\overline{WR}$	A_1	A_0	D_6 to D_0	DIG_3	DIG_2	DIG_1	DIG_0
0	0	0	ASCII code "A"	NC	NC	NC	A
0	0	1		NC	NC	A	NC
0	1	0		NC	A	NC	NC
0	1	1		A	NC	NC	NC
1	X	X	X	previously written data			

Character Set

BITS	D_3 D_2 D_1 D_0	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
D_6 D_5 D_4	HEX	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0 1 0	2	space	!	"	#	$	%	&	'	(	)	*	+	,	−	.	/
0 1 1	3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
1 0 0	4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1 0 1	5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_

The HPDL-1414 is a "smart," 4-character, 16-segment display. This device is complete with LEDs, on-board 4-word ASCII memory, a 64-word character generator, 17-segment drivers, 4-digit drivers, and scanning circuitry necessary to multiplex the four LED characters. It is TTL-compatible and relatively easy to use. The seven data inputs D_0 to D_6 accept a 7-bit ASCII code, while the digital select inputs A_0 and A_1 accept a 2-bit binary code that is used to specify which of the four digits is to be lighted. The WRITE ($\overline{WR}$) input is used to load new data into memory. After a character has been written to memory, the IC decodes the ASCII data, drives the display, and refreshes it without the need for external hardware or software.

FIGURE 12.155

12.11.2 Liquid-Crystal Displays

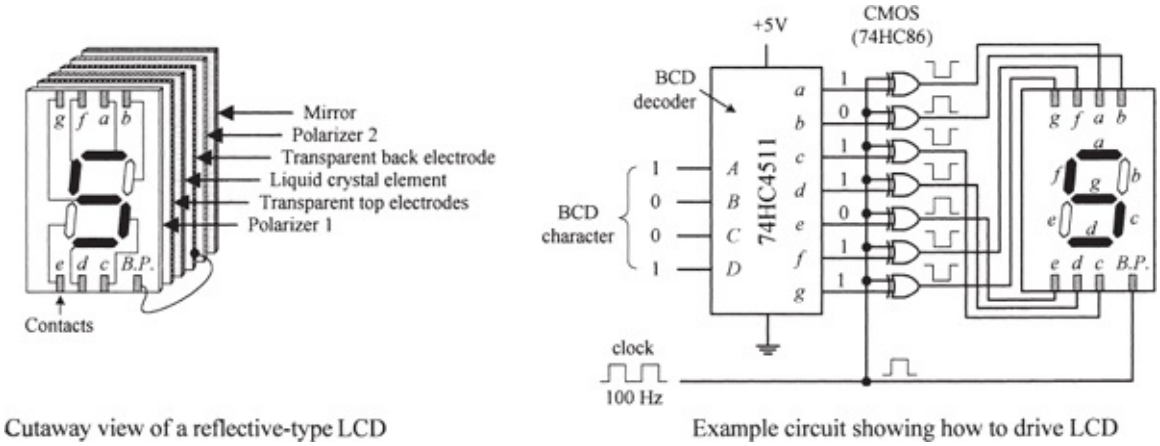
In low-power CMOS digital systems (for example, battery- or solar-powered electronic devices), the dissipation of an LED display can consume most of a system's power requirements, which is something you want to avoid, especially since you are looking to save power when using CMOSs. LCDs, on the other hand, are ideal for low-power applications.

Unlike an LED display, an LCD is a passive device. This means that instead of using electric current to generate light, it uses light that is already externally present (such as sunlight, room lighting). For the LCD's optical effects to occur, the external light source needs to supply only a minute amount of power (within the mW/cm^2 range).

One disadvantage with LCDs is their slow switching speeds (the time it takes for a new digit/character to appear). Typical switching speeds for LCDs range from around 40 to 100 ms. At low temperatures, the switching speeds get even worse. Another problem with LCDs is the requirement that external light be present. Though there are LCD displays that come with backlighting (such as an LED behind the display), obviously, this will increase power consumption.

Basic Explanation of How an LCD Works

An LCD consists of a number of layers that include a polarizer, a set of transparent electrodes, a liquid-crystal element, a transparent back electrode, a second polarizer, and a mirror (see the leftmost illustration in Fig. 12.156).



Cutaway view of a reflective-type LCD

Example circuit showing how to drive LCD

FIGURE 12.156

The transparent top electrodes are used to generate the individual segments of a digit, character, and so on, while the transparent back electrode forms a common plane, often referred to as the *back plane* (BP). The top electrode segments and the back electrode are wired to external contacts. With no potential difference between a given top electrode and the back electrode, the region where the top electrode is located appears silver in color against a silver background. However, when a potential is applied between a given top electrode and back electrode, the region where the top electrode is located appears dark against a silver background.

The circuit in Fig. 12.156 shows a basic way to drive a seven-segment LCD. It uses a 74HC4511 BCD decoder and XOR gates to generate the prior drive signals for the LCD. A very important thing to note in this circuit is the clock. As it turns out, an LCD actually requires ac drive signals (for example, squarewaves) instead of dc drive signals. If dc were used, the primary component of the display—namely, the liquid crystal—would undergo electrochemical degradation (more on the liquid crystal in a moment). The optimal frequency of the applied ac drive signal is typically from around 25 Hz to a couple hundred hertz. Now that we understand that, it is easy to see why we need the XOR gates.

As the clock delivers squarewaves to the back electrode (back plane, or BP), the XOR gates act as enable gates that pass and invert a signal and apply it to a given top electrode segment. For example, if a BCD code of 1001 (5) is applied to the decoder, the decoder's outputs *a*, *c*, *d*, *f*, and *g* go high, while outputs *b* and *e* go low. When a positive clock pulse arrives, XOR gates attached to the outputs that are high invert the high levels. XOR gates attached to outputs that are low pass on the low levels. During the same pulse duration, the back plane is set high. Potentials now are present between *a*, *c*, *d*, *f*, and *g* segments and the back plane, and therefore these segments appear dark. Segments *b* and *e*, along with the background, appear silverish because no potential exists between them and the back plane. Now, when the clock pulse goes low, the display remains the same (provided the BCD input hasn't changed), since all that has occurred is a reverse in polarity. This has no effect on the optical properties of the display.

Detailed Explanation of How an LCD Works (the Physics)

Figure 12.157 shows how an LCD generates a clear (silverish) segment. When control signals sent to the transparent top and back electrodes are in phase, no potential exists between the two electrodes. With no potential present, the cigar-shaped organic liquid crystals (nematic crystals) arrange themselves in spiral state, as shown in the figure.

The upper crystal aligns itself horizontal to the page, while the lowest crystal aligns itself perpendicular to the page. The upper crystal and the lower crystal are held in place by tiny grooves that are etched into the inner surfaces of the glass surfaces of the cell. Crystals in between the upper crystal and the lower crystal progressively spiral 90° due to electrostatic forces that exist between neighboring crystals. When polarized light passes through a region of the display that contains these spirals, the

polarization angle of the light is rotated 90°.

Now, looking at the display as a whole, when incident unpolarized light passes through polarizer 1 (as shown in the figure), the light becomes polarized in the same direction of the plane of polarization of the first polarizer. The polarized light then passes through the transparent top electrode and enters the liquid-crystal cell. As it passes through the cell, its polarization angle is rotated 90°. The polarized light that exits the cell then passes through the transparent back electrode and the second polarizer without problems. (If we were to remove the liquid-crystal cell, all polarized light that passed through the first polarizer would be absorbed, since we would have crossed polarizers.) The light that passes through the second polarizer then reflects off the mirror, passes through the second polarizer, on through the liquid-crystal cell (getting rotated 90°), through the first polarizer, and finally reaches the observer's eye. This reflected light appears silver in color. Note that the background of LCDs constantly appears silver because no potential exists across the liquid-crystal cell in the background region.

Figure 12.158 shows how an LCD generates a dark segment. When control signals sent to the top and back electrodes are out of phase, a potential difference exists between the two electrodes. This causes the crystals to align themselves in a parallel manner, as shown in the figure. When the polarized light from the first polarizer passes through the cell region containing these parallel crystals, nothing happens—the polarization angle stays the same. However, when the light comes in contact with the second polarizer, it is absorbed because the angle of polarization of the light and the plane of polarization of the second polarizer are perpendicular to each other. Since light reaches the mirror, no light is reflected back to the observer's eye, and hence the segment appears dark.

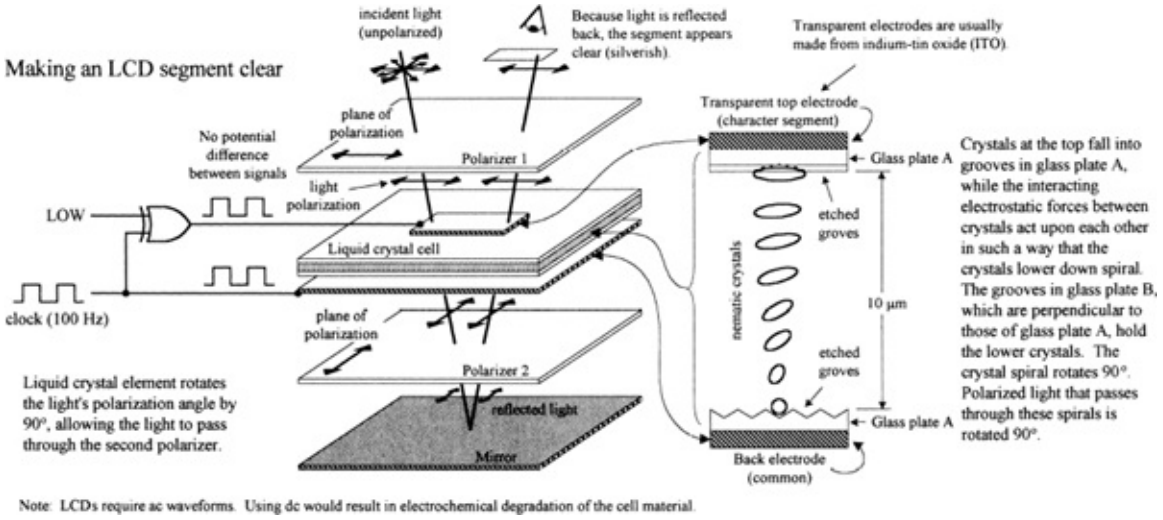


FIGURE 12.157

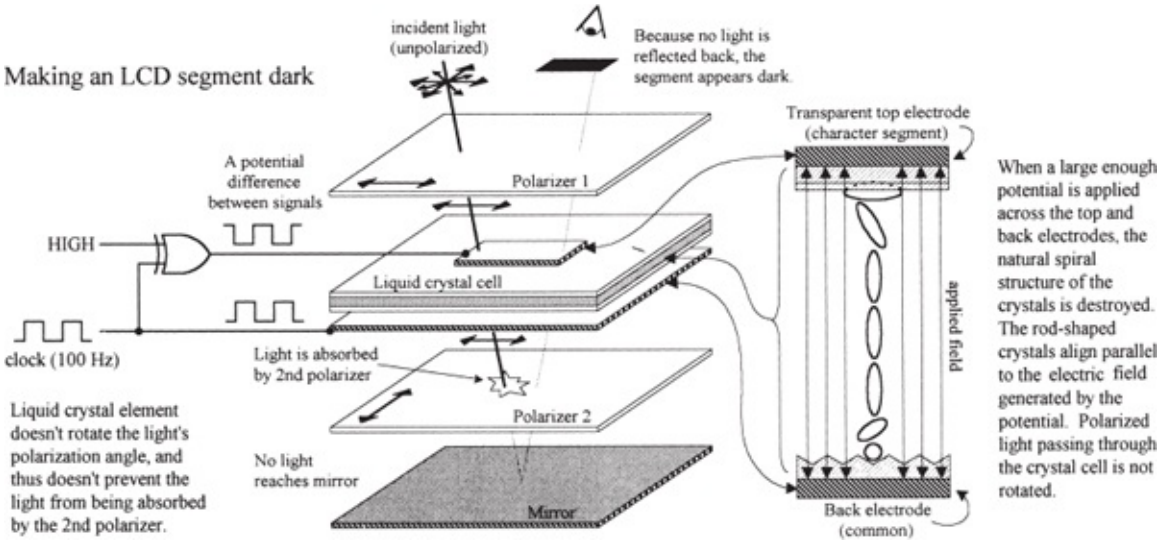


FIGURE 12.158

The LCD shown in Fig. 12.157 represents what is referred to as a *standard twisted nematic display*. Another common LCD is the *supertwist nematic display*. Unlike the standard twisted display, this display's nematic crystals rotate 270° from top to bottom. The extra 180° twist improves the contrast and viewing angle.

Driving LCDs

CD4543B CMOS BCD-TO-SEVEN-SEGMENT LATCH/DECODER/DRIVER

The CD4543B (Texas Instruments), shown in Fig. 12.159, is a BCD-to-seven-segment latch/decoder/driver that is designed for LCDs, as well as for LED displays. When used to drive LCDs, a squarewave must be applied simultaneously to the CD4543B's Phase (*Ph*) input and to the LCD's back plane. When used to drive LED displays, a high is required at the Phase input for common cathode displays, while a low is required for common anode displays. To blank the display (set outputs *a*–*g* low), the *BL* input is set high. The CD4543B also comes with a Latch Disable input (*LD*), which can be used to latch onto input data, preventing new input data from altering the display.

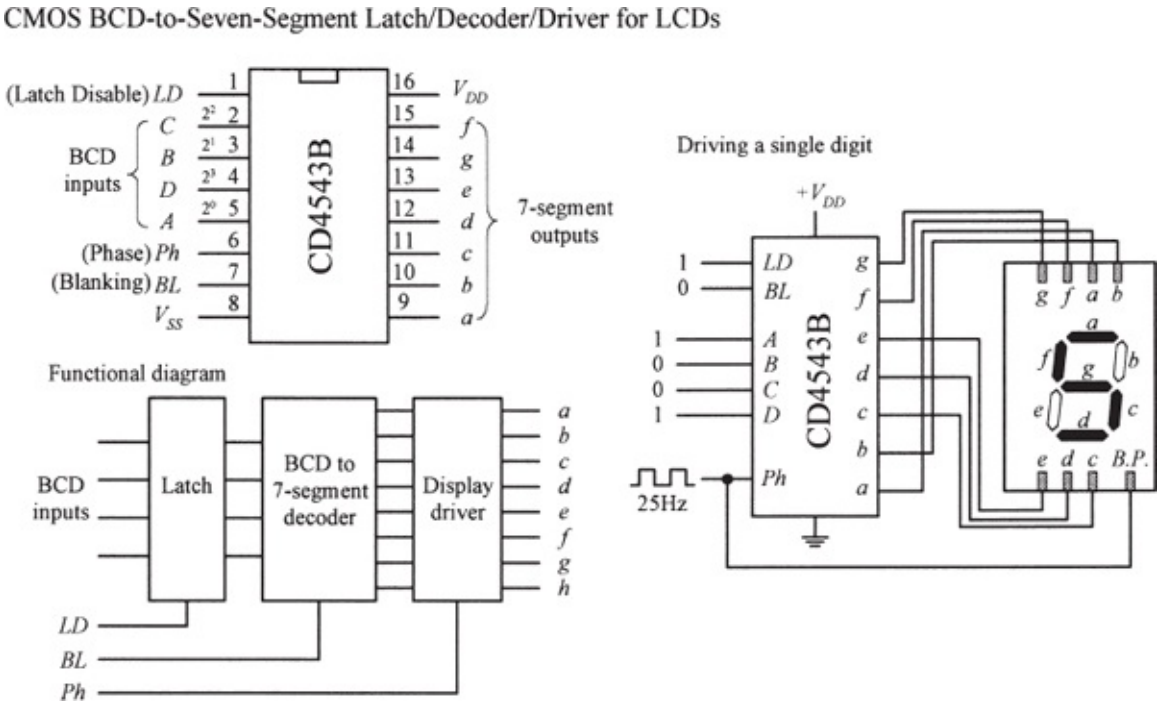


FIGURE 12.159

MM5453 LCD DRIVER

The MM5453 (National Semiconductor) is a 40-pin IC that can drive up to 33 segments of an LCD, which can be used to drive 4½-digit seven-segment displays. It houses an internal oscillator section (requiring an external RC circuit) that generates the necessary squarewaves used to drive the LCD. To activate given segments within the display, a serial code is applied to the data input. The code first starts out with a start bit (high) followed by data bits that specify which outputs should be driven high or low. Figure 12.160 shows an example display circuit, along with corresponding data format required to drive a 4½-digit display.

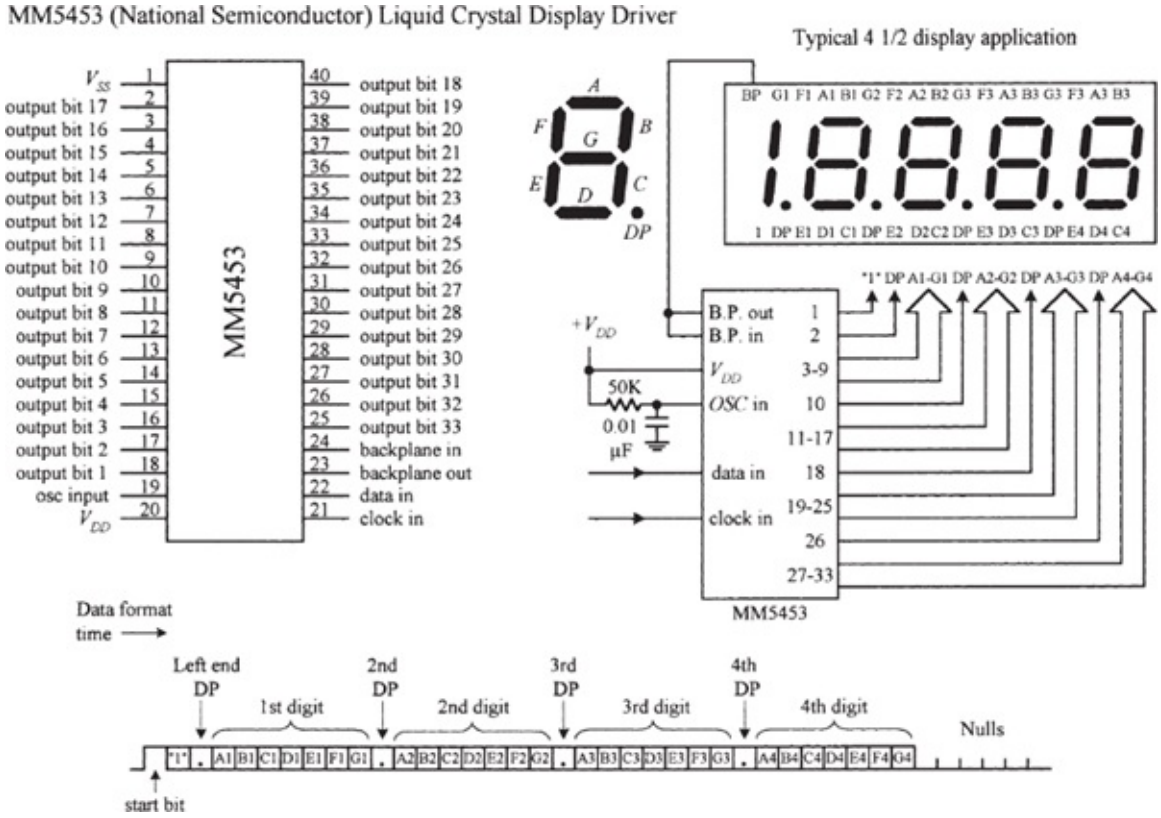


FIGURE 12.160

VI-322-DP LCD AND ICL7106 3½-DIGIT LCD, ADC DRIVER

There are a number of specialized LCDs that can be found in the electronics catalogs. An example is Varitronix's VI-322-DP 3½-digit (plus ~, +, BAT, Δ) LCD, shown in Fig. 12.160. This display is configured in a static drive arrangement (each segment has a separate lead) and is found in many test instruments. To drive this display, you first check to see what kind of driver the manufacturer suggests. In this case, the manufacturer suggests using Intersil's ICL7106. This IC is a 3½-digit LCD/LED display driver as well as an ADC. This dual-purpose feature makes it easy to interface transducers directly to the same IC that is driving the display. To learn how to use the ICL7106, check out Intersil's data sheet at http://www.intersil.com.

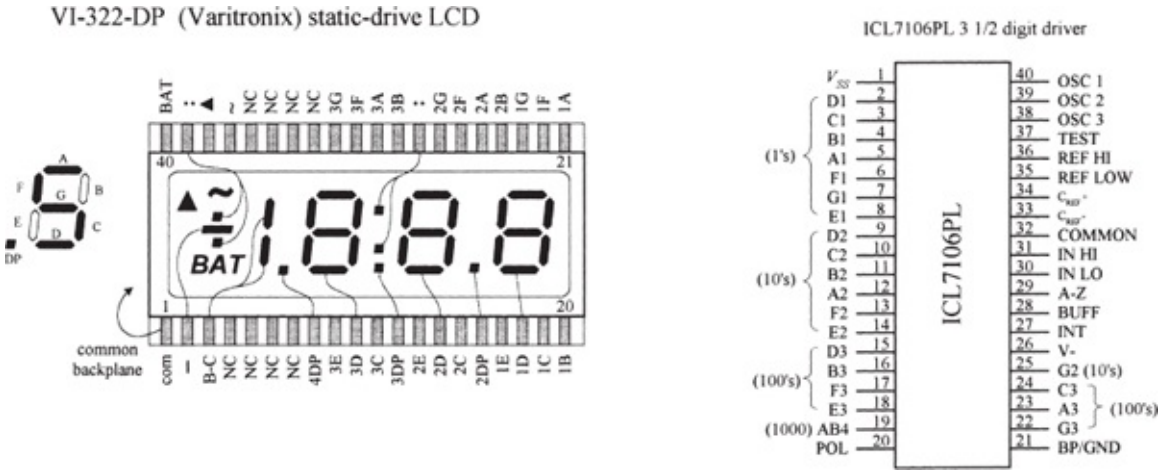


FIGURE 12.161

Multiplexed LCDs

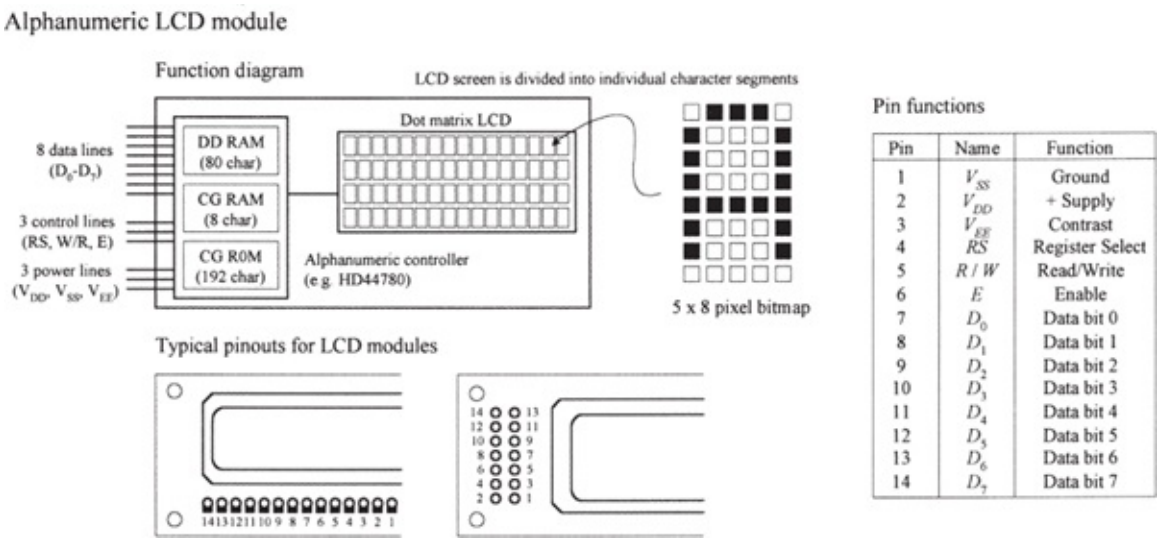
We have just seen examples of static-drive-type LCDs, where each segment (to electrode) had its own lead, and a single common plane was used as the back electrode. Another type of LCD is designed with multiplexing in mind and is referred to as *dynamic drive* or *multiplexed display*.

As with the multiplexed LED display, multiplexed LCDs can greatly reduce the number of external connections required between the display and driver. However, they require increased complexity in drive circuitry (or software) to drive. In a multiplexed LCD, appropriate segments are connected together to form groups that are sequentially addressed by means of multiple back-plane electrodes.

"Intelligent" Dot-Matrix LCD Modules

Dot-matrix LCDs are used to display alphanumeric characters and other symbols. These displays are used in cell phones, calculators, vending machines, and many other devices that provide the user with simple textual information. Dot-matrix LCDs are also used in laptop computer screens; however, these displays incorporate special filters, multicolor back lighting, and so on. For practical purposes, we'll concentrate on the simple alphanumeric LCDs.

An alphanumeric LCD screen is usually divided into a number of 5 × 8 pixel blocks, with vertical and horizontal spaces separating each block. Figure 12.162 shows a display with 20 columns and 4 rows of 5 × 8 pixel blocks. Other standard configurations come with 8, 16, 20, 24, 32, or 40 columns and 1, 2, or 4 rows. To generate a character within a given block requires that each pixel within the block be turned on or off. As you can imagine, to control so many different pixels (electrode segments) requires a great deal of sophistication. For this reason, an intelligent driver IC is required.



Pin	Name	Function
1	V_{SS}	Ground
2	V_{DD}	+ Supply
3	V_{EE}	Contrast
4	RS	Register Select
5	R/W	Read/Write
6	E	Enable
7	D_0	Data bit 0
8	D_1	Data bit 1
9	D_2	Data bit 2
10	D_3	Data bit 3
11	D_4	Data bit 4
12	D_5	Data bit 5
13	D_6	Data bit 6
14	D_7	Data bit 7

FIGURE 12.162

Almost all alphanumeric LCD modules are controlled by Hitachi's HD44780 (or equivalent) driver IC. This driver contains the following:

- A permanent memory (CG ROM) that stores 192 alphanumeric characters
- A random access memory (DD RAM) used to store the display's contents
- A second random access memory (CG RAM) used to hold custom symbols
- Input lines for data and instruction control signals
- Multiplexed outputs for driving LCD pixels
- Additional outputs for communicating with expansion chips to drive more LCD pixels

This driver is built right into the LCD module. (You could attempt to construct your own module by interfacing the driver with an LCD, but it would not be worth the effort—the numerous tiny connections would drive you nuts.) From now on, all modules described in this section are assumed to be HD44780-driven.

BASIC OVERVIEW OF THE PINS

The standard LCD module comes with a 14-pin interface: eight data lines (D_0–D_7), three control lines (*RS*, *W*/*R*, and *E*), and three power lines ($V_{DD,}$ V_{SS}, and V_{EE}).

V_{DD} (pin 2) and V_{SS} (pin 1) are the module's positive and negative power supply leads. Usually, V_{DD} is set to +5 V, while V_{SS} is grounded. V_{EE} (pin 3) is the display's contrast control. By changing the voltage applied to this lead, the contrast of the display increases or decreases. A potentiometer placed between supply voltages, with its wiper connected to V_{EE}, allows for manual adjustment.

D_0–D_7 (pins 7–14) are the data bus lines. Data can be transferred to and from the display either as a single 8-bit byte or as two 4-bit nibbles. In the latter case, only the upper four data lines (D_4–D_7) are used.

RS (pin 4) is the Register Select line. When this line is low, data bytes transferred to the display module are interpreted as commands, and data bytes read from the display module indicate its status. When the *RS* line is set high, character data can be transferred to and from the display module.

R/W (pin 5) is the Read/Write control line. To write commands or character data to the module, *R/W* is set low. To read character data or status information from the module, *R/W* is set high.

E (pin 6) is the Enable control input, which is used to initiate the actual transfer of command or character data to and from the module. When writing to the display, data on the D_0–D_7 lines is transferred to the display when the enable input receives a high-to-low transition. When reading from the display, data become available to the D_0–D_7 lines shortly after a low-to-high transition occurs at the enable input and will remain available until the signal goes low again.

Figure 12.163 shows the instruction set and standard set of characters for an LCD module. Next, we'll go through some examples illustrating how to use the instructions and how to write characters to the display.

LCD Instruction Set

INSTRUCTION	R/S	R/W	D_7	D_6	D_5	D_4	D_3	D_2	D_1	D_0
Clear Display	0	0	0	0	0	0	0	0	0	1
Display & Cursor Home	0	0	0	0	0	0	0	0	1	X
Character Entry Mode	0	0	0	0	0	0	0	1	I/D	S
Display & Cursor On/Off	0	0	0	0	0	0	1	D	C	B
Display/Cursor Shift	0	0	0	0	0	1	D/C	R/L	X	X
Function Set	0	0	0	0	1	DL	N	F	X	X
Set CGRAM Address	0	0	0	1	A	A	A	A	A	A
Set Display Address	0	0	1	A	A	A	A	A	A	A
Poll the "Busy Flag"	0	0	BF	X	X	X	X	X	X	X
Write Character to Display [a]	1	0	D	D	D	D	D	D	D	D
Read Character on Dsplay [b]	1	1	D	D	D	D	D	D	D	D

I/D = Increment (I/D = 1)*/Decrement (I/D = 0) each byte written to display
S = Display shift on (S =1), Display shift off (S = 0)*
D = Turn display on (D = 1), Turn display off (D = 0)*
C = Show cursor (C = 1), Hide cursor (C = 0)
B = Underline cursor (B = 0, C = 1), Blink cursor (B = 1, C = 1)
D/C = Move display (D/C = 1), Move cursor (D/C = 0)
R/L = Direction of shift: Shift right (R/L = 1), Shift left (R/L =0)
DL = Set data interface length: 8-bit interface (DL = 1)*, 4-bit interface (DL = 0)
N = Number of display lines: 2 line mode (N = 1), 1 line mode (N = 0)*
F = Character font format: 5 x 10 dot (F = 1), 5 x 7 dot (F = 0)*
BF = Poll the Busy Flag: controller not busy (BF = 0), controller busy (BF = 1)
A = CGRAM or display address bit
D = Character data bit
a = Write character to display at the current cursor position
b = Read character on display at the current cursor position
X = Don't care
* = Initialization settings

Standard LCD Character Table

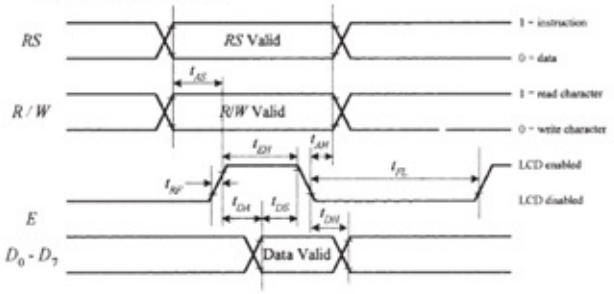


Steps used to Read and Write data to and from LCD module (HD44780 controlled)

Steps for displaying a character
Set to Write Character to Display mode: $R/W = 0$, $RS = 1$.
Apply data bits (character code) to D_7 - D_0.
Breifly set $E = 1$, then set $E = 0$

Steps for reading data from display
Set to Read Character on Display mode: $R/W = 1$, $RS = 1$.
Set $E = 1$.
Read data from D_7 - D_0.
Set $E = 0$

HD44780 Timing Diagram



t_{AS} (Address setup time) - For data inputs to be interpreted correctly, they must be held for a minimum time of t_{AS} (~140ns) prior to the E-table signal.
t_{EH} (Enable high time) - E must be held HIGH for a minimum of t_{EH} (~ 450ns) for proper operation.
t_{DS} (Data setup time) - Data inputs must be held stable for a time t_{DS} (~ 200ns) prior to the E signal for proper operation.
t_{AH} (Address hold time) - Control lines RS and R/W lines must not change for a duration of t_{AH} (~10ns) after the E line goes LOW for proper operation.
t_{DH} (Data hold time) - Data lines D_0 - D_7 must not change for a duration of t_{DH} (~20ns) after the E line goes LOW for prior operation.
t_{EL} (enable low time) - E line must not be set HIGH again (for the next command), for at least t_{EL} (~500ns) for proper operation.
t_{Er} (rise and fall time) - Rise and fall times are each ~ 25ns each.

FIGURE 12.163

TEST CIRCUIT USED TO DEMONSTRATE HOW TO CONTROL THE LCD MODULE

Figure 12.164 shows a simple test circuit that is quite useful for learning how to send commands and character data to the LCD module. (In reality, the LCD module is connected to a microprocessor or microcontroller, as shown to the left in the figure.) In this circuit, switches connected to data inputs use pullup resistors in order to supply a high (1) when the switch is open or supply a low (0) when the switch is closed.

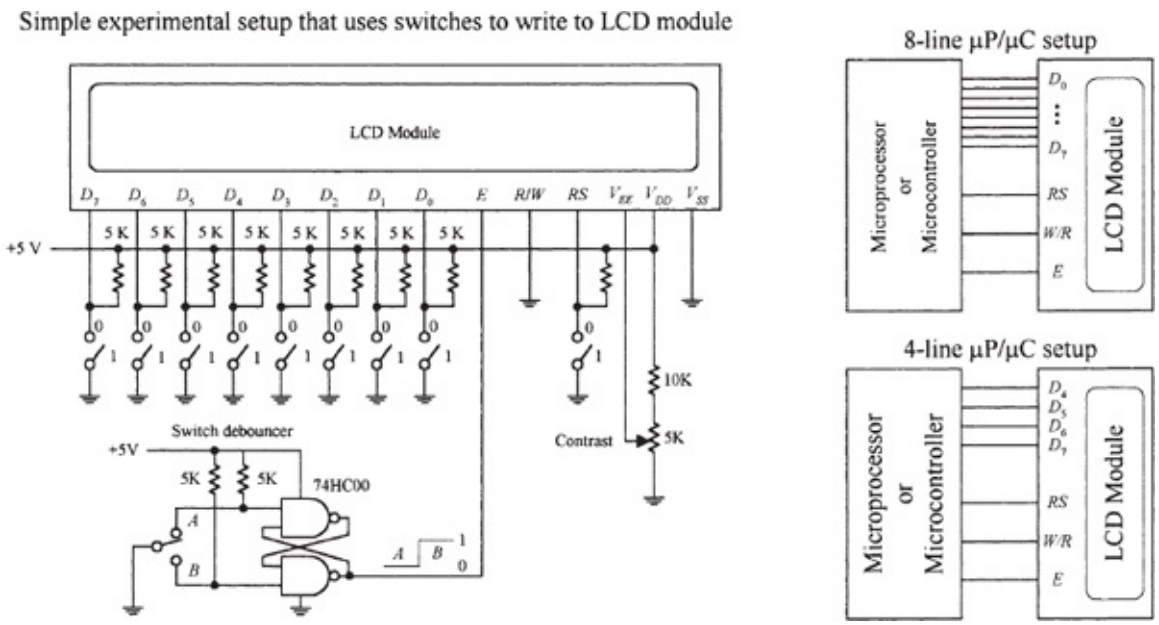


FIGURE 12.164

The enable input receives its high and low levels from a debounced toggle switch. Debouncing the enable switch prevents the likelihood of multiple enable signals being generated. Multiple enable signals tend to create unwanted effects, such as generating the same character over and over again across the display. The 5-kΩ pot is used for contrast. Note that in this circuit, we've grounded the *R*/*W* line, which means we'll only deal with writing to the display.

WHEN POWER IS FIRST APPLIED

When power is first applied to the display, the display module resets itself to its initial settings. Initial settings are indicated in the LCD instruction set with an asterisk. As indicated, the display is actually turned off during the initial setting condition. If we attempt to write character data to the display now, nothing will show up. In order to show something, we must issue a command to the module telling it to turn on its display.

According to the instruction set, the Display & Cursor On/Off instruction can be used to turn on the display. At the same time, this instruction also selects the cursor style. For example, if we apply the command code 0000 1111 to D_7–D_0, making sure to keep *RS* low so the module will interpret data as a command, a blinking cursor with an underline should appear at the top leftmost position on the display. But before this command can take effect, it must be sent to the module by momentarily setting the Enable (*E*) line low.

Another important instruction that should be implemented after power-up is the Function Set command. When a two-line display is used, this command tells the module to turn on the second line. It also tells the module what kind of data transfer is going to be used (8-bit or 4-bit), and whether a 5 × 10 or 5 × 7 pixel format will be used (5 × 10 is found in some one-line displays). Assuming that the display used in our example circuit is a two-line display, we can send the command 0011 1000 telling the display to turn on both lines, use an 8-bit transfer, and provide a 5 × 7 pixel character format. Again, to send this command, we set *RS* low, then supply the command data to D_7–D_0, and finally pulse *E* low.

Now that the module knows what format to use, we can try writing a character to the display. To do this, we set the module to character mode by setting *RS* high. Next, we apply one of the 8-bit codes listed in the standard LCD character set table to the data inputs D_7–D_0. For example, if we want to display the letter *Q*, we apply 01010001 (hex 51 or 51_H). To send the character data to the LCD module, we pulse *E* low. A *Q* should then appear on the display. To clear the screen, we use the Clear Display command 0000 0001, remembering to keep *RS* low and then pulsing *E* low.

ADDRESSING

After power-up, the module's cursor is positioned at the far-left corner of the first line of the display. This display location is assigned a hexadecimal address of 00_H. As new characters are entered, the cursor automatically moves to the right to a new address of 01_H, then 02_H, and so on. Although this automatic incrementing feature makes life easy when entering characters, there are times when it is necessary to set the cursor position to a location other than the first address location.

To set the cursor to another address location, a new starting address must be entered as a command. There are 128 different addresses to choose from, although not all these addresses have their own display location. In fact, there are only 80 display locations laid out on a single line in one-line mode or 40 display locations laid out on each line in two-line mode. Now, as it turns out, not all display locations are necessarily visible on the screen at one time. This will be made more apparent in a moment. Let's first try a simple address example with the LCD module set to two-line mode (provided that two lines are actually available).

To position the cursor to a desired location, we use the Set Address command. This command is specified with the binary code 1000 0000 + (binary value of desired hex address). For example, to send a command telling the cursor to jump to the 07_H address location, we apply (1000 0000 + 0000 0111) = 1000 0111 to the D_7–D_0 inputs, remembering to hold *RS* low and then pulsing *E* low. The cursor should now be located at the eighth position over from the left.

It is important to realize that the relationship between addresses and display locations varies from module to module. Most displays are configured with two lines of characters, with the first line starting at address 00_H and the second line at address 40_H. Figure 12.165 shows the relationship between the address and display locations for various LCD modules. Note that the four-line module is really a two-line type with the two lines split, as shown in the figure.

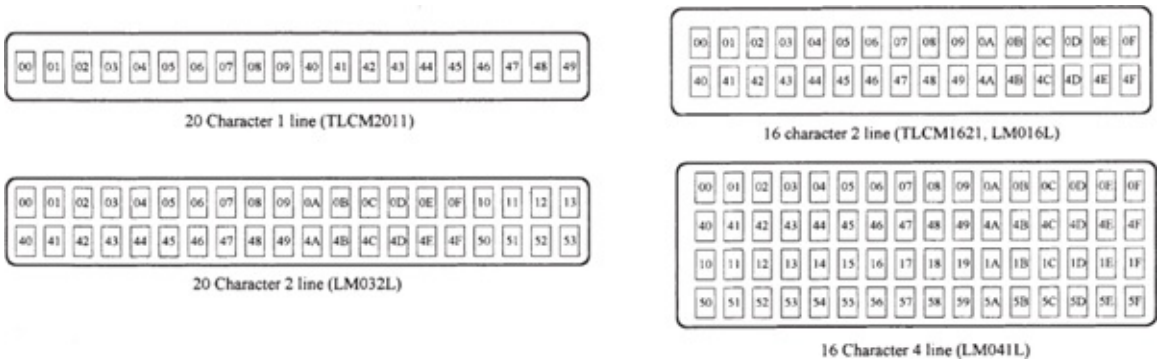


FIGURE 12.165

SHIFTING THE DISPLAY

Regardless of their size, LCD modules have 80 display locations that can be written to. With smaller displays, not all 80 locations can be displayed at once on the screen. For example, if we were to enter all the letters of the alphabet onto the first line of a 20-character display, only letters *A* through *T* would appear on the screen. Letters *S* through *Z*, along with the cursor, would be "pushed off" to the right of the screen, hidden from view.

To bring these hidden characters into view, we can apply the Cursor/Display Shift command to shift all display locations to the left. The command for shifting to the left is 0001 1000. Every time this command is issued, the characters shift one step to the left. In our example, it would take seven of these commands to bring T through Z and the cursor into view.

To shift things to the right, we apply the command 0001 1100. To bring the cursor back to address 00_H and shift the display address 00_H back to the left-hand side of the display, a Cursor Home command (0000

0010) can be issued. Another alternative is to use the Clear Display command 0000 0001. However, this command also clears all display locations.

CHARACTER ENTRY MODE

If you do not want to enter characters from left to right, you can use the Character Entry Mode command to enter characters from right to left. To do this, the cursor must first be sent to the rightmost display location on the screen. After that, the Character Entry Mode command 0000 0111 is entered into the module. This sets the entry mode to autoincrement/display shift left. Now, when characters are entered, they appear on the right-hand side, while the display shifts left for each character entered.

USER-DEFINED GRAPHICS

Commands 0100 0000 to 0111 1111 are used to program user-defined graphics. To program these graphics on-screen, the display is cleared, and the module is sent a Set Display Address command to position the cursor at address 00H. At this point, the contents of the eight user character locations can be viewed by entering binary data 0000 0000 to 0000 0111 in sequence. These characters will appear initially as garbage.

To start defining the user-defined graphics, a Set CGRAM command is sent to the module. Any value between 0100 0000 (40H) and 0111 1111 (7F) will work. Data entered from now on will be used to construct the user-defined graph, row by row. For example, to create a light bulb, the following data entries are made: 0000 1110, 0001 0001, 0001 0001, 0001 0001, 0000 1110, 0000 1010, 0000 1110, 0000 0100. Notice that the first three most significant bits are always 0 because there are only 5 pixels per row. Other user-defined graphics can be defined by entering the 8-byte sequence, and so on. Figure 12.166 shows how the CGRAM address corresponds to the individual pixels of the user-defined graphic.

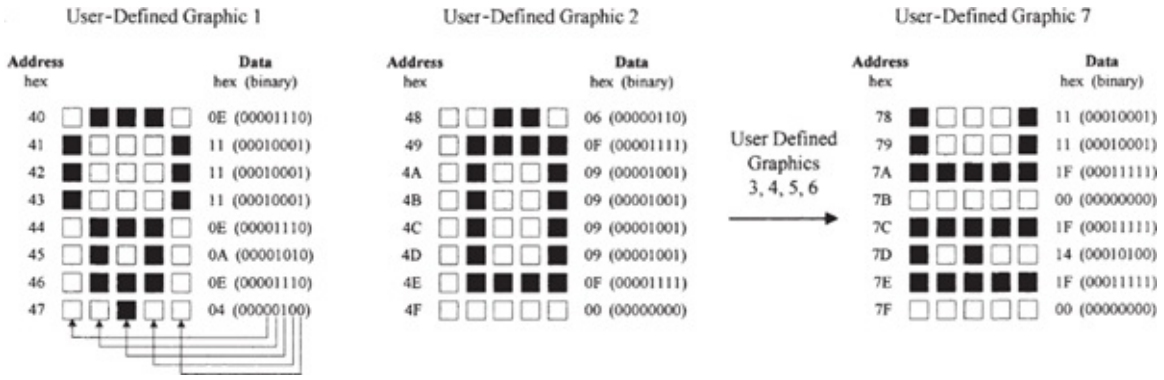


FIGURE 12.166

There are up to eight user-defined graphics that can be programmed. These then become part of the character set and can be displayed by using codes 0000 0000 to 0000 1111 or 0000 1000 to 0000 1111, both of which produce the same result.

One problem when creating user-defined graphics is they will be lost when power is removed from the module—a result of the volatile CGRAM. Typically, the user-defined graphic data is actually stored in an external nonvolatile EPROM or EEPROM, where the data is copied by a microprocessor and loaded into the display module sometime after power-up.

4-BIT DATA TRANSFER

As indicated in the Function Set command, the LCD module is capable of both 8-bit and 4-bit data transfer. In 4-bit mode, only data lines D_4–D_7 are used. The other four lines, D_0–D_3, are left either floating or tied to the power supply. To send data to the display requires sending two 4-bit chunks instead of one 8-bit word.

When power is first applied, the module is set up for 8-bit transfer. To set up 4-bit transfer, the Function Set command with binary value 0010 0000 is sent to the display. Note that since there are only four data lines in use, all 8 bits cannot be sent. However, this is not a problem, since the 8-bit/4-bit selection is on data bit D_4. From now on, 8-bit character and command bits must be sent in two halves, the first 4 most significant bits and then the remaining 4 bits. For example, to write character data 0100 1110 to the display requires setting *RS* high, applying 0100 to the data lines, pulsing *E* low, then applying 1110 to the data lines, and pulsing *E* low again.

The 4-bit transfer is frequently used when the LCD module is interfaced with a microcontroller that has limited I/O lines. See Fig. 12.166.

12.12 Memory Devices

Memory devices provide a means of storing data on a temporary or permanent basis for future recall. The storage medium used in a memory device may be a semiconductor-based IC (primary memory), a magnetic tape, a magnetic disk, or an optical disk (secondary memories). In most cases, the secondary memories are capable of storing more data than primary memories because their surface areas are larger. However, secondary memories take much longer to access (read or write) data because memory locations on a disk or tape must be physically positioned to the point where they can be read or written to by the read/write mechanism. Within a primary memory device, memory locations are arranged in tiny regions within a large matrix, where each memory location can be accessed quickly (matter of nanoseconds) by applying the proper address signals to the rows within the matrix.

Figure 12.167 shows an overview of primary and secondary memories. In this section, we'll discuss only the primary memories, since these devices are used more frequently in designing gadgets than secondary memories. Secondary memories are almost exclusively used for storing large amounts of computer data, audio data, or video data.

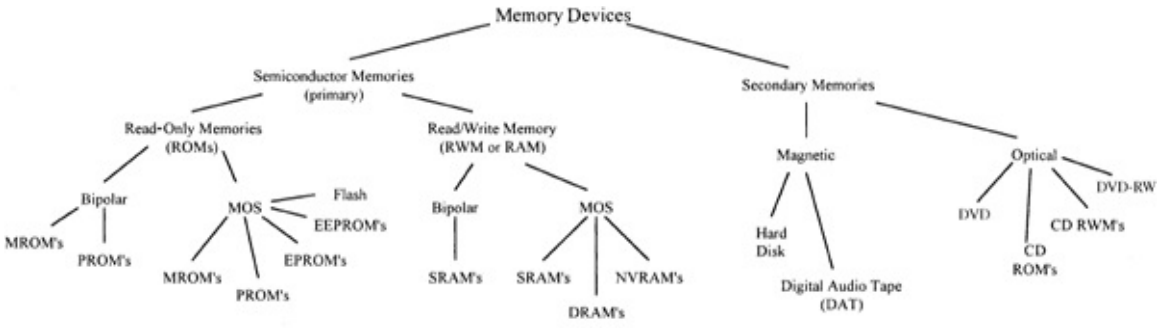


FIGURE 12.167

Today, the technology used in the construction of primary memory devices is almost exclusively based on MOSFET transistors. Bipolar transistors are also used within memory ICs. However, these devices are less popular because the amount of data they can store is significantly smaller than that of a memory IC built with MOSFET transistors. At one time, bipolar memories had a significant edge in speed over MOSFET memories, but today the speed gap has almost disappeared.

Memory devices consist of two basic subfamilies: *read-only memory* (ROM) and *read/write memory* (RWM), which is more commonly referred to as *random-access memory* (RAM). Within each of these subfamilies exist more subfamilies, as shown in Fig. 12.167. Let's start out by discussing the ROM devices.

12.12.1 Read-Only Memory

ROM is used to store data on a permanent basis. These devices are capable of random access, like RAM devices, but unlike RAM devices, they do not lose stored data when power is removed from the IC.

ROM is used in nearly all computers to store boot-up instructions (such as stack allocation, port and interrupt initializations, and instructions for retrieving the operating system from disk storage) that are enacted when the computer is first turned on.

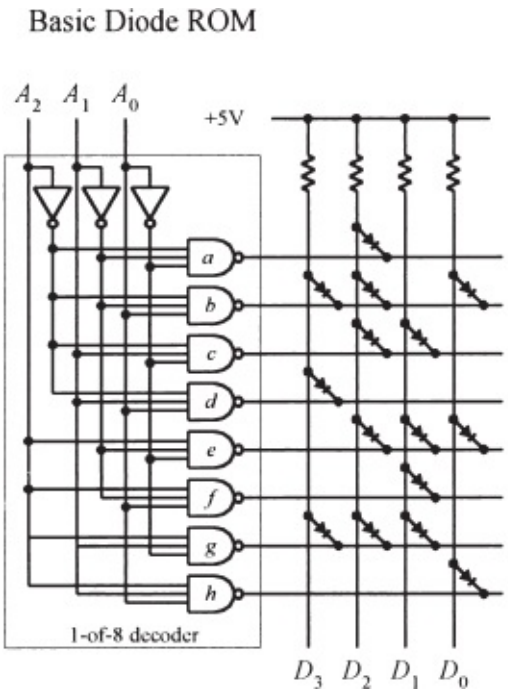
In some microcontroller applications (simple-function gadgets, appliances, toys, and so on), the entire stand-alone program resides in ROM. The microcontroller's central processing unit (CPU) retrieves the program instructions and uses volatile RAM for temporary data storage as it runs through the ROM's stored instructions.

In some instances, you find ROM within discrete digital hardware, where it is used to store lookup tables or special code-conversion routines. For example, digital data from an ADC could be used to address stored words that represent, say, a binary equivalent to a temperature reading in Celsius or Fahrenheit. This also can be used to replace a complex logic circuit, where, instead of using a large number of discrete gates to get the desired function table, you simply program the ROM to provide the designed output response when input data is applied. The last few applications mentioned, however, are becoming a bit obsolete—the microcontroller seems to be taking over everything.

ROM is generally used for read-only operations and not written to after initially programmed. However, some ROM-like devices, such as EPROM, EEPROM, and flash memory, are capable of erasing stored data and rewriting data to memory. Before we take a look at these erasable ROM-like devices, let's first cover some memory basics.

12.12.2 Simple ROM Made Using Diodes

To get a general idea of how ROM works, let's consider the simple circuit in Fig. 12.168.



A_2 A_1 A_0	D_3 D_2 D_1 D_0	Active gate (sink)
0 0 0	1 0 1 1	*a*
0 0 1	0 0 1 0	*b*
0 1 0	1 0 0 1	*c*
0 1 1	0 1 1 1	*d*
1 0 0	1 0 0 0	*e*
1 0 1	1 1 0 1	*f*
1 1 0	0 0 0 1	*g*
1 1 1	1 1 1 0	*h*

This is a simple ROM device that uses an address decoder IC to access eight different 4-bit words stored in a diode matrix. Data to be read is output via the D_3–D_0 lines. The diode matrix is broken up into rows and columns. The intersection of a row and column represents a bit location. When a given row and column are linked together with a diode, the corresponding data output line goes low (0) when the corresponding column is selected by the address decoder via the A_2–A_0 inputs. When a specific row is addressed, the NAND gate sinks current, so the current from the supply passes through the diode and into the NAND gate's output. This makes the corresponding data line low. When no diode is placed between a given column and row, the

corresponding data line goes high (0) when the corresponding row is selected by the address decoder. (There is no path to ground in this case.) In this particular example, we have an 8 × 4 ROM (eight different 4-bit words). By increasing the width of the matrix (adding more columns), it is possible to increase the word size. By increasing the height of the matrix (adding more rows—more addresses), it is possible to store more words. In other words, we could make an *m* × *n* ROM.

FIGURE 12.168

In reality, today's ROM devices rarely use diode memory cells. Instead, they typically use transistor-like memory cells formed on silicon wafers. Also, a more realistic ROM device comes with three-state output buffers that can be enabled or disabled (placed in a high *Z* state) by applying a control signal. The three-state buffers make it possible to effectively disconnect the memory from a data bus to which it is attached. (In our simple diode memory circuit, the data is always present on the output lines.) The basic layout, with address decoder and memory cells, is pretty much the same for all memory devices. There are additional features, however, and we'll discuss these in a minute. First, let's cover some memory nomenclature.

12.12.3 Memory Size and Organization

A ROM that is organized in an $n \times m$ matrix can store n different m-bit words; in other words, it can store $n \times m$ bits of information. To access n different words requires $\log_2 n$ address lines. For example, our simple ROM in Fig. 12.168 requires $\log_2 8 = 3$ address inputs (this may look more familiar: $2^3 = 8$). Note that within multiplexed memories and memories that come with serial inputs, the actual physical number of address inputs is ether reduced or the address information is entered serially, along with data and other protocol information.

In terms of real memory ICs, the number of address inputs is typically eight or higher (for parallel input devices at any rate). Common memory sizes are indicated in Table 12.2. Note that in the table, 1K is used to represent 1024 bits, not 1000 bits, as the k (kilo) would lead you to believe. By digital convention, we say that $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, … $2^8 = 256$, $2^9 = 512$, $2^{10} = 1{,}024$ (or 1 K), $2^{11} = 2{,}048$ (or 2 K), … $2^{18} = 262{,}144$ (256K), $2^{19} = 524{,}288$ (540 K), $2^{20} = 1{,}048{,}576$ (or 1 M, for mega), $2^{21} = 2{,}097{,}152$ (2 M), … $2^{30} = 1{,}073{,}741{,}824$ (or 1 G, for giga), and so on.

TABLE 12.2 Common Memory Sizes

NO. OF ADDRESS LINES	NO. OF MEMORY LOCATIONS	NO. OF ADDRESS LINES	NO. OF MEMORY LOCATIONS	NO. OF ADDRESS LINES	NO. OF MEMORY LOCATIONS
8	256	14	16,384 (16 K)	20	1,048,576 (1 M)
9	512	15	32,768 (32 K)	21	2,097,152 (2 M)
10	1,024 (1 K)	16	65,536 (64 K)	22	4,194,304 (4 M)
11	2,048 (2 K)	17	131,072 (128 K)	23	8,388,608 (8 M)
12	4,096 (4 K)	18	262,144 (256 K)	24	16,777,216 (16 M)
13	8,192 (8 K)	19	524,288 (540 K)	25	33,554,432 (32 M)

If this convention confuses you, it should. It is not exactly obvious, and leaves you scratching your head. Also, when a data sheet says 64 K, you need to read further to figure out what the actual organization is, say, 2048 × 32 (2 *K* × 32), 4096 × 16 (4 *K* × 16), 8192 × 8 (8 *K* × 8), 16,384 × 4 (16 *K* × 4), or another system.

In Table 12.2, watch out for terms such as kB, MB, and GB. These terms refer to bytes not bits; the *B* signifies 1 byte, or 8 bits. This means that a memory that stores 1 kB actually stores 1 *K* × 8 (8 K) bits of data. Likewise, memories that store 1 MB and 1 GB actually store 1 *M* × 8 (8 M) and 1 *G* × 8 (8 G) bits of data, respectively.

12.12.4 Simple Programmable ROM

Figure 12.169 shows a more accurate representation of ROM-type memory. Unlike the diode ROM, each memory cell contains a transistor and fusible link. Initially, the ROM has all programmable links in place. With every programmable link in place, every transistor is biased on, causing high voltage levels (logic 1s) to be stored throughout the array. When a programmable link is broken, the corresponding memory cell's transistor turns off, and the cell stores a low voltage level (logic 0). Note that this ROM contains three-state output buffers that keep the output floating until a low is applied to the Chip Enable ($\overline{CE}$) input. This feature allows the ROM to be interfaced with a data bus.

A basic ROM circuit schematic is shown in Fig. 12.169, along with the appropriate address and chip-enable waveforms needed to enact a read operation. To read data stored at a given address location, the Chip Enable input is set high to disable the chip (remove old data from data outputs)—see time t_0. At time t_2, a new address is placed on the 3-bit address bus (A_2, A_1, and A_0). At time t_2, the Chip Enable input is set low, which allows addressed data stored in memory to be output via D_3, D_2, D_1, and D_0.

In reality, the stored data is not output immediately but is delayed for a very short time (from t_2 to t_3) due to the propagation delay that exists between the initial chip enable signal and the signal that reaches the enable leads of the output buffers. In memory lingo, the time from t_1 to t_4 is referred to as the *access time*, which is between around 10 ns and a couple hundred nanoseconds, depending on the specific technology used.

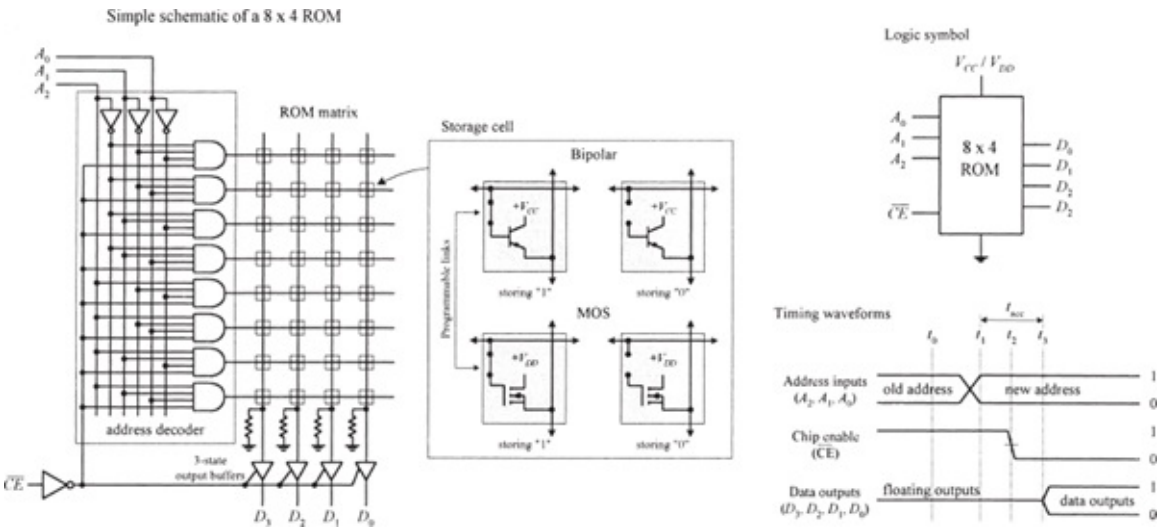


FIGURE 12.169

Now two important questions need addressing:

- How does one "break" a programmable link? In other words, how do we program the ROM?
- Is it possible to restore a broken programmable link back to its unbroken" state? In other words, is it possible to reprogram the ROM?

These lead to the next topic.

12.12.5 ROM Devices

There are basically two kinds of ROMs: those that can be programmed only once and those that can be reprogrammed any number of times. One-time programmable memories include the mask ROM (MROM) and the programmable ROM (PROM). ROMs that can be reprogrammed include the erasable programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), and flash

memory.

MROM

An MROM is a custom memory device that is permanently programmed by the manufacturer simply by adding or leaving out diodes or transistors within a memory matrix. In order to create a desired memory configuration, you must supply the manufacturer with a truth table stating which data configuration is desired. Using the truth table, the manufacturer then generates a *mask* that is used to create the interconnections within the memory matrix during the fabrication process.

As you can imagine, producing a custom MROM is not exactly cheap; in fact, it is rather costly (more than $1,000). It is only worthwhile using an MROM if you plan to mass produce some device that requires the same data instructions (for example, program instructions) over and over again—no upgrades to memory needed in the future. In this case, the cost for each IC—after the initial mask is made—is relatively cheap, assuming you need more than a couple thousand chips.

MROMs are commonly found in computers, where they are used to store system operating instructions and data that is used to decode keyboard instructions.

PROM

PROMs are fusible-link programmable ROMs. Unlike the MROM, with PROM devices, data is not etched in stone. Instead, the manufacturers provide you with a memory IC whose matrix is clean (full of 1s). The number of bits and the configuration (n × m) of the matrix vary depending on specific ROM. To program the memory, each fusible link must be blown with a high-voltage pulse (such as 21 V).

The actual process of blowing individual fuses requires a PROM programming unit. This PROM programmer typically includes a hardware unit (where the actual PROM IC is attached), along with programming cable that is linked to a computer (such as via a serial or parallel port). Using software provided by the manufacturer, you enter the desired memory configuration in the program running on the computer and then press a key, which causes the software program to instruct the external programming unit to blow the appropriate links within the IC.

PROMs are relatively easy to program once you have figured out how to use the software, but as with MROMs, once the device is programmed; the memory cannot be altered. In other words, if you mess things up, you must begin afresh with a new chip. These devices were popular some years ago, but today they are considered obsolete.

The most popular ROM-type devices used today are EPROM, EEPROM, and flash memory. These devices, unlike MROM and PROM devices, can be erased and reprogrammed—a very useful feature when prototyping or designing a gadget that requires future memory alterations.

EPROM

An EPROM is a device whose memory matrix consists of a number of specialized MOSFET transistors. Unlike a conventional MOSFET transistor, the EPROM transistor has an additional floating gate that is buried beneath the control gate—insulated from both the control gate and drain-to-source channel by an oxide layer (see Fig. 12.170).

In its erased (unprogrammed) state, the floating gate is uncharged and does not affect the normal operation of the control gate (which when addressed results in a high voltage or logic 1 being passed through to the data lines). To program an individual transistor, a high-voltage pulse (around 12 V) is applied between the control gate and the drain terminal. This pulse, in turn, forces energetic electrons through the insulating layer and onto the floating gate (referred to as *hot electron injection*). After the high voltage is removed, a negative charge remains on the floating gate and will stay there for decades under normal operating conditions.

With the negative charge in place, the normal operation of the control gate is inhibited; when the control gate is addressed, the charge on the floating gate prevents a high voltage from reaching the data line—the addressed data appears as a low, or logic 0.

In order to reprogram (erase) an EPROM, you must first remove the device from the circuit and then remove a sticker covering its quartz window. After that, you remove all stored charges on the floating gates by shining ultraviolet (UV) light through the window onto the interior transistor matrix. The UV light liberates the stored electrons within the floating gate region by supplying them with enough energy to force them through the insulation. It usually takes 20 minutes of UV exposure for the whole memory matrix to be erased. The number of times an EPROM can be reprogrammed is typically limited to a couple hundred cycles. After that, the chip degrades considerably.

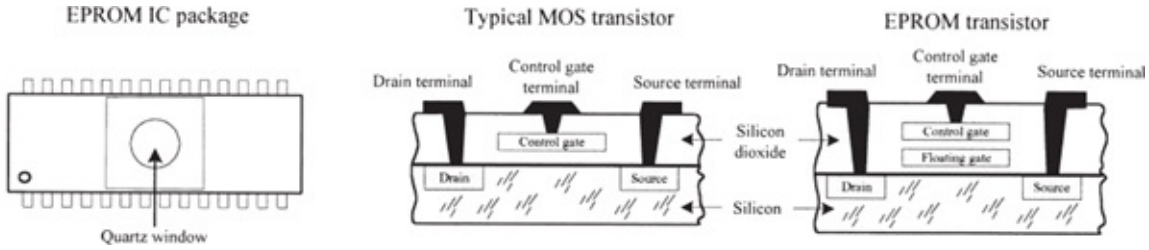


FIGURE 12.170

EPROM is often used as nonvolatile memory within microprocessor-based devices that require the provision for future reprogramming. They are frequently used in prototyping and then substituted with MROMs during the mass-production phase. EPROMs are also integrated within microcontroller chips where their sole purpose is to store the microcontroller's main program (more on this in Chapter 13).

EEPROM

An EEPROM device uses a technology somewhat related to the EPROM, but it does not require out-of-circuit programming or UV erasing. Instead, an EEPROM device is capable of selective memory cell erasure by means of controlled electrical pulses.

In terms of architecture, an EEPROM memory cell consists of two transistors: one transistor resembles the EPROM transistor and is used to store data, and the other transistor is used to clear charge from the first transistor's floating gate. By supplying the appropriate voltage level to the second transistor, it is possible to selectively erase individual memory cells instead of having to erase the entire memory matrix, as is the case with EPROM. The only major disadvantage with EEPROM over EPROM is size—due to the two transistors. However, today, with the introduction of new fabrication processes, size is becoming less of an issue.

In terms of applications, EEPROM is ideal for remembering configuration and calibration settings of a device when the power is turned off. For example, EEPROM is found within TV tuners, where it is used to remember the channel, volume setting of the audio amplifier, and so on when the TV is turned off. EEPROM is also found on microcontrollers, where it can be used to store the main program or to hold other nonvolatile data

Flash Memory

Flash memory is generally regarded as the next evolutionary step in ROM technology that combines the best features of EPROM and EEPROM. These devices have the advantage of both in-circuit programming (like EEPROM) and high storage density (like EPROM).

Some variants of flash memory are electrically erasable, like EEPROM, but must be erased and reprogrammed on a device-wide basis, similar to EPROM. Other devices are based on a dual transistor cell and can be erased and reprogrammed on a word-by-word basis. Flash devices are noted for their fast

write and erase times, which exceed those of EEPROM devices.

Flash memories are becoming very popular as mass-storage devices. They are found in digital cameras, where a high-capacity flash memory card is inserted directly into a digital camera and can store hundreds or thousands of high-resolution images. They are also used in digital music players, cellular phones, tablets, and so on.

Microcontrollers often include flash memory to contain their program.

Serial Access Memory

So far, we have seen only memories that incorporated parallel access. These devices sit directly on the address and data buses, making it easy for processors to quickly access the memory. Serial access memory is easy to use in principle; however, since all their address lines are typically tied to an address bus within a microprocessor-based system, it is not uncommon for the data to be inadvertently destroyed when the processor runs amuck (issues an undesired write).

Another type of memory that can "hide" the memory from the processor, as well as reduce the total number of pins, uses a serial access format. To move data to and from memory and the processor, a serial link is used. This serial link imposes a strict protocol on data transfers that practically eliminates the possibility that the processor can destroy data accidentally.

Figure 12.171 shows a few serial EPROM and EEPROM devices from Microchip. The SDA pin found in the EEPROM devices acts as a bidirection data lead used to transfer address and data information into the memory IC, as well as transfer data out to the processor. The SCL pin is the serial clock input used to synchronize the data transfer from and to the device. The 24xx64 and 24LC01B/02B EEPROMs also come with special device address inputs A_0, A_1, and A_2, which are used for multiple device operation. WP is used to enable normal memory operation (read/write entire memory) or inhibit write operations.

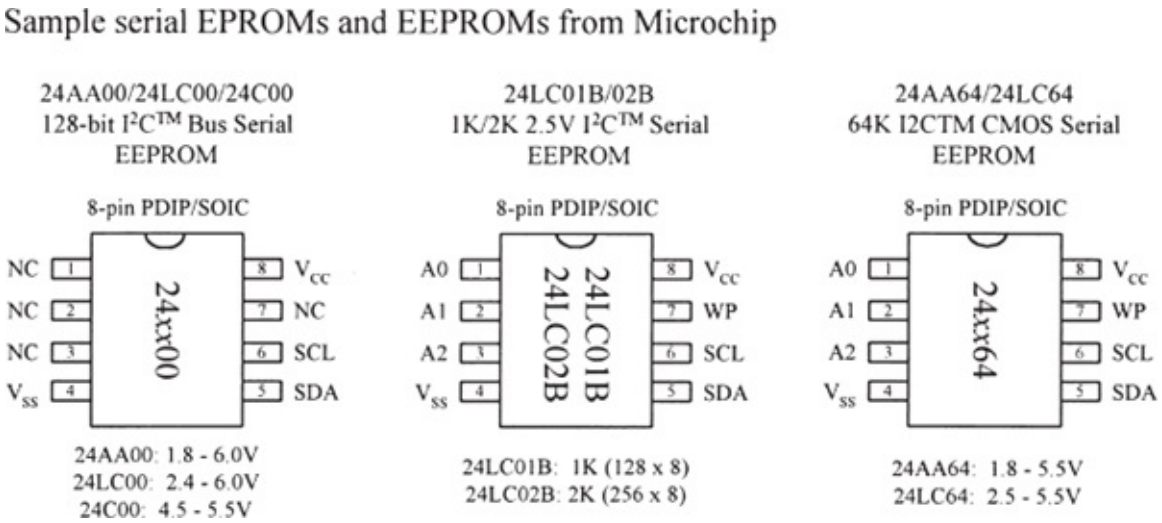


FIGURE 12.171

Controlling a serial memory device is a bit complex, due to the serial protocol and variations in protocol from IC to IC. If you want to learn more about these serial memories (and you should—they are very handy in microcontroller applications for logging data and storing programs and similar tasks), check out the various manufacturers' websites and read through their data sheets.

12.12.6 RAM

The erasable programmable ROM devices, like EEPROM, have limited read/write endurance—around 100,000 cycles—and take considerable time to write to memory. For applications that require constant and quick read and write cycles, it is necessary to use RAM. This type of memory is used for temporary

storage of data and program instructions in microprocessor-based applications. Unlike ROM devices, however, RAM devices are volatile, which means they lose their data if power to the IC is interrupted.

There are two basic types of RAM:

- *Static RAM (SRAM)*: In an SRAM device, data is stored in memory cells that consist of flip-flops. A bit that is written into an SRAM memory cell stays there until overwritten or until the power is turned off.
- *Dynamic RAM (DRAM)*: In a DRAM device, a bit written to the memory cell will disappear within milliseconds if not refreshed, or supplied with periodic clocking to replenish capacitor charge lost to leakage.

In general, the major practical differences between SRAM and DRAM include overall size, power consumption, speed, and ease of use. In terms of size, DRAM devices can hold more data per unit area than SRAM devices, since a DRAM's capacitor takes up less space than an SRAM's flip-flop. In terms of power consumption, SRAMs are more energy-efficient because they do not require constant refreshing. In terms of speed and ease of use, SRAMs are superior because they do not require refresh circuitry.

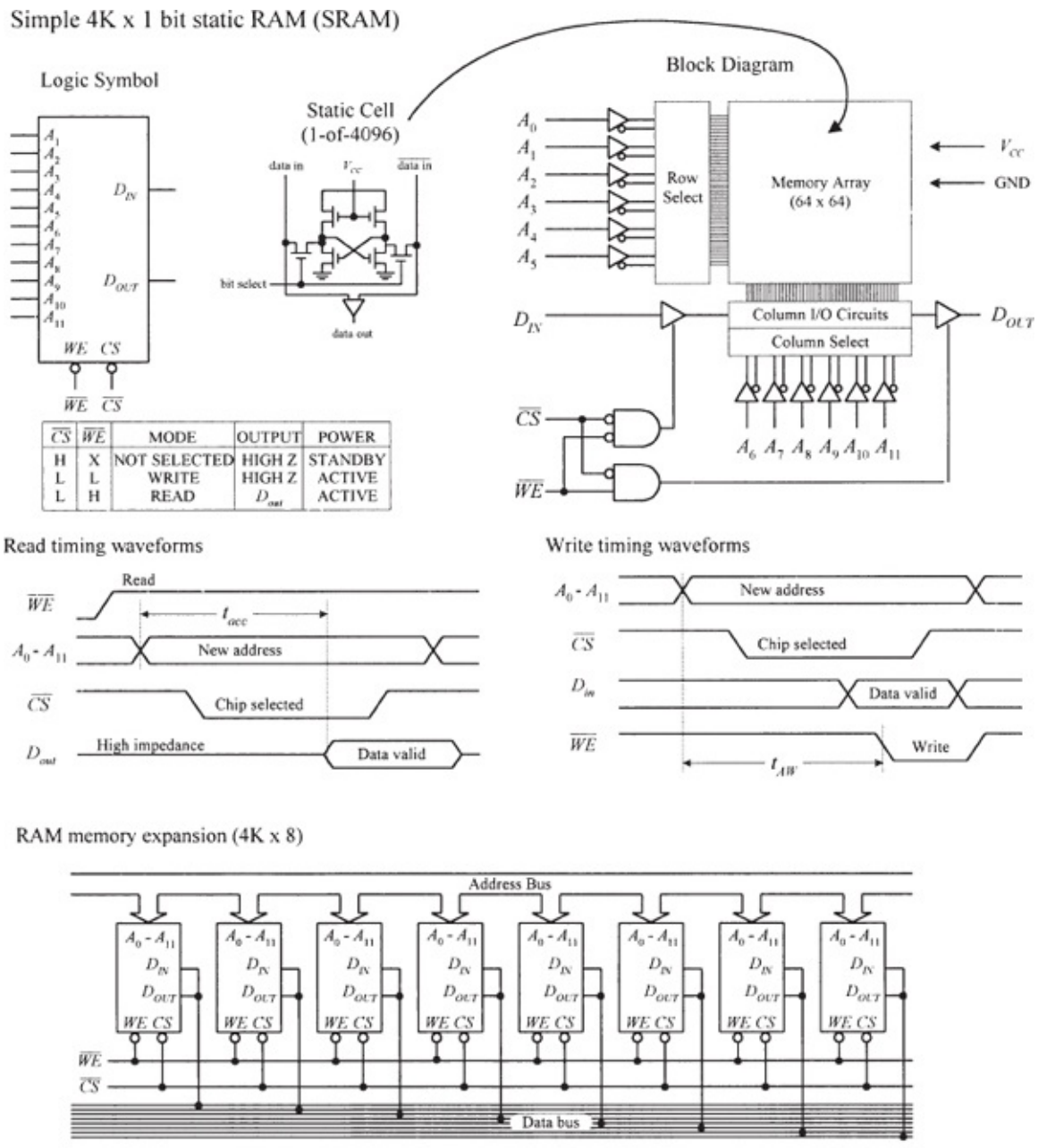
In terms of applications, SRAMs are used when relatively small amounts of read/write memory are needed and are typically found within application-specific ICs that require extremely low standby power. For example, they are frequently used within portable equipment such as pocket calculators. SRAM is also integrated into all modern microprocessors, where it acts as on-chip cache memory that provides a high-speed link between the processor and memory. On the other hand, DRAM is used in applications where a large amount of read/write memory (within the megabyte range) is needed, such as within computer memory modules.

In most situations, you do not need to worry about dealing with discrete RAM memory ICs. Most of the time, RAM is already built into a microcontroller or conventionally housed on PCB memory modules that simply plug into a computer's memory banks. In both these cases, you really do not need to know how to use the memory, because you can let the existing hardware and software take care of the addressing, refreshing, and so on. For this reason, we will not discuss the finer details of the various discrete SRAM and DRAM ICs out there. Instead, we will take a look at some SRAM and DRAM block diagrams that illustrate the basics, and then we will discuss some memory packages, such as SIMMs and DIMMs, that are used within computers.

Very Simple SRAM

Figure 12.172 shows a very elementary SRAM that is set up with a 4096 (4 K) × 1-bit matrix. It uses 12 address lines to address 4096 different memory locations; each location contains a flip-flop. The memory matrix is set up as a 64 × 64 array, with A_0 to A_5 identifying the row and A_6 to A_{11} identifying the column to pinpoint the specific location to be used. The box labeled "Row Select" is a 6-to-64 decoder for identifying the appropriate 1-of-64 row. The box labeled "Column Select" is also a 6-to-64 decoder for identifying the appropriate 1-of-64 column.

To write a new bit of data to memory, the bit is applied to D_{IN}, the address lines are set, the Chip Select input ($\overline{CS}$) is set low (to enable the chip), and the Write Enable input ($\overline{WE}$) is set low (to enable the D_{IN} buffer). To read a bit of data from memory, the address lines are set, $\overline{CS}$ is set low, and $\overline{WE}$ is set high (to enable the D_{OUT} buffer). See the timing waveforms in Fig. 12.172.



$\overline{CS}$	$\overline{WE}$	MODE	OUTPUT	POWER
H	X	NOT SELECTED	HIGH Z	STANDBY
L	L	WRITE	HIGH Z	ACTIVE
L	H	READ	D_{out}	ACTIVE

FIGURE 12.172

By combining eight 4 K × 1 SRAM ICs together, as shown in the lower circuit in Fig. 12.172, the memory can be expanded to form a 4 K × 8 configuration, which is useful in simple 8-bit microprocessor systems. When an address is applied to the address bus, the same address locations within each memory IC are accessed at the same time. Therefore, each data bit of an 8-bit word applied to the data bus is stored in the same corresponding address locations within the memory ICs.

There are other SRAM ICs that come with configurations larger than n × 1. For example, they may come in, say, an n × 4 or n × 8 configuration. As with the n × 1 devices, these SRAMs can be expanded (two n × 8 devices could be combined to form an n × 16 expanded memory, four n × 8 devices could be combined to form an n × 32 expanded memory, and so on).

Serial SRAM with a similar interface to serial EEPROM is also available. For an example of the types of serial SRAM available see http://ww1.microchip.com/downloads/en/DeviceDoc/22127a.pdf.

Note on Nonvolatile SRAMs

In many applications, it would be ideal to have a memory device that combines both the speed and cycle endurance of an SRAM with the nonvolatile characteristics of ROM devices. To solve this problem, manufacturers have created what are called *nonvolatile SRAMs*. One such device incorporates a low-

power CMOS SRAM together with a lithium battery and power-sensory circuitry. When the power is removed from the chip, the battery kicks in, providing the flip-flops with sufficient voltage to keep them set (or reset). SRAMs with battery backup, however, have limited lifetimes due to the life expectancies of the lithium batteries—around ten years.

Another nonvolatile SRAM that requires no battery backup is referred to as *non-volatile RAM* (NOVRAM). These chips incorporate a backup EEPROM memory array in parallel with an ordinary SRAM array. During normal operation, the SRAM array is written to and read from just like an ordinary SRAM. When the power supply voltage drops, an onboard circuit automatically senses the drop and performs a store operation that causes all data within the volatile SRAM array to be copied to the nonvolatile EEPROM array. When power to the chip is turned on, the NOVRAM automatically performs a recall operation that copies all the data from the EEPROM array back into the SRAM array. A NOVRAM has essentially unlimited read/write endurance, like a conventional SRAM, but has a limited number of store-to-EEPROM cycles—around 10,000.

DRAM

Figure 12.173 shows a very basic 16 $K \times 1$ DRAM. Normally, to access all 16,384 memory locations (capacitors) would require 14 address lines. However, in this DRAM (as within most large-scale DRAMs), the number of address lines is cut in half by multiplexing.

To address a given memory location is a two-step process. First, a 7-bit row address is applied to A_0–A_6, and then Row Address Strobe ($\overline{RAS}$) is sent low. Second, a 7-bit column address is applied to A_0–A_6, and then Column Address Strobe ($\overline{CAS}$) is sent low. At this point the memory location is latched and can now be read or written to by using the $\overline{WE}$ input. When $\overline{WE}$ is low, data is written to the RAM via D_{in}. When $\overline{WE}$ is high, data is read from the RAM via D_{out}. See the timing waveforms in Fig. 12.173.

Simple DRAM devices like this must be refreshed every 2 ms or sooner to replenish the charge on the internal capacitors. For our simple device, there are three ways to refresh the cells: use a Read cycle, use a Write cycle, or use an $\overline{RAS}$-only cycle. Unless you are reading or writing to and from all 128 rows every 2 ms, the $\overline{RAS}$-only cycle is the preferred technique. To perform this cycle, $\overline{CAS}$ is set high, A_0–A_6 are set up with the row address 000 0000, $\overline{RAS}$ is pulsed low, the row address is then incremented by 1, and the last two steps are repeated until all 128 rows have been accessed.

As you can see, needing to come up with the timing waveforms to refresh the memory is a real pain. For this reason, manufacturers produce DRAM controllers or actually incorporate automatic refreshing circuitry within the DRAM IC. In other words, today's DRAMs have all the "housekeeping" functions built in. Practically speaking, this makes the DRAM appear static to the user.

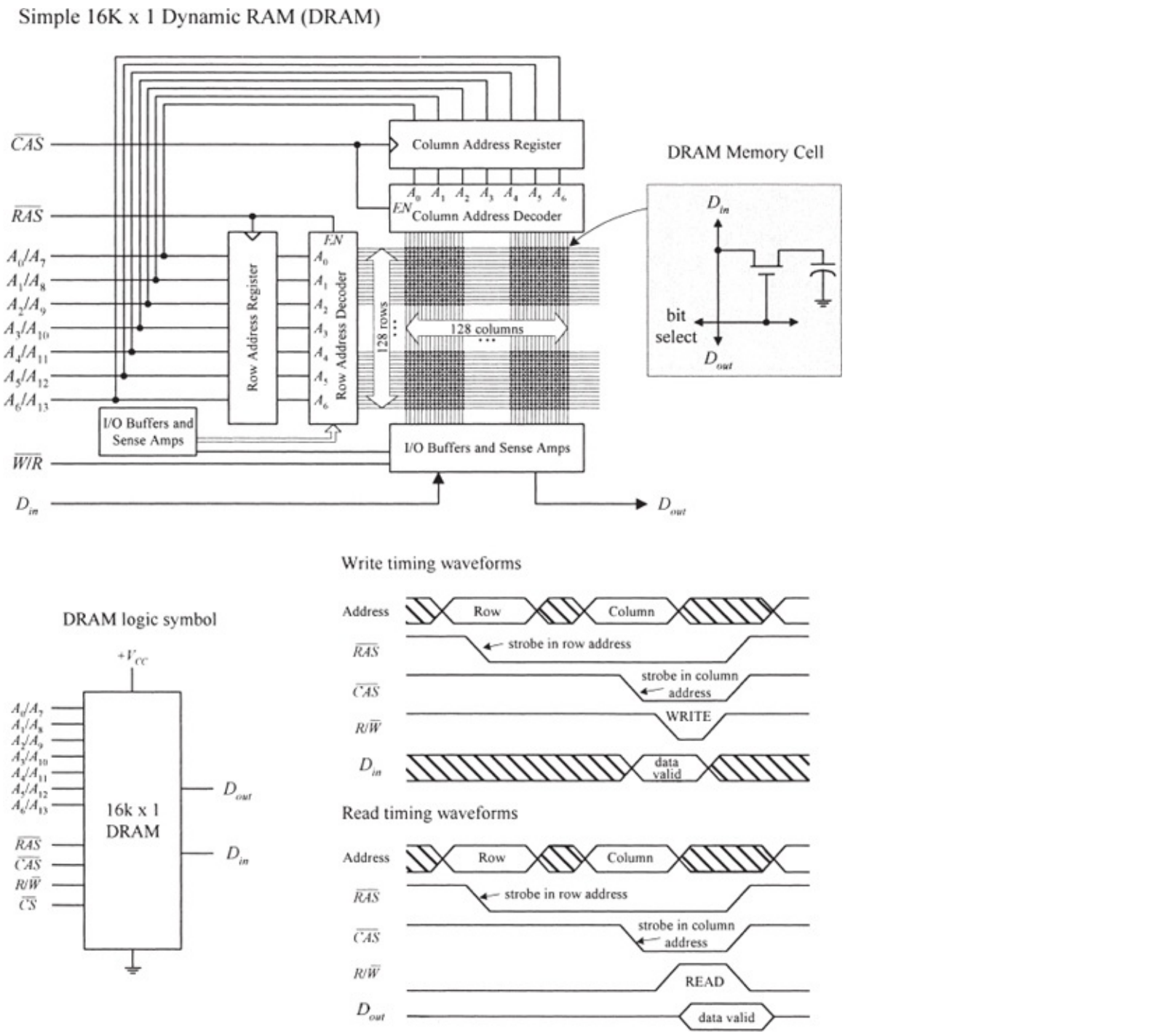


FIGURE 12.173

DRAM technology is changing very rapidly. Today, there are a number of DRAM-like devices that go by such names as ECC DRAM, EDO DRAM, SDRAM, SDRAM II, RDRAM, and SLDRAM (we will discuss some of these at the end of this chapter).

Computer Memory

As mentioned, you typically do not need to worry about RAM. (The only real exception would be NOVRAM that is used in many EEPROM-like applications.) RAM is usually either already integrated into a chip, such as a microcontroller, or is placed in reduced pin devices like single in-line memory modules (SIMMs) or dual in-line memory modules (DIMMs) that slide (snap) into a computer's memory bank sockets. In both cases, not much thought is needed—assuming you are not trying to design a microcontroller or computer from scratch. The main concern nowadays is figuring out what kind of RAM module to buy for your computer.

Within computers, RAM is used to hold temporary instructions and data needed to complete tasks. This

enables the computer's CPU to access instructions and stored data in memory very quickly. For example, when the CPU loads an application, such as a word processor or page layout program, into memory, the CPU can quickly find what it needs, instead of needing to search for bits and pieces from, say, the hard drive or external drive. For RAM to be quick, it must be in direct communication with the computer's CPU. Early on, memory was soldered directly onto the computer's system board (motherboard). However, over time, as memory requirements increased, having fixed memory onboard became impractical. Today, computers house expansion slots arranged in memory banks. The number of memory banks and the specific configuration vary, depending of the computer's CPU and how the CPU receives information.

Historically, computers initially used either SIMM or DIMM memory modules. Both types of modules use dynamic RAM ICs as the core element. The actual SIMM or DIMM module resembles a PCB and houses a number of RAM ICs that are expanded onboard to provide the necessary bit width required by the CPU using the module. To install a SIMM or DIMM module, simply insert the module into one of the computer memory banks sockets found on the motherboard. Many computer systems use 168-pin DIMMs. Older Pentium and later 486 PCs commonly use 72-pin SIMMs, while still older 486 PCs commonly use 30-pin SIMMs.

A variation on standard sizes of DIMM memory is the SODIMM. SODIMM (Small Outline DIMM) are electrically the same as standard DIMMs but smaller in size and intended for use in laptop computers.

DIMMs

DIMMs have opposing pins electrically isolated to form two separate contacts. DIMMs are often used in computer configurations that support a 64-bit or wider memory bus.

Each new generation of memory brings with it new formats of DIMM, with strategically placed slots to prevent accidental insertion of the wrong type of memory.

DDR3 memory comes in 240-pin DIMMs. Figure 12.174 shows a somewhat simpler sample package from an old 16 *M* × 64-bit synchronous DRAM that comes in a 168-pin DIMM package.

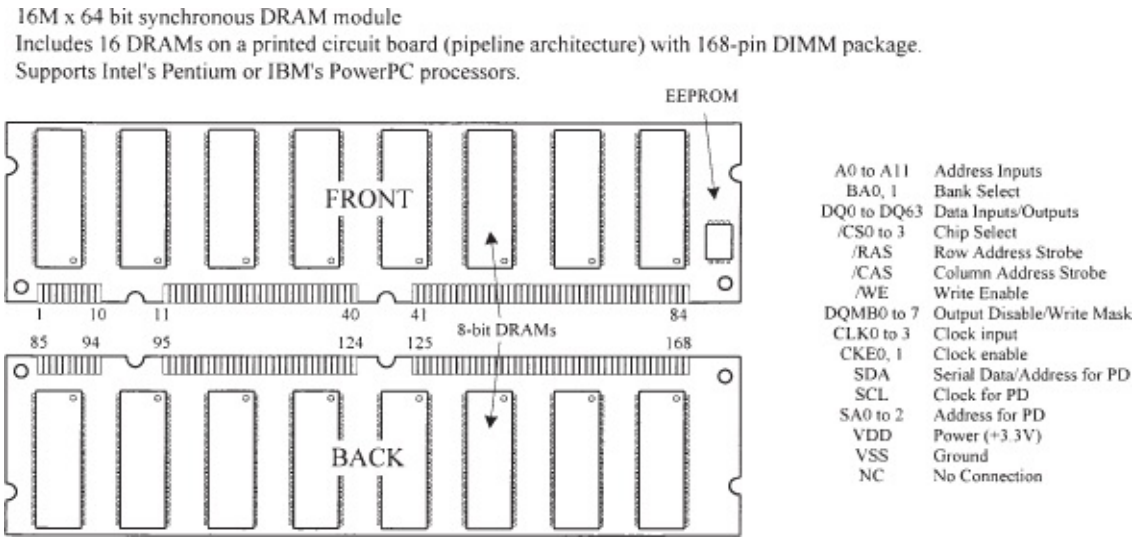


FIGURE 12.174

DRAM Technology Used in Computer Memories

A number of DRAM technologies are incorporated into computer memory modules these days. Extended data out (EDO) memory is a technology that allows the CPU (ones that support EDO) to access memory 10 to 20 percent faster than standard DRAM chips.

Another variation of DRAM is the synchronous DRAM (SDRAM), which uses a clock to synchronize signals input and output on a memory chip. The clock is coordinated with the CPU clock so the timing of

the memory chips and the timing of the CPU are in synch. Synchronous DRAMs save time in executing commands and transmitting data, thereby increasing the overall performance of the computer. SDRAM allows the CPU to access memory approximately 25 percent faster than EDO memory.

Double-data-rate SRAM (DDR or SDRAMM II) is a faster version of SDRAM that is able to read data on both the rising and falling edges of the system clock, thus doubling the data rate of the memory chip. Rambus DRAM (RDRAM) is an extremely fast DRAM technology that uses a high-bandwidth “channel” to transmit data at speeds about ten times faster than a standard DRAM.

CHAPTER 13

Microcontrollers

The microcontroller is essentially a computer on a chip. It contains a processing unit, ROM, RAM, serial communications ports, ADCs, and so on. In essence, a microcontroller is a computer, but without the monitor, keyboard, and mouse. These devices are called *microcontrollers* because they are small (micro) and because they control machines, gadgets, and so on.

With one of these devices, you can build an "intelligent" machine. You write a program on a host computer; download the program into the microcontroller via the USB, parallel, or serial port of the PC; and then disconnect the programming cable and let the program run the machine. For example, in the microwave oven, a single microcontroller has all the essential ingredients to read from a keypad, write information to the display, control the heating element, and store data such as cooking time.

There are literally thousands of different kinds of microcontrollers available. Some are one-time-programmable (OTP), meaning that once a program is written into its ROM (OTP-ROM), no changes can be made to the program. OTP microcontrollers are used in devices such as microwaves, dishwashers, automobile sensor systems, and many application-specific devices that do not require changing the core program. Other microcontrollers are reprogrammable, meaning that the microcontroller's program stored in ROM (which may be EPROM, EEPROM, or flash) can be changed if desired, which is a useful feature when prototyping or designing test instruments that may require future I/O devices.

Microcontrollers are found in bicycle light flashers, data loggers, toys such as model airplanes and cars, antilock braking systems, VCRs, microwave ovens, alarm systems, fuel injectors, exercise equipment, and many other items. They also can be used to construct robots, where the microcontroller acts as the robot's brain, controlling and monitoring various input and output devices, such as light sensors, stepper and servo motors, temperature sensors, and speakers. With a bit of programming, you can make the robot avoid objects, sweep the floor, and generate various sounds to indicate that it has encountered difficulties (such as being low on power or tipped over) or has finished sweeping. The list of applications for microcontrollers is endless, and because they are so widely used, their cost is low.

13.1 Basic Structure of a Microcontroller

Figure 13.1 shows the basic ingredients found within many microcontrollers. These include a CPU, ROM (OTP-ROM, EPROM, EEPROM, of flash), RAM, I/O ports, timing circuitry/leads, interrupt control, a serial port adapter (such as UART or USART), and an ADC/DAC.

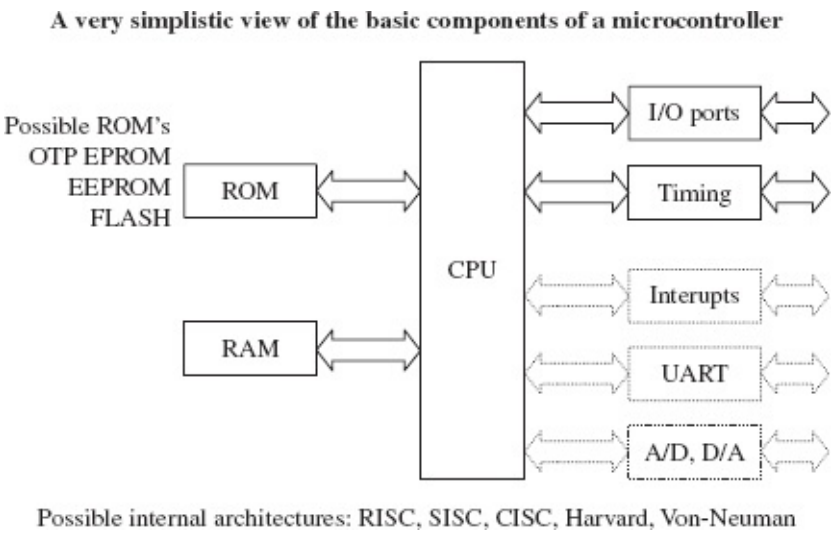


FIGURE 13.1

The CPU retrieves program instructions that the user programs into ROM, while using RAM to store temporary data needed during program execution. The I/O ports are used to connect external devices that send or receive instructions to or from the CPU.

The serial port adapter is used to provide serial communications between the microcontroller and a PC or between two microcontrollers. It is responsible for controlling the different rates of data flow common between devices. Example serial port adapters found within microcontrollers are the universal asynchronous receiver transmitter (UART) and universal synchronous/asynchronous receiver transmitter (USART). The UART can handle asynchronous serial communications, while the USART can handle either asynchronous or synchronous serial communications. Some microcontrollers take this a step further and include an interface for a Universal Serial Bus (USB) interface on the chip.

An interrupt system is used to interrupt a running program in order to process a special routine called the *interrupt service routine*. This boils down to the ability of a microcontroller to respond to external data that requires immediate attention, such as data conveyed by an external sensor indicating important shutdown information, say, when things get too hot or objects get too close. A timer/counter is used to "clock" the device—to provide the driving force needed to move bits around. Most microcontrollers that come with built-in ADCs and DACs can be used to interface with analog transducers, such as temperature sensors, strain gauges, and position sensors.

13.2 Example Microcontrollers

There are many different families of microcontrollers. Two of the most popular are made by the manufacturers Atmel and Microchip. In this section, we will take a close look at microcontrollers from these two manufacturers.

13.2.1 The ATtiny85 Microcontroller

The Atmel ATtiny85 microcontroller is an 8-pin IC available in both surface-mount and through-hole DIL packages. The device is designed to operate with a minimum of external components. Figure 13.2 shows what is inside one of these little packages.

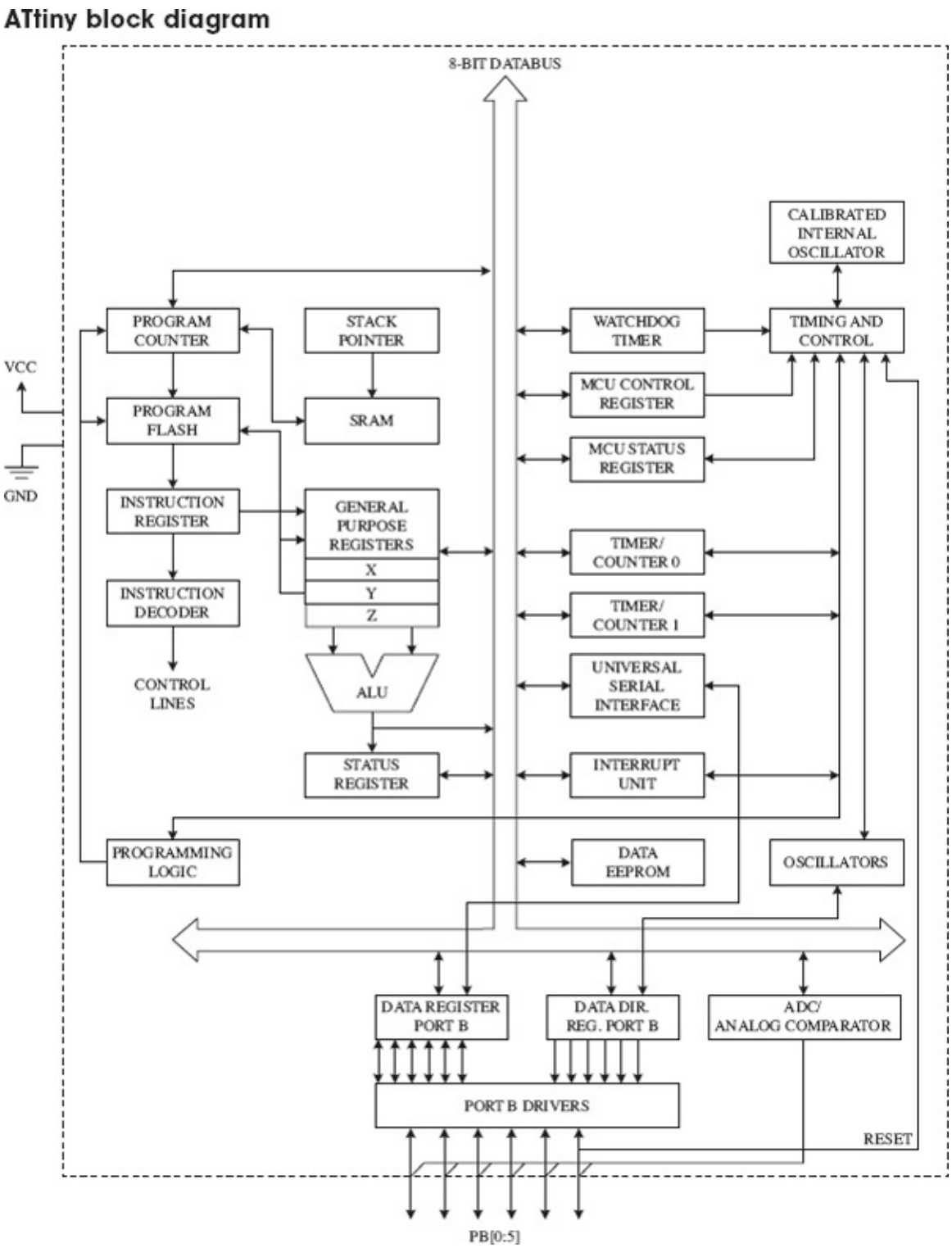


FIGURE 13.2

The ATtiny 85 has three different types of memory:

- 8 kB of flash memory in which the program instructions are stored
- 256 bytes of SRAM, which is used to contain data during execution of instructions
- 512 bytes of EEPROM that is used to store nonvolatile data that needs to be retained after a loss of power

The watchdog timer allows the microcontroller to be put into a sleep mode, in which it consumes negligible power. The watchdog timer will wake up the microcontroller after a certain amount of time has passed.

The device can either use the inaccurate internal oscillator or two of the pins that would otherwise be

used for inputs or outputs can be sacrificed to use an external crystal oscillator.

Two timers can be used to generate internal interrupts; that is, to trigger some code to be executed periodically. External interrupts that are triggered by a change in the level at a pin are also possible.

All of the I/O pins can also be used with the internal ADC.

The ATtiny also has a Universal Serial Interface, which can communicate with a number of different types of serial buses including USB, Inter-Integrated Circuit (I^2C), and serial. There is more about these serial protocols in Section 13.5.

Minimizing External Components

Figure 13.3 illustrates just how few components you need to make something with an ATtiny. The potentiometer is connected to a pin that will be used as an analog input, which could, for example, be used to control the rate at which the LED flashes.

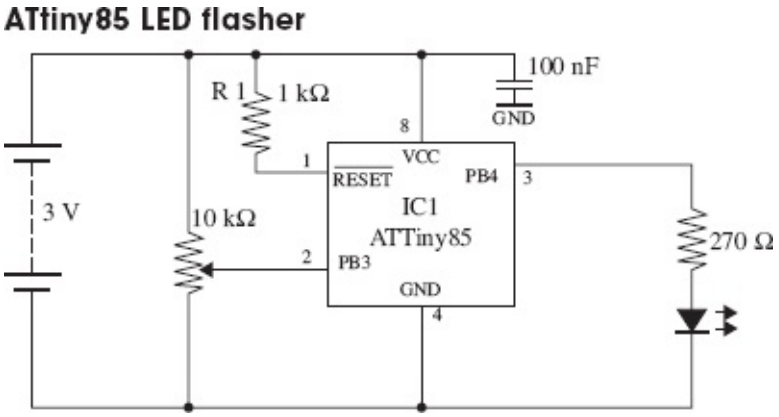


FIGURE 13.3

The resistor R1 could be replaced by a direct connection from the RESET pin to VCC, but by using a resistor here, it becomes possible to have the RESET pin low force a reset—something that is necessary during programming.

The chip will operate from a supply voltage of between 2.7 V and 5.5 V at a clock frequency of 10 MHz or below, making it suitable for running from a 3 V lithium cell or a pair of AA batteries. The clock frequency can be set during programming, and it can also be changed from program code while the ATtiny is actually running. The main reason for controlling the clock frequency is to reduce the power consumption. At 1 MHz, the power can be reduced to just 300 µA and in power-down mode, waiting for an interrupt from the watchdog timer, power consumption is just 0.1 µA.

You may be wondering why we used a microcontroller to create something that we could have made with a 555 timer. Well, why not? The microcontroller is more expensive than a 555 timer, but only around a dollar, and also we can use slightly fewer components with a microcontroller. We also have a great deal more flexibility. Using this same hardware, we could do clever tricks such as turning the LED completely off if the potentiometer is at its most counterclockwise position. And there are three other unused I/O pins that we could do something with.

The catch is that to use a microcontroller, you need to program. But, if you are serious about developing something that will eventually become a product, chances are it will have a microcontroller in it.

Programming the ATtiny with AVR Studio

Atmel, the makers of the ATtiny, supply an integrated development environment (IDE) called AVR Studio (see Fig. 13.4) that takes away some of the pain of microprocessor programming.

AVR Studio with a Blink program

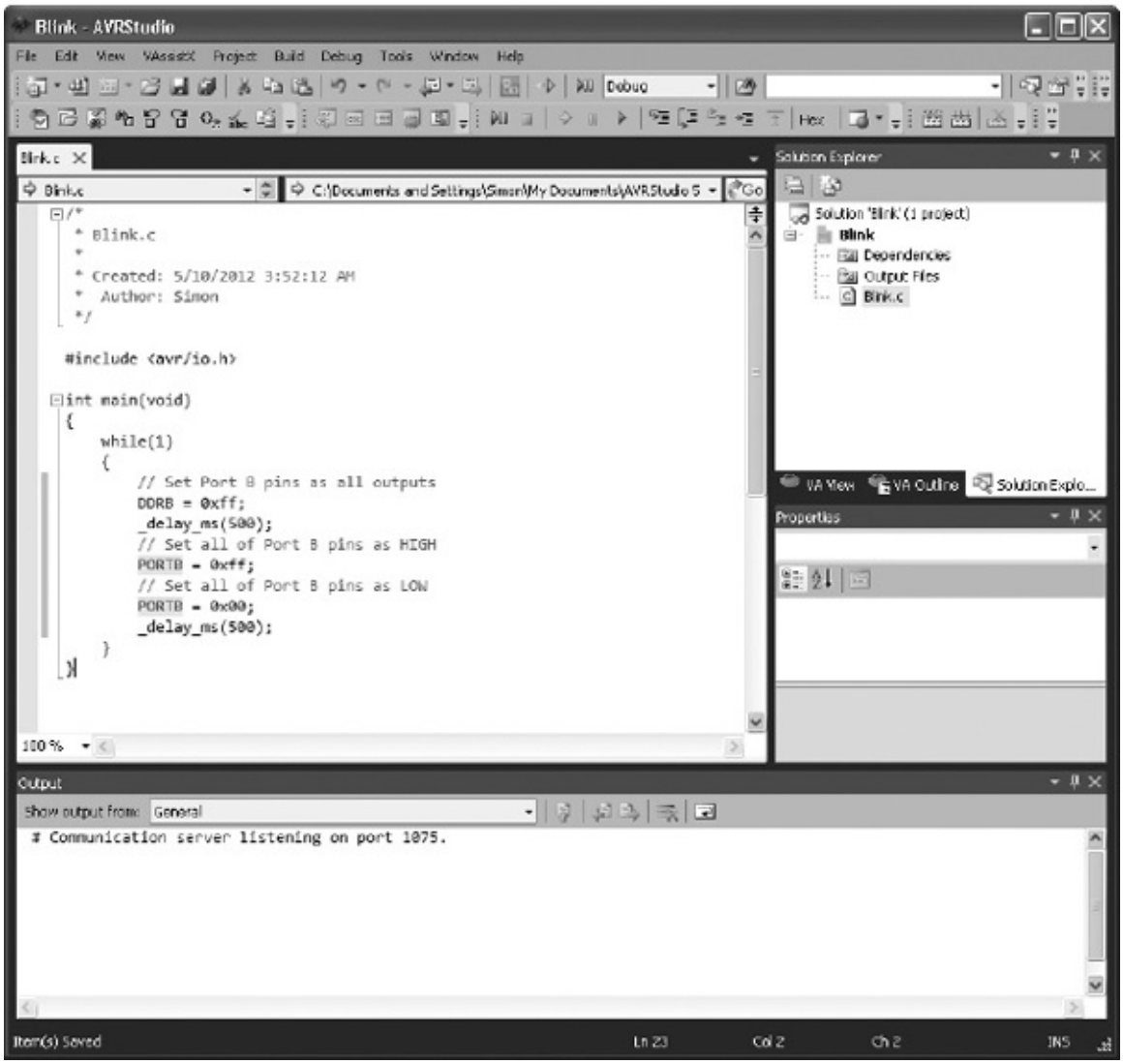


FIGURE 13.4

The ATtiny is usually programmed in C, which offers a good compromise between performance and readability. We define *readability* as the property of a program that allows it to be understood by someone other than the person who wrote it.

The standard AVR Studio way of doing things is powerful and flexible, but the C that you need to write is at a fairly low level, and not nearly as accessible as the BASIC Stamp language.

Programming the ATtiny with Arduino

Many people use the Arduino library to simplify the writing of code. This library includes all sorts of utility functions, rather like the commands found in the BASIC Stamp language. The Arduino library was developed from a project called Wiring, which is a useful library for AVR microcontrollers. It is used mostly with the Arduino development boards (see Section 13.4). However, the Arduino IDE (different from AVR Studio) can also be used to write the programs for most processors in the AVR 8-bit range, including the ATtiny85.

Figure 13.5 shows the program that would control the hardware of Fig. 13.4 within the Arduino IDE.



FIGURE 13.5

The Arduino IDE is open source and runs on Mac, Linux, and Windows systems. You need to install some extra configuration and library files to make the Arduino IDE work with the ATtiny microcontrollers. These instructions were developed at MIT. and details for installing and downloading can be found at http://hlt.media.mit.edu/?p=1695.

The advantages of using an ATtiny microcontroller here rather than a full Arduino board are reduced cost (an ATtiny costs $1 or less) and lower power consumption. We will meet the Arduino in Section 13.4.

Whether you are programming the microcontroller with AVR Studio or Arduino, you will need USB programmer hardware that is connected to a USB port of your computer. The programmers use a technique called in-circuit system programming (ICSP), which uses a six pin header connected to pins on the ATtiny. This is one of the reasons that you use a resistor to tie up the RESET pin, as it is one of the pins used by the programmer. It is quite common to design this header into the PCB for the circuit, during development, allowing changes to be easily made to the firmware.

Other ATtiny Microcontrollers

If you find that you need more I/O pins, or you can make do with less memory and save cost, you may want to consider one of the other microcontrollers in the ATtiny range. The Atmel website contains comparison tables for all its microcontrollers at http://www.atmel.com/products/microcontrollers/avr/tinyAVR.aspx?tab=parameters.

13.2.2 The PIC16Cx Microcontrollers

Having briefly explored the ATtiny85 microcontroller, we turn our attention to some rival devices from Microchip. Like the ATtiny range, these microcontrollers are 8-bit.

Figure 13.6 shows Microchip's PIC16Cx range of microcontrollers. As you can see in the internal

architecture diagram, both microcontrollers house on-chip CPU, EPROM, RAM, and I/O circuitry. The architecture is based on a register file concept that uses separate buses and memories for programs and data (Harvard architecture). This allows execution to occur in parallel. As an instruction is being "pre-fetched," the current instruction is executing on the data bus.

Microchip's PIC16C57 microcontrollers

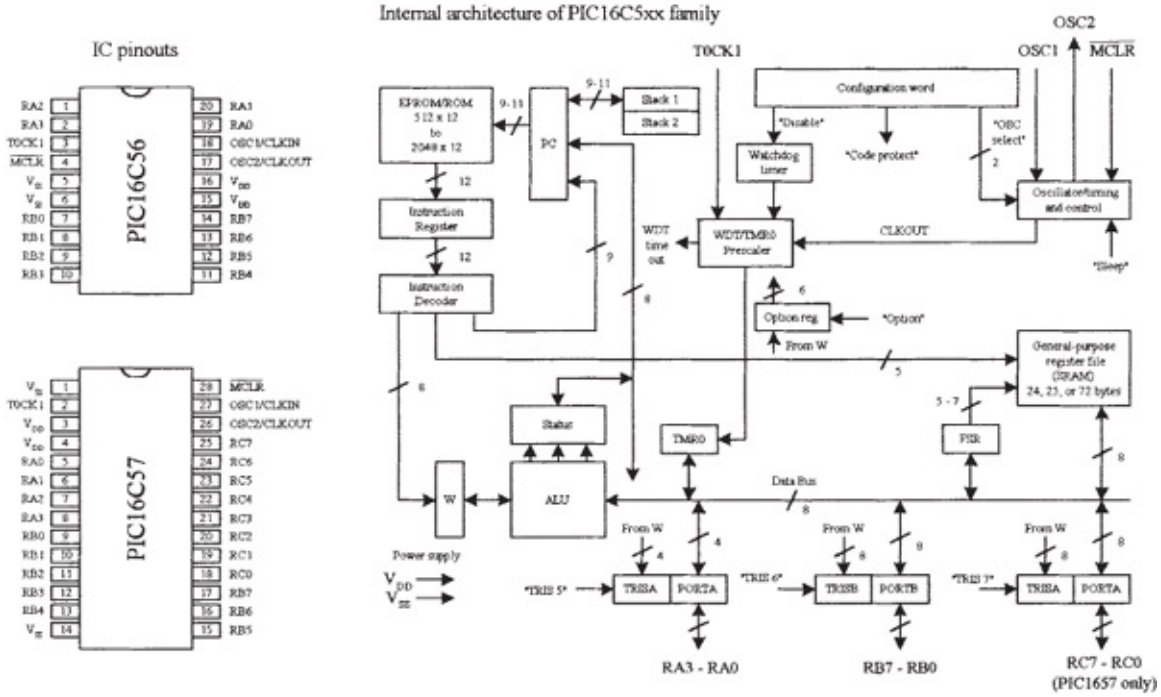


FIGURE 13.6

The PIC16C56's program memory (EPROM) has space for 1024 words, while the PIC16C57 has space for 2048 words. An 8-bit-wide ALU contains one temporary working register, and performs arithmetic and Boolean functions between data held in the working register and any file register. The ALU and register file are composed of up to 80 addressable 8-bit registers, and the I/O ports are connected via the 8-bit-wide data bus. Thirty-two bytes of RAM are directly addressable, while the access to the remaining bytes work through bank switching.

In order for bit movement to occur (clock generation), the PIC controllers require a crystal or ceramic resonator connected to pins OSC1 and OSC2. The PIC microcontrollers reach a performance of 5 million instructions per second (5 MIPS) at a clock frequency of 20 MHz. A watchdog timer is also included, which is a free-running on-chip RC oscillator that works without external components. It continues running when the clock has stopped, making it possible to generate a reset independent of whether the controller is working or sleeping.

These chips also come with a number of I/O pins that can be linked to external devices such as light sensors, speakers, LEDs, or other logic circuits. The PIC16C56 comes with 12 I/O pins that are divided into three ports: port A (RA3–RA0), port B (RB7–RB0), and port C (RC7–RC0). The PIC16C57 comes with eight more I/O pins than the PIC16C56.

Programming the PIC Microcontroller

Microcontrollers use a set of machine-code instructions (1's and 0's) to perform various tasks such as adding, comparing, sampling, and outputting data via I/O ports. These machine-code instructions are typically programmed into onboard ROM (EPROM, EEPROM, or flash) via a programming unit linked to a PC. The actual programming, however, isn't written out in machine code. It is written in a high-level language within an editor program running on the PC. The high-level language used may be a popular

language such as C or a specially tailored language that the manufacturer has created to optimize all the features present within its microcontrollers. For ultimate performance, the user can revert to assembly language, which could also reduce memory usage and program size but at the expense of readability of the code.

Using a manual and software you get from the manufacturer, you learn to write statements that tell the microcontroller what to do. You type the statements in an editor program, and then use the compile/run option to check for syntax errors. Once you think the program is ready, you save it and run a compiler program to translate it into machine language. If there is an error in your program, the compiler may refuse to perform the conversion. In this case, you must return to the text editor and fix the bugs before moving on.

Once the bugs are eliminated and the program is compiled successfully, a third piece of software is used to load the program into the microcontroller. This may require physically removing the microcontroller from the circuit and placing it into a special programmer unit linked to the host PC or may use ICSP, as described in the previous section.

Another way to do things involves using an interpreter instead of a compiler. An *interpreter* is a high-level language translator that resides within the microcontroller's ROM, rather than in the host PC. This often means that an external ROM (EPROM, EEPROM, or flash) is needed to store the actual program. The interpreter receives the high-level language code from the PC and, on the spot, interprets the code and places the translated code (machine code) into the external ROM, where it can be used by the microcontroller.

The interpreter approach may seem like a waste of memory, since the interpreter consumes valuable on-chip memory space. Also, using an interpreter significantly slows things down—a result of needing to retrieve program instructions from external memory. However, using an interpreter provides a very important advantage. By having the interpreter onboard to translate on the spot, an immediate, interactive relationship between the host program and microcontroller is created. This allows you to build your program, immediately try out small pieces of code, test the code by downloading the chunks into the microcontroller, and then see if the specific chunks of code work. The host programs used to create the source code often come with debugging features that let you test to see where possible programming or hardwiring errors may result by displaying the results (such as the logic state at a given I/O pin) on the computer screen while the program is executing within the microcontroller. This allows you to perfect specific tasks within the program, such as a sound-generation routine, a stepper motor control routine, and so forth.

A halfway house between using interpreted code (BASIC Stamp, which we'll look at next) and compiled machine code (AVR Studio) is to use a microcontroller with a boot loader installed into its EEPROM (Arduino). In this approach, a small boot loader program is installed once into the flash memory of the microcontroller. The boot loader then runs after every reset of the microcontroller and quickly checks for incoming programming commands on the serial port. If it finds them, it reads the serial data into the flash memory of the device so that it can then be run. This removes the need for special programming hardware.

Programming the PIC with BASIC Stamp

The BASIC Stamp is essentially a microcontroller with interpreter software built in. These devices also come with additional support circuitry, such as an EEPROM, voltage regulator, ceramic oscillator, and so on. BASIC Stamps are ideal for beginners because they are easy to program, quite powerful, and relatively cheap—a whole startup package costs around $80 or so. These devices are also very popular among inventors and hobbyists, and you'll find a lot of helpful literature, application notes, and fully tested projects on the Internet.

The original Stamp was introduced in 1993 by Parallax, Inc. It got its name from the fact that it resembled a postage stamp. The early version of the BASIC Stamp was the REV D. Later improvements led to the BASIC Stamp I (BSI) and then BASIC Stamp II (BSII).

Both the BSI and BSII have a specially tailored BASIC interpreter firmware built into the microcontroller's EPROM. For both Stamps, a PIC microcontroller is used. The actual program that is to be run is stored in an onboard EEPROM. When the battery is connected, the Stamps run the BASIC program in memory. Stamps can be reprogrammed at any time by temporarily connecting them to a PC running a simple host program. The new program is typed in, a key is hit, and the program is loaded into the Stamp. I/O pins can be connected with other digital devices such as sense switches, LED displays, LCDs, servos, and stepper motors.

Here, we'll focus on the BSII. To get started with the BSII, you will need programming software, programming cable, the manual, the BASIC Stamp module, and an appropriate carrier board (optional). These all come in the BSII startup kit, at a lower cost than purchasing each part separately.

For more information about the BASIC Stamp family, visit http://www.parallax.com/tabid/436/Default.aspx.

Note: To fully understand all the finer details needed to program BASIC Stamps, it is necessary to read through the user's manual. However, reading the user's manual alone tends not to be the best learning strategy, as it is easy to lose your place within all the technical terms, especially if you are a beginner. A good source to learn more about BASIC Stamps is the book *Programming and Customizing the Basic Stamp Computer* by Scott Edwards (McGraw-Hill, 2001). This book is geared toward beginners and is easy reading.

BASIC STAMP II

BSII is a module that comes in a 24-pin DIL package (see Fig. 13.7). The brain of the BSII is the PIC16C57 microcontroller that is permanently programmed with a PBASIC2 instruction set within its internal one-time programmable EPROM (OTP-EPROM). When programming the BSII, you tell the PIC16C57 to store symbols, called *tokens*, in external EEPROM memory. When the program runs, the PIC16C57 retrieves tokens from memory, interprets them as PBASIC2 instructions, and carries out those instructions. The PIC16C57 can execute its internal program at a rate of 5 MIPS. However, each PBASIC2 instruction takes up many machine instructions, so the PBASIC2 executes more slowly, around 3 to 4 MIPS.

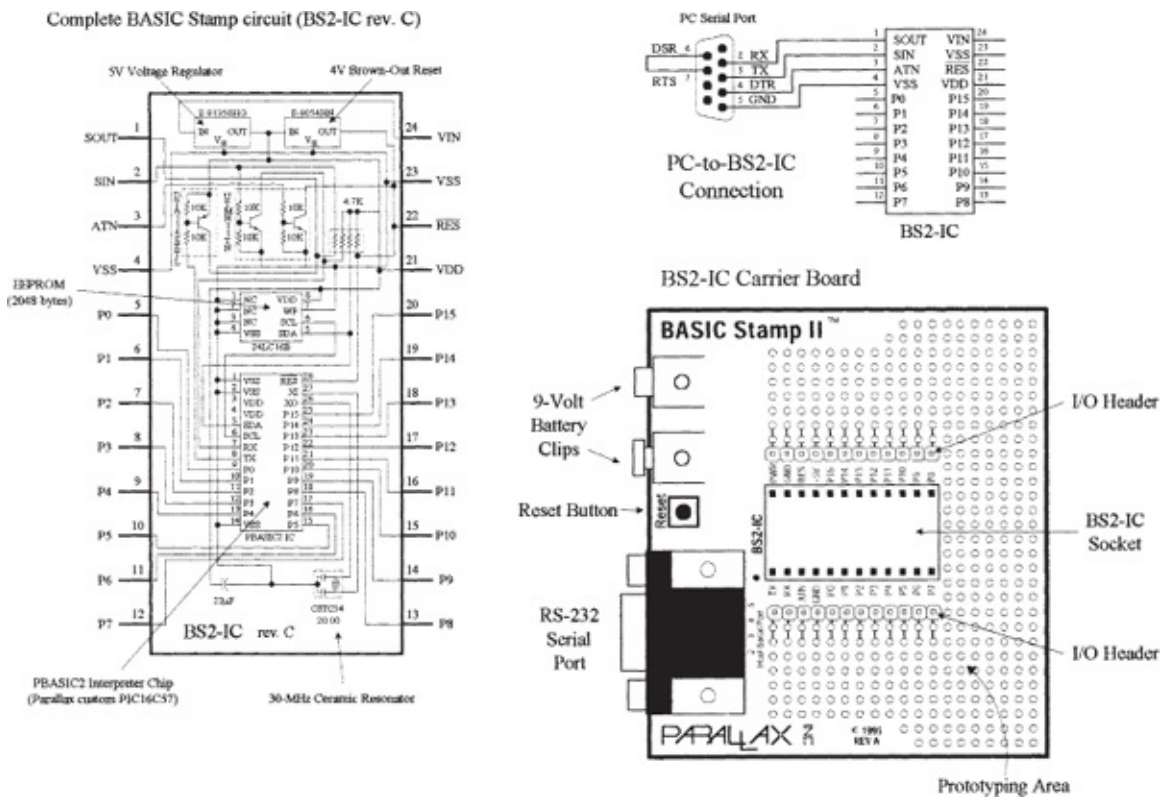


FIGURE 13.7

The BSII comes with 16 I/O pins (P0–P15) that are available for general use by your programs. These pins can be interfaced with all modern 5-V logic, from TTL through CMOS (technically, they have characteristics like the 74HCT logic series). The direction of a pin—either input or output—is set during the programming phase. When a pin is set as an output pin, the BSII can send signals to other devices, like LEDs and servos. When a pin is set as an input pin, it can receive signals from external devices, such as switches and photosensors. Each I/O pin can source 20 mA and sink 25 mA. Pins P0–P7 and pins P8–P15, as groups, can each source a total of 40 and sink 50 mA per group.

2048-BYTE EEPROM

The BSII's PIC's internal OTP-EPROM is permanently programmed at the factory with Parallax's firmware, which turns this memory into a PBASIC2 interpreter chip. Because they are interpreters, the Stamp PICs have the entire PBASIC language permanently programmed into their internal program memory. This memory cannot be used to store your PBASIC2 program. Instead, the main program must be stored in the EEPROM, which retains data without power and can be reprogrammed easily. At runtime, the PBASIC2 program created on the host computer is loaded into the BSII's EEPROM starting at the highest address (2047) and working downward. Most programs do not use the entire EEPROM, which means that PBASIC2 lets you store data in the unused lower portion of the EEPROM. Since programs are stored from the top of the memory downward, data is stored in the bottom of the memory working upward. If there is an overlap, the Stamp host software will detect this problem and display an error message.

RESET CIRCUIT

The BSII comes with a reset circuit. When power is first connected to the Stamp, or if it falters due to a weak battery, the power supply voltage can fall below the required 5 V. During such brownouts, the PIC is in a voltage-deprived state and will have the tendency to behave erratically. For this reason, a reset chip is incorporated into the design, forcing the PIC to reset to the beginning of the program and hold until the supply voltage is within acceptable limits.

POWER SUPPLY

To avoid supplying the BSII with unregulated supply power, a 5-V regulator is incorporated into the BSII. This regulator accepts a voltage range from slightly over 5 V up to 15 V and regulates it to a steady 5 V. It provides up to 50 mA. The regulated 5 V is available at output V_{DD}, where it can be used to power other parts of your circuits, as long as no more than 50 mA is required.

CONNECTING BSII TO A HOST PC

To program a Stamp requires connecting it to a PC that runs host software to allow you to write, edit, download, and debug PBASIC2 programs. The PC communicates with the BSII through an RS-232 (COM port) interface consisting of pins S_{IN}, S_{OUT}, and *ATM* (serial in, serial out, and attention, respectively).

During programming, the BSII host program pulses *ATM* high to reset the PIC and then transmits a signal to the PIC through S_{IN} indicating that it wants to download a new program. PC-to-BSII connector hookup is shown in Fig. 13.7. This connection allows the PC to reset the BSII for programming, download programs, and receive debug data from the BSII. The additional pair of connections, pin 6 and 7 of the DB9 socket, lets the BSII host software identify the port to which the BSII is connected.

Usually, when programming a BSII, you use a special BSII carrier board, which comes with a prototyping area, I/O header, BSII-IC socket, 9-V battery clips, and an RS-232 serial port connector, as shown in Fig. 13.7. These boards, along with programming cable and software, can be purchased as startup packages.

THE PBASIC LANGUAGE

Even though the BASIC Stamp has BASIC in its name, it cannot be programmed in Visual BASIC. It does not have a graphical user interface, a hard drive, or a lot of RAM. The BASIC Stamp must be programmed only with Parallel's BASIC, PBASIC, which has been specifically designed to exploit all the BASIC Stamp's capabilities.

PBASIC is a hybrid form of the BASIC programming language, with which many people are familiar. PBASIC is called a *hybrid* because, while it contains some simplified forms of normal BASIC control constructs, it also has special commands to efficiently control I/O pins. PBASIC is an easy language to master and includes familiar instructions such as GOTO, FOR … NEXT, and IF … THEN. It also contains Stamp-specific instructions, such as PULSOUT, DEBUG, and BUTTON, which will be discussed shortly.

The actual program to be downloaded into the Stamp is first written using BSII editor software running on a Microsoft Windows PC, or on a Linux or Mac system running Windows using virtualization software. After you write the code for your application, you simply connect the Stamp to a serial port or USB-to-serial adapter connected to your computer, provide power to the Stamp, and download the code into the Stamp. As soon as the program has been downloaded successfully, it begins executing its new program from the first line of code.

The size of the program that can be stored in a Stamp is limited. For the BSII, 2048 bytes worth of program space are available, which is enough for around 500 to 600 lines of PBASIC code. The amount of program memory for the Stamps cannot be expanded, since the interpreter chip (PIC) expects the memory to be specific and fixed in size. However, in terms of data memory, expansion is possible. You can interface EEPROM or other memory devices to the Stamp's I/O pins to gain more data storage area. This requires that you supply the appropriate code within your PBASIC program to make communication between the Stamp and external memory device you choose possible. Additional data memory is often available with Stamp-powered applications that monitor and record data (such as from an environmental field instrument).

The PBASIC language, like other high-level computer languages, involves defining variables and constants and using address labels, mathematical and binary operators, and various instructions (including

branching, looping, numerics, digital I/O, serial I/O, analog I/O, sound I/O, EEPROM access, time, power control, and so on). Here's a quick rundown of the elements of the PBASIC2 language.

Comments: Comments can be added within the program to describe what you're doing. They begin with an apostrophe (') and continue to the end of the line.

Variables: These are locations in memory that your program can use to store and recall values. These variables have limited range. Before a variable can be used in a PBASIC2 program, it must be declared. The common way used to declare variables is to use a directive VAR:

```
symbolvar  size
```

where the symbol can be any name that starts with a letter; can contain a mixture of letters, numbers, and underscore; and must not be the same as PBASIC keywords or labels used in the program. The size establishes the number of bits of storage the variable is to contain. PBASIC2 provides four sizes: bit (1 bit), nib (4 bits), byte (8 bits), and word (16 bits). Here are some examples of variable declarations:

```
'Declare variables.
sense_invar bit     'Value can be 0 or 1.
speedvar nib    'Value in range 0 to 15.
lengthvar byte    'Value in range 0 to 255.
nvar word   'Value in range 0 to 65535.
```

Constants: Constants are unchanging values that are assigned at the beginning of the program and may be used in place of the numbers they represent within the program. Defining constants can be accomplished by using the CON directive:

```
beeps         con  5        'number of beeps
```

By default, PBASIC2 assumes that numbers are in decimal (base 10). However, it is possible to use binary and hexadecimal numbers by defining them with prefixes. For example, when the prefix % is placed in front of a binary number (for example, %0111 0111), the number is treated as a binary number, not a decimal number. To define a hexadecimal number, the prefix $ is used (as in $EF). Also, PBASIC2 will automatically convert quoted text into the corresponding ASCII codes. For example, defining a constant as A will be interpreted as the ASCII code for *A* (65).

Address labels: The editor uses address labels to refer to addresses (locations) within the program. This is different from other versions of BASIC, which use line numbers. In general, an address label name can be any combination of letters, numbers, and underscores. However, the first character in the label name cannot be a number, and the label name must not be the same as a reserved word, such as a PBASIC instruction or variable. The program can be told to go to the address label and follow whatever instructions are listed after. Address labels are indicated with a terminating colon (for example, loop:).

Mathematical operators: PBASIC2 uses two types of operators: unary and binary. Unary operators take precedence over binary operators. Also, unary operations are always performed first. For example, in the expression 10 – SQR 16, the BSII first takes the square root of 16 and then subtracts it from 10. The unary operators are as follows:

ABS Returns absolute value

SQR Returns square root of value
DCD 2^n-power decoder
NCD Priority encoder of a 16-bit value
SIN Returns 2's complement sine
COS Returns 2's complement cosine

The binary operators are as follows:

+ Addition
− Subtraction
/ Division
// Remainder of division
* Multiplication
** High 16 bits of multiplication
*/ Multiplies by 8-bit whole and 8-bit part
MIN Limits a value to specified low
MAX Limits a value to specified high
DIG Returns specified digit of number
<< Shifts bits left by specified amount
>> Shifts bits right by specified amount
REV Reverses specified number of bits
& Bitwise AND of two values
| Bitwise OR of two values
^ Bitwise XOR of two values

Table 13.1 shows the PBASIC instructions used by BSII.

TABLE 13.1 PBASIC Instructions

INSTRUCTION	DESCRIPTION
Branching	
IF *condition* THEN *addressLabel*	Evaluate condition and, if true, go to the point in the program marked by *addressLabel*. (Conditions: =, <> not equal, >, <, >=, and <=)
BRANCH *offset, [address0, address1, . . . addressN]*	Go to the address specified by offset (if in range).
GOTO *addressLabel*	Go to the point in the program specified by *addressLabel*.
GOSUB *addressLabel*	Store the address of the next instruction after GOSUB, then go to the point in the program specified by *addressLabel*, with the intention of returning to the point at which the subroutine was called once the subroutine has returned.
RETURN	Return from subroutine.
Looping	
FOR *variable=start to end{STEP stepVal}* . . . NEXT	Create a repeating loop that executes the program lines between FOR and NEXT, incrementing or decrementing *variable* according to *stepVal* until the value of the variable passes the *end* value.
Numerics	
LOOKUP *index, [value0, value1, . . . valueN], resultVariable*	Look up the value specified by the index and store it in a variable. If the index exceeds the highest index value of the items in the list, the variable is unaffected. A maximum of 256 values can be included in the list.
LOOKDOWN *value, {comparisonOp,} [value0, value1, . . . valueN], resultVariable*	Compare a value to a list of values according to the relationship specified by the comparison operator. Store the index number of the first value that makes the comparison true in *resultVariable*. If no value in the list makes the comparison true, *resultVariable* is unaffected.
RANDOM *variable*	Generate a pseudo-random number using a byte or word variable where the bits are scrambled to produce a random number.
Digital I/O	
INPUT *pin*	Make the specified pin an input.
OUTPUT *pin*	Make the specified pin an output.
REVERSE *pin*	If the pin is an output, make it an input. If the pin is an input, make it an output.
LOW *pin*	Make the specified pin's output low.
HIGH *pin*	Make the specified pin's output high.
TOGGLE *pin*	Invert the state of a pin.
PULSIN *pin, state, resultVariable*	Measure the width of a pulse in 2-µs units.
PULSOUT *pin, time*	Output a timed pulse by inverting a pin for some time (× 2 µs).
BUTTON *pin, downstate, delay,rate,bytevariable, targetstate, address*	Debounce button input, perform auto-repeat, and branch to the address if the button is in target state. Button circuits may be active-low or active-high.
SHIFTIN *dpin, cpin, mode, [result{\bits}{,result{\bits}* . . . *}]*	Shift data in from a synchronous serial device.

SHIFTOUT *dpin, cpin, mode, [data{\bits}{,data{\ bits} . . . }]*	Shift data out to a synchronous serial device.
COUNT *pin, period, variable*	Count the number of cycles (0-1-0 or 1-0-1) on the specified pin during *period* number of milliseconds and store that number in variable.
XOUT *mpin, zpin, [house\ keyORCommand{\cycles} {,house\keyOrCommand{\ cycles} . . . }]*	Generate X-10 powerline control codes.
Serial I/O	
SERIN *rpin{\fpin}, baudmode, {plabe}{timeout,tlabe,}[input Data]*	Receive asynchronous serial transmission.
SEROUT *tpin, baudmode, {pace,} [outputData]*	Send data serially with optional byte pacing and flow control.
Analog I/O	
PWM *pin, duty, cycles*	Output fast pulse-width modulation, then return pin to input. This can be used to output an analog voltage (0–5 V) using a capacitor and resistor (see Section 13.5).
RCTIME *pin, state, resultVariable*	Measure an RC charge/discharge time. This can be used to measure the position of a potentiometer or capacitance in capacitative sensing (see Chap. 6).
Sound	
FREQOUT *pin, duration, freq1{,freq2}*	Generate one or two sine-wave tones of a specified frequency for a specified duration.
DTMFOUT *pin, {ontime,offtime,} {,tone . . . }*	Generate dual-tone, multifrequency tones (DTMF, i.e., telephone touch tones).
EEPROM Access	
DATA	Store data in EEPROM before downloading the PBASIC program.
READ *location, variable*	Read EEPROM location and store the value in a variable.
WRITE *address,byte*	Write a byte of data to the EEPROM at the appropriate address.
Time	
PAUSE *milliseconds*	Pause the program (do nothing) for the specified number of milliseconds. Pause execution for 0–65,535 ms.
Power Control	
NAP *period*	Enter sleep mode for a short period. Power consumption is reduced to about 50 µA assuming no loads are being driven. The duration is $(2^{period}) \times$ 18 ms.
SLEEP *seconds*	Sleep from 1–65,535 seconds to reduce power consumption by ~50 µA.
END	Sleep until the power cycles or the PC connects ~50 µA.
Program Debugging	
DEBUG *outputData{,outputData . . . }*	Display variables and messages on the PC screen within the BSII host program. *outputData* consists of one or more of the following: text strings, variables, constants, expressions, formatting modifiers, and control characters.

DEBUGGING

To debug PBASIC programs, the BASIC Stamp editor comes with two handy features: syntax checking and a DEBUG command.

Syntax checking alerts you to any syntactical error and is automatically performed on your code the moment you try to download to the BASIC Stamp. Any syntax errors will cause the download process to abort and will cause the editor to display an error message, pointing out the error in the source code.

The DEBUG command, unlike syntax checking, is an instruction that is written into the program to find logical errors—ones that the Stamp does not find, but ones that the designer had not intended. DEBUG operates similar to the PRINT command in the BASIC language and can be used to print the current status of specific variables within your PBASIC program as it is executed within the BASIC Stamp. If your PBASIC code includes a DEBUG command, the editor opens a special window at the end of the download process to display the result for you.

Making a Robot Using BSII

To demonstrate how easy it is to make interesting gadgets using BSII, let's take a look at a robot application. In this application, the main objective is to prevent the robot from running into objects. The robot aimlessly moves around, and when it comes close to an object, the robot stops and then backs up and moves off in another direction. In this example, the robot is constructed as follows:

- A BSII acts as the robot's brain.
- Two servos connected to wheels act as its legs.
- A pair of infrared transmitters and sensors acts as its eyes.
- A piezoelectric speaker acts as it voice.

Figure 13.8 shows the completed robot, along with the various individual components.

Components and connections used to create object-avoiding robot

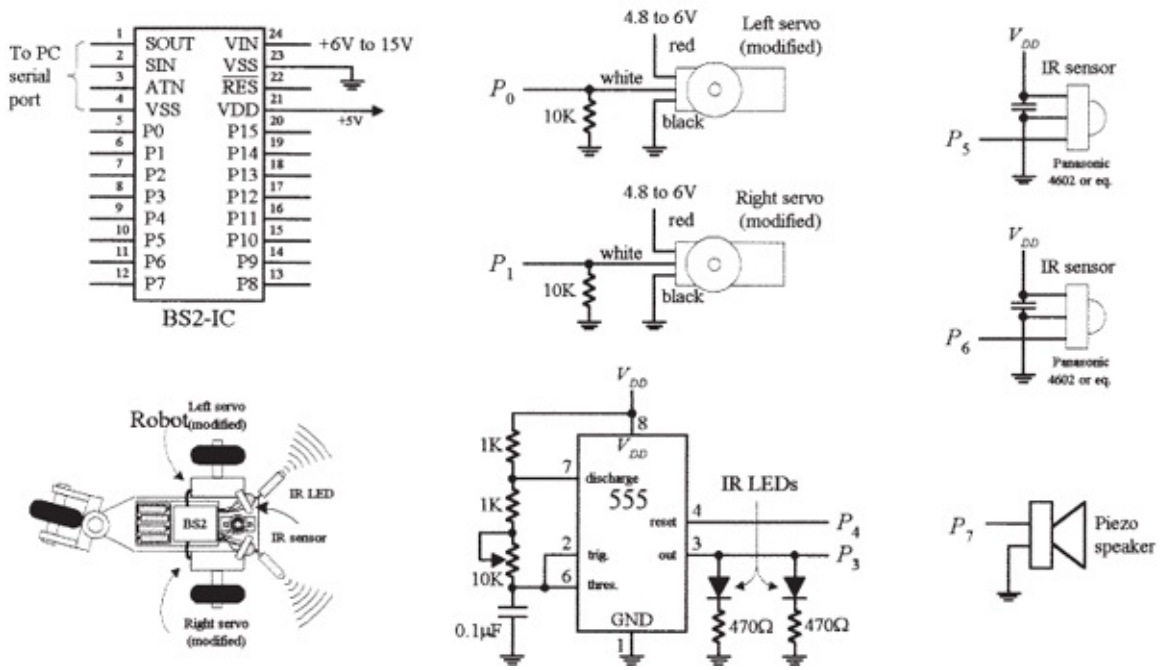


FIGURE 13.8

THE SERVOS

The directional movement of the robot is controlled by right and left servo motors that have been modified so as to provide a full 360 degrees worth of rotation (modifying a servo is discussed in Chap. 14). To control a servo requires generating pulses ranging from 1000 to 2000 µs in width at intervals of approximately 20 ms. With one of the servos used in our example, when the pulse width sent to the servo's control line is set to 1500 µs, the servo is centered—it doesn't move. However, if the pulse width is shortened to, say, 1300 µs, the modified servo rotates clockwise. Conversely, if the pulse width is lengthened to, say, 1700 µs, the modified servo rotates counterclockwise.

The actual control pulses used to drive one of the servos in the robot are generated by the BSII using the PULSOUT *pin*, *time1* and the PAUSE *time2* instructions. The *pin* represents the specific BSII pin that is linked to a servo's control line, and *time1* represents how long the pin will be pulsed high. Note that for the PULSOUT instruction, the decimal placed in the *time1* slot actually represents half the time, in microseconds (s), that the pin is pulsed high. For example, PULSOUT 1, 1000 means that the BSII will pulse pin 1 high for 2000 µs, or 2ms. For the PAUSE instruction, the decimal placed in the *time2* slot represents a pause in milliseconds. For example, PAUSE 20, represents a 20-ms pause. Figure 13.9 shows sample BSII code used to generate desired output waveforms to control a servo.

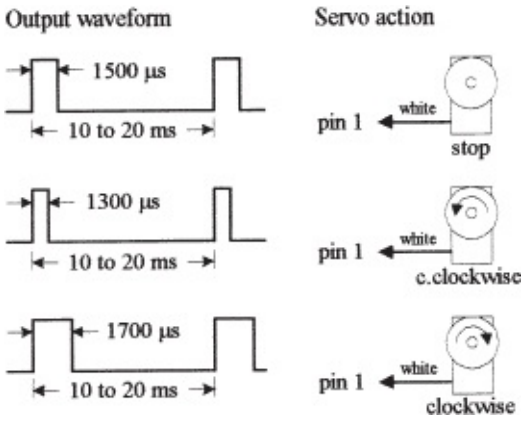


FIGURE 13.9

```
 BS2 code              Comments
pulsout 1, 750       'pulse width of 1500us on pin 1

pause 20             'pause for 20 ms
pulsout 1, 650       'pulse width of 1300us on pin 1
pause 20             'pause for 20 ms
pulsout 1, 850       'pulse width of 1700us on pin 1
pause 20             'pause for 20 ms
```

The pulse widths need to be repeated for them to have a practical effect on the motors. It may be worth expanding this fragment to show loops of a second or more for each of the pulse widths.

Note that this sequence of pulses must be repeated every 20 ms or so for the servo to maintain its position.

IR TRANSMITTERS AND RECEIVERS

The robot's object-detection system consists of a right and left set of infrared (IR) LED transmitters and IR detector modules. The IR LEDs are flashed via a 555 timer at a high frequency, which in this example happens to be 38 kHz, 50 percent duty cycle. This frequency is used to avoid interference from other household sources of IR light, primarily incandescent lights, and to match the IR sensor shown in the figure. (Many types of IR LED transmitters and sensors could be used in this robot, and they may work best using a different frequency.) It is also possible to generate these pulses using the BASIC Stamp, but we have chosen to use external hardware to keep the program simple.

The IR photons emitted by the LED rebound off objects in the path of the robot and reflect back to the IR detector module. When a detector module receives photons, the I/O pin of the BSII connected to the module goes low. Note that the BSII can execute only around 4000 instructions per second, while the number of pulses generated by the detector module is 38,000. In this case, the actual number of pulses received by the BSII will be less—around 10 or 20.

PIEZOELECTRIC SPEAKER

A piezoelectric speaker is linked to one of the BSII I/O terminals and is used to generate different sounds when the robot is moving forward or backing up. To provide the piezoelectric speaker with a sinusoidal waveform to generate sound, the FREQOUT *pin, time, frequency* instruction is used. The instruction FREQOUT 7, 1000, 440 creates a 440-Hz sinusoidal frequency on pin 7 that lasts for 1000 ms.

THE PROGRAM

The following is a program used to control the robot. It is first created using the PBASIC2 host software, and then downloaded into the BSII during runtime.

```
'Program for object-avoiding robot
'Define variables and constants
'-----------------------------------------------
nvar word     'n acts as a variable that changes.
right_IRvar in5      'Sets pin 5 as an input for right IR detector.
left_IRvar in6      'Sets pin 6 as an input for left IR detector.
right_servo  con 0        'Assigns 0 which will be used to identify right servo.
left_servo   con 1        'Assigns 1 to identify left servo.
IR_out       con 3        'Assigns 3 to identify IR output.
delay        con 10       'A constant that will be used in the program.
speed        con 100      'Used to set servo speed.
turn_speed   con 50       'Used to set turn speed of robot.
'Main program
'-----------------------------------------------
highIR_out                                    'Sets pin 3 "high"
pause 50                                       'Pauses for 50 milliseconds
sense:                                         'Label used to specify IR-sense routine.
ifleft_IR = 0 and right_IR = 0 then backup  'Object in front, jump to back_up routine.
ifleft_IR = 0 then turn_right               'Object on left side, jump to turn_right routine.
ifright_IR = 0 then turn_left               'Object on right, jump to "turn_left" routine.
'Sound Routines
'-----------------------------------------------
forward_sound:                        'Label
freqout 7,1000, 440                   'Generate 1000ms, 440 Hz tone on pin 7
back_sound:                           'Label
freqout 7,1000,880                    'Generate 1000ms, 880 Hz tone on pin 7
'Motion routines
'-----------------------------------------------
forward:                              'Label used to specify forward routine.
gosubforward_sound                    'Tells program to jump to forward sound subroutine.
debug "forward"                       'Tells stamp to display the word "forward" on debug window.
pause 50                              'Pause for 50ms.
for n = 1 to delay*2                  'For…Next loop that starts x = 1 and repeats until x = 20.
pulsoutleft_servo, 750-speed          'Make left servo spin to make robot move forward.
pulsoutright_servo, 750+speed         'Make right servo spin to make robot move forward.
pause 20                              'Pauses for 20ms, path of servo control.
next                                  'End of For…Next loop.
goto sense                            'Once forward routine is finished go back to sense routine.
backup:                               'Label used to specify back-up routine.
gosubbackup_sound                     'Tells program to jump to back-up sound subroutine.
debug "backward"                      'Displays "backward" on the debug window.
pause 50                              'Pause for 50ms to ensure
for n = 1 to delay*3                  'For…Next loop that starts x = 1 and repeats until x = 60
pulsoutleft_servo, 750+speed         'Makes left servo spin to make robot move backward.
pulsoutright_servo, 750-speed        'Makes right servo spin to make robot move backward.
pause 20                              'Pauses for 20ms, part of servo control.
next                                  'End of For…Next loop.
turn_left:                            'Label used to specify turn-left routine.
debug "left"                          'Displays "left" on the debug window.
pause 50                              'Pause for 50ms.
for x = 1 to delay*1                  'For…Next loop that starts x = 1 and repeats until x = 10.
pulsoutleft_servo, 750-turn_speed  'Makes left servo spin to make robot turn left.
pulsoutright_servo, 750-turn_speed 'Makes right servo spin to make robot turn left.
pause 20                              'Pause for 20ms, part of servo control.
next                                  'End of For…Next loop.
goto sense                            'Once left-turn routine is finished, jump back to sense.
turn_right:                           'Label used to specify turn-right routine.
debug "right"                         'Displays "right" on debug window.
pause 50                              'Pause for 50 ms.
for x = 1 to delay*1                  'For…Next loop.
pulsoutleft_servo, 750+turn_speed  'Makes left servo spin to make robot turn right.
pulsoutright_servo, 750+turn_speed 'Makes right servo spin to make robot turn right.
pause 20                               'Pause for 20 ms, part of servo control.
next                                   'End of For…Next loop
goto sense                             'Once right-turn is finished, jump back to sense.
```

Note: Due to production tolerances, the motors may not stop rotating when pulsing at 1500 µs, so a small "fudge factor" may need to be added to or subtracted from the PULSOUT values.

Thinking about Mass Production

Recall that the major components of the BASIC Stamp circuit are the PIC (houses the CPU and ROM for storing PBASIC interpreter), external EEPROM (stores the program), and the resonator. In large-scale runs, it would be nice to get rid of the external memory and remove the interpreter program, and simply download a compiled PBASIC code directly into the PIC. This would save space and money. As it turns

out, the BASIC Stamp editor software includes a feature to program PBASIC code directly into a PIC microcontroller using Parallax's PIC16Cxx programmer.

The major benefit of starting out with the Stamp is that you can easily fine-tune your code, test chunks, and immediately see if it works, which is important when creating prototypes. When prototyping with a PIC, checking for errors is much harder because you must compile everything at once—you can't test out chunks of code.

13.2.3 32-Bit Microcontrollers

The microcontrollers that we have explored in the previous sections use an 8-bit data bus with a clock frequency in the tens of megahertz and a few kilobytes of storage. Anyone familiar with the microprocessor-based home computers of the 1980s and 1990s would recognize the specifications. In comparison to a modern smartphone with a 32-bit processor, clock frequencies in the gigahertz, and hundreds of megabytes of RAM, these microcontrollers offer lamentable performance. However, the important thing here is what they are being used for. The adage of "not using a sledgehammer to crack a nut" was never more appropriate.

Atmel, Microchip, and most of the other microcontroller manufacturers all produce high-performance microcontrollers that use a 32-bit data bus and have more memory and processing performance than most desktop computers had ten years ago. These are useful for some high-performance applications. If you find that you need this kind of performance, then it is worth considering using one from a manufacturer whose 8-bit devices you are already familiar with, as they generally use the same or similar software tools and can be programmed in much the same way. They just cost a lot more and work a lot faster.

13.2.4 Digital Signal Processing

If you have a microcontroller that has an ADC input and a DAC output, you can digitize an audio signal (let's say music), process the data in some way, and then send it back out through the DAC. You might create a graphics equalizer or a dynamic voice changer that raises the pitch of your voice by an octave. This is called *digital signal processing* (DSP).

While you can do simple low-quality DSP with a standard 8-bit microcontroller, the ADCs are often too slow and the algorithms such as Fourier transforms that you apply to the audio signal must work in real time, and therefore benefit greatly from a fast CPU.

Microchip (among others) has variants of its standard microcontroller lines specifically designed for DSP. The dsPIC from Microchip is one such device that is frequently used in low-cost DSP applications. It has a 16-bit internal data bus, a 40 MHz clock, and 2kB of RAM.

Note: DSP is a complex area, and there are many good books devoted to this topic. *Understanding Digital Signal Processing* by Richard G. Lyons (Pearson, 1996) is one such book.

13.3 Evaluation/Development Boards

The microcontroller manufacturers are very keen to get their products into your products, so most will offer low-cost development and evaluation boards for their microcontrollers. These will often take the form of a PCB that contains the microcontroller, supporting components such as a crystal oscillator and voltage regulator, and a prototyping area where you can add your own components. They can also usually be programmed from a USB or sometimes RS-232 serial port, for which you will need to find an ancient

PC, or more likely, use a USB-to-serial convertor. These boards include the use of the manufacturers' preferred software development tools, although these are sometime limited in some way unless you buy the professional version of the software.

As well as boards produced by the microcontroller manufacturers, other boards are offered by third parties. These boards can be useful during the development process, as they make prototyping much easier than starting from scratch.

Table 13.2 lists some of the most popular development boards available at the time of writing.

TABLE 13.2 Popular Microcontroller Evaluation Boards

MANUFACTURER	BOARD NAME	MICROPROCESSOR	URL	NOTES
Atmel	AVR Butterfly	ATMega169	http://www.atmel.com/tools/AVRBUTTERFLY.aspx	Includes LCD screen Serial programmer
Freescale (Motorola)	DEMO908JL16	MC68HC08JL16 family	http://www.freescale.com/files/microcontrollers/doc/user_guide/DEMO908JL16UM.pdf	USB programmed
Microchip	PICkit 1 Flash Starter Kit	PIC12F675	http://www.microchip.com/stellent/idcplg?IdcService=SS_GET_PAGE&nodeId=1406&dDocName=en010053	Includes USB programming hardware
Microchip	MPLAB Starter Kit for dsPIC	dsPIC33FJ256GP506	http://www.microchip.com/stellent/idcplg?IdcService=SS_GET_PAGE&nodeId=1406&dDocName=en534506	USB Intended for audio Includes amplifier
Arduino	Arduino Uno	ATmega328	http://www.arduino.cc	See Section 13.4

13.4 Arduino

Arduino is an open source hardware platform for microcontroller prototyping. It encompasses both a microcontroller development board and an IDE. The IDE is simple to use and available for Mac, Linux, and Windows computers.

Arduino boards are extremely popular as a starting point for using microcontroller technologies. Their popularity is due to a number of factors, including the following:

- Low cost (around $30)
- Open source design
- Easy-to-use and cross-platform IDE
- Availability of plug-in shields (expansion hardware)

13.4.1 A Tour of Arduino

The most popular Arduino board is the Arduino Uno (see Fig. 13.10). This board is based on an Atmel microcontroller, similar to the ATtiny called ATmega328 (although the ATtiny can also be programmed using the Arduino IDE).

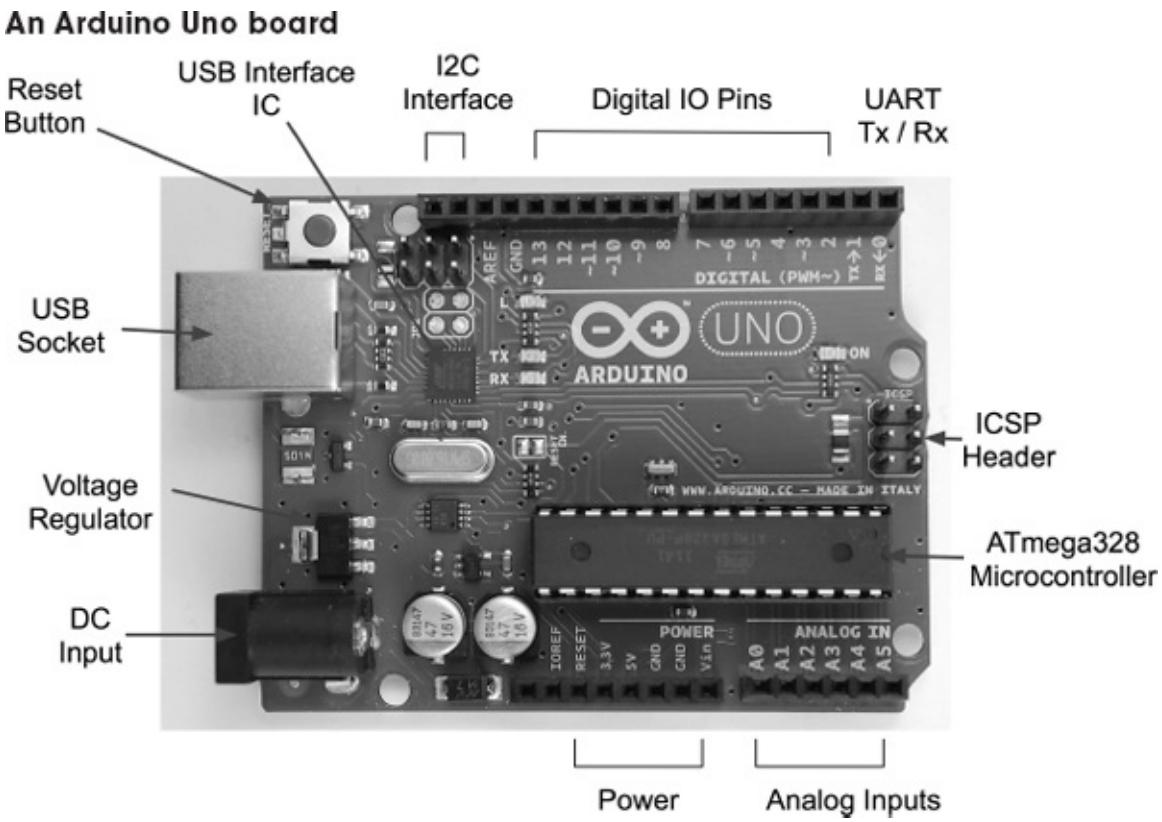


FIGURE 13.10

The ATmega328 microcontroller has 32kB of flash memory for storing programs, 2kB of RAM, and 1kB of EEPROM. It also has a hardware serial interface, or UART, as well as the usual timers and interrupt capabilities.

The microcontroller itself is the large 28-pin IC in the bottom right of the board, as shown in Fig. 13.10. Beneath this are six analog pins that can also be used as digital I/O pins, and then a block of power connections.

The Arduino can either be powered through the DC input socket (7–12 V dc) or from USB, and will switch over automatically to whichever is supplied.

The connectors on the top side of the board offer an I^2C interface, which actually uses two of the analog pins (A4 and A5) in the Arduino Uno, but they are repeated here for future boards that may have a separate I^2C interface. There is also a row or digital I/O pins, some marked as being pulse-width modulation (PWM) capable. Two of these, D0 and D1, double as the Rx and Tx pins on the UART.

13.4.2 The Arduino IDE

The Arduino IDE provides a simple-to-use editor into which you can type your programs and upload them onto your Arduino board over USB (see Fig. 13.11). As well as the program editing area, the Arduino IDE also provides the following features:

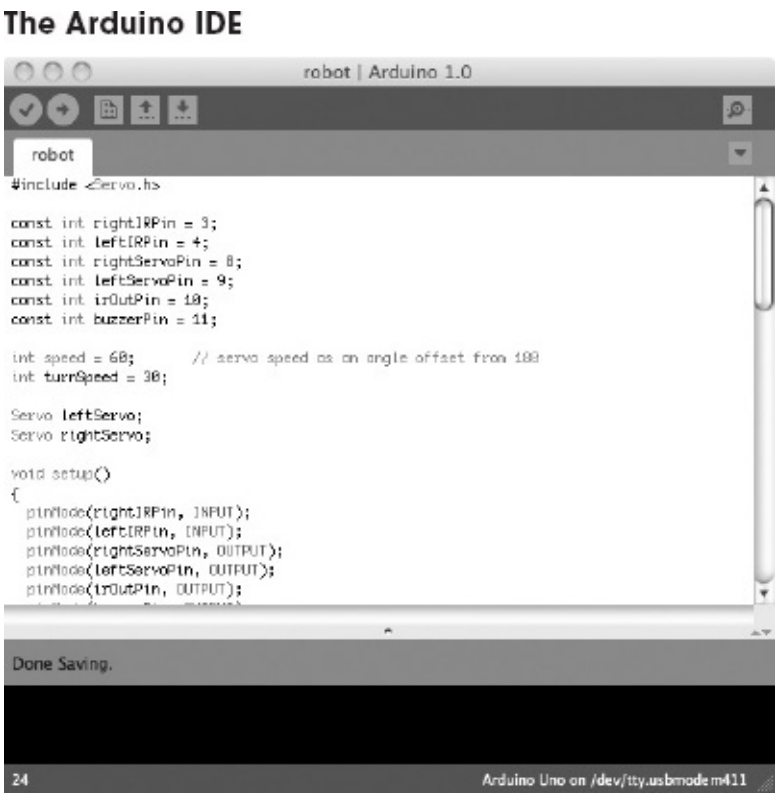


FIGURE 13.11

- Color syntax highlighting
- A status area where the memory usage of your completed program is shown
- Links to the Arduino library documentation
- A serial monitor that allows two-way communication with the Arduino's USB port

Most of the Arduino boards, including the Arduino Uno, have a USB connector through which they can be programmed. So, after writing your program, or "sketch" as it is called in Arduino parlance, you select the type of board and click the Upload button. Your program will be compiled and loaded into the flash memory of the microcontroller.

13.4.3 Arduino Board Models

Along with the Arduino Uno, there are many other Arduino boards to suit different uses. They are all programmed in the same way, but have different sizes, costs, and numbers of I/O pins available.

New Arduino models are released quite often. As an open source project, different manufacturers frequently take a basic model and add some different features to it.

Some of the most used official Arduino models are listed in Table 13.3.

TABLE 13.3 Kinds of Arduino Boards

MODEL	FEATURES	NOTES
Uno R3	At the time of writing, the latest version of Arduino Uno	Almost identical to the original Arduino Uno, but with extra header sockets for I²C and power status
Uno	14 digital, 6 analog or digital I/O pins, 32kB of flash memory, 2kB of SRAM, 1kB of EEPROM	The most popular Arduino board and a good all-rounder to start with New USB interface means no USB drivers needed
Leonardo	Same I/O and memory spec as an Arduino Uno	Lower cost than Uno, but has a nonremovable SMD microcontroller with programmable USB features
Duemilanove	Either the same spec as Uno, or on some models using the ATMega168, half of all the memory capacities of the Uno	Predecessor to the Uno Uses an FTDI-based USB interface that requires drivers to be installed on Windows
Lilypad	14 digital, 6 analog or digital I/O pins, 16kB of flash memory, 1kB of SRAM, and 512 bytes of EEPROM running at 8 MHz	Intended to be stitched into clothing using conductive thread to connect to other Lilypad devices such as LEDs and accelerometers Requires a separate USB-to-serial converter available from SparkFun (SKU: DEV-09716)
Mega 2560	54 digital, 16 analog or digital I/O pins, 4 UARTs, 256kB of flash memory, 8kB of SRAM, 4kB of EEPROM	The device for you if you need a *lot* of I/O pins You can fit Arduino Uno style shields on it, but there are occasional compatibility problems
Mini	Same I/O and memory spec as an Arduino Uno	Much smaller than an Arduino Uno. Requires a USB-to-serial converter to program it
Nano	Same I/O and memory spec as an Arduino Uno	Much smaller than an Arduino Uno, and fits directly onto the breadboard Includes a mini USB socket for programming
Fio	Similar spec to Arduino, but running at 8 MHz XBee wireless socket	Intended for mobile wireless applications Includes a LiPo battery charger IC
Ethernet	Same I/O and memory spec as an Arduino Uno	Arduino Uno with built-in Ethernet

Along with the boards listed in Table 13.3, other manufacturers offer boards that simply replicate the features of the official designs with minor differences. More interesting are the special-purpose Arduino boards. Some of these are listed in Table 13.4.

TABLE 13.4 Unofficial Arduino Variants

MODEL	FEATURES	NOTES	URL
DFRobot-ShopRover	Built-in motor driver	Intended for robots	http://www.dfrobot.com/
Electric Sheep	Arduino Mega with built-in USB host connection	Often used to link to Android phones that support the Open Accessory standard	http://www.sparkfun.com/products/10745
EtherTen	Arduino Uno with built-in Ethernet connection		http://www.freetronics.com/
Lightuino	LED drivers	70 constant current LED channels	http://www.toastedcircuits.com
USBDroid	Arduino Uno with built-in USB host connection		http://www.freetronics.com/
Teensyduino	Similar spec to Leonardo	Tiny breadboard-friendly device with USB capabilities	http://www.pjrc.com/teensy/teensyduino.html

13.4.4 Shields

The success of Arduino had been in no small part due to the wide range of plug-in shields that add useful

features to a basic Arduino board. A shield is designed to fit into the header sockets of the main Arduino board (see Fig. 13.12). Most shields will then pass through these connections in another row of header sockets, making it possible to construct stacks of shields with an Arduino at the bottom. Shields that have a display on them will not normally pass through in this way. You also need to be aware that if you stack shields in this way, you need to make sure that there are no incompatibilities, such as two of the shields using the same pin. Some shields get around this problem by providing jumpers to add some flexibility to pin assignments.

Arduino Ethernet shield on an Arduino

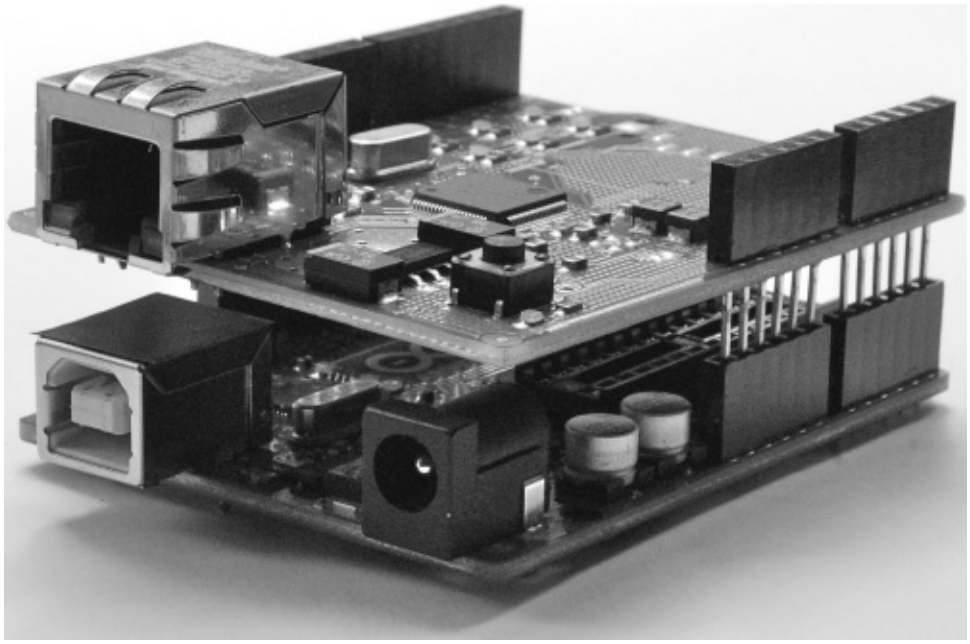


FIGURE 13.12

There are shields available for almost anything you could want an Arduino to do. They range from relay control to LED displays and audio file players. Most of these are designed with the Arduino Uno in mind, but are also usually compatible with the Arduino Mega.

An encyclopedic list that includes useful technical details about the pin usage of these shields can be found at http://shieldlist.org/. Some of the author's favorite shields are listed in Table 13.5.

TABLE 13.5 Common Arduino Shields

SHIELD	DESCRIPTION	URL
Motor	Ardumoto shield Dual H-bridge bidirectional motor control at up to 2 A per channel	http://www.sparkfun.com/products/9815
Ethernet	Ethernet and SD card shield	http://arduino.cc/en/Main/ArduinoEthernetShield
Relay	Controls four relays Screw terminals for relay contacts	http://www.robotshop.com/seeedstudio-arduino-relay-shield.html
LCD	16 × 2 character alphanumeric LCD shield with joystick	http://www.freetronics.com/products/lcd-keypad-shield

13.4.5 The Arduino C Library

You may hear people refer to the "Arduino language," but Arduino is actually just programmed in the C programming language, which has been around for many years. But Arduino provides a set of Arduino

core functions that you can use in your programs, or sketches.

There are a large number of commands available in the Arduino library. A selection of the most commonly used commands are listed in Table 13.6.

TABLE 13.6 Arduino Library Functions

COMMAND	EXAMPLE	DESCRIPTION
Digital I/O		
pinMode	pinMode(8, OUTPUT);	Sets pin 8 to be an output. The alternative is to set it to INPUT.
digitalWrite	digitalWrite(8, HIGH);	Sets pin 8 high. To set it low, use the constant LOW instead of HIGH.
digitalRead	inti; i = digitalRead(8);	Sets the value of i to HIGH or LOW depending on the voltage at the pin specified (in this case, pin 8).
pulseIn	i = pulseIn(8, HIGH)	Returns the duration in microseconds of the next HIGH pulse on pin 8.
tone	tone(8, 440, 1000);	Makes pin 8 oscillate at 440 Hz for 1000 ms.
noTone	noTone();	Cuts short the playing of any tone that was in progress.
Analog I/O		
analogRead	int r; r = analogRead(0);	Assigns a value to r of between 0 and 1023. 0 for 0 V 1023 if pin 0 is 5 V (3.3 V for a 3 V board).
analogWrite	analogWrite(9, 127);	Outputs a PWM signal (see Section 13.5). The duty cycle is a number between 0 and 255, 255 being 100%. This must be used by one of the pins marked as PWM on the Arduino board (3, 5, 6, 9, 10, and 11).
Time Commands		
millis	unsigned long l; l = millis();	The variable type long in Arduino is represented in 32 bits. The value returned by millis() will be the number of milliseconds since the last reset. The number will wrap around after approximately 50 days.
micros	long l; l = micros();	Like millis, except this is microseconds since the last reset. It will wrap after approximately 70 minutes.
delay	delay(1000);	Delay for 1000 ms, or 1 second.
delayMicroseconds	delayMicroseconds(100000);	Delay for 100,000 microseconds. Note the minimum delay is 3 microseconds; the maximum is around 16 ms.
Interrupts		
attachInterrupt	attachInterrupt(1, myFunction, RISING);	Associates the function myFunction with a rising transition on interrupt 1 (D3 on an Uno).
detachInterrupt	detachInterrupt(1);	Disables any interrupt on interrupt 1.

The main Arduino core, including all the commands listed in Table 13.6, is automatically included in every sketch that you write. However, there are a number of other libraries that come bundled with the Arduino IDE that are added to your code only when you use them. To include them, use the include command followed by the name of the library, like this:

```
#include <Servo.h>
```

This command includes the Servo library that we will use in the example Arduino project that follows.

The libraries that are included in the Arduino IDE are listed in Table 13.7.

TABLE 13.7 Standard Arduino Libraries

LIBRARY	DESCRIPTION
EEPROM	Reading and write to EEPROM from your sketches
Ethernet	TCP/IP communications when using an Ethernet board or shield, including DNS, DHCP, HTTP, and UDP
Fermata	A protocol for turning pins on and off, reading analog values, etc., using serial commands
LiquidCrystal	Interface to the de facto standard alphanumeric LCD modules based on the HD44780 IC (most alphanumeric LCD modules)
SD	Read and write to an SD card; shields that combine an SD card socket with an Ethernet or real-time clock interface are available
Servo	Control a number of servos simultaneously (see Section 13.4.6)
SoftwareSerial	Use any two pins to receive and transmit data; the Arduino has one hardware serial port (UART)
SPI	Serial Peripheral Interface bus library
Stepper	Control stepper motors
Wire	I^2C library

As well as the official Arduino libraries, as an open system, anyone can write a library and contribute it to the community, and many of these are extremely useful. Some of these other libraries are listed in Table 13.8.

TABLE 13.8 Contributed Arduino Libraries

LIBRARY	SOURCE	DESCRIPTION
Android Accessory	http://developer.android.com/guide/topics/usb/accessory.html	For serial communications between an Android phone and an Arduino
Bounce	http://www.arduino.cc/playground/Code/Bounce	Software debouncing of switches
Dallas Temperature Control	http://milesburton.com/index.php?title=Dallas_Temperature_Control_Library	Not simply "HOT!" but a library to interface with the DS18B20 family of temperature sensors (see Chap. 6)
Handbag	http://rancidbacon.com/p/android-arduino-handbag/	An alternative mechanism for communicating with Android devices
IRRemote	https://github.com/shirriff/Arduino-IRremote	Sending and receiving IR remote commands using an IR LED sender and IR receiver
Keypad	http://arduino.cc/playground/Code/Keypad	Decodes keypresses from matrix keypads
OneWire	http://arduino.cc/playground/Learning/OneWire	1-Wire Interface library
RTC library	http://jeelabs.org/2010/02/05/new-date-time-rtc-library/	Interfaces with various real-time clock (RTC) ICs
Si4703_Breakout	http://www.doctormonk.com/2011/09/sparkfun-si4703-fm-receiver-breakout.html	Allows easy control of the Si4703 radio receiver IC
USB Host Shield	http://www.circuitsathome.com	Allows the use of USB devices such as keyboards; also used for Android accessories
VirtualWire	http://www.open.com.au/mikem/arduino/VirtualWire.pdf	Provides serial communication between two Arduinos over a 433 MHz FM radio link
xbee	http://code.google.com/p/xbee-arduino/	Communicates with XBee data modules

For more information about Arduino, the official Arduino website (http://www.arduino.cc) should be your first port of call.

13.4.6 Arduino Example Project

We are going to repeat the BASIC Stamp example project, using the same external electronics, but

controlled by an Arduino rather than a BSII. The following listing shows the code needed to control the robot.

```
#include <Servo.h>

constintrightIRPin = 3;
constintleftIRPin = 4;
constintrightServoPin = 8;
constintleftServoPin = 9;

constintirOutPin = 10;
constintbuzzerPin = 11;

int speed = 60;       // servo speed as an angle offset from 180
intturnSpeed = 30;

Servo leftServo;
Servo rightServo;

void setup()
{
pinMode(rightIRPin, INPUT);
pinMode(leftIRPin, INPUT);
pinMode(irOutPin, OUTPUT);
pinMode(buzzerPin, OUTPUT);
digitalWrite(irOutPin, HIGH);
leftServo.attach(leftServoPin);
rightServo.attach(rightServoPin);
}

void loop()
{
   if (digitalRead(leftIRPin) == LOW &&digitalRead(rightIRPin) == LOW)
   {
     backup();
   }
   else if (digitalRead(leftIRPin) == LOW)
   {
     turnRight();
   }
   else if (digitalRead(rightIRPin) == LOW)
   {
     turnLeft();
   }
   else
   {
     forward();
   }
}

void forward()
{
  tone(buzzerPin, 440, 1000);   // play 440Hz for 1 second
  leftServo.write(180 - speed);
  rightServo.write(180 + speed);
}

void backup()
{
  tone(buzzerPin, 880, 1000);   // play 440Hz for 1 second
  leftServo.write(180 + speed);
  rightServo.write(180 - speed);
}

void turnLeft()
{
  leftServo.write(180 - turnSpeed);
  rightServo.write(180 - turnSpeed);
}
```

```
void turnRight()
{
  leftServo.write(180 + turnSpeed);
  rightServo.write(180 + turnSpeed);
}
```

There are many obvious similarities with the BASIC Stamp version. However, the C language used by Arduino provides a little more structure to the program, and seasoned programmers will probably feel happier using C rather than BASIC.

As noted in the previous section, a feature of the Arduino platform is the large number of libraries, both official and user-supplied, that can be included in the program. The first line includes a library for controlling servos. Strictly speaking, this is C++, the object-oriented extension to C. So, to control a servo, we first need to create an instance (well, two in this case):

```
Servo leftServo;
Servo rightServo;
```

Then we need to associate it with a particular pin using the following:

```
leftServo.attach(leftServoPin);
rightServo.attach(rightServoPin);
```

From now on, all we need to do is give the servo an angle in degrees, using the write command, like this:

```
leftServo.write(180 - speed);
```

There is no need to worry about pulse lengths; all this is handled for us automatically.

13.4.7 Taking the Arduino Offboard

For all its convenience and ease of use, an Arduino Uno is just a microcontroller with the necessary support components to provide it with a regulated voltage and allow it to be programmed over USB. When the time comes to create a product, if it is a one-off, you will probably just use the Arduino as it is, and buy yourself a replacement for the next project. If however, you will be producing a number of devices, you will probably want to lose the Arduino board and just use the programmed microcontroller in your project.

Taking the Arduino offboard means that you will probably design your own PCB for the microcontroller that includes just those features of the Uno that you need in your project, along with those extra components that may have been provided by a shield or connections to other electronics on, say, a breadboard.

The ATmega328 is not as easy to run using an internal oscillator as the ATtiny.

The IDE expects an external clock (crystal or ceramic resonator). You will probably also want to use a voltage regulator IC to provide a stable voltage to the microcontroller.

The schematic shown in Fig. 13.13 was produced in Eagle CAD, along with the corresponding board shown in Fig. 13.14.

EagleCAD schematic for offboard Arduino project

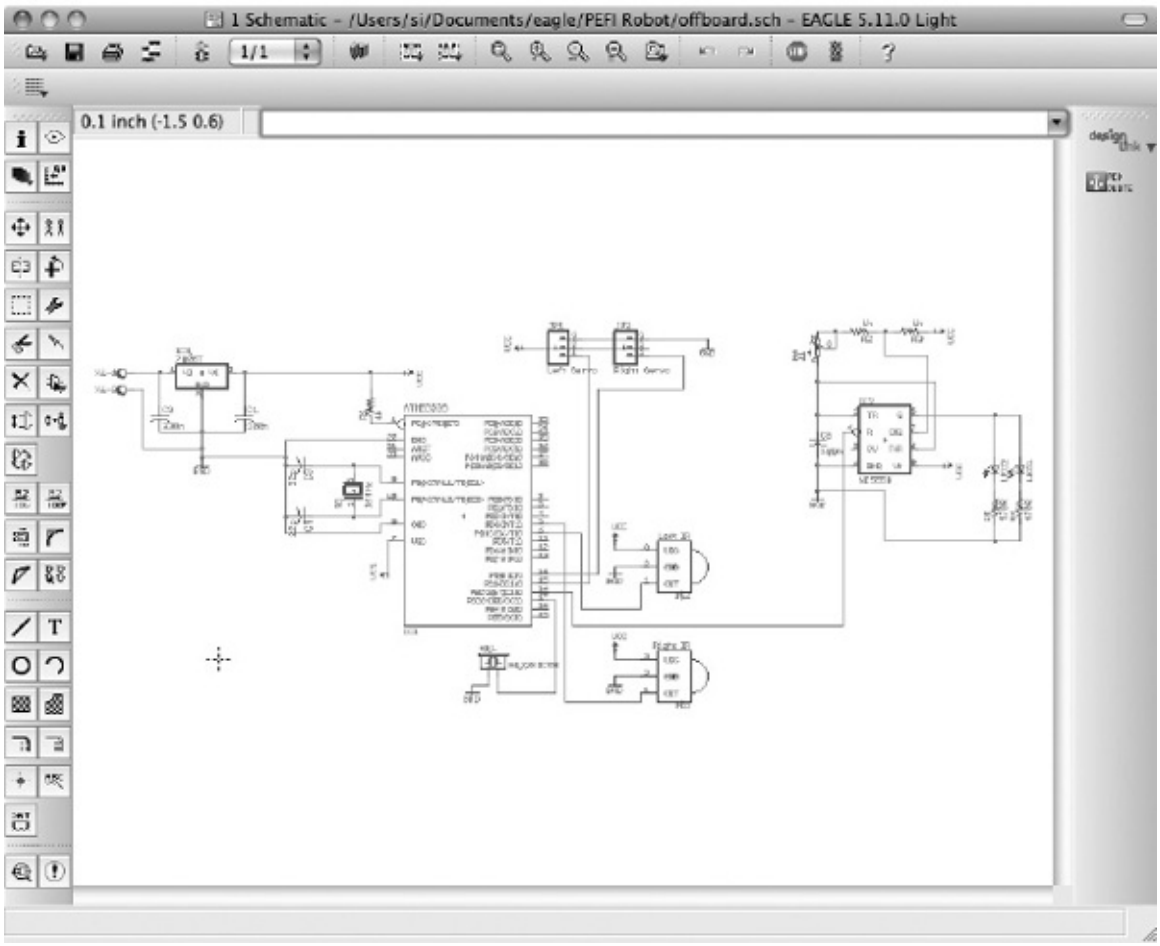


FIGURE 13.13

PCB layout for offboard Arduino project

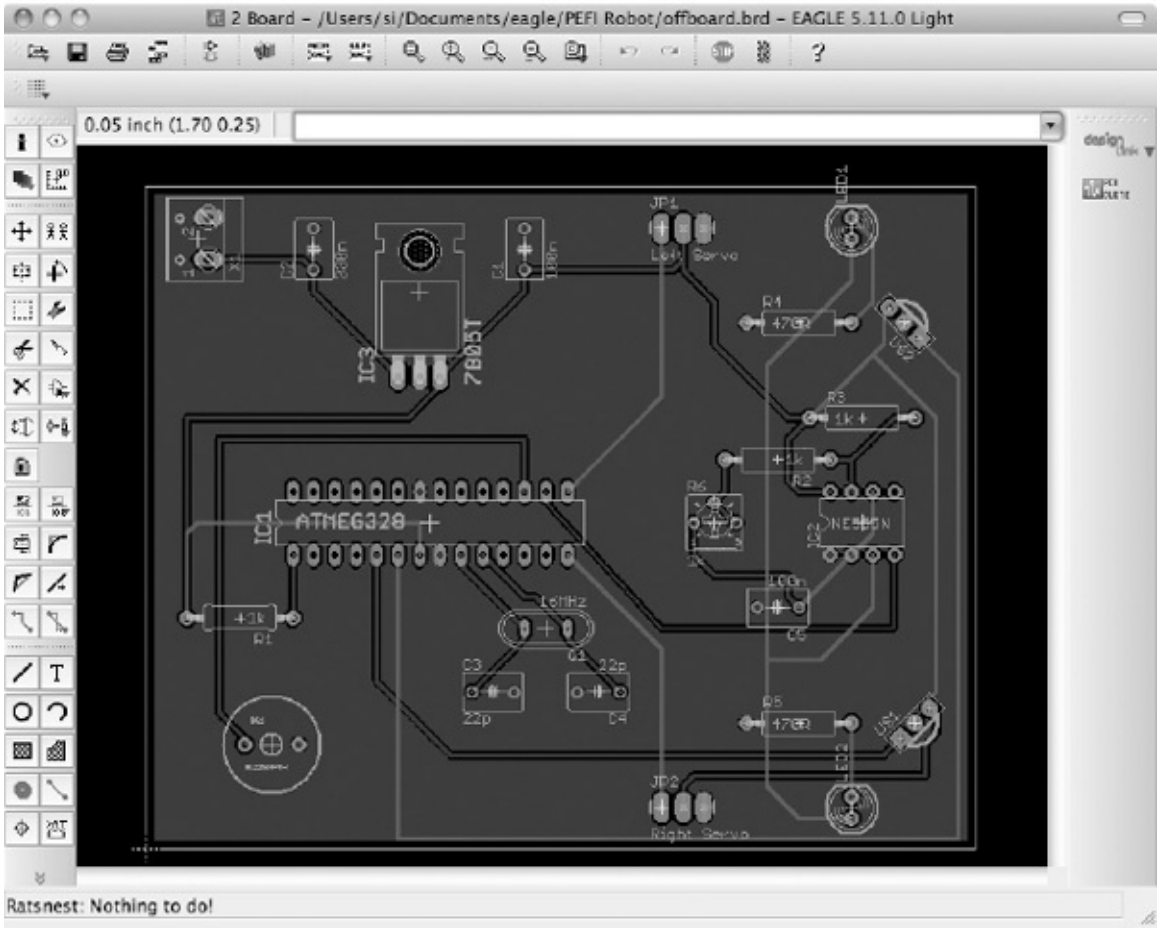


FIGURE 13.14

13.5 Interfacing with Microcontrollers

Whether you are using an ATtiny, a PIC, or an Arduino, you can be fairly certain that you will need to connect some components to it. At the very least, there will probably be a switch or two.

You can use three types of interfaces to your microcontroller:

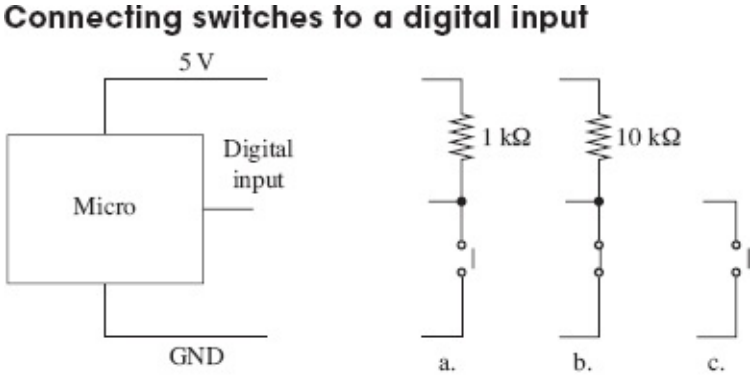
- *Digital:* Switches as inputs, LEDs or similar as outputs
- *Analog:* Sensors of various types (see Chap. 6)
- *Serial:* A serial communications protocol of which there are four main types: TTL serial, I^2C, 1-Wire, and Serial Peripheral Interface (SPI).

In the following sections, we will assume that your microcontroller has both analog and digital inputs, as well as digital and PWM outputs. It is also assumed that the microcontroller is operating at 5V. This may not always be the case, as many microcontrollers can be used at lower voltages, and 3.3 V is another common choice. If this is the case, you will need to adapt some of the schematics.

13.5.1 Switches

Single Switches

Switches are easy to connect to a digital input (see Fig. 13.15). Note the use of pullup resistors that keep the pin high until it is closed. When the switch is of the normally closed variety, then that current will be drawn continually, so you may want to use a high value of resistor there—say, 10 kΩ. However, for a normally open switch, current will flow only when the switch is pressed, so 1 kΩ is fine.



a. N.O.
b. N.C.
c. N.O. internal pullup resistor

FIGURE 13.15

With the switch to GND, as shown in all the examples in Fig. 13.15, when the switch is closed, the digital input will go low. This means that the logic of a button being pressed is inverted, as shown in the following sample Arduino C code.

```
if (digitalRead(4) == LOW)
{
  // the key was pressed, do something
}
```

It is also possible to swap over the switch and resistor, so that the resistor is now a pulldown resistor, and the switch being closed will result in a logical HIGH at the input.

The choice of pullup resistor depends on how electrically noisy your environment is and how long the leads are from the microcontroller to the switch. Essentially, it is a compromise between immunity to noise and current consumption. Given that for a normally open switch, current will flow only when the button is pressed, 5 mA using a 1 kΩ resistor is not normally a problem. In fact, some would advocate a lower value, such as 270 Ω.

Many microcontrollers include internal pullup resistors that can be turned on and off for a particular digital input. On an ATmega and ATtiny microcontroller, this resistor typically has a value of 20 k to 40 kΩ, so in a noisy environment or if the lead to the switch is long, an external pullup resistor may be better.

Multiple Switches to One Analog Input

If you have a lot of switches and do not want to tie up a load of digital inputs, then a common technique is to use an analog input and a number of resistors. The voltage at the analog input will then depend on the switches that are pressed (see Fig. 13.16).

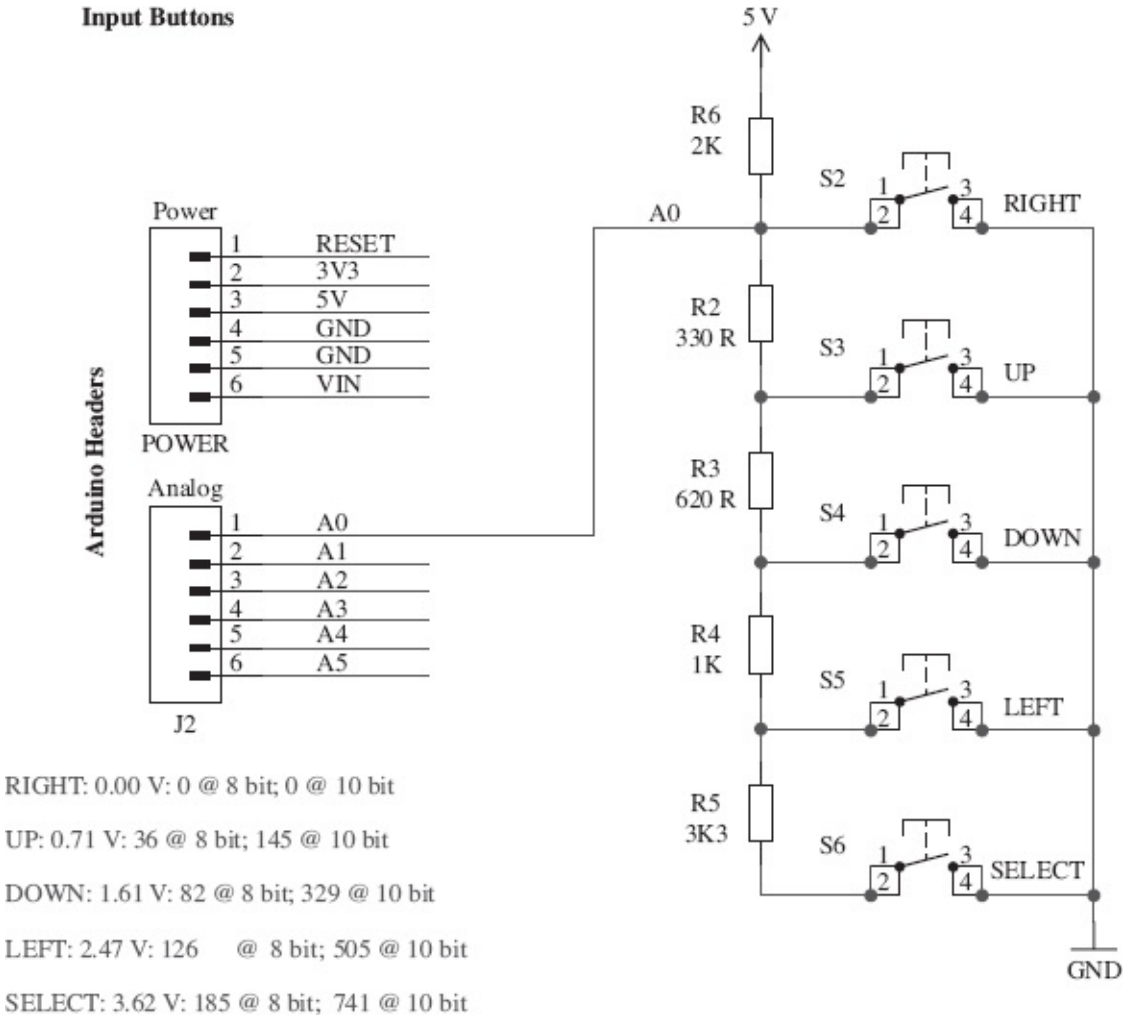


FIGURE 13.16

Figure 13.16 is taken from the schematic diagram for the Freetronics Arduino LCD shield, where it is used for the five switches of a joystick type arrangement of push buttons (thanks to Freetronics for permission to use this diagram). Note how the decimal values for 10-bit A to D for each button are given as a table.

Multiple switches and an analog input

The analog reading will not normally be exactly the value required due to resistor tolerances and power supply voltage changes, so in the code that interprets this, you would normally specify a band that would indicate a certain button, rather than just one value.

Using a Matrix Keypad

Keypads use switches arranged in a matrix, as shown in Fig. 13.17. The 4 × 3 keypad shown in the figure has a key at the intersection of each row and column. To determine which keys are pressed, the microcontroller will take each of the output pins Q0 to Q2 high in turn, and see what value is presented at each of the inputs I0 to I3. Note that if the microprocessor does not support internal pullup resistors, then these pullup resistors would be required on each input.

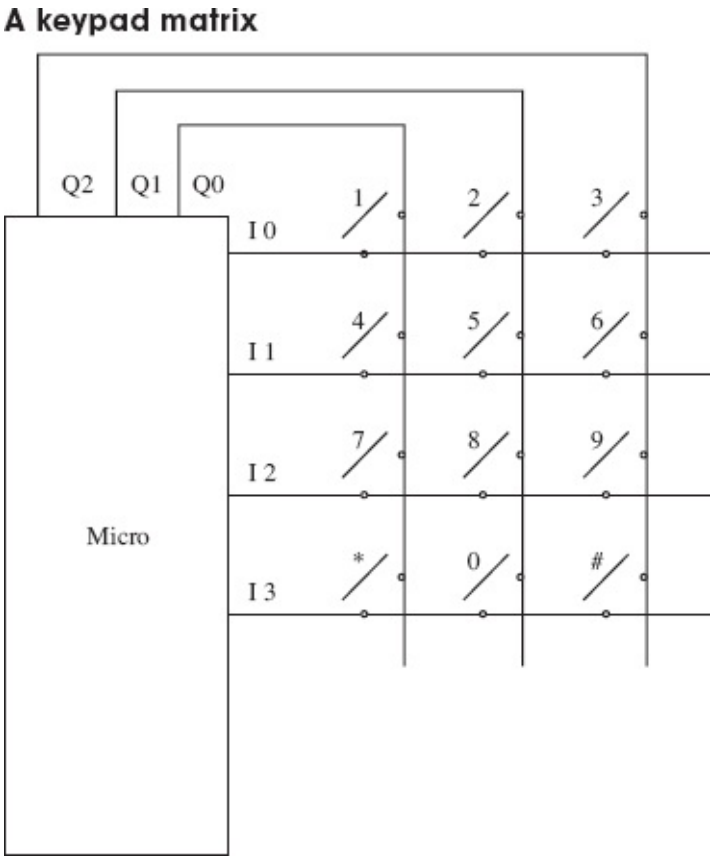


FIGURE 13.17

In practice, this is such a common component for microcontrollers that to write your own code for it would be needlessly reinventing the wheel. As an example, the following is the Arduino code for this that uses a library.

```
#include <Keypad.h>

char keys[4][3] = {
  {'1','2','3'},
  {'4','5','6'},
  {'7','8','9'},
  {'*','0','#'}
};

byte rowPins[4] = {2, 7, 6, 4};
byte colPins[3] = {3, 8, 5};

Keypad keypad = Keypad(makeKeymap(keys), rowPins, colPins, 4, 3);
```

```
void setup()
{
  Serial.begin(9600);
}

void loop()
{
  char key = keypad.getKey();
  if (key != null)
  {
    Serial.println(key);
  }
}
```

This example code will send any key that is pressed out through the Arduino serial monitor.

Debouncing

Attach an oscilloscope to the output of any of the circuits in Fig. 13.15, and you are likely to see an output something like what is shown in Fig. 13.18 when the switch is closed. This is called *bouncing*.

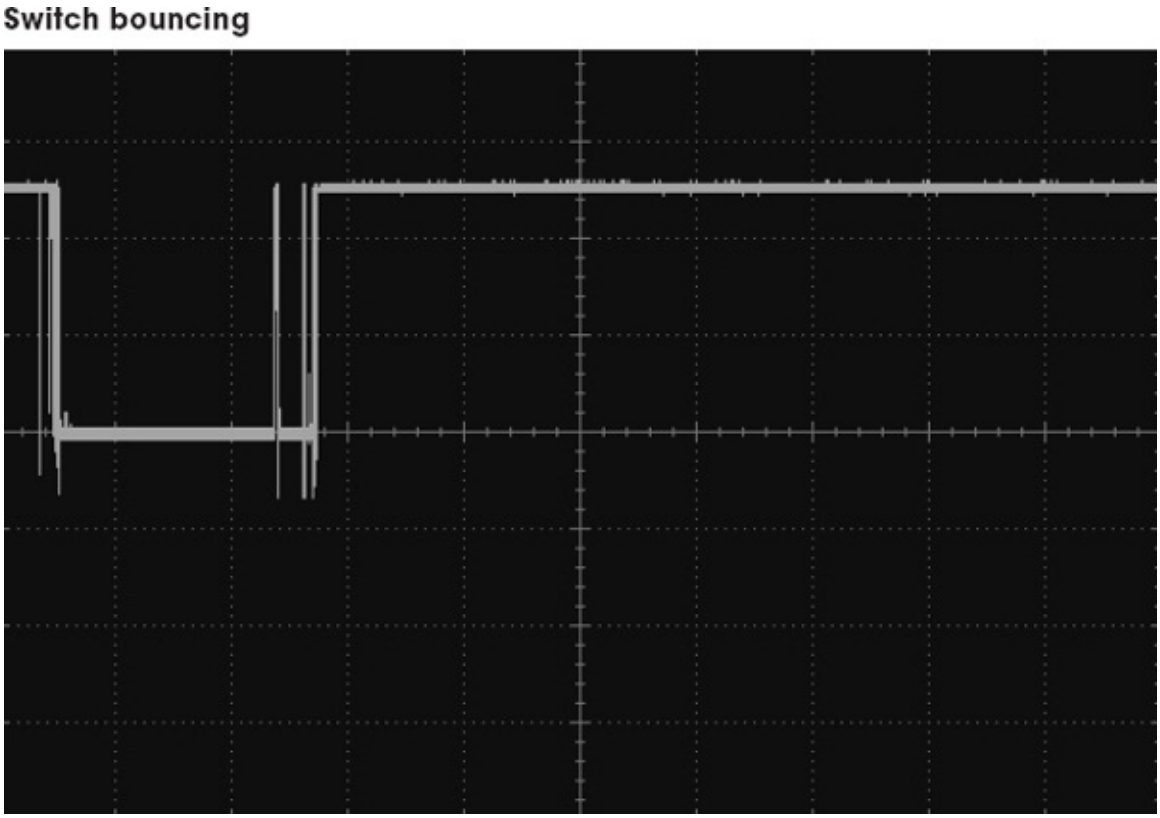


FIGURE 13.18

Switch bouncing can cause problems. Imagine the situation where pressing a button toggles an LED on and off. If there are an even number of bounces, then the LED will toggle on and then immediately off again, giving the impression that nothing happened.

It is therefore a good idea to *debounce* any switches that are connected to a microcontroller input. Although it is perfectly possible to do this with hardware—say, with a monostable that ignores subsequent pulses from the switch after triggering—it reduces the component count if you do the debouncing in software.

As with keyboard matrices, debouncing is a common problem that has been solved many times. The essence of software debouncing is the same as hardware debouncing, and that is to take action on the first

transition and then ignore any subsequent transitions of the output until a safe debounce period has elapsed. Depending on what else the microcontroller has to do, this can be as simple as inserting a delay for the debounce period in the code that immediately follows the detection of the first transition. However, sometimes this is not possible, such as when the microcontroller has other responsibilities (like refreshing an LED display). In these cases, a common approach is to set a variable to the milliseconds tick after the first transition and make a condition of actioning the button press that sufficient debounce time has elapsed. The Arduino code for this is as follows.

```
constintdebouncePeriod = 100;
long lastKeyPressTime = 0;

void loop()
{
  long timeNow = millis();
  if (digitalRead(5) == LOW &&lastKeyPressTime>timeNow + debouncePeriod)
  {
    // button pressed and enough time elapsed since last press
    // do what you need to do
    lastKeyPressTime = timeNow;
  }
}
```

13.5.2 Analog Inputs

Many of the sensors described in Chap. 6 provide an analog output to indicate the property that they are reading. For example, the TMP36 temperature sensor IC would typically be connected directly to an analog input of a microcontroller, as shown in Fig. 13.19.

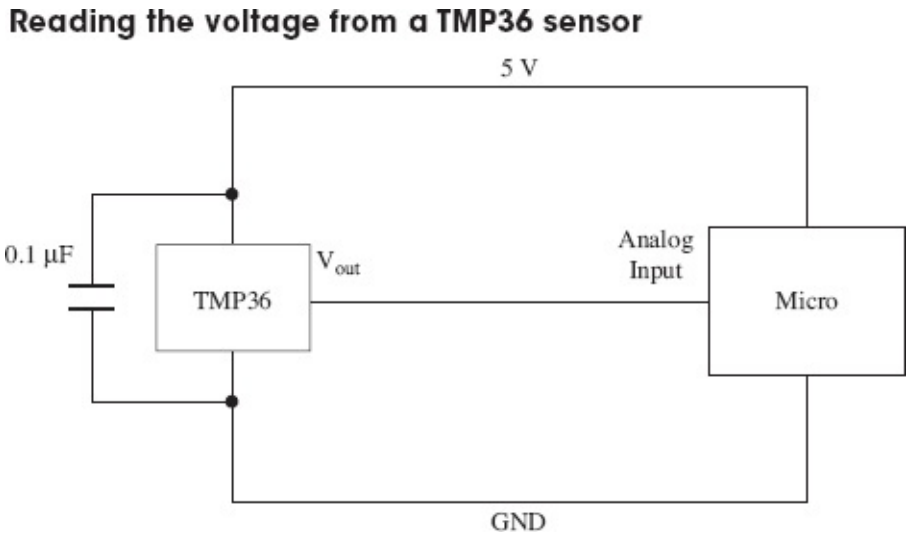


FIGURE 13.19

If you are measuring a voltage that is outside the range of the microcontroller's analog input (say 0 to 10 V), then you can just use two resistors as a voltage divider to reduce the voltage appropriately. If there is a risk that the voltage may exceed the expected range, you can protect the microcontroller's analog input by adding a zener diode (see Fig. 13.20).

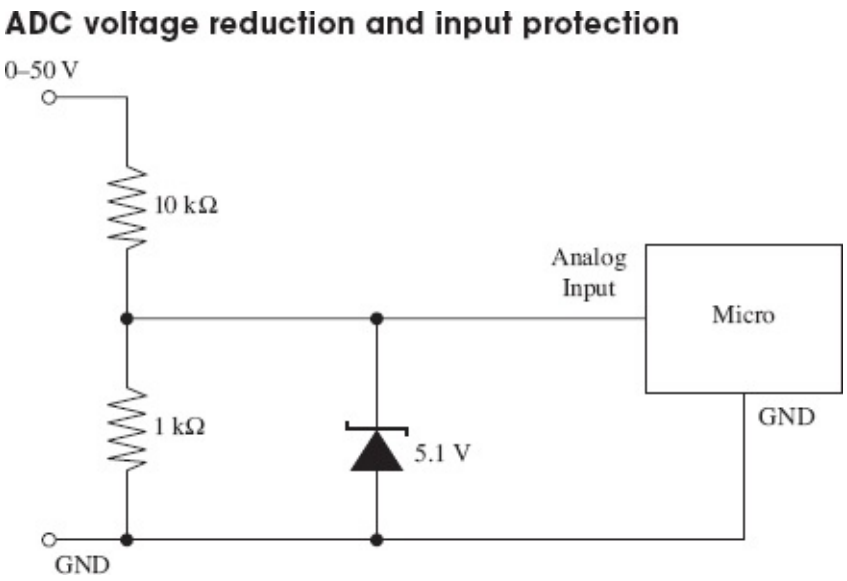


FIGURE 13.20

In practice, the zener diode will start to conduct before 5.1 V, to the detriment of the linearity of the readings, which is why the input range is labeled as 0 to 50 V, rather than 0 to 55 V. Remember the voltage divider is 1:11, not 1:10.

If switching behavior is required, then this should be implemented in code rather than through hardware. This approach allows more flexibility, for instance, for adding hysteresis or changing the set temperature. The following Arduino C code illustrates a simple temperature-control algorithm with 4 (±2) degrees of hysteresis.

```
const float hysteresis = 2.0;
const float setPoint = 20.0;
float temp = readTemperature();
if (temp <setPoint - hysteresis)
{
  digitalWrite(heaterControlPin, HIGH);
  }
else if (temp >setPoint + hysteresis)
{
  digitalWrite(heaterControlPin, LOW);
}
```

This example assumes that there is a user-supplied function called read Temperature.

For measuring resistance, sensors that are resistive, like LDRs and thermistors, will usually simply be used as one leg of a potential divider to produce a voltage that can be read. This is discussed in the relevant sensor sections in Chap. 6.

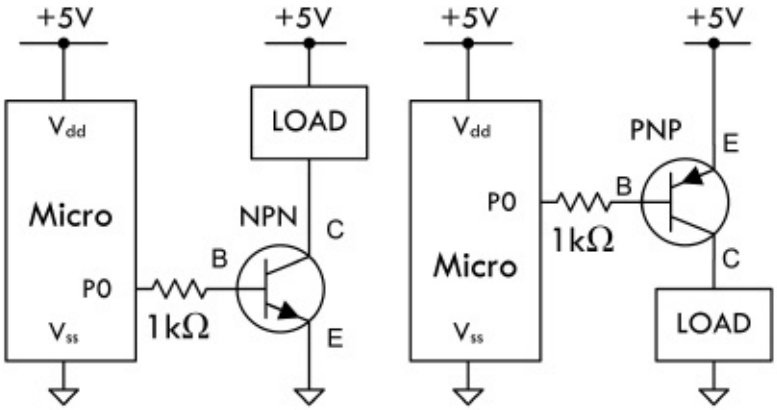
13.5.3 High-Power Digital Outputs

Most microcontrollers will reliably provide us with only around 20 mA of source or sink current as a direct digital output. If you want to drive a higher power load, such as a relay or a high-power LED, then you need to use a transistor.

Note that Arduino will handle up to 40 mA per pin with a maximum per chip of 200 mA. These figures should be derated by 25 percent for a production product, but Atmel says that the chip can comfortably cope with the absolute maximum current ratings, as these figures are already derated.

Figure 13.21 shows how to accomplish this with bipolar transistors, and Fig. 13.22 shows how to use MOSFETs.

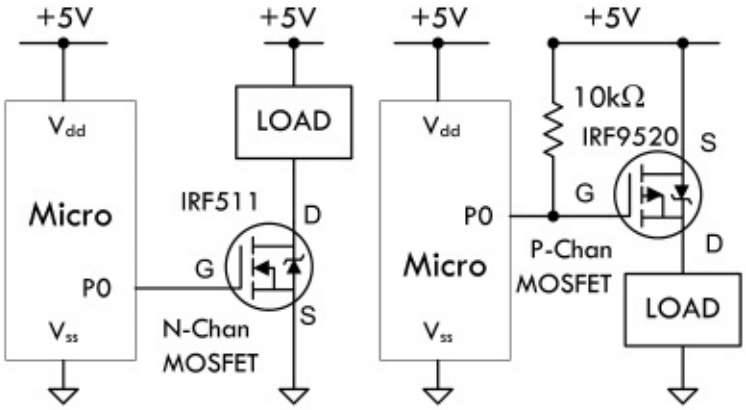
Handling high-power outputs with bipolar transistors



A bipolar transistor can be used to switch on and off loads. In the circuit to the left, when P0 is set HIGH, the NPN transistor is turned on—a low resistance is now present between C and E. In the circuit to the right, when P0 is set LOW, the PNP transistor is turned on. Microcontroller I/O current levels are usually large enough to provide enough base current for a bipolar transistor. Select a bipolar transistor with the appropriate current/voltage ratings for your load.

FIGURE 13.21

Handling high-power outputs with MOSFETs



MOSFETs have a much lower on-state resistance (in milliohms) than bipolar transistors (10's to 100's of milliohms). This means that a MOSFET driver will experience a smaller voltage drop, and in general, can handle much larger currents. MOSFETs also have very high input impedances; they draw very little gate current from a microcontroller's I/O pins. Some MOSFETs are capable of handling 60A or more. In the circuit to the left, an N-Chan MOSFET is triggered with a HIGH on P0. In the right circuit, a P-Chan MOSFET is triggered with a LOW on P0. Separate load supplies can be used, and are recommended if there are inductive loads present.

FIGURE 13.22

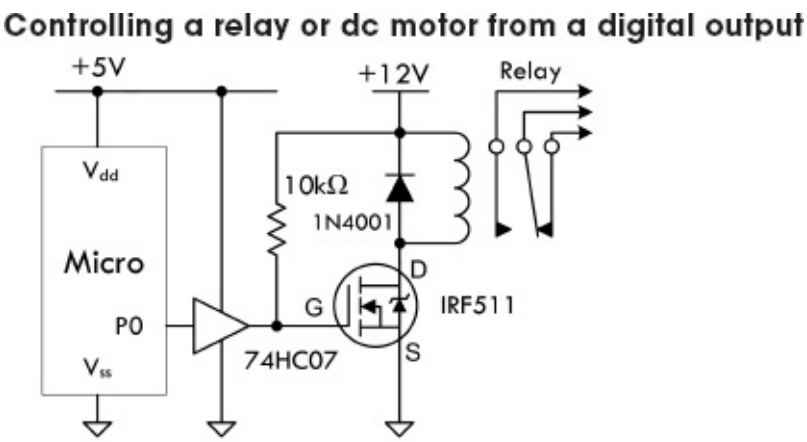
MOSFETs have a number of advantages over bipolar transistors. One is that for most applications, they do not need a gate resistor. However, MOSFETs are a capacitative load, so the inrush current when the pin changes state can be very high, albeit for a very short duration. Microcontrollers will generally cope with this, but for ultimate adherence to design rules, use a gate resistor of around 1 kΩ.

Another advantage of using MOSFETs as switches is their exceptionally low drain-source on resistance and high off resistance. This makes a small MOSFET capable of controlling quite big loads. However, you should check the gate threshold voltage to make sure that it is not above the logic level. For

instance, an N-channel MOSFET with a gate threshold voltage of 6 V is not going to turn on when the gate goes to just 5 V. This is more of a problem with high-power MOSFETs. When using high-power MOSFETs, look for those described as “logic-level” MOSFETs, meaning that they have a gate threshold significantly less than 5 V.

Relays and Other Inductive Loads

With the exception of some reed relays, very few relays will switch with a current less than 50 mA, and therefore you will nearly always need to use a transistor as just described. You also need to remember to use a reverse-biased diode across the relay’s coil, to prevent voltage spikes damaging the transistor during switching. Figure 13.23 shows this arrangement.

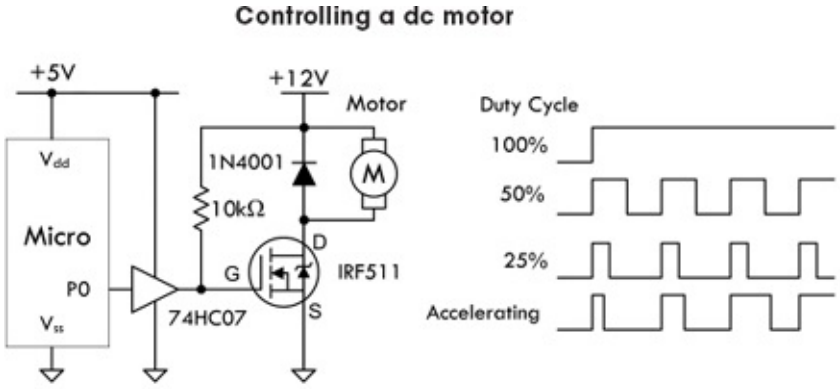


For high-current applications, say a 12V relay, a MOSFET transistor is a better choice than a bipolar transistor. Here an N-Chan MOSFET is driven by the microcontroller through a 74HC07 buffer. The diode is used to stomp out inductive spikes generated by coil. When P0 is set HIGH, MOSFET is turned on, and relay switches states.

FIGURE 13.23

Pulse-Width Modulation

The schematic shown in Fig. 13.23 is also suitable for controlling inductive loads like dc motors. If the digital output is driven as PWM, then this circuit can also be used to control the power going to the motor, and hence its speed (see Fig. 13.24).




```
Example pseudo code segments:
Declare variables x
'100% duty cycle
Set P0 = 1    '100% duty cycle

'50% duty cycle
For x = 1 to 200
  Set P0 = 1
  Pause 5   'ON for 5 milliseconds
  Set P0 = 0   'Set pin 0 LOW
  Pause 5   'OFF for 5 milliseconds

'25% duty cycle
For x = 1 to 100
  Set P0 = 1   'Set pin0 HIGH
  Pause 5   'ON for 5 milliseconds
  Set P0 = 0   'Set pin 0 LOW
  Pause 15   'OFF for 15 milliseconds
Next

'Accelerating
For x = 100 to 1
  Set P0 = 1   'Set pin0 HIGH
  Pause 15   'ON for 15 milliseconds
  Set P0 = 0   'Set pin0 LOW
  Pause x
Next
```

FIGURE 13.24

The waveforms on the right in Fig. 13.24 show how you can control the motor speed by adjusting the duty-cycle (proportion of time the power is on). The pseudocode on the right shows how this is

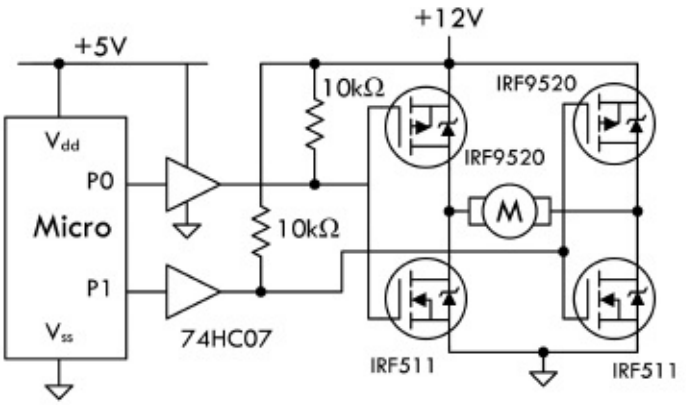
accomplished. Note that some microcontrollers also have dedicated hardware support to simplify the process of generating PWM signals.

If your microcontroller has fairly robust output drivers, then there is probably little point in using the 74HC07 buffer.

Directional Motor Control

Directional control of motors can be achieved using an H-bridge, as shown in Fig. 13.25.

Bidirectional motor control



The H-bridge circuit built using MOSFETs provides forward and reverse directional control of a dc motor. The H-bridge provides built-in dynamic breaking action useful for applications requiring greater control. To make motor go in one direction, P0 is set HIGH while P1 is set LOW. To switch directions, P0 is set LOW while P1 is set HIGH. Buffer stage (74HC07) could be replaced with a optoisolator to provide greater electrical isolation from motor section of circuit.

FIGURE 13.25

Most useful for controlling motor currents of less than a couple of amps are IC H-bridges such as the TB6612FNG, which combine all the transistors into one package. They often also have features such as thermal shutdown, to protect against overloading.

Servo Motor Control

We have already touched on controlling servo motors in our robot example project. Since the servo uses a control signal, this can be provided directly from a digital output (see Fig. 13.26).

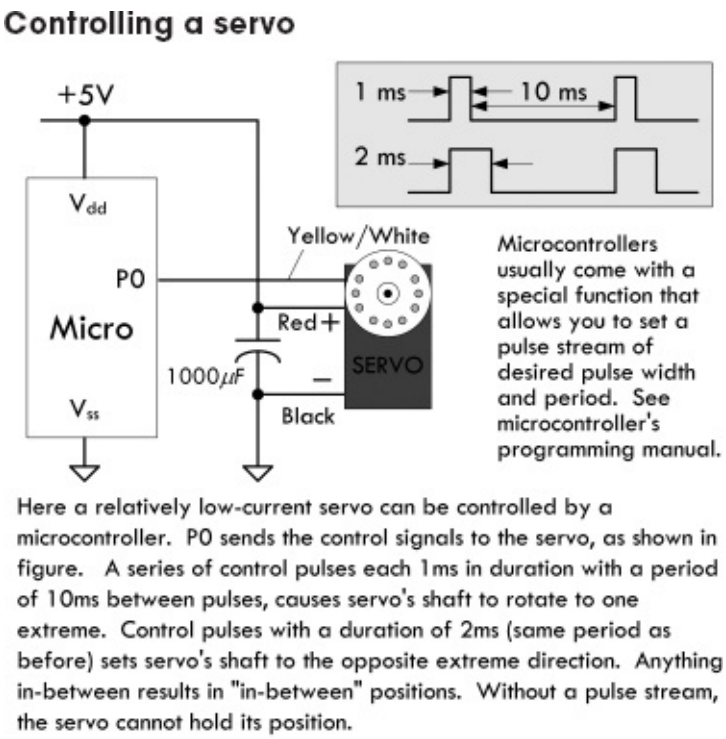


FIGURE 13.26

Stepper Motor Control

A stepper motor has a number of coils that must be energized in the correct sequence to move the rotor. The arrangement shown in Fig. 13.27 can be used to accomplish this.

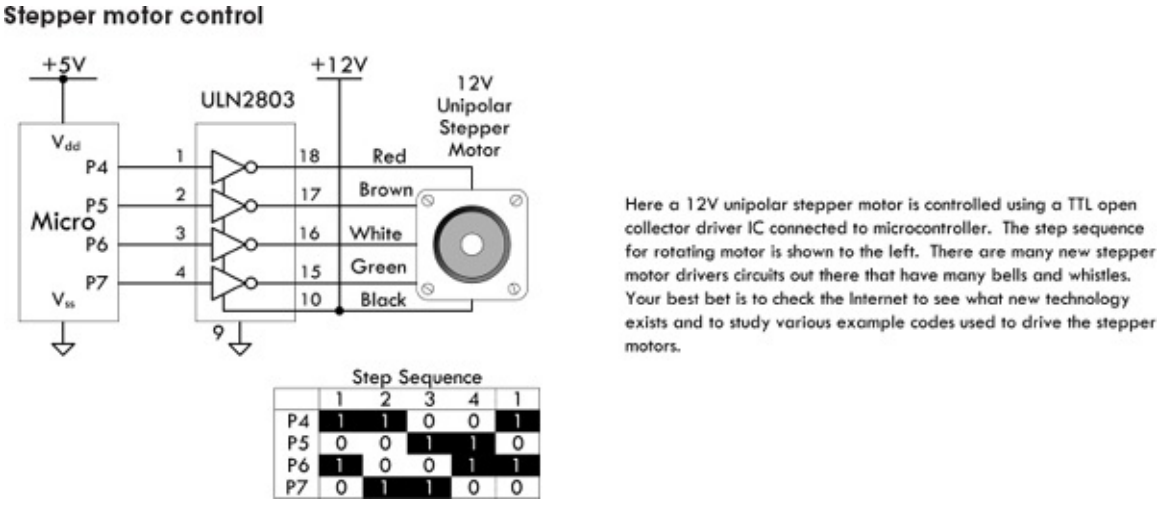


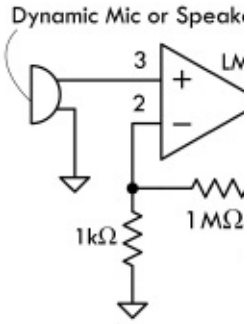
FIGURE 13.27

See Chap. 14 for more information about motors.

13.5.4 Sound Interfaces

Figure 13.28 shows a schematic for detecting sound. The second comparator stage is optional, and the output of the first stage could be fed directly into an analog input, allowing the sound to be sampled. Most microcontroller ADCs are not terribly fast, but even so, they should be able to sample at above 10 kHz, allowing some primitive digital signal processing.

Detecting sound



This circuit uses an LM324 comp
microphone or speaker. When
by pot, the output suddenly ch
microcontroller's input.

FIGURE 13.28

When it comes tc
microcontroller than
wait again, and kee
library and the BASI
can be driven directl
be beyond the drive
audio amplifier circ
quality amplification
loudspeaker, will wc
collector current, as

FIGURE 13.31

Generating a sine
of one of the pins
microcontrollers is
squarewave. A bette
voltage proportional
need are resistors.

Figure 13.29 shov
so *R* might be 5 KΩ
output. The four digit
Higher-resolution D/
convenient to use, ar

CHAPTER 14

Motors

Perhaps one of the most entertaining things to do with electronics is make some mechanical device move. Three very popular devices used to “make things move” include dc motors, RC servos, and stepper motors.

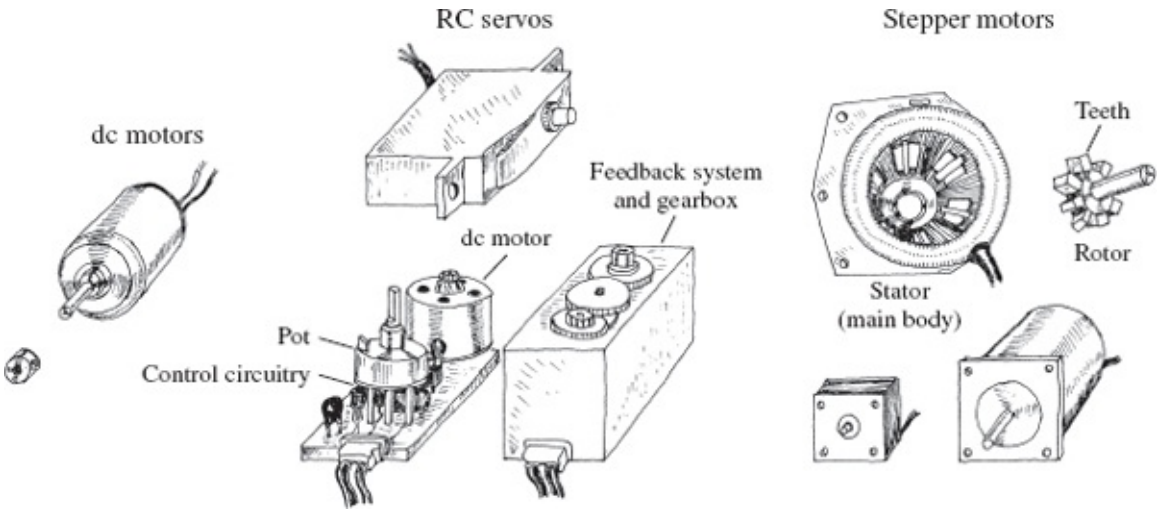


FIGURE 14.1

14.1 DC Continuous Motors

A *dc motor* is a simple two-lead, electrically controlled device that comes with a rotary shaft on which wheels, gears, propellers, etc., can be mounted. A dc motor generates a considerable amount of revolutions per minute (rpm) for its size and can be made to rotate clockwise or counterclockwise by reversing the polarity applied to the leads. At low speeds, dc motors provide little torque and minimal position control, making them impractical for pointlike position-control applications.

dc motors are available in many different shapes and sizes. Most dc motors provide rotational speeds anywhere between 3000 and 8000 rpm at a specific operating voltage typically set between 1.5 and 24 V. The operating voltage provided by the manufacturer tells you at what voltage the motor runs most efficiently. Now, the actual voltage applied to a motor can be made slightly lower to make the motor slower or can be elevated to make the motor faster. However, when the applied voltage drops to below around 50 percent of the specified operating voltage, the motor usually will cease to rotate. Conversely, if the applied voltage exceeds the operating voltage by around 30 percent, there is a chance that the motor will overheat and become damaged. In practice, as you will see in a second, the speed of a dc motor is most efficiently controlled by means of pulse-width modulation, whereby the motor is rapidly turned on and off. The width of the applied pulse, as well as the period between pulses, controls the speed of the motor. Also, it is worth noting that a freely running motor (no load) may draw little current (power). However, if a load is applied, the amount of current drawn by the motor’s inner coils goes up immensely (up to 1000 percent or more). Manufacturers usually will provide what is called a *stall current* rating for

their motors. This rating specifies the amount of current drawn at the moment the motor stalls. If your motor's stall current rating is not listed, it is possible to determine it by using an ammeter; slowly apply a force to the motor's shaft, and note the current level at the point when the motor stalls. Another specification given to dc motors is a torque rating. This rating represents the amount of force the motor can exert on a load. A motor with a high torque rating will exert a larger force on a load placed at a tangent to its rotational arm than a motor with a lower torque rating. The torque rating of a motor is usually given in lb/ft, g/cm, or oz/in.

14.2 Speed Control of DC Motors

Bad Designs

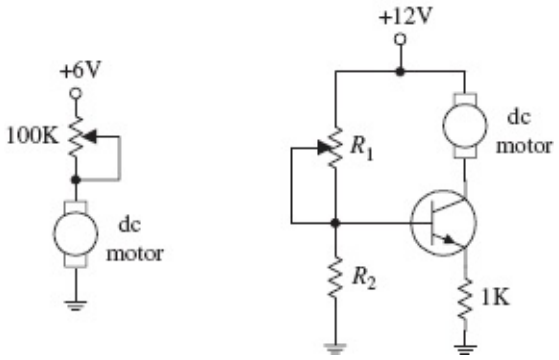
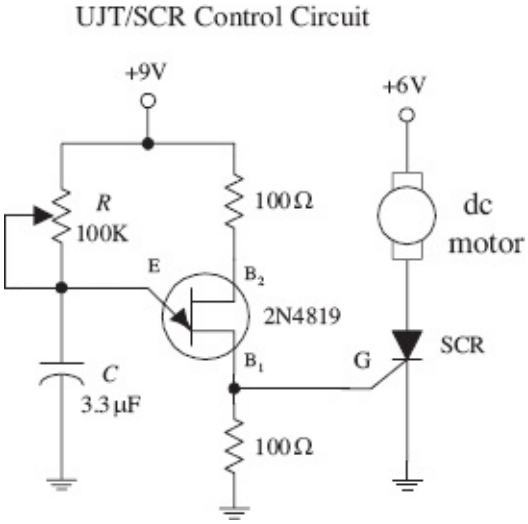


FIGURE 14.2

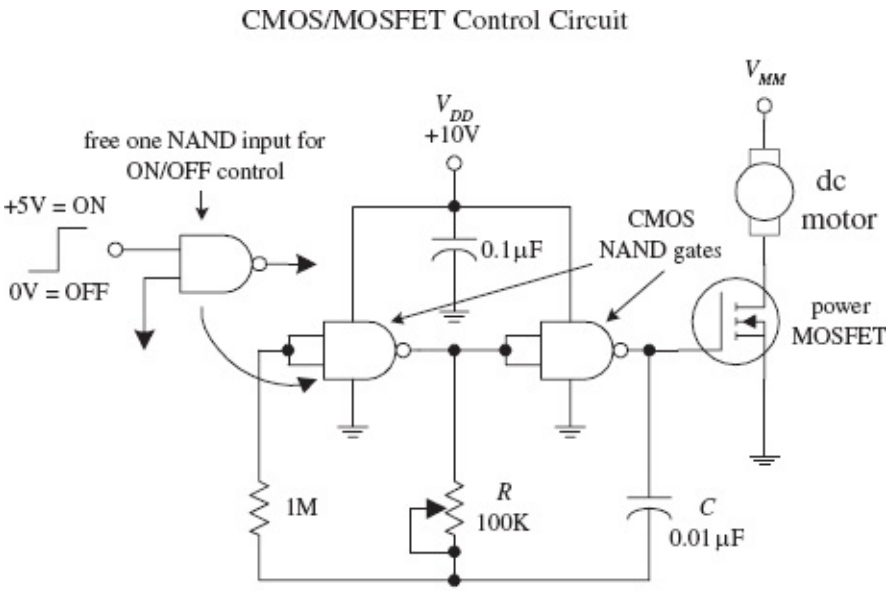
A seemingly obvious approach to control the speed of a dc motor would be simply to limit the current flow by using a potentiometer, as shown in the circuit to the left in the figure. According to Ohm's law, as the resistance of the pot increases, the current decreases, and the motor will slow down. However, using a pot to control the current flow is inefficient. As the pot's resistance increases, the amount of current energy that must be converted into heat increases. Producing heat in order to slow a motor down is not good—it consumes supply power and may lead to potentiometer meltdown. Another seemingly good but inefficient approach to control the speed of a motor is to use a transistor amplifier arrangement like the one shown to the right in the figure. However, again, there is a problem. As the collector-to-emitter resistance increases with varying base voltage/ current, the transistor must dissipate a considerable amount of heat. This can lead to transistor meltdown.

Better Designs

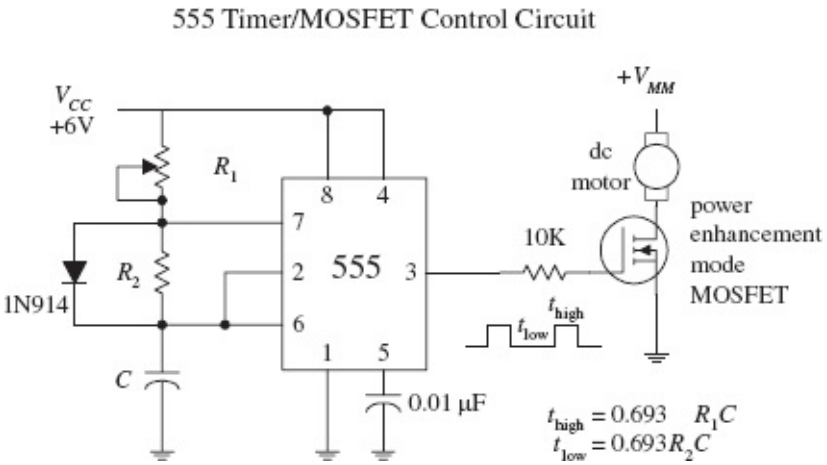


In order to conserve energy and prevent component meltdown, an approach similar to what was used in switching power supplies is used to control the speed of the motor. This approach involves sending the motor short pulses of current. By varying the width and frequency of the applied pulses, the speed of the motor can be controlled. Controlling a motor's speed in this manner prevents any components from experiencing continuous current stress. Figure 13.3 shows three simple circuits used to provide the desired motor-control pulses.

In the first circuit, a UJT relaxation oscillator generates a series of pulses that drives an SCR on and off. To vary the speed of the motor, the UJT's oscillatory frequency is adjusted by changing the RC time constant.



In the second circuit, a pair of NAND gates make up the relaxation oscillator section, while an enhancement-type power MOSFET is used to drive the motor. Like the preceding circuit, the speed of the motor is controlled by the oscillator's *RC* time constant. Notice that if one of the input leads of the left NAND gate is pulled out, it is possible to create an extra terminal that can be used to provide on/off controls that can be interfaced with CMOS logic circuits.



The third circuit is a 555 timer that is used to generate pulses that drive a power MOSFET. By inserting a diode between pins 7 and 6, as shown, the 555 is placed into low-duty cycle operation. R_1, R_2, and C set the frequency and on/off duration of the output pulses. The formulas accompanying the diagram provide the details.

A microcontroller-based dc control circuit with speed control is found in Chap. 13.

FIGURE 14.3

14.3 Directional Control of DC Motors

To control the direction of a motor, the polarity applied to the motor's leads must be reversed. A simple manual-control approach is to use a DPDT switch (see leftmost circuit in Fig. 14.4). Alternately, a transistor-driven DPDT relay can be used (see middle circuit). If you do not like relays, you can use a push-pull transistor circuit (see leftmost circuit). This circuit uses a complementary pair of transistors (similar betas and power rating)—one is an *npn* power Darlington, and the other is a *pnp* power Darlington. When a high voltage (e.g., +5 V) is applied to the input, the upper transistor (*npn*) conducts,

allowing current to pass from the positive supply through the motor and into ground. If a low voltage (0 V) is applied to the input, the lower transistor (*pnp*) conducts, allowing current to pass through the motor from ground into the negative supply terminal.

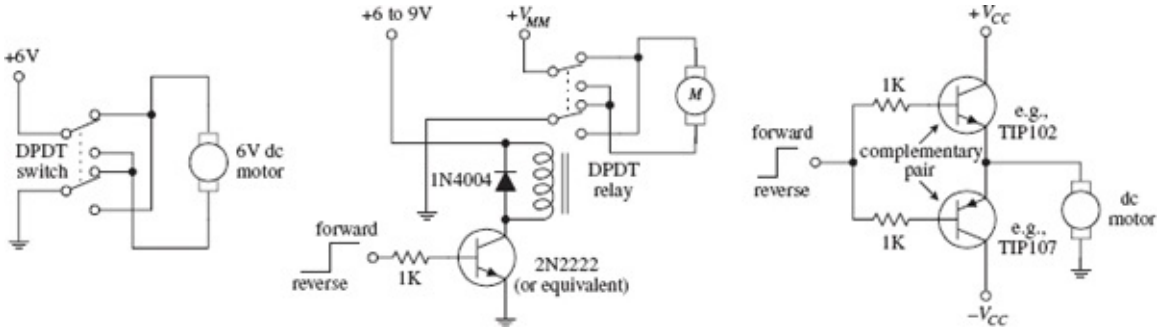


FIGURE 14.4

Another very popular circuit used to control the direction of a motor (as well as the speed) is the H-bridge. Figure 14.5 shows two simple versions of the H-bridge circuit. The left H-bridge circuit is constructed with bipolar transistors, whereas the right H-bridge circuit is constructed from MOSFETs. To make the motor rotate in the forward direction, a high (+5-V) signal is applied to the forward input, while no signal is applied to the reverse input (applying a voltage to both inputs at the same time is not allowed). The speed of the motor is controlled by pulse-width modulating the input signal. Here is a description of how the bipolar H-bridge works: When a high voltage is applied to Q_3's base, Q_3 conducts, which in turn allows the *pnp* transistor Q_2 to conduct. Current then flows from the positive supply terminal through the motor in the right-to-left direction (call it the *forward direction* if you like). To reverse the motor's direction, the high voltage signal is removed from Q_3's base and placed on Q_4's base. This sets Q_4 and Q_1 into conduction, allowing current to pass through the motor in the opposite direction. The MOSFET H-bridge works in a similar manner. The diodes within the H-bridge circuits help dampen transient spikes that are generated by the motor's coils so that they do not damage the other components within the circuit. All transistors (except the bipolar within the MOSFET circuit) should have high power ratings.

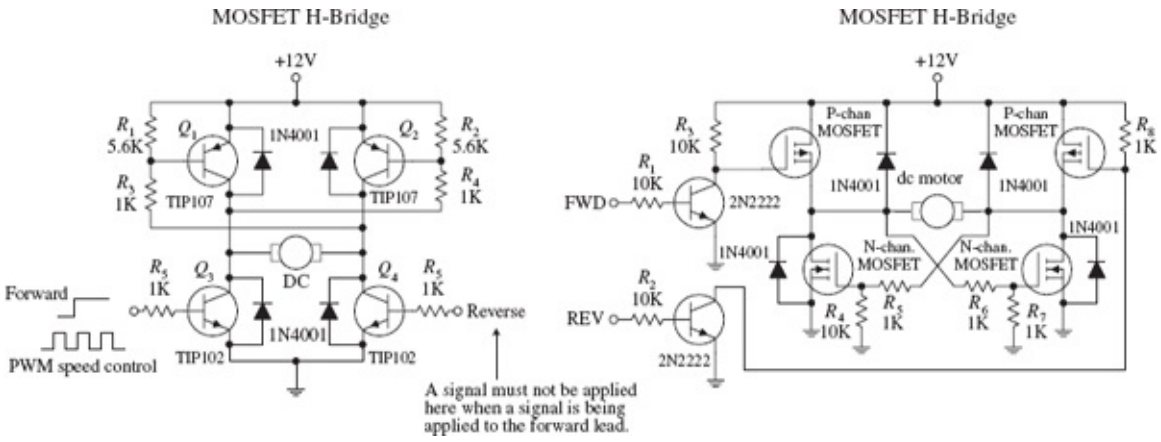


FIGURE 14.5

Now, it is possible to construct these H-bridge circuits from scratch, but it is far easier and usually cheaper to buy a motor-driven IC. For example, National Semiconductor's LMD18200 motor-driver IC is a high-current, easy-to-use H-bridge chip that has a rating of 3 A and 12 to 55 V. This chip is TTL and CMOS compatible and includes clamping diodes, shorted load protection, and a thermal warning interrupt output lead. The L293D (Unitrode) is another popular motor-driver IC. This chip is very easy to use and is cheaper than the LMD18200, but it cannot handle as much current and does not provide as

many additional features. There are many other motor-driver ICs out there, as well as a number of prefab motor-diver boards that are capable of driving a number of motors. Check the electronics catalogs and Internet to see what is available.

14.4 RC Servos

Remote control (*RC*) *servos*, unlike dc motors, are motorlike devices designed specifically for pointerlike position-control applications. An RC servo uses an external pulse-width-modulated (PWM) signal to control the position of its shaft to within a small fraction of its maximum range of rotation. To alter the position of the shaft, the pulse width of the modulated signal is varied. The amount of angular rotation of an RC servo's shaft is limited to around 180 or 210° depending on the specific brand of servo. These devices can provide a significant amount of low-speed torque (due to an internal gearing system) and provide moderate full-swing displacement switching speeds. RC servos frequently are used to control steering in model cars, boats, and airplanes. They are also used commonly in robotics as well as in many sensor-positioning applications.

The standard RC servo looks like a simple box with a drive shaft and three wires coming out of it. The three wires consist of a power supply wire (usually black), a ground wire (usually red), and the shaft-positioning control wire (color varies based on manufacturer). Within the box there is a dc motor, a feedback device, and a control circuit. The feedback device usually consists of a potentiometer whose control dial is mechanically linked to the motor through a series of gears. When the motor is rotated, the potentiometer's control dial is rotated. The shaft of the motor is usually limited to a rotation of 180° (or 210°)—a result of the pot not being able to rotate indefinitely. The potentiometer acts as a position-monitoring device that tells the control circuit (by means of its resistance) exactly how far the shaft has been rotated. The control circuit uses this resistance, along with a pulse-width-modulated input control signal, to drive the motor a specific number of degrees and then hold. (The amount of holding torque varies from servo to servo.) The width of the input signal determines how far the servo's shaft will be rotated.

By convention, when the pulse width is set to 1.5 ms, the servo rotates its shaft to neutral position (e.g., 90° if the servo is constrained within a 0 to 180° range). To rotate the shaft a certain number of degrees from neutral position, the pulse width of the control signal is varied. To make the shaft go counterclockwise from neutral, a pulse wider than 1.5 ms is applied to the control input. Conversely, to make the shaft go clockwise from neutral, a pulse narrower than 1.5 ms is applied (see figure). Knowing exactly how much wider or narrower to make the pulse to achieve exact angular displacements depends largely on what brand of servo you are using. For example, one brand of servo may provide maximum counterclockwise rotation at 1 ms and maximum clockwise rotation at 2 ms, whereas another brand of servo may provide maximum counterclockwise rotation at 1.25 ms and maximum clockwise rotation at 1.75 ms. The supply voltage used to power servos is commonly 4.8 V but may be 6.0 V or so depending on the specific brand of servo. Unlike the supply voltage, the supply current drawn by a servo varies greatly, depending on servo's power output.

A simple 555 timer circuit like the one shown in Fig. 14.6 can be used to generate the servo control signal. In this circuit, R_2 acts as the pulse-width control. Servos also can be controlled by a microprocessor or microcontroller. See Chap. 13 for two microcontroller-based servo control circuits.

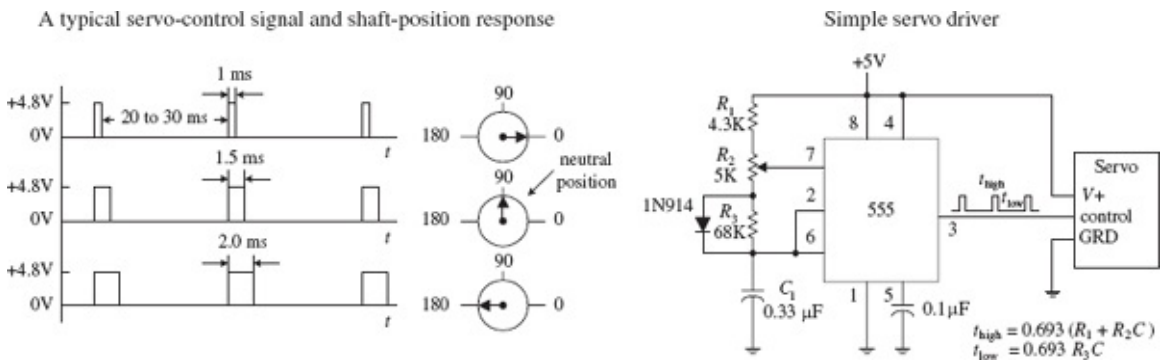


FIGURE 14.6

Now, when controlling servos within model airplanes, an initial control signal (generated by varying position-control potentiometers) is first sent to a radiowave modulator circuit that encodes the control signal within a carrier wave. This carrier wave is then radiated off as a radiowave by an antenna. The radiowave, in turn, is then transmitted to the model's receiver circuit. The receiver circuit recovers the initial control signal by demodulating the carrier. After that, the control signal is sent to the designated servo within the model. If there is more than one servo per model, more channels are required. For example, most RC airplanes require a four-channel radio set; one channel is used to control the ailerons, another channel controls the elevator, another controls the rudder, and another controls the throttle. More complex models may use five or six channels to control additional features such as flaps and retractable landing gear. The FCC sets aside 50 frequencies in the 72-MHz band (channels 11–60) dedicated to aircraft use only. No license is needed to operate these radios. However, with an amateur (ham) radio operator's license, it is possible to use a radio within the 50-MHz band. Also, there are frequencies set aside within the 27-MHz band that are legal for any kind of model use (surface or air). If you are interested in radio-controlled RC servos, a good starting point is to check out an RC model hobby shop. These shops carry a number of transmitter and receiver sets, along with the servos.

As a final note, with a bit of rewiring, a servo can be converted into a drive motor with unconstrained rotation. A simple way to modify the servo is to break the feedback loop. This involves removing the three-lead potentiometer (and unlinking the gear system so that it can rotate 360°) and replacing it with a pair of voltage-divider resistors (the output of the voltage divider replaces the variable terminal of the potentiometer). The voltage divider is used to convince the servo control circuit that the servo is in neutral position. The exact values of the resistors needed to set the servo in neutral position can be determined by using the old potentiometer and an ohmmeter. Now, to turn the motor clockwise, a pulse wider than 1.5 ms is applied to the control input. As long as the control signal is in place, the motor will keep turning and not stop—you have removed the feedback system. To turn the motor counterclockwise, a pulse narrower than 1.5 ms is applied to the control input.

14.5 Stepper Motors

Stepper motors, or *steppers*, are digitally controlled brushless motors that rotate a specific number of degrees (a step) every time a clock pulse is applied to a special translator circuit that is used to control the stepper. The number of degrees per step (resolution) for a given stepper motor can be as small as 0.72° per step or as large as 90° per step. Common general-purpose stepper resolutions are 15 and 30° per step. Unlike RC servos, steppers can rotate a full 360° and can be made to rotate in a continuous manner like a dc motor (but with a lower maximum speed) with the help of proper digital control circuitry. Unlike dc motors, steppers provide a large amount of torque at low speeds, making them suitable in applications where low-speed and high-precision position control is needed. For example, they are used in printers to control paper feed and are used to help a telescope track stars. Steppers are

also found in plotter-and sensor-positioning applications. The list goes on. To give you a basic idea of how a stepper works, take a look at Fig. 14.7.

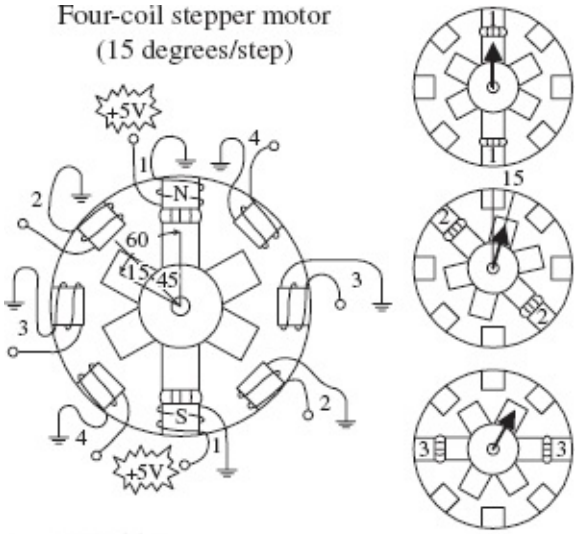


FIGURE 14.7

Here is a simple model depicting a 15° per step variable-reluctance stepper. The stationary section of the motor, called the *stator*, has eight poles that are spaced 45° apart. The moving section of the motor, called the *rotor*, is made from a ferromagnetic material (a material that is attracted to magnetic fields) that has six teeth spaced 60° apart. To make the rotor turn one step, current is applied, at the same time, through two opposing pole pairs, or coil pairs. The applied current causes the opposing pair of poles to become magnetized. This in turn causes the rotor's teeth to align with the poles, as shown in the figure. To make the rotor rotate 15° clockwise from this position, the current through coil pair 1 is removed and sent through coil pair 2. To make the rotor rotate another 15° clockwise from this position, the current is removed from coil pair 2 and sent through coil pair 3. The process continues in this way. To make the rotor spin counterclockwise, the coil-pair firing sequence is reversed.

14.6 Kinds of Stepper Motors

The model used in the last example was based on a variable-reluctance stepper. As it turns out, this model is incomplete—it does not show how a real variable-reluctance stepper is wired internally. Also, the model does not apply to a class of steppers referred to as *permanent-magnet steppers*. To make things more realistic, let's take a look at some real-life steppers.

Variable-Reluctance Steppers

Figure 14.8 shows a physical model and schematic diagram of a 30° per step variable-reluctance stepper. This stepper consists of a six-pole (or three-coil pair) stator and a four-toothed ferromagnetic rotor. Variable-reluctance steppers with higher angular resolutions are constructed with more coil pairs and/or more rotor teeth. Notice that in both the physical model and the schematic, the ends of all the coil pairs are joined together at a common point. (This joining of the coil ends occurs internally within the motor's case.) The common and the coil pair free ends are brought out as wires from the motor's case. These wires are referred to as the *phase wires*. The common wire is connected to the supply voltage, whereas the phase wires are grounded in sequence according to the table shown in Fig. 14.8.

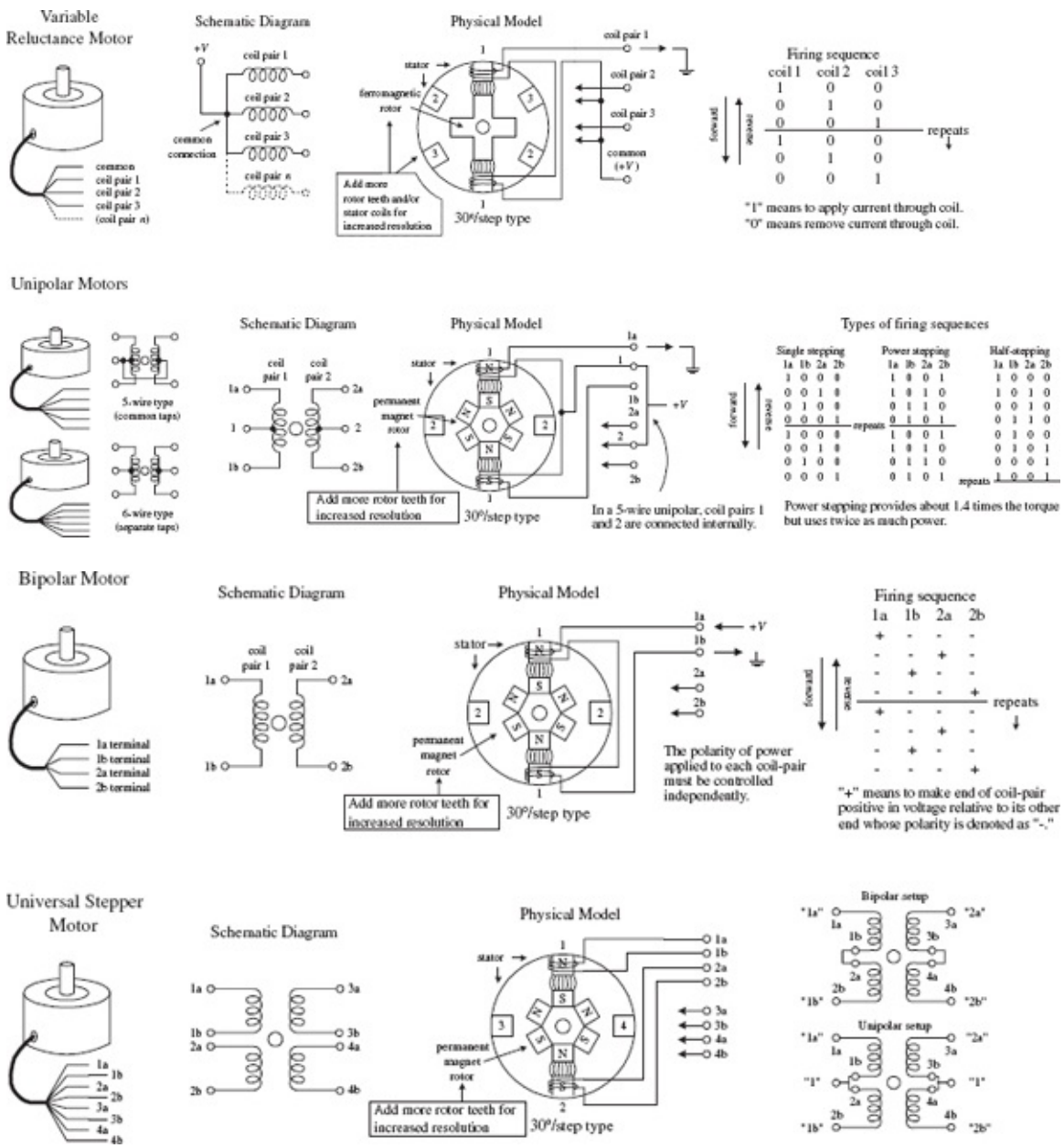


FIGURE 14.8

Permanent-Magnet Steppers (Unipolar, Bipolar, Universal)

UNIPOLAR STEPPERS

These steppers have a similar stator arrangement as the variable-reluctance steppers, but they use a permanent-magnet rotor and different internal wiring arrangements. Figure 14.8 shows a 30° per step unipolar stepper. It consists of a four-pole (or two-coil pair) stator with center taps between coil pairs and a six-toothed permanent-magnetic rotor. The center taps may be wired internally and brought out as one wire or may be brought out separately as two wires. The center taps typically are wired to the positive supply voltage, whereas the two free ends of a coil pair are alternately grounded to reverse the direction of the field provided by that winding. As shown in the figure, when current flows from the center tap of winding 1 out terminal 1a, the top stator pole "goes north," while the bottom stator pole "goes south." This causes the rotor to snap into position. If the current through winding 1 is removed, sent through winding 2, and out terminal 2a, the horizontal poles will become energized, causing the rotor to turn 30°, or one step. In Fig. 14.8, three firing sequences are shown. The first sequence provides full stepping action (what I just discussed). The second sequence, referred to as the *power stepping sequence*, provides full stepping action with 1.4 times the torque but twice the power consumption. The third sequence provides half stepping (e.g., 15° instead of the rated 30°). Half stepping is made possible by

energizing adjacent poles at the same time. This pulls the rotor in-between the poles, thus resulting in one-half the stepping angle. As a final note, unipolar steppers with higher angular resolutions are constructed with more rotor teeth. Also, unipolars come in either five-or six-wire types. The five-wire type has the center taps joined internally, while the six-wire type does not.

BIPOLAR STEPPERS

These steppers resemble unipolar steppers, but their coil pairs do not have center taps. This means that instead of simply supplying a fixed supply voltage to a lead, as was the case in unipolar steppers (supply voltage was fixed to center taps), the supply voltage must be alternately applied to different coil ends. At the same time, the opposite end of a coil pair must be set to the opposite polarity (ground). For example, in Fig. 14.8, a 30° per step bipolar stepper is made to rotate by applying the polarities shown in the firing sequence table to the leads of the stepper. Notice that the firing sequence uses the same basic drive pattern as the unipolar stepper, but the "0" and "1" signals are replaced with "+" and "-" symbols to show that the polarity matters. As you will see in the next section, the circuitry used to drive a bipolar stepper requires an H-bridge network for every coil pair. Bipolar steppers are more difficult to control than both unipolar steppers and variable-reluctance steppers, but their unique polarity-shifting feature gives them a better size-to-torque ratio. As a final note, bipolar steppers with higher angular resolutions are constructed with more rotor teeth.

UNIVERSAL STEPPERS

These steppers represent a type of unipolar-bipolar hybrid. A universal stepper comes with four independent windings and eight leads. By connecting the coil windings in parallel, as shown in Fig. 14.8, the universal stepper can be converted into a unipolar stepper. If the coil windings are connected in series, the stepper can be converted into a bipolar stepper.

14.7 Driving Stepper Motors

Every stepper motor needs a driver circuit that can control the current flow sent through the coils within the stepper's stator. The driver, in turn, must be controlled by a logic circuit referred to as a *translator*. I will discuss translator circuits after I have covered the driver circuits.

Figure 14.9 shows driver networks for a variable-reluctance stepper and for a unipolar stepper. Both drivers use transistors to control current flow through the motor's individual windings. In both driver networks, input buffer stages are added to protect the translator circuit from the motor's supply voltage in the event of transistor collector-to-base breakdown. Diodes are added to both drivers to protect the transistors and power supply from inductive kickback generated by the motor's coils. (Notice that the unipolar driver uses extra diodes because inductive kickback can leak out on either side of the center tap. As you will see in a moment, a pair of diodes within this driver can be replaced with a single diode, keeping the diode count to four.) The single driver section shown in Fig. 14.9 provides a general idea of what kinds of components can be used within the driver networks. This circuit uses a high-power Darlington transistor, a TTL buffer, and a reasonably fast protection diode (the extra diode should be included in the unipolar circuit). If you do not want to bother with discrete components, transistor-array ICs, such as the ULN200x series by Allegro Microsystems or the DS200x series by National Semiconductor, can be used to construct the driver section. The ULN2003, shown in Fig. 14.9, is a TTL-compatible chip that contains seven Darlington transistors with protection diodes included. The 7407 buffer IC can be used with the ULN2003 to construct a full-stepper driver. Other ICs, such as Motorola's MC1414 Darlington array IC, can drive multiple motor winding directly from logic inputs.

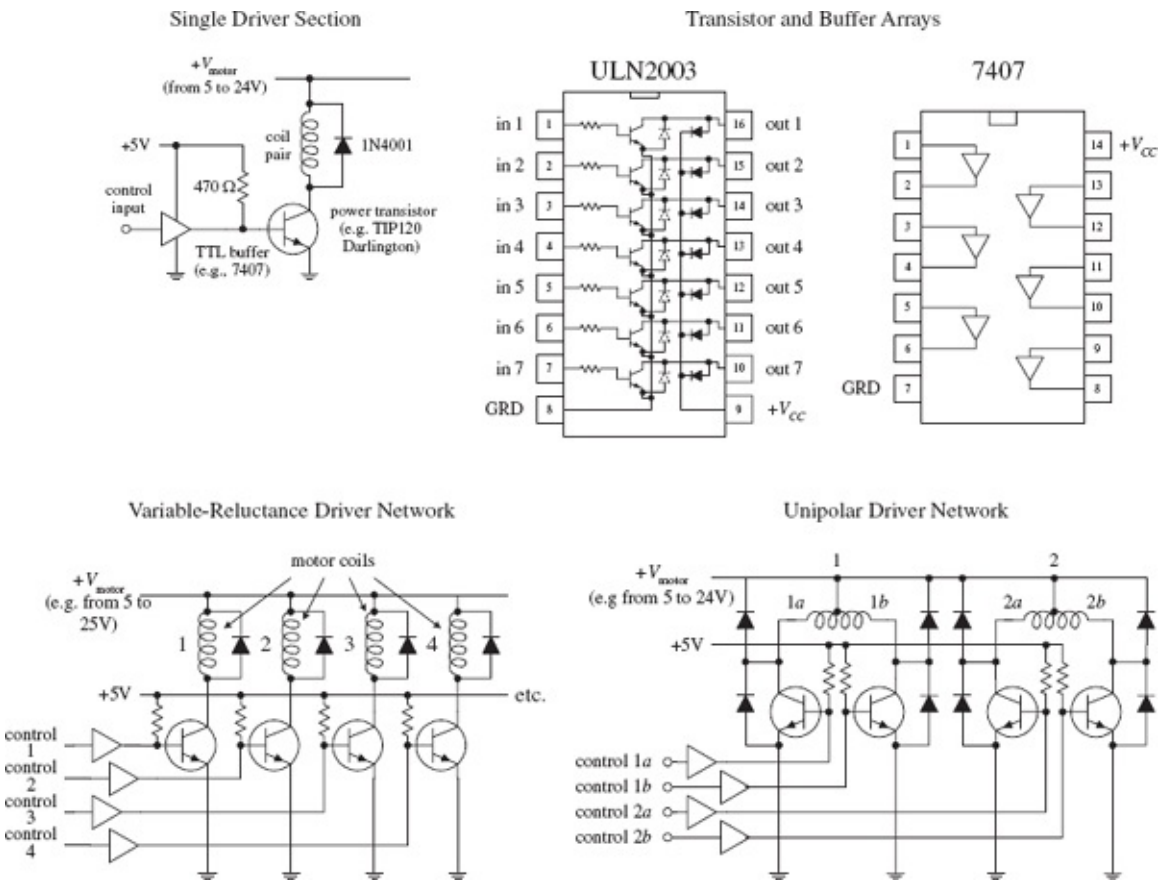
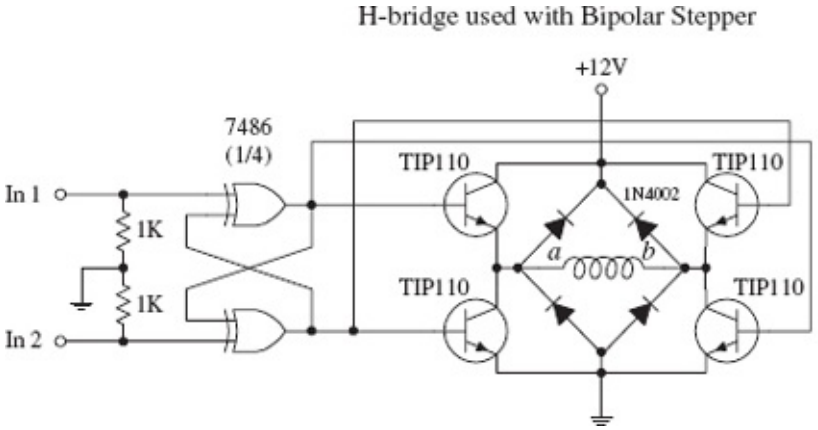


FIGURE 14.9

The circuitry used to drive a bipolar stepper requires the use of an H-bridge circuit. The H-bridge circuit acts to reverse the polarity applied across a given coil pair within the stepper. (Refer back to the section on dc direction control for details on how H-bridges work.) For each coil pair within a stepper, a separate H-bridge is needed. The H-bridge circuit shown in Fig. 14.10 uses four power Darlington transistors that are protected from the coil's inductive kickback by diodes. An XOR logic circuit is added to the input to prevent two high (1's) signals from being applied to the inputs at the same time. [If two high signals are placed at both inputs (assuming that there is no logic circuit present), the supply will short to ground. This is not good for the supply.] The table in Fig. 14.10 provides the proper firing sequence needed to create the desired polarities.



Digital input		Polarity	
In 1	In 2	*a* end	*b* end
0	0	0	0
0	1	-	+
1	0	+	-
1	1	0	0

FIGURE 14.10

As mentioned in the dc motor section of this chapter, H-bridges can be purchased in IC form. SGS Thompson's L293 dual H-bridge IC is a popular choice for driving small bipolar steppers drawing up to 1 A per motor winding at up to 36 V. The L298 dual H-bridge is similar to the L293 but can handle up to 2 A per winding. National Semiconductor's LMD18200 H-bridge IC can handle up to 3 A, and unlike the

L293 and L298, it has protection diodes built in. More H-bridge ICs are available, so check the catalogs.

14.8 Controlling the Driver with a Translator

A translator is a circuit that enacts the sequencing pulses used to drive a driver. In some instances, the translator may simply be a computer or programmable interface controller, with software directly generating the outputs needed to control the driver leads. In most cases, the translator is a special IC that is designed to provide the proper firing sequences from its output leads when a clock signal is applied to one of its input leads; another input signal may control the direction of the firing sequence (the direction of the motor). There are a number of stepper translator ICs available that are easy to use and fairly inexpensive. Let's take a look at one of these devices in a second. First, let's take a look at some simple translator circuits that can be built from simple digital components.

A simple way to generate a four-phase drive pattern is to use a CMOS 4017 decade counter/divider IC (or a 74194 TTL version). This device sequentially makes 1 of 10 possible outputs high (others stay low) in response to clock pulses. Tying the fifth output (Q_4) to ground makes the decade counter into a quad counter. To enact the drive sequence, a clock signal is applied to the clock input (see Fig. 14.11). Another four-phase translator circuit that provides power stepping control as well as direction control can be constructed with a CMOS 4027 dual JK flip-flop IC (or a 7476 TTL version). The CMOS 4070 XOR logic (or 7486 TTL XOR logic) is used to set up directional control.

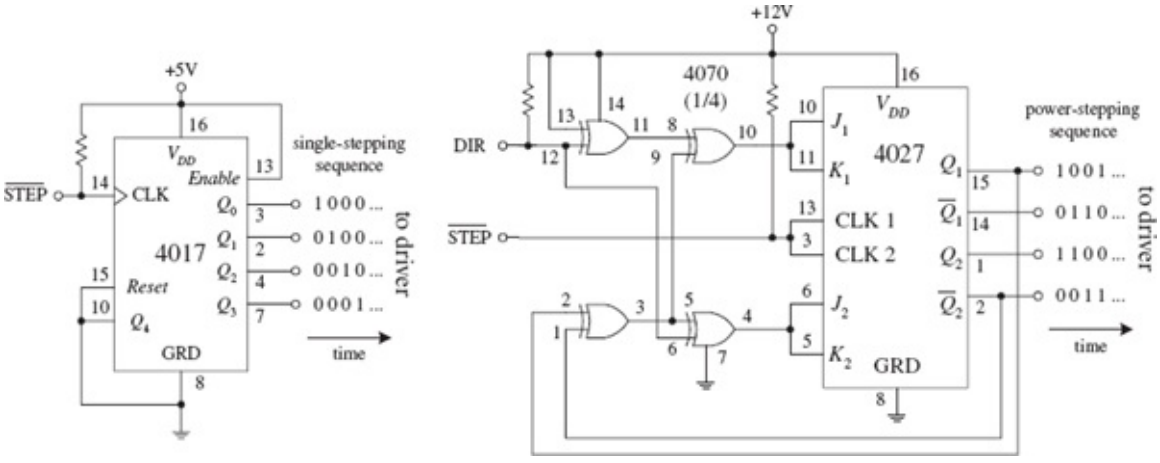


FIGURE 14.11

Figure 14.12 shows a circuit that contains the translator, driver, and stepper all in one. The motor, in this case, is a unipolar stepper, while the translator is a TTL 74194 shift counter. The 555 timer provides clock signals to the 74194, while the DPDT switch acts to control the direction of the motor. The speed of the motor is dependent on the frequency of the clock, which in turn is dependent on R_1's resistance. The translator in this circuit also can be used to control a variable reluctance stepper. Simply use the variable-reluctance driver from Fig. 14.9 and the firing sequence shown in Fig. 14.8 as your guides.

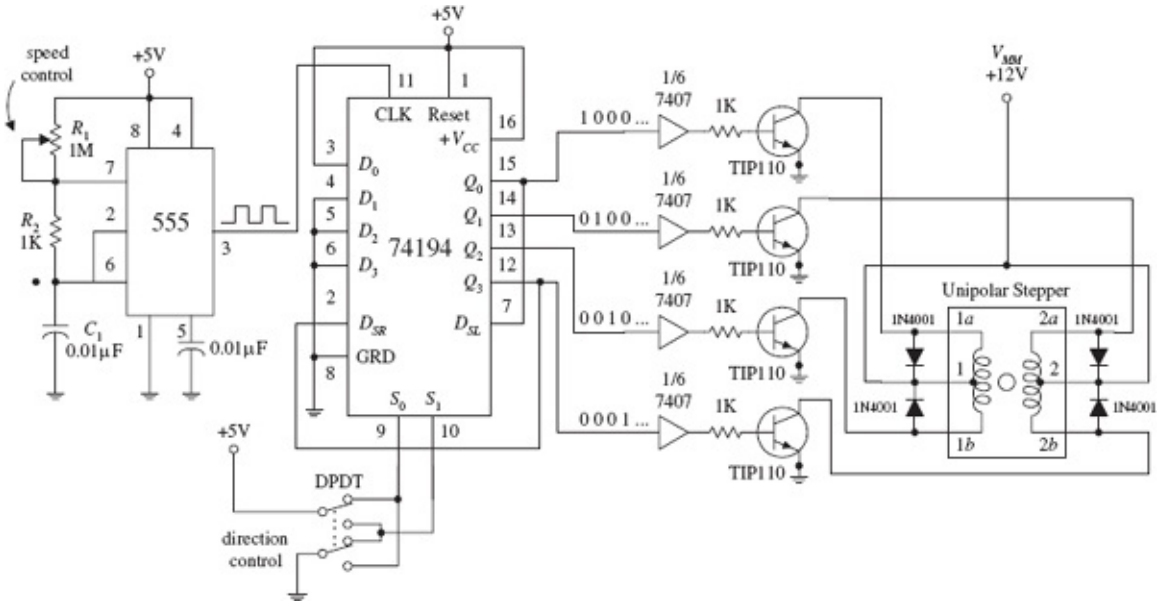
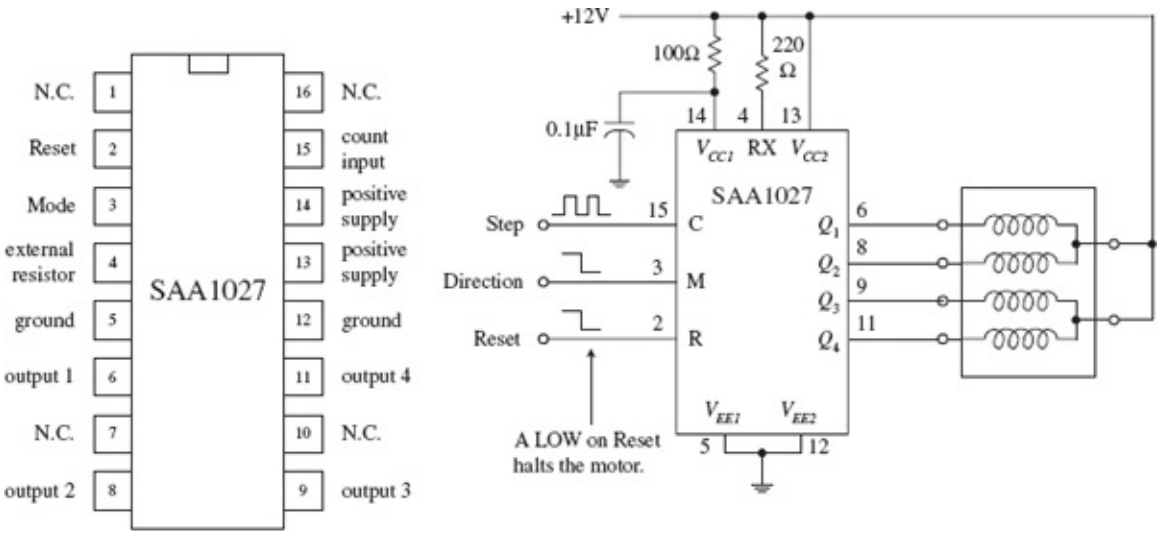


FIGURE 14.12

Perhaps the best translator circuits you can hope for come in integrated packages. A number of manufacturers produce stepper motor controller ICs that house both the translator and driver sections. These chips are fairly simple to use and inexpensive. A classic stepper controller chip is the Philips SAA1027. The SAA1027 is a bipolar IC that is designed to drive four-phase steppers. It consists of a bidirectional four-state counter and a code converter that are used to drive four outputs in sequence. This chip has high-noise-immunity inputs, clockwise and counterclockwise capability, a reset control input, high output current, and output voltage protection. Its supply voltage runs from 9.5 to 18 V, and it accepts input voltages of 7.5 V minimum for high (1) and 4.5 V maximum for low (0). It has a maximum output current of 500 mA. Figure 14.13 will paint the rest of the picture.



M = 0				M = 1			
Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
0	1	0	1	0	1	0	1
1	0	0	1	0	1	1	0
1	0	1	0	1	0	1	0
0	1	1	0	1	0	0	1
0	1	0	1	0	1	0	1

FIGURE 14.13

***Count input C* (*pin 15*)—A low-to-high transition at this pin causes the outputs to change states.**

***Mode input M* (*pin 3*)—Controls the direction of the motor. See table to the left.**

***Reset input R* (*pin 2*)—A low (0) at the *R* input resets the counter to zero. The outputs take on the levels shown in the upper and lower line of the table to the left.**

***External resistor RX* (*pin 4*)—An external resistor connected to the RX terminal sets the base current of the transistor drivers. Its value is based on the required output current.**

***Outputs Q_1 through Q_4* (*pins 6, 8, 9, 11*)—Output terminals that are connected to the stepper motor.**

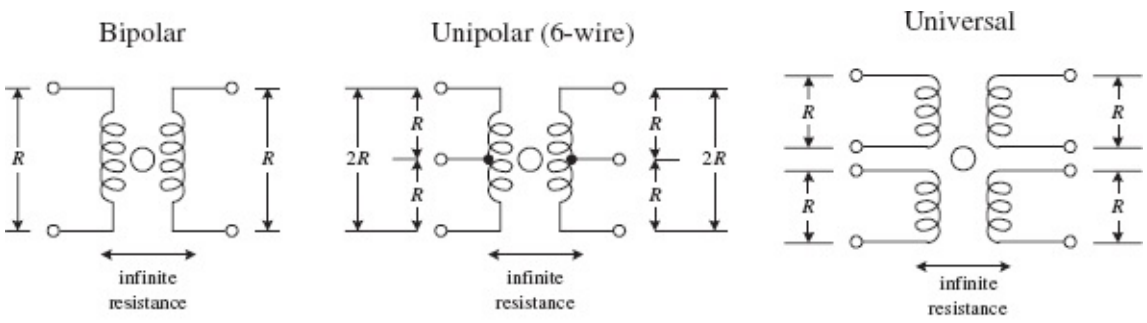
As mentioned, the SAA1027 is a classic chip (old chip). Newer, better stepper control ICs are available from a number of manufacturers. If you are interested in learning more about these chips, try searching the Internet. You will find some useful websites that discuss stepper controller ICs in detail. Also, these websites often will provide links to manufacturers and distributors of stepper motors and controller ICs.

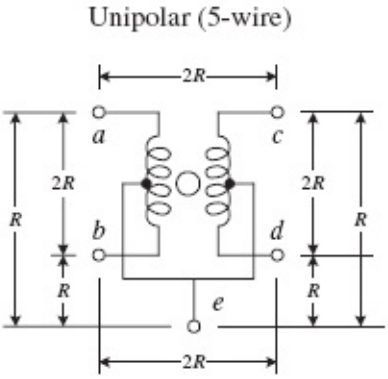
14.9 A Final Word on Identifying Stepper Motors

When it comes to identifying the characteristics of an unknown stepper, the following suggestions should help. The vast majority of the steppers on the market today are unipolar, bipolar, or universal types. Based on this, you can guess that if your stepper has four leads, it is most likely a bipolar stepper. If the stepper has five leads, then the motor is most likely a unipolar with common center taps. If the stepper has six leads, it is probably a unipolar with separate center taps. A motor with eight leads would most likely be a universal stepper. (If you think your motor might be a variable-reluctance stepper, try spinning the shaft. If the shaft spins freely, the motor is most likely a variable-reluctance stepper. A coglike resistance indicates that the stepper is a permanent-magnet type.)

Once you have determined what kind of stepper you have, the next step is to determine which leads are which. A simple way to figure this out is to test the resistance between various leads with an ohmmeter.

Decoding the leads of a bipolar stepper is easy. Simply use an ohmmeter to determine which wire pair yields a low resistance value. A low resistance indicates that the two wires are ends of the same winding. If the two wires are not part of the same winding, the resistance will be infinite. A universal stepper can be decoded using a similar approach. Decoding a six-wire unipolar stepper requires isolating two three-wire pairs. From there, you figure out which wire is the common center tap by noticing which measured pair among the isolated three wires gives a unit R worth of resistance and which pair gives a unit of $2R$ worth of resistance (see Fig. 14.14). Now, decoding a five-wire unipolar (with common center tap) is a bit more tricky than the others because of the common, but hidden, center tap. To help decode this stepper, you can use the diagram and table shown in Fig. 14.14. (The dots within the table represent where the ohmmeter's two probes are placed within the diagram.) With the table you isolate e (common tap wire) by noting when the ohmmeter gives a resistance of R units. Next, you determine which of the two wires in your hand is actually e by testing one of the two with the rest of the wires. If you always get R, then you are holding e, but if you get $2R$, you are not holding e. Once the e wire is determined, any more ohmmeter deducing does not work—at least in theory—because you will always get $2R$. The best bet now is to connect the motor to the driver circuitry and see if the stepper steps. If it does not step, fiddle around with the wires until it does.





a	*b*	*c*	*d*	*e*	Resistance
•	•				2*R*
•		•			2*R*
•			•		2*R*
•				•	*R*
	•	•			2*R*
	•		•		2*R*
	•			•	*R*
		•	•		2*R*
		•		•	*R*
			•	•	*R*

FIGURE 14.14

CHAPTER 15

Audio Electronics

Audio electronics, in part, deals with converting sound signals into electrical signals. This conversion process typically is accomplished by means of a microphone. Once the sound is converted, what is done with the corresponding electrical signal is up to you. For example, you can amplify the signal, filter out certain frequencies from the signal, combine (mix) the signal with other signals, transform the signal into a digitally encoded signal that can be stored in memory, modulate the signal for the purpose of radiowave transmission, use the signal to trigger a switch (e.g., transistor or relay), etc.

Another aspect of audio electronics deals with generating sound signals from electrical signals. To convert electrical signals into sound signals, you can use a speaker. (If you are not interested in retaining frequency response—say, you are only interested in making a warning alarm—you can use an audible sound device such as a dc buzzer or compression washer.) The electrical signals used to drive a speaker may be sound-generated in origin or may be artificially generated by special oscillator circuits.

15.1 A Little Lecture on Sound

Before you start dealing with audio-related circuits, it is worthwhile reviewing some of the basic concepts of sound. Sound consists of three basic elements: *frequency, intensity* (*loudness*), and *timbre* (*overtones*).

The frequency of a sound corresponds to the vibrating frequency of the object that produced the sound. In terms of human physiology, the human ear can perceive frequencies from around 20 to 20,000 Hz; however, the ear is most sensitive to frequencies between 1000 and 2000 Hz.

The intensity of a sound corresponds to the amount of sound energy transported across a unit area per second (or W/m^2) and depends on the amplitude of oscillation of the vibrating object. As you move further away from the vibrating object, the intensity drops in proportion to one over the distance squared. The human ear can perceive an incredible range of intensities, from 10^{-12} to 1 W/m^2. Because this range is so extensive, it is usually more convenient to use a logarithmic scale to describe intensity. For this purpose, decibels are used. When using decibels, sound intensity is defined as $dB = 10 \log_{10} (I/I_0)$, where I is the measured intensity in watts per meter, and $I_0 = 10^{-12}$ W/m^2 is defined as the smallest intensity that is perceived as sound by humans. In terms of decibels, the audio intensity range for humans is between 0 and 120 dB. Figure 15.1 shows a number of sounds, along with their frequency and intensity ranges.

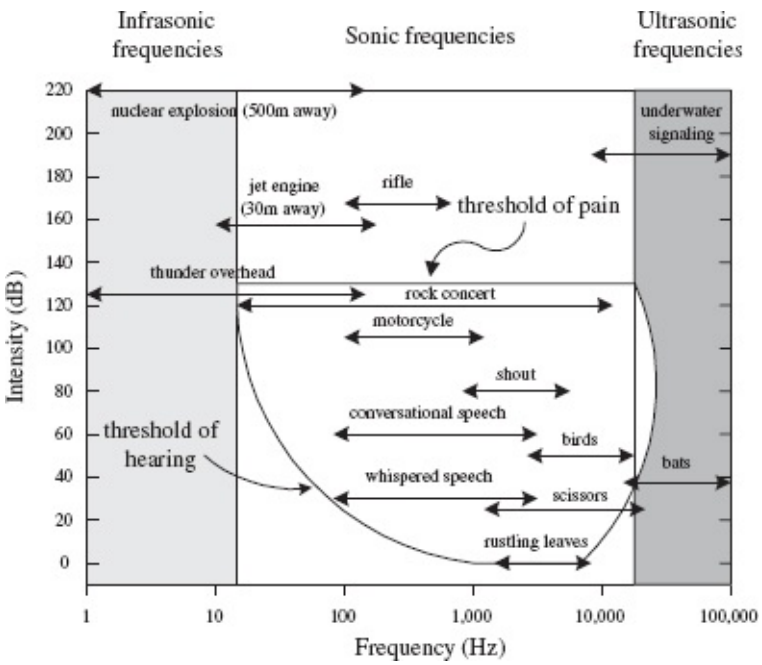


FIGURE 15.1

Tonal quality, or timbre, represents the complex wave pattern that is generated when the overtones of an instrument, voice, etc., are present along with the fundamental frequency. To demonstrate what overtones mean, consider a simple tuning fork that has a resonant frequency of 261.6 Hz (middle C). If you treat the fork as an ideal vibrator, when it is hit, it will vibrate off soundwaves with a frequency of 261.6 Hz. In this case, you have no overtones—you only get one frequency. But now, if you play middle C on a violin, you get an intensely sounding 261.1 Hz, along with a number of other higher, typically less intense frequencies called *overtones* (or *harmonics*). The most intense frequency sounded is typically referred to as the *fundamental frequency*. The overtones of importance have frequencies that are integer multiples of the fundamental frequency (e.g., 2 × 261.1 Hz is the first harmonic, 3 × 261.1 Hz is the second harmonic, and n × 261.1 Hz is the nth harmonic). It is the specific intensity of each overtone within the harmonic spectrum of an instrument, voice, etc., that is largely responsible for giving the instrument, voice, etc. its unique tonal quality. (The reason for an instrument's unique set of overtones depends on the construction of the instrument.) Figure 15.2*a* shows a harmonic spectrum (spectral plot) for an oboe that is tuned to middle C—the fundamental frequency.

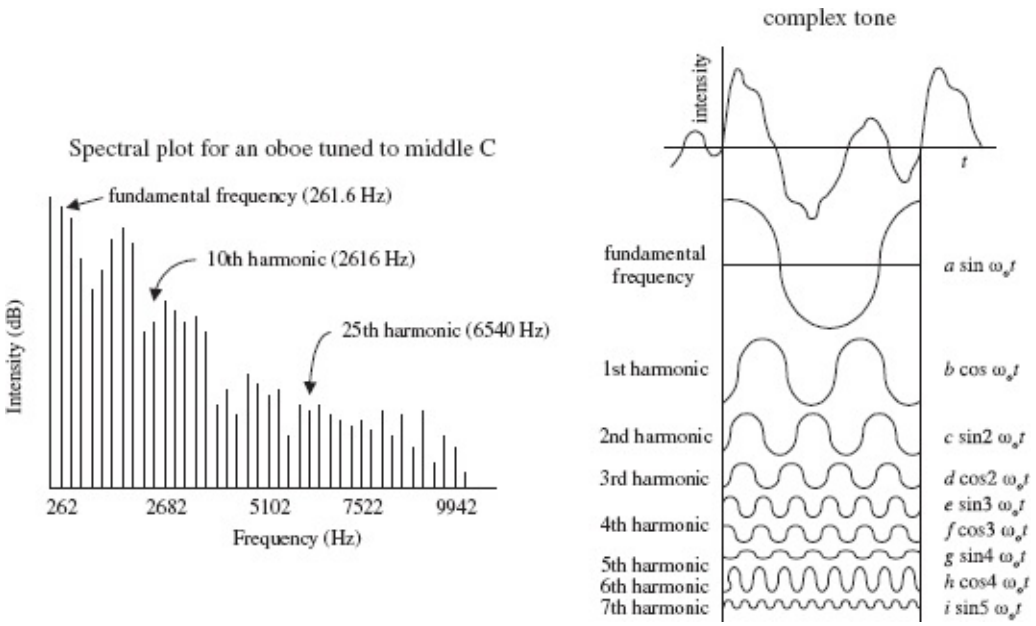


FIGURE 15.2

In theory, you can create the sound from any type of instrument (e.g., violin, tuba, banjo, etc.) by examining the harmonic spectrum of that instrument. To illustrate how this can be done, pretend that you have a number of ideal tuning forks. One fork represents the fundamental frequency; the other forks represent the various overtone frequencies. Using the harmonic spectrum of an instrument as a guide, you can mimic the sound of the instrument by varying the intensity of each "overtone fork." (In reality, to accurately mimic an instrument, you also must consider rise and decay times of certain overtones—controlling the intensities of the overtones is not enough.) Mathematically, you can express a complex sound as the sum of all its overtones:

$$\text{Signal} = a \sin \omega_0 t + b \cos \omega_0 t + c \sin 2\omega_0 t + d \cos 2\omega_0 t + e \sin 3\omega_0 t + f \cos \omega_0 t + \ldots$$

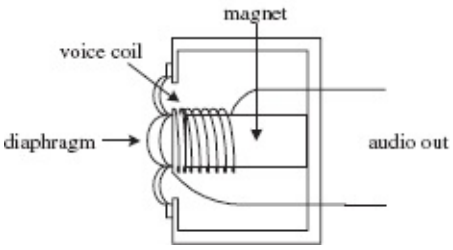
The coefficients *a*, *b*, *c*, *d*, etc., are the intensities of the overtones, and the fundamental frequency $f_0 = \omega_0/2\pi$. This expression is referred to as a *Fourier series*. The coefficients must be calculated from the given waveform or the data from which it is plotted—although an instrument called a *harmonic analyzer* can automatically compute the coefficients. Figure 15.2*b* shows a complex sound made up of seven of its harmonics.

The art of synthesizing sounds via electric circuits is fairly complex business. To accurately mimic an instrumental sound, train whistle, bird chirp, etc., you must design circuits that can generate complex waveforms that contain all the overtones and decay and rise time information. For this purpose, special oscillator and modulator circuits are needed.

15.2 Microphones

A microphone converts variations in sound pressure into corresponding variations in electric current. The amplitude of the ac voltage generated by a microphone is proportional to the intensity of the sound, while the frequency of the ac voltage corresponds to the frequency of the sound. (Note that if overtones are present within the sound signal, these overtones also will be present in the electrical signal.) Three commonly used microphones are listed next.

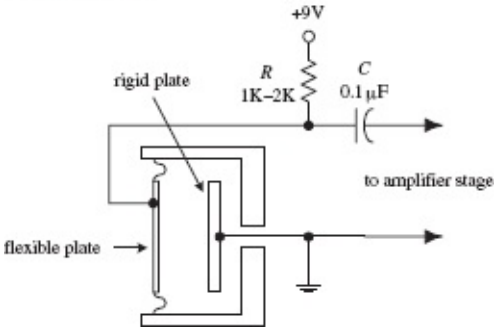
Dynamic



This type of microphone consists of a plastic diaphragm, voice coil, and a permanent magnet. The diaphragm is connected to one end of the voice coil, while the other end of the coil is loosely supported around (or within) the magnet. When an alternating pressure is applied to the diaphragm, the voice coil alternates in response. Since the voice coil is accelerating through the magnet's magnetic field, an induced voltage is set up across the leads of the voice coil. You can use this voltage to power a very small load, or you can use an amplifier to increase the strength of the signal so as to drive a larger load. Dynamic microphones are extremely rugged, provide smooth and extended frequency response, do not require an external dc source to drive them, perform well over a wide range of temperatures, and have a low impedance output. Some dynamic microphones house internal transformers within their bodies, which give them the ability to have either a high- or low-impedance output—a switch is used to select between the two. Dynamic microphones are widely used in public address, hi-fi, and recording applications.

FIGURE 15.3

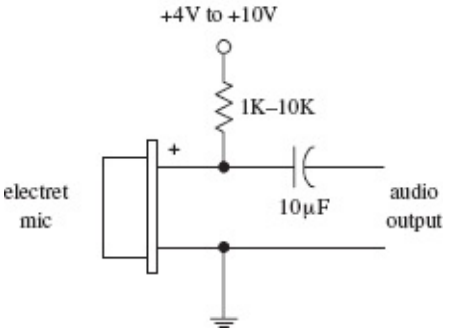
Condenser



This type of microphone consists of a pair of charged plates that can be forced closer or further apart by variations in air pressure. In effect, the plates act like a sound-sensitive capacitor. One plate is made of a rigid metal that is fixed in place and grounded. The other plate is made of a flexible metal or metal-coiled plastic that is positively charged by means of an external voltage source. A very low-noise, high-impedance amplifier is required to operate this type of microphone and to provide low output impedance. Condenser microphones offer crisp, low-noise sound and are used for high-quality sound recording.

FIGURE 15.4

Electret



An electret microphone is a variation of the condenser microphone. Instead of requiring an external voltage source to charge the diaphragm, it uses a permanently charged plastic element (electret) placed in parallel with a conductive metal backplate. Most electret microphones have a small FET amplifier built into their cases. This amplifier requires power to operate—typically a voltage between +1.5 and +10 V is needed. This voltage is fed into the microphone through a resistor (1–10 K) (see figure). Electret microphones used to suffer from poor performance, but modern designs can achieve results comparable to those of a condenser microphone.

FIGURE 15.5

15.3 Microphone Specifications

A microphone's *sensitivity* represents the ratio of electrical output (voltage) to the intensity of sound input. Sensitivity is often expressed in decibels with respect to a reference sound pressure of 1 dyn/cm^2.

The *frequency response* of a microphone is a measure of the microphone's ability to convert different acoustical frequencies into ac voltages. For speech, the frequency response of a microphone need only cover a range from around 100 to 3000 Hz. However, for hi-fi applications, the frequency response of the microphone must cover a wider range, from around 20 to 20,000 Hz.

The *directivity characteristic* of a microphone refers to how well the microphone responds to sound coming from different directions. *Omnidirectional* microphones respond equally well in all directions, whereas *directional* microphones respond well only in specific directions.

The *impedance* of a microphone represents how much the microphone resists the flow of an ac signal. Low-impedance microphones are classified as having an impedance of less than 600 Ω. Medium-impedance microphones range from 600 to 10,000 Ω, whereas high-impedance microphones extend above 10,000 Ω. In modern audio systems, it is desirable to connect a lower-impedance microphone to a higher-impedance input device (e.g., 50-Ω microphone to 600-Ω mixer), but it is undesirable to connect a high-impedance microphone to a low-impedance input. In the first case, not much signal loss will occur,

whereas in the second case, a significant amount of signal loss may occur. The standard rule of thumb is to allow the load impedance to be 10 times the source impedance. We'll take a closer look at impedance matching later on in this chapter.

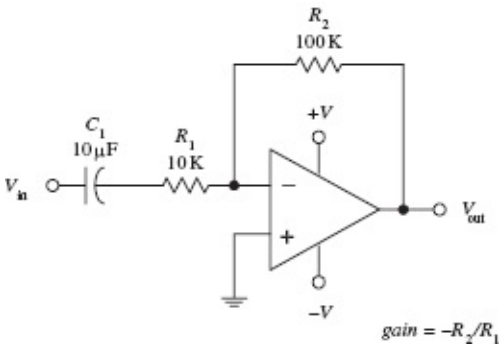
15.4 Audio Amplifiers

Electrical signals within audio circuits often require amplification to effectively drive other circuit elements or devices. Perhaps the easiest and most efficient way to amplify a signal is to use an op amp. General-purpose op amps such as the 741 will work fine for many noncritical audio applications, but they may cause distortion and other undesirable effects when audio signals get complex. A better choice for audio applications is to use an audio op amp especially designed to handle audio signals. Audio amplifiers have high slew rates, high gain-bandwidth products, high input impedances, low distortion, high voltage/power operation, and very low input noise. There are a number of good op amps produced by a number of different manufacturers. Some high-quality op amps worth mentioning include the AD842, AD847, AD845, AD797, NE5532, NE5534, NE5535, OP-27, LT1115, LM833, OPA2604, OP249, HA5112, LM4562, OPA134, OPA2134, and LT1057.

15.4.1 Inverting Amplifier

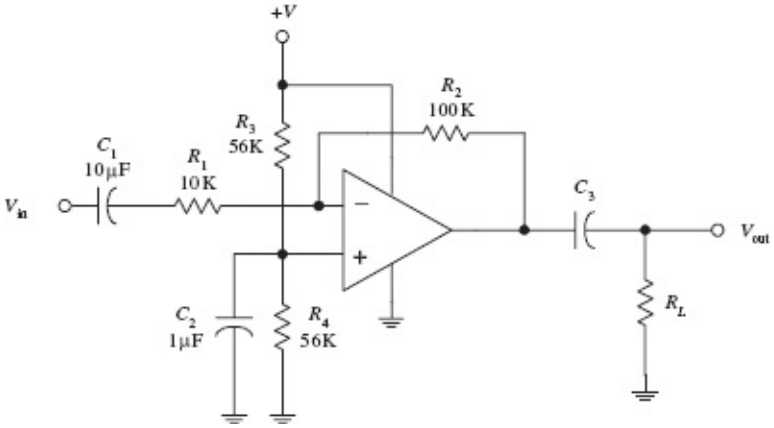
The following two circuits act as inverting amplifiers. The gain for both circuits is determined by $-R_2/R_1$ (see Chap. 8 for the theory), while the input impedance is approximately equal to R_1. The first op amp circuit uses a dual power supply, while the second op amp circuit uses a single power supply.

Inverting amplifier (dual power supply)



In both amplifier circuits, C_1 acts as an ac coupling capacitor—it acts to pass ac signals while preventing unwanted dc signals from passing from the previous stage. Without C_1, dc levels would be present at the op amp's output, which in turn could lead to amplifier saturation and distortion as the ac portion of the input signal is amplified. C_1 also helps prevent low-frequency noise from reaching the amplifier's input.

Inverting amplifier (single power supply)



In the single-power-supply circuit, biasing resistors R_3 and R_4 are needed to prevent the amplifier from clipping during negative swings in the audio input signals. They act to give the op amp's output a dc level on which the ac signal can safely fluctuate. Setting $R_3 = R_4$ sets the dc level of the op amp's output to 1/2 (+V). For reliable results, the biasing resistors should have resistance values between 10 and 100 k. Now, to prevent passing the dc level onto the next stage, C_3 (ac coupling capacitor) must be included. Its value should be equal to $1/(2\pi f_C R_L)$, where RL is the load resistance, and f_C is the cutoff frequency. C^2 acts as a filtering capacitor used to eliminate power-supply noise from reaching the op amp's noninverting input.

Notably, many audio op amps are especially designed for single-supply operation—they do not require biasing resistors.

FIGURE 15.6

15.4.2 Noninverting Amplifier

The preceding inverting amplifier works fine for many applications, but its input impedance is not incredibly large. To achieve a larger input impedance (useful when bridging a high-impedance source to the input of an amplifier), you can use one of the following noninverting amplifiers. The left amplifier circuit uses a dual power supply, whereas the right amplifier circuit uses a single power supply. The gain for both circuits is equal to $R_2/R_1 + 1$.

15.4.3 Digital Amplifiers

Digital power amplifiers are also known as class-D amplifiers or PWM amplifiers. They are extremely efficient and so run quite cool. This makes them ideal for very high-power amplifiers of powers in the hundreds of watts or even the kilowatts range.

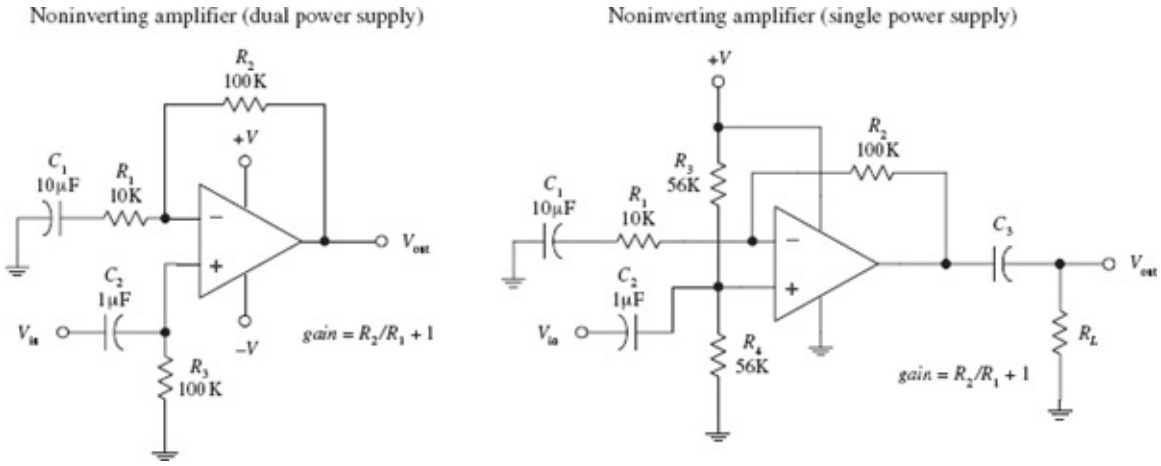


FIGURE 15.7 Components R_1, C_1, R_2, and the biasing resistors serve the same function as was seen in the inverting amplifier circuits. The noninverting input offers an exceptionally high input impedance and can be matched to the source impedance more readily by adjusting C_2 and R_3 (dual-supply circuit) or R_4 (single-supply circuit). The input impedance is approximately equal to R_3 (dual-supply circuit) or R_4 (single-supply circuit).

Figure 15.8 shows the block diagram for a class-D amplifier. The input signal is converted into a PWM signal by comparing it with a triangle wave. This is a very neat trick. As the triangle wave rises, at some point, it will become greater than the signal voltage. If the signal voltage is high, this will take longer; if the voltage is low, it will happen sooner. In this way, the pulse length produced will be proportional to the instantaneous input voltage.

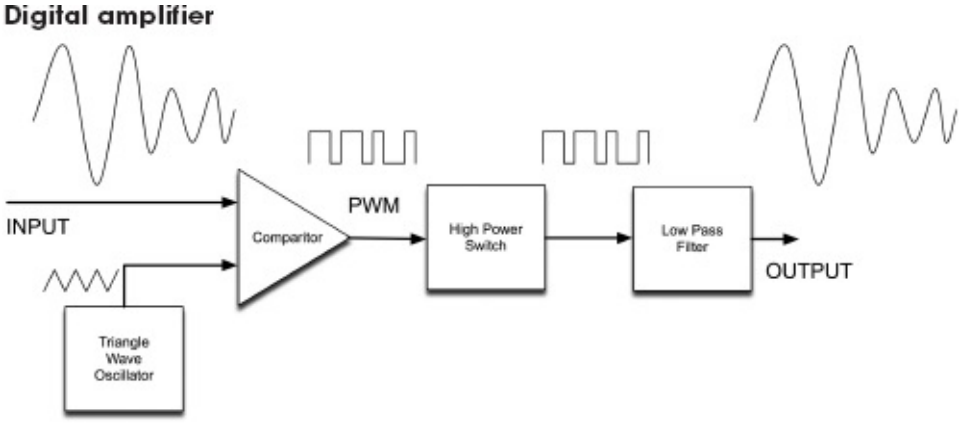


FIGURE 15.8

Figure 15.9 shows the result of simulating the comparator with a triangle waveform of 10 kHz sampling a sine wave input signal at 1 kHz. The squarewave is the PWM output.

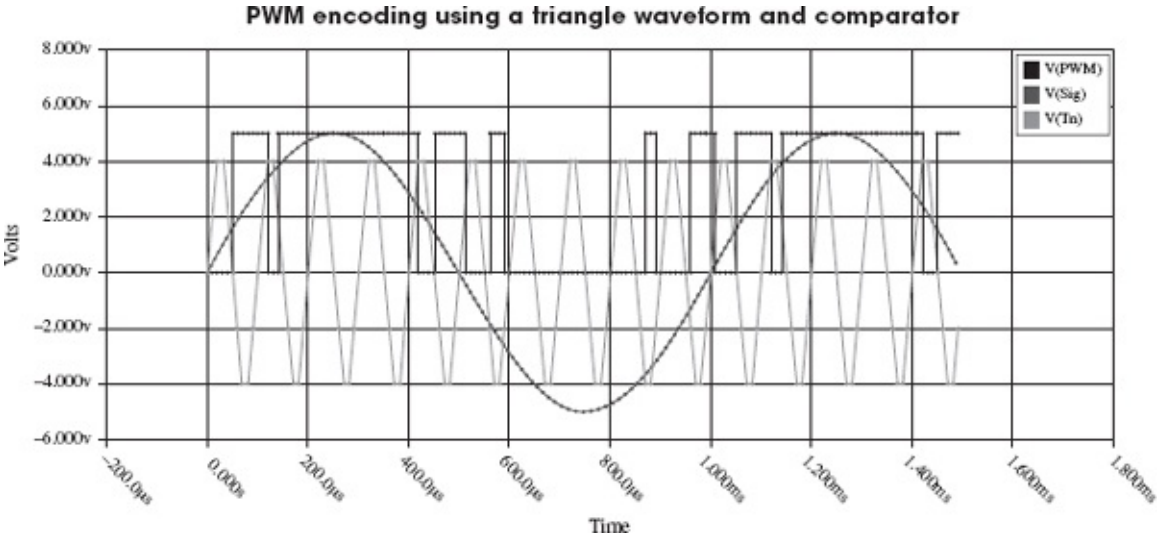


FIGURE 15.9

We have digitized our audio signal. It is now just on or off. So we can use that to switch as high a current as we like, using big MOSFETs in a complementary arrangement or some other switching transistor. This high-power PWM signal will be conveying the right energy at the right time, but it is being carried in the high-frequency switching squarewave. This needs low-pass filtering to remove this carrier, leaving just the original signal in amplified form.

The quality of digital amplifiers can be very variable, but their efficiency makes them an attractive proposition. There are a number of ICs available that either encapsulate the whole amplifier in one device or provide outputs suitable for driving a complementary pair of MOSFETs. A couple devices to look at are the NCP2704 and LX1720.

15.4.4 Reducing Hum in Audio Amplifiers

If your project includes an audio amplifier that you are designing yourself, then you need to consider *hum*. Domestic power lines and devices will easily induce an annoying 60 Hz hum in an audio amplifier. Some of this signal can arrive through the amplifier's power supply. Hence, design of power supplies for hi-fi amplifiers is almost as important as the designs for the amplifiers themselves. The main thing is to add as much smoothing capacitance to the power supply as possible.

Other 60 Hz signals will arrive in your audio amplifier circuit by mutual induction with wires and tracks on your PCB. You should make sure that you keep all PCB tracks and wires as short as possible.

Wires should also be screened with the screening layer grounded.

15.5 Preamplifiers

In most audio applications, the term *preamplifier* refers to a control amplifier that is used to control features such as input selection, level control, gain, and impedance levels. Here are a few simple microphone preamplifier circuits to get you started. (Note that “high Z” refers to a microphone with a high input impedance—one that is greater than ~600 Ω.)

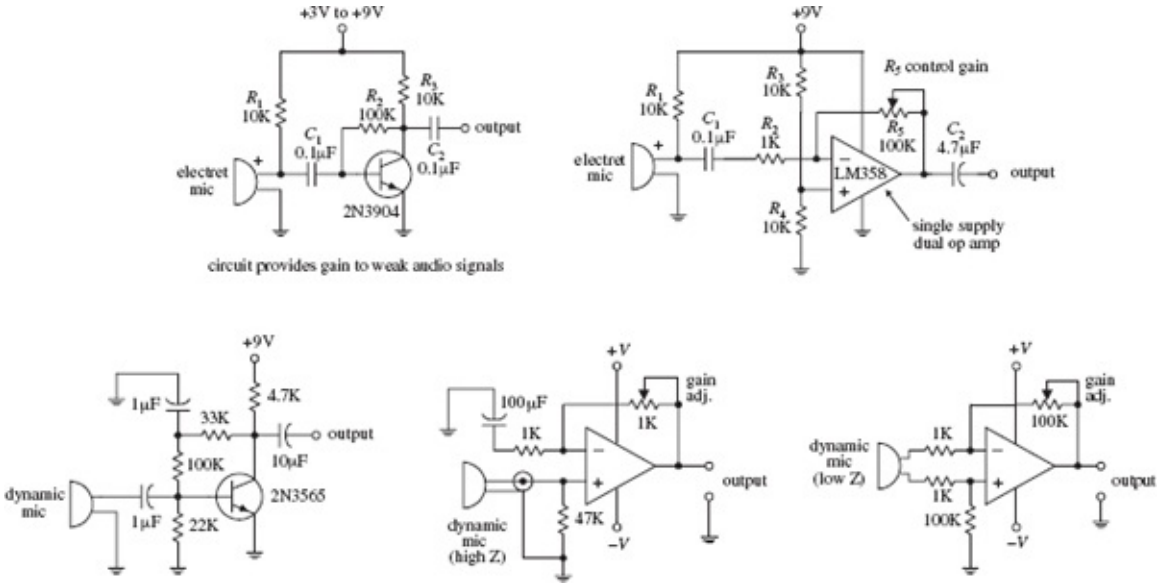


FIGURE 15.10

15.6 Mixer Circuits

Audio mixers are basically summing amplifiers—they add a number of different input signals together to form a single superimposed output signal. The two circuits below are simple audio mixer circuits. The left circuit uses a common-emitter amplifier as the summing element, while the right circuit uses an op amp. The potentiometers are used as independent input volume controls.

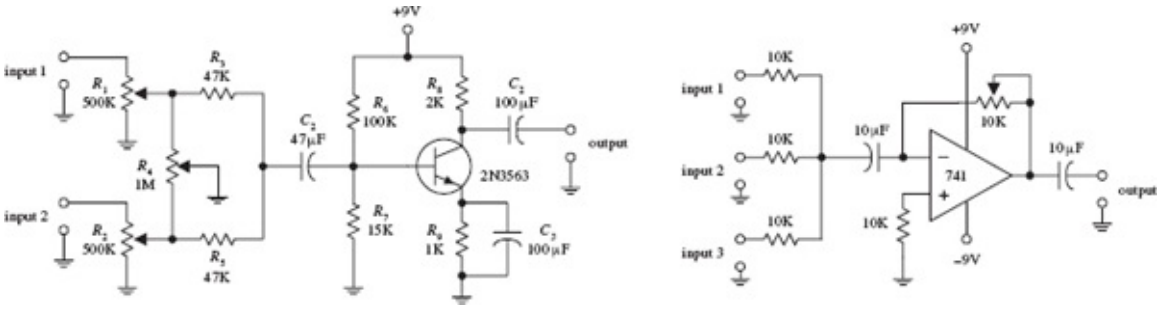


FIGURE 15.11

15.7 A Note on Impedance Matching

Is matching impedances between audio devices necessary? Not any more, at least when it comes to connecting a low-impedance source to a high-impedance load. In the era when vacuum-tube amplifiers were the standard, it was important to match impedances to achieve maximum power transfer between two devices. Impedance matching reduced the number of vacuum-tube amplifiers needed in circuit design

(e.g., the number of vacuum-tube amplifiers needed along a telephone transmission line). However, with the advent of the transistor, more efficient amplifiers were created. For these new amplifiers, what was important—and still is important—was maximum voltage transfer, not maximum power transfer. (Think of an op amp with its extremely high input impedance and low output impedance. To initiate a large-output-current response from an op amp, practically no input current is required into its input leads.) For maximum voltage transfer to occur, it was found that the destination device (called the *load*) should have an impedance of at least 10 times that of the sending device (called the *source*). This condition is referred to as *bridging*. (Without applying the bridging rule, if two audio devices with the same impedances are joined, you would see around a 6 dB worth of attenuation loss in the transmitted signal.) Bridging is the most common circuit configuration used when connecting modern audio devices. It is also applied to most other electronic source-load connections, with the exception of certain radiofrequency circuits where matching impedance is usually desired and in cases where the signal being transmitted is a current rather than a voltage. If the transmitted signal is a current, the source impedance should be larger than the load impedance.

Now, if you consider a high-impedance source connected to a low-impedance load (e.g., a high-impedance microphone connected to a low-impedance mixer), voltage transfer can result in significant signal loss. The amount of signal loss in this case would be equal to

$$\text{dB} = 20 \log_{10} \frac{R_{\text{load}}}{R_{\text{load}} + R_{\text{source}}}$$

As a rule of thumb, a loss of 6 dB or less is acceptable for most applications.

15.8 Speakers

Speakers convert electrical signals into audible signals. The most popular speaker used today is the dynamic speaker. The dynamic speaker operates on the same basic principle as a dynamic microphone. When a fluctuating current is applied through a moving coil (voice coil) that surrounds a magnet (or that is surrounded by a magnet), the coil is forced back and forth. A large paper cone attached to the coil responds to the back-and- forth motion by “drumming off” sound waves.

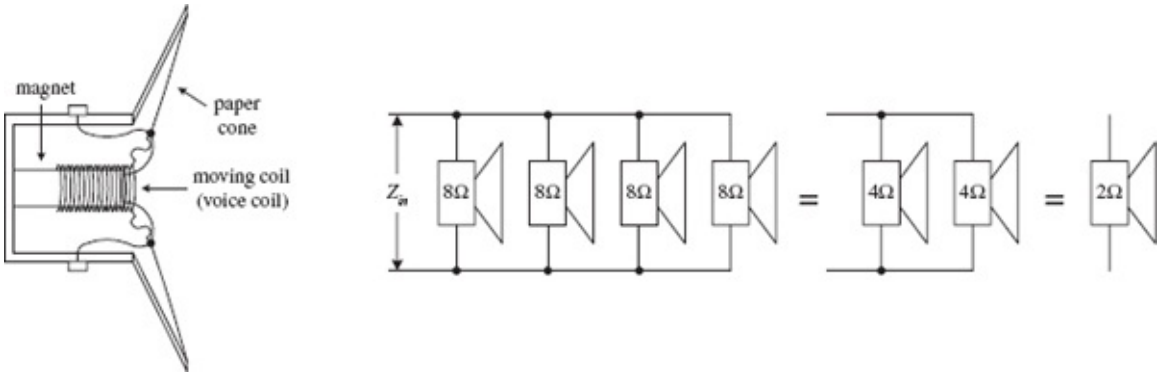


FIGURE 15.12

Every speaker is given a *nominal impedance Z* that represents the average impedance across its leads. (In reality, a speaker’s impedance varies slightly with frequency, above and below the nominal level.) In terms of applications, you can treat the speaker like a simple resistive load of impedance *Z*. For example, if you attach an 8-Ω speaker to an amplifier’s output, the amplifier will treat the speaker as an 8-Ω load. The amount of current drawn from the amplifier will be $I = V_{\text{out}}/Z_{\text{speaker}}$. However, if you replace the 8-Ω speaker with a 4-Ω speaker, the current drawn from the amplifier will double.

Driving two 8-Ω speakers in parallel is equivalent to driving a 4-Ω speaker. Driving two 4-Ω speakers in parallel is equivalent to driving a 2-Ω speaker. By using high-power resistors, it is possible to change the overall impedance sensed by the amplifier. For example, by placing a 4-Ω resistor in series with a 4-Ω speaker, you can create a load impedance of 8 Ω. However, using a series resistor to increase the impedance may hurt the sound quality. There are speaker-matching transformers that can change from 4 to 8 Ω, but a high-quality transformer like this can cost as much as a new speaker and can add slight frequency response and dynamic range errors.

Another important characteristic of a speaker is its frequency response. *The frequency response* represents the range over which a speaker can effectively vibrate off audio signals. Speakers that are designed to respond to low frequencies (typically less than 200 Hz) are referred to as *woofers*. *Midrange speakers* are designed to handle frequencies typically between 500 and around 3000 Hz. A *tweeter* is a special type of speaker (typically dome or horn type) that can handle frequencies above midrange. Some speakers are designed as full-range units that are capable of reproducing frequencies from around 100 to 15,000 Hz. A full-range speaker's sound quality tends to be inferior to a speaker system that incorporates a woofer, midrange, and tweeter speaker all together.

15.9 Crossover Networks

To design a decent speaker system, it would be best to incorporate a woofer, midrange speaker, and tweeter together so that you get good sound response over the entire audio spectrum (20 to 20,000 Hz). However, simply connecting these speakers in parallel will not work because each speaker will be receiving frequencies outside its natural frequency-response range. What you need is a filter network that can divert high-frequency signals to the tweeter, low-frequency signals to the woofer, and midrange-frequency signals to the midrange speaker. The filter network that is used for this sort of application is called a *crossover network*.

There are two types of crossovers: passive or active. Passive crossover networks consist of passive filter elements (e.g., capacitors, resistors, inductors) that are placed between the power amplifier and the speaker—they are placed inside the speaker cabinet. Passive crossover networks are cheap to make, easy to make foolproof, and can be tailored for a specific speaker. However, they are nonadjustable and always use up some amplifier power. Active crossover networks consist of a set of active filters (op amp filters) that are placed before the amplifier section. The fact that active crossover networks come before the power amplifier section makes it easier to manipulate the signal because the signal is still tiny (not amplified). Also, a single active crossover network can be used to control a number of different amplifier-speaker combinations at the same time. Since active crossover networks use active filters, the audio signal will not suffer as much attenuation loss as it would if it were applied through a passive crossover network.

Figure 15.13 shows a simple passive crossover network used to drive a three-speaker system. The graph shows typical frequency-response curves for each speaker. To produce an overall flat response from the system, you use a low-pass, bandpass, and high-pass filter. C_1 and R_t form the low-pass filter, L_1, C_1, and R_m form the bandpass filter, and L_2 and R_w form the low-pass filter (R_t, R_m, and R_w are the nominal impedances of the tweeter, midrange, and woofer speakers).

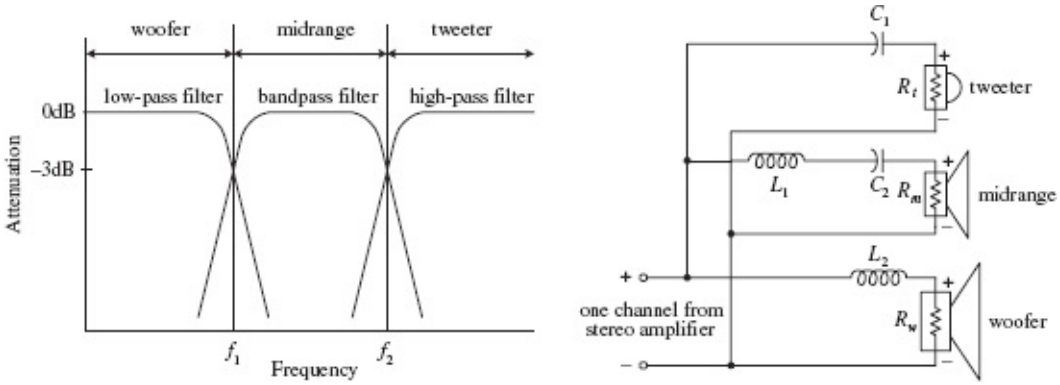


FIGURE 15.13

To determine the component values needed to get the desired response, use the following: $C_1 = 1/(2\pi f_2 R_t)$, $L_1 = R_m/(2\pi f_2)$, $C_2 = 1/(2\pi f_1 R_m)$, and $L_2 = R_w/2\pi f_1$, where f_1 and f_2 represent the 3-dB points shown in the graph. Usually, passive crossover networks are a bit more sophisticated than the one presented here. They often incorporate higher-order filters along with additional elements, such as an impedance compensation network, attenuation network, series notch filter, etc., all of which are used to achieve a flatter overall response.

Here is a more practical passive crossover network (crossover frequency 1.8 kHz) used to drive a two-speaker system consisting of an 8-Ω tweeter and an 8-Ω woofer. An 18 × 12 × 8 in fiberboard box acts as a good resonant cavity for this system.

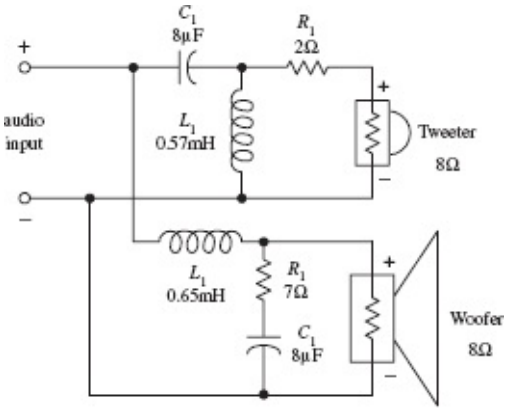


FIGURE 15.14

The following is an active crossover network that is used to drive a two-speaker system that has a crossover frequency (3-dB point) around 600 Hz and 18 dB per octave response. A LF356 high-performance op amp is used as the active element. Remember that for active filters the output signals must be amplified before they are applied to the speaker inputs.

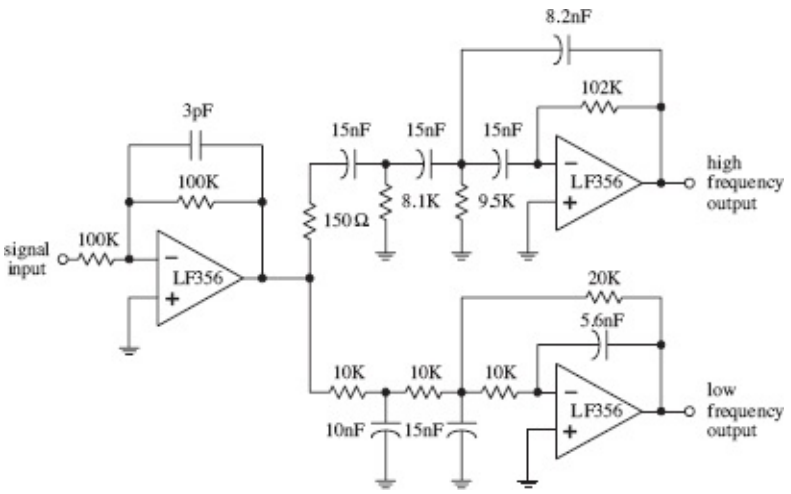
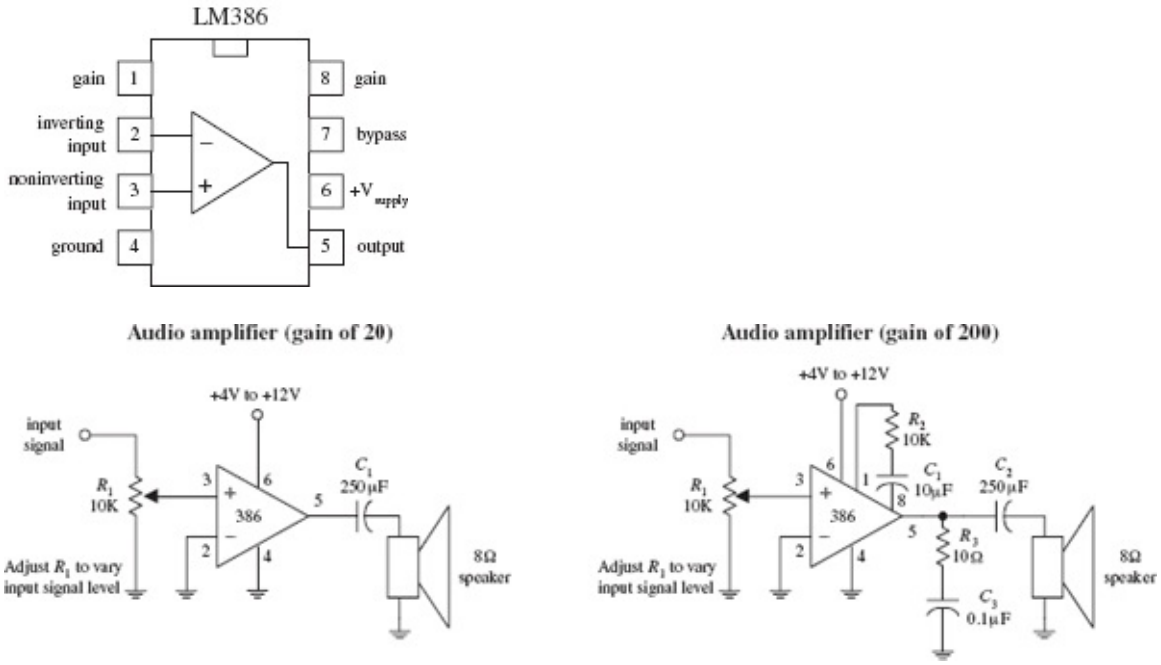


FIGURE 15.15

15.10 Simple ICs Used to Drive Speakers

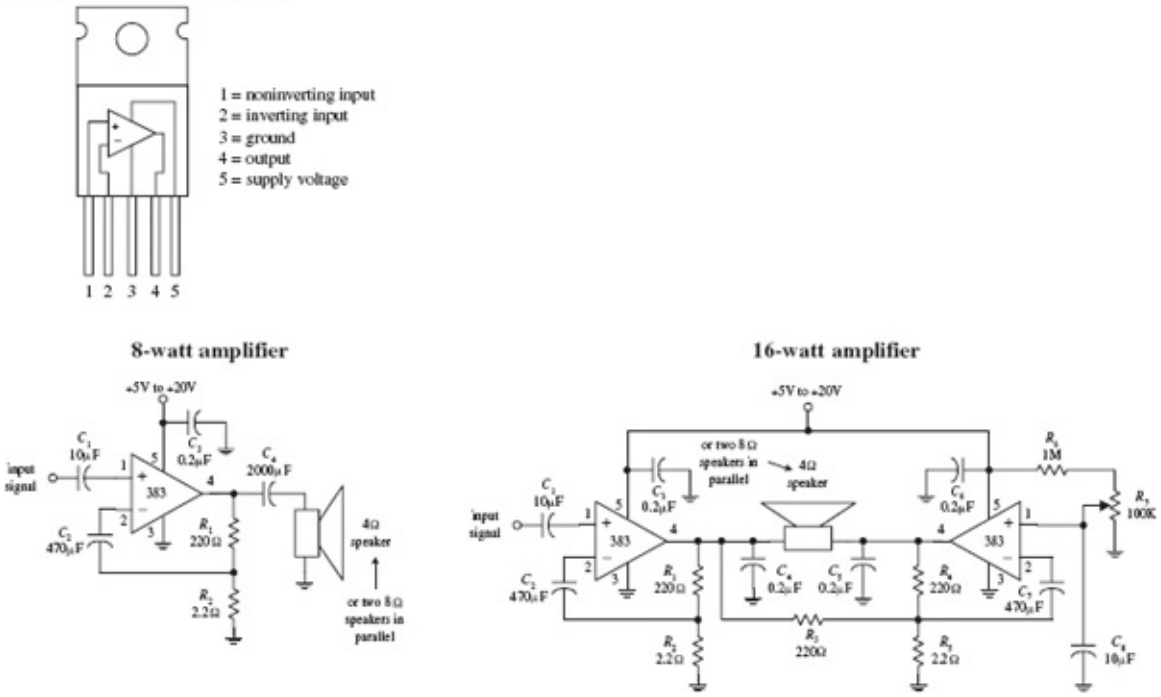
Audio Amplifier (LM386)



The LM386 audio amplifier is designed primarily for low-power applications. A +4- to +15-V supply voltage is used to power the IC. Unlike a traditional op amp, such as the 741, the 386's gain is internally fixed to 20. However, it is possible to increase the gain to 200 by connecting a resistor-capacitor network between pins 1 and 8. The 386's inputs are referenced to ground, while internal circuitry automatically biases the output signal to one-half the supply voltage. This audio amplifier is designed to drive an 8-Ω speaker.

FIGURE 15.16

Audio Amplifier (LM383)



The LM383 is a power amplifier designed to drive a 4-Ω speaker or two 8-Ω speakers in parallel. It contains thermal shutdown circuitry to protect itself from excessive loading. A heat sink is required to avoid shutdown at levels below the maximum power output.

FIGURE 15.17

15.11 Audible-Signal Devices

There are a number of unique audible-signal devices that are used as simple warning signal indicators. Some of these devices produce a continuous tone, others produce intermittent tones, and still others are capable of generating a number of different frequency tones, along with various periodic on/off cycling characteristics. Audible-signal devices come in both dc and ac types and in various shapes and sizes. Some of these devices are extremely small—no bigger than a dime. A good electronics catalog will provide a listing of audio-signal devices, along with their size, sound type, dB ratings, voltage ratings, and current-drain specifications.

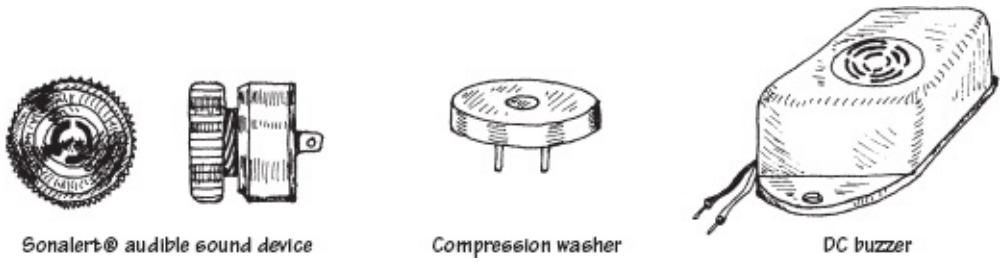


FIGURE 15.18

15.12 Miscellaneous Audio Circuits

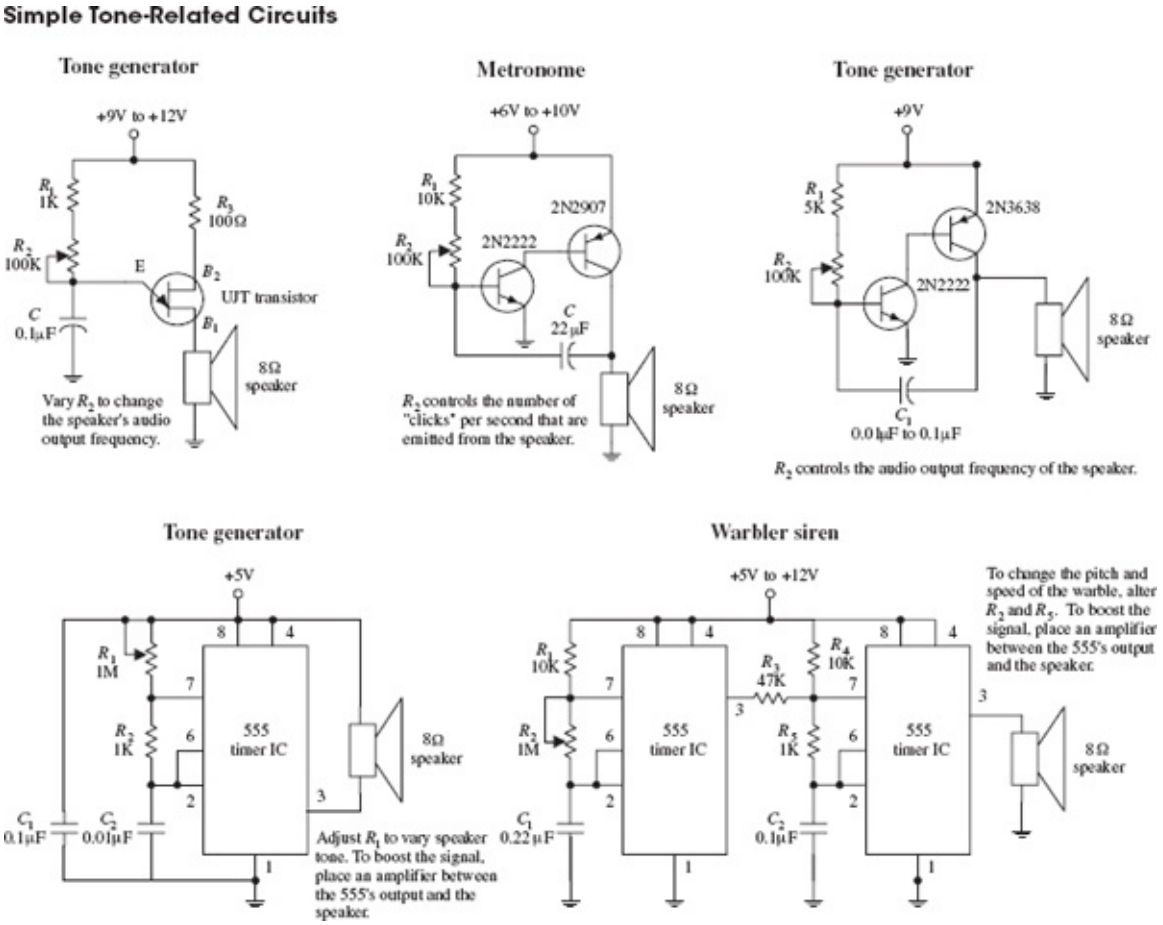
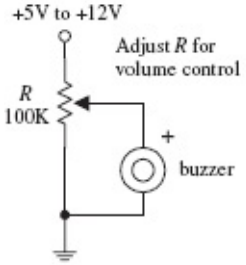


FIGURE 15.19

Simple Buzzer Circuits

Buzzer volume control



Digitally actuated buzzers

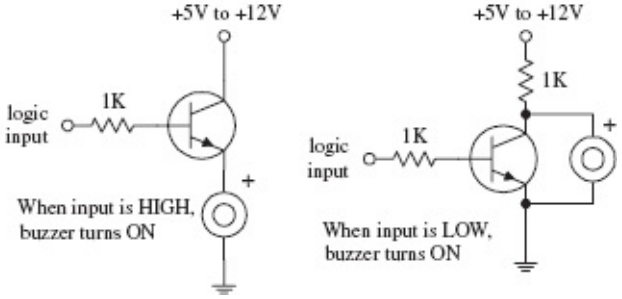


FIGURE 15.20

Megaphone

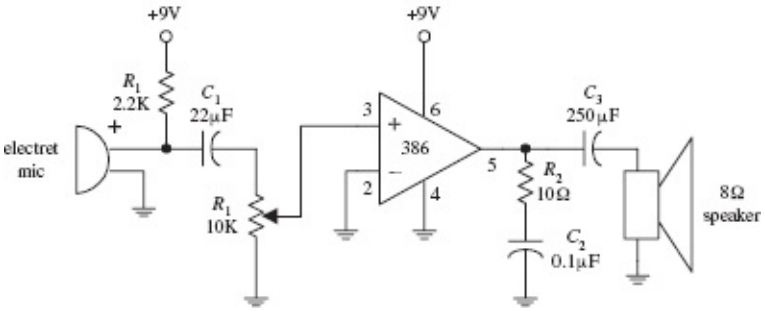


FIGURE 15.21

Sound-Activated Switch

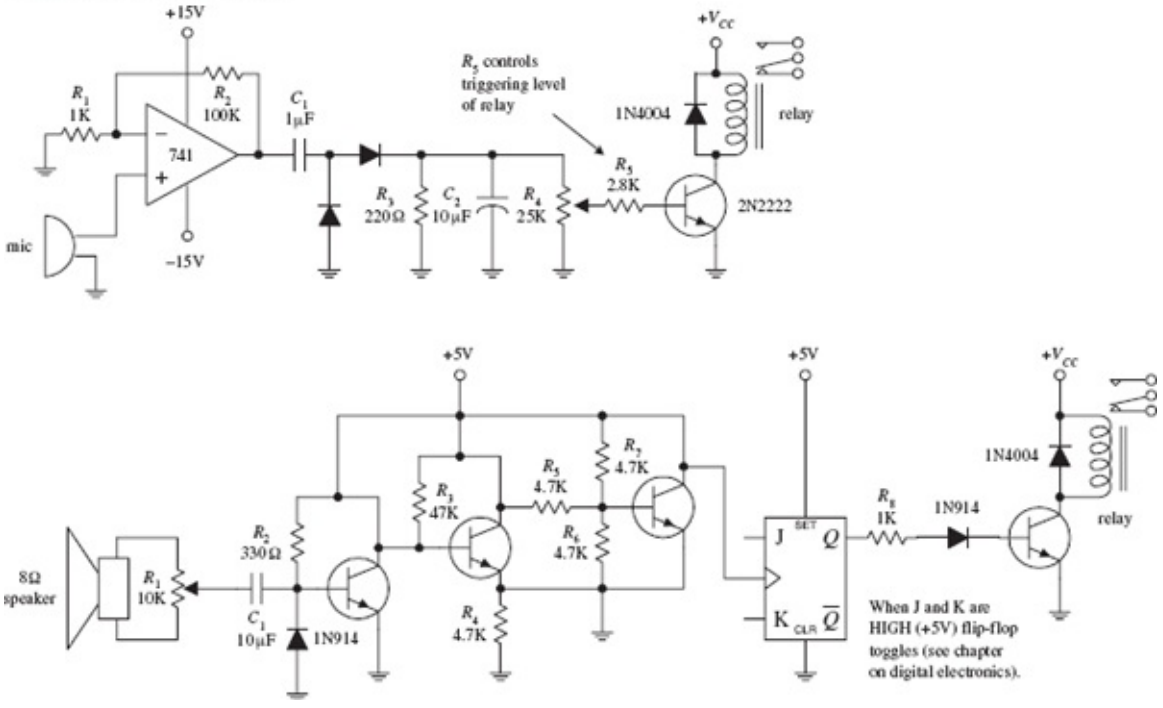


FIGURE 15.22

CHAPTER 16

Modular Electronics

Electronics has changed over the past few years, and more and more people have an invention, but don't necessarily want to learn degree-level electronics to make it a reality. Suppliers such as SparkFun, Seeed Studio, Pololu, and others support this by supplying modules and breakout boards that simplify the process of using complex devices. In addition, there are ICs for most of what you might want to do in electronics, and these will greatly simplify your project build. Also, there are whole systems, such as Arduino (see Chap. 13), .NET Gadgeteer, and Netduino, that provide plug-together modules for pretty much anything that you might want to build that has a microcontroller at its heart.

16.1 There's an IC for It

Although plenty of useful, general-purpose ICs are available, there are also some very specialized devices. Before designing anything complex from discrete devices, you should always check that you are not reinventing the wheel. There may be an IC that you can use that will reduce the component count and cost of your project.

Table 16.1 lists some of the ICs that you may find useful for your projects. Some of them are general purpose; others fill a very narrow niche. This is not intended to be any exhaustive list, but rather to provide inspiration. We have not listed part URLs. You can find places to buy components using the Octopart parts search engine (http://www.octopart.com), where you will also be able to track down data sheets.

TABLE 16.1 ICs for Electronics Projects

IC	DESCRIPTION
Audio	
HT9200	DTMF code generator, for use in applications such as phone auto-dialers; easy connection to microcontroller
ICL7611	Low-voltage, single power supply op amp
LM358	Low-cost dual op amp
RTS0072	Voice changer IC that distorts or transposes audio signals
ISD1932	Voice recorder IC
SAE800	Door chime generator
TDA7052	1W audio power amp
TDA2003	10 W audio power amp
DG201B	Quad analog switch
Power Control	
L298	Two-channel 2A H-bridge motor controller
S202T01F	Solid-state relay, 2A 600V
MAX1551	Lithium polymer battery charger
L297	Stepper motor controller
LED Drivers	
MAX6958	4 × 9 segment LED driver with I²C interface
LM3914	Bar graph LED display driver (10 LED outputs, analog input)
LM3404	1A constant current LED driver
Miscellaneous	
NE555	Timer IC
DS1302	Real-time clock, I²C interface
24C1024	128 k × 8 bit I2C EEPROM
SST25VF010A	1 M bit SPI flash

16.2 Breakout Boards and Modules

You do not always need to start your design from scratch. Many ready-made modules and breakout boards are available to inventors to incorporate into their designs.

The difference between a breakout board and a module is often a little blurred. The intention of a breakout board is that it simply “breaks out” the difficult-to-access pins of a surface-mounted device (SMD) IC into 0.1-in connections that are much easier to use. However, breakout boards often add a few extra components like a decoupling capacitor, voltage regulator, or level conversion. So the point at which a breakout board becomes a module is a bit arbitrary.

Whatever you chose to call them, breakout boards and modules can be very useful in prototyping when you want to get something working to prove a concept using prebuilt modules, and then advance to a final design without the module.

There are many sensor modules (see Table 16.2), and some of these are covered in more detail in Chap. 6. Table 16.2 lists some of the more interesting modules that we have come across.

TABLE 16.2 Some Modules and Breakout Boards

MODULE	USAGE	SOURCES
RF Modules		
I^2C FM radio receiver	FM radio	SparkFun: BOB-10344
433/315 MHz transmitters/ receivers	Data link between microcontrollers, sensors, and actuators	SparkFun: WRL-10533, WRL-10535 Seeed Studio: WLS105B5B
Bluetooth	Microcontroller/cellular phone/computer link	SparkFun: WRL-10269, WRL-10253 Seeed Studio: WLS123A1M
WiFi	Wireless network to microcontroller	SparkFun: WRL-10004 Seeed Studio: WLS48188P
XRF modules	Data link between microcontrollers, sensors, and actuators	http://shop.ciseco.co.uk/wireless/
XBee	Data link between microcontrollers, sensors, and actuators; medium range	SparkFun: WRL-10414
XBee Pro	Long-range data link	SparkFun: WRL-09085
RFID tag reader	Security	SparkFun: SEN-08419 Seeed Studio: RFR101A1M
GSM modem	GPS tracking, telemetry	SparkFun: CEL-09533
Audio Modules		
Audio power amps	Driving speakers	SparkFun: BOB-11044
MP3 encoder/decoder/ player	Sound file playing	SparkFun: DEV-10628
Microphone with pre-amp	Sound detection; sound recording or DSP	SparkFun: BOB-09964
MIDI decoder	Musical instruments	SparkFun: BOB-08953
MP3 player	Playing sound samples	SparkFun: BOB-10608
Power Control		
H-bridge	Bidirectional motor control	SparkFun: ROB-09457, DEV-10182
Relay (wireless)	Wireless power control, home automation	Seeed Studio: WLS120B5B
Display Modules		
LCD alphanumeric	2 and 4 line by 16 or 20 character displays	SparkFun: LCD-00255 Seeed Studio: LCD108B6B
LCD graphical	128 × 64 up to 320 × 240 pixel displays, color and grayscale	SparkFun: LCD-00569, LCD-10089 Seed Studio: LCD101B6B, LCD105B6B
OLED displays	Bright color displays	Spark Fun: LCD-09678 Seeed Studio: OLE42178P
LED Matrix Displays (serial interface)	Large color displays and signs	SparkFun: COM-00760
Humidity sensor	Weather stations, humidity control	SparkFun: SEN-10239 Seeed Studio: SEN111A2B
Compass	Direction finding	SparkFun: SEN-07915 Seeed Studio: SEN101D1P
Magnetometer	Measurement of magnetic field strength	SparkFun: SEN-00244
Color sensor	Measurement of light color, for example in industrial control	SparkFun: SEN-10904 Seeed Studio: SEN60256P
Temperature sensor	Digital thermometers and thermostats	SparkFun: SEN-09418 Seeed Studio: SEN01041P

16.2.1 Radio Frequency Modules

Radio frequency (RF) electronics is a specialized part of electronics that is a discipline in its own right. At high frequencies, PCB layout becomes critical, and design is not as straightforward as with more typical analog and digital design. For this reason, you will often find ready-to-use RF modules that can either be soldered onto your own PCB as part of a bigger design or plugged into socket headers. Figure 16.1 shows a variety of RF modules:

RF modules

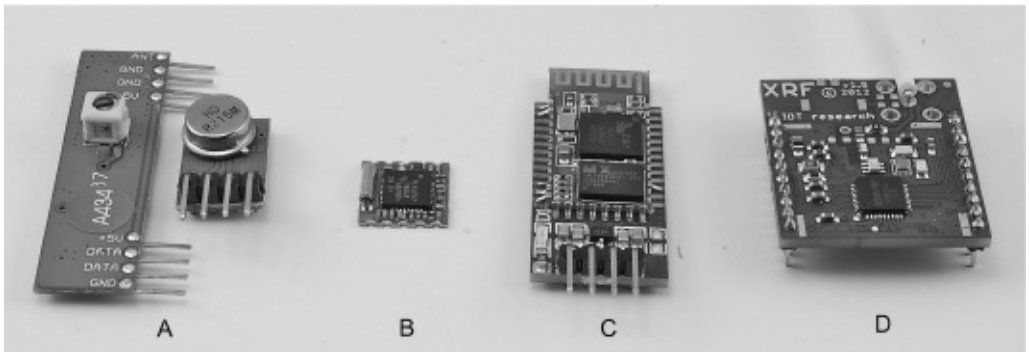


FIGURE 16.1

- A 433MHz receiver and transmitter pair (A)
- A TEA5767 FM receiver board (B)
- A Bluetooth modem, which is actually one module on top of another, with the larger one providing level conversion (C)
- An XRF wireless serial module that fits a standard XBee socket (D)

These represent just a small subset of the wide range of RF modules available to the inventor (see Table 16.2 for a more complete list).

433 MHz and 315 MHz Modules

The 433 MHz and 315 MHz modules (Fig. 16.1A) are very low cost. They find their way into all sorts of consumer electronics requiring remote control, such as wireless door bells, remote car unlocking, smart meters, and remote-control toys.

The data rates are generally low (8 kb/s is usually the top end, and 2 kb/s is quite common), but power consumption is also low. Their maximum range is normally about 100 yards, but can be much less indoors.

These modules are usually separate transmitters and receivers, rather than a combined transceiver, and so they are generally used in situations where data is flowing in only one direction.

Bluetooth Modules

Bluetooth modules provide a great means of interfacing your electronics to a mobile phone. The module in Fig. 16.1C is actually a basic 3.3 V Bluetooth module designed to be surface mounted onto another board. In this case, it is mounted on a level changing board that allows the module to operate at 5 V TTL serial levels, but 3.3 V modules like this often find themselves incorporated into products. Again, economies of scale mean that these modules can be bought for just a few dollars if you look around.

XBee Modules

XBee is a proprietary standard of Digi International, but the sockets have been adopted for a range of radio link modules from many manufacturers. These modules, such as the XRF module from Ciseco Plc

(see Fig. 16.2), act as a transparent serial interface between two devices. For example, one of the devices might be an Arduino with an XBee socket shield with an XRF module installed communicating with a low-power remote sensor with an XBee socket and another XRF module, as shown on the left in Fig. 16.2.

XRF radio module and wireless temperature sensor module

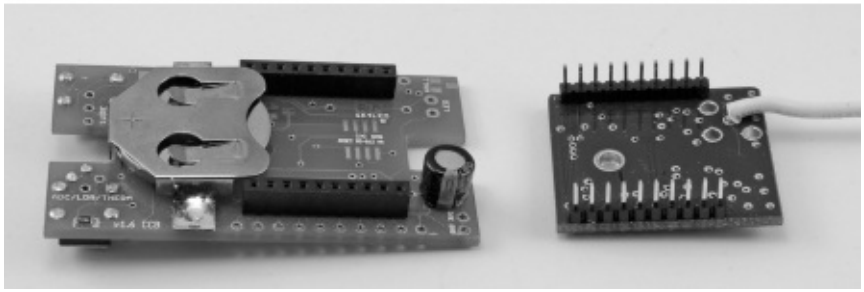


FIGURE 16.2

GSM/GPRS Modem Modules

The GSM/GPRS modem modules are essentially cellular phone modules that provide most of the features of a cellular phone, including sending Short Message Service (SMS) text messages and Global System for Mobile Communications (GSM)/General Packet Radio Service (GPRS) data. Serial commands are sent to the module from a microcontroller to control its operation.

16.2.2 Audio Modules

Although audio electronics is not as difficult as RF electronics, you still need to be careful to avoid earth loops and the dreaded 60 Hz hum that all too easily finds its way on to the signal path given half a chance.

You will find a good selection of audio amplifier modules such as the class D amplifier based on the XMA2012, which is a 2 x 3 W power amp IC (see Fig. 16.3).

Audio power amp module



FIGURE 16.3

This kind of module is very convenient, as it comes with screw terminals for power and speakers and a socket for audio input. Economies of scale make the cost of such modules low.

As well as power amp modules, you will also find audio modules for pre-amps, MP3 players, MIDI interfaces, and tone controls.

16.3 Plug-and-Play Prototyping

When taken to its extreme, the modular approach can result in a system such as .NET Gadgeteer that allows complete plug-and-play development of a prototype (see Fig. 16.4). This system is based on attaching a wide range of sensors and other modules to a microcontroller main board using plug-in cables.

The .NET Gadgeteer system

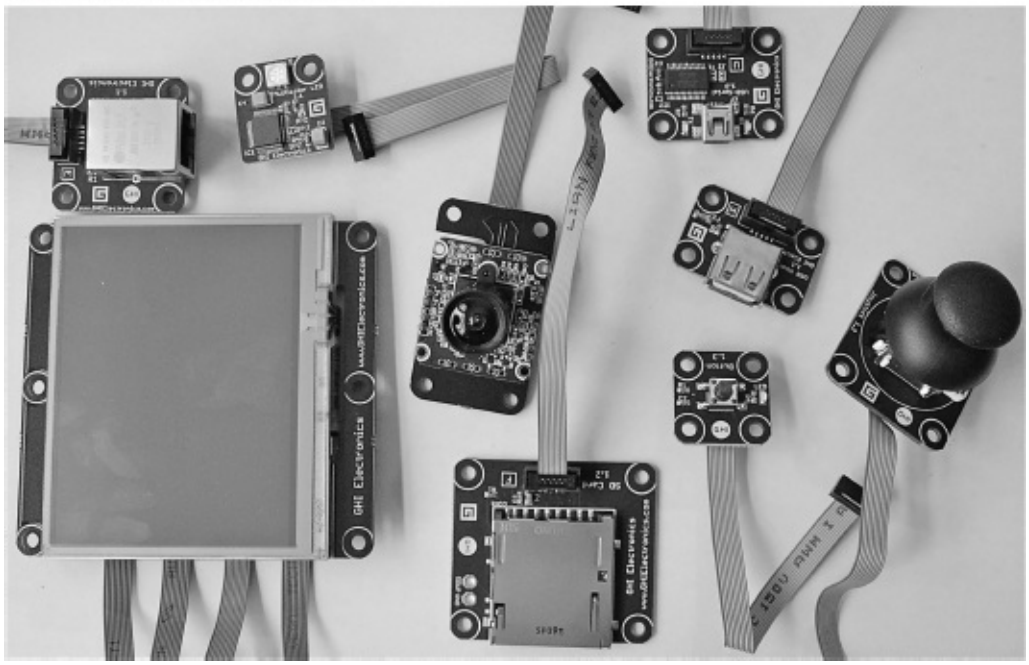


FIGURE 16.4

.NET Gadgeteer is programmed using Microsoft Visual Studio. When used with .NET Gadgeteer, this IDE includes a graphical editor that generates much of the boilerplate code, simply by connecting modules in the design window (see Fig. 16.5).

Connecting modules in Visual Studio

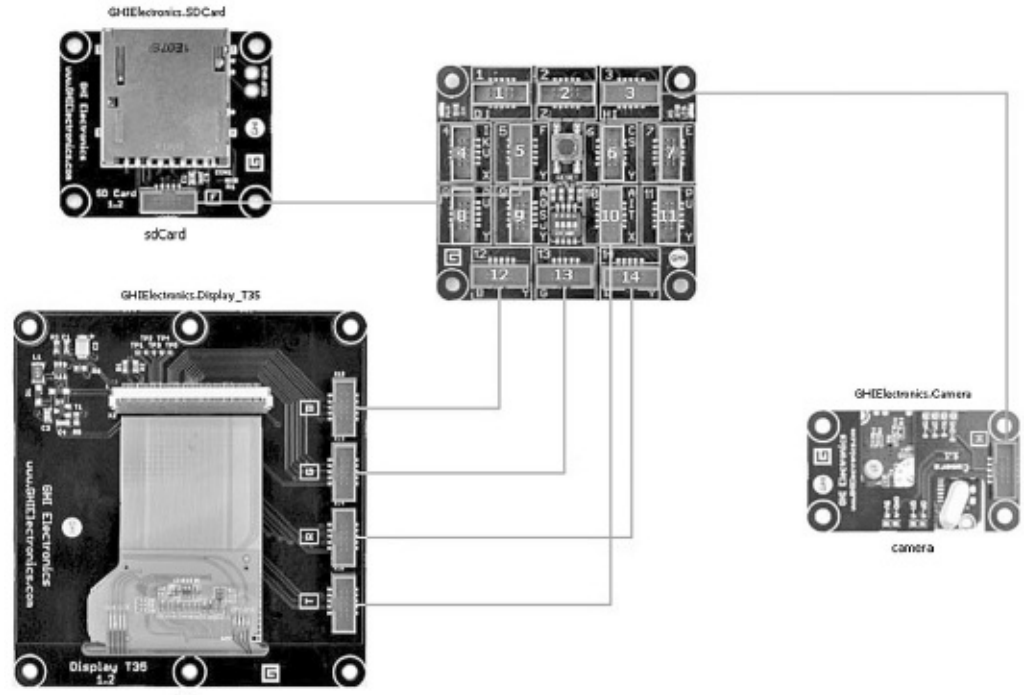


FIGURE 16.5

New modules for the .NET Gadgeteer are being developed all the time. The following are most of the modules available at the time of writing:

- Accelerometer

- Barometer
- Bluetooth
- Buttons
- Camera
- CAN (vehicle engine management unit)
- Cellular radio/GSM/GPRS modem
- Compass
- Current measurement (for smart meters and energy monitors)
- Display, touch screen LCD
- GPS
- Gyroscope
- Joystick
- LED, multicolor
- Light sensor
- Moisture sensor
- Motor driver
- MP3 music player
- OLED display
- Potentiometer
- Pulse oximeter (heart pulse rate measurement)
- Relay
- SD card
- USB interface
- Video output
- Wi-Fi
- XBee

The module manufacturers also supply main boards of different sizes and specifications. For up-to-date lists, see the websites of the main suppliers of .NET Gadgeteer hardware: GHI Electronics, Seeed Studio, Sytech Designs, and DFRobot.

If you want to learn more about using .NET Gadgeteer, the book *Getting Started with .NET Gadgeteer* by Simon Monk (Make, 2012) is a good place to start.

16.4 Open Source Hardware

Open source software has been with us for many years now, but open source hardware is a relatively new concept. The term *open source* really applies to the design files for the hardware, and means that the schematic and PCB design are made publicly available (usually in EAGLE CAD format). The implications of this are that anyone is free to take those design files and produce their own boards using the design. Often, the originator of a new piece of open source hardware will also manufacture the boards and sell them directly or through distributors. The originator will generally be known in the community, and these will be considered to be the “original” boards and therefore the most valuable. Copies and

clones may appear if the design is successful, but the original boards are likely to remain the most used.

Perhaps the most successful open source hardware design is the Arduino (see Chap. 13). All the Arduino boards and most of the shields available for the Arduino are released under an open source or creative commons license. Other well-known open source hardware projects include the following:

- **BeagleBoard** A single-board computer
- **MIDIbox** MIDI music hardware
- **Monome** A button and LED grid for controlling virtual synthesizers
- **Ultimaker, RepRap, and MakerBot** 3D printer designs
- **Chumby** An embedded computer
- **Open** EEG A medical device

There are also many small modules that are released as open source designs.

So, why would anyone want to give away their ideas for free for the world to use? Well, for one thing, many creative people make things for the fun of making them and not to get rich. There is a world of difference between having a good idea and building a business. If you want people to see and use what you have done, then why not release it into the wild? As with open source software, there are also opportunities for businesses to develop with the so called “halo effect,” providing consultation and support for the hardware.

Power Distribution and Home Wiring

A.1 Power Distribution

Figure A.1 shows a typical power-distribution system found in the United States (region in California). The voltages listed are sinusoidal and are represented by their rms values. Note that the system in your area may look a bit different from the system shown here. Contact your local utility to learn about how things are set up in your region.

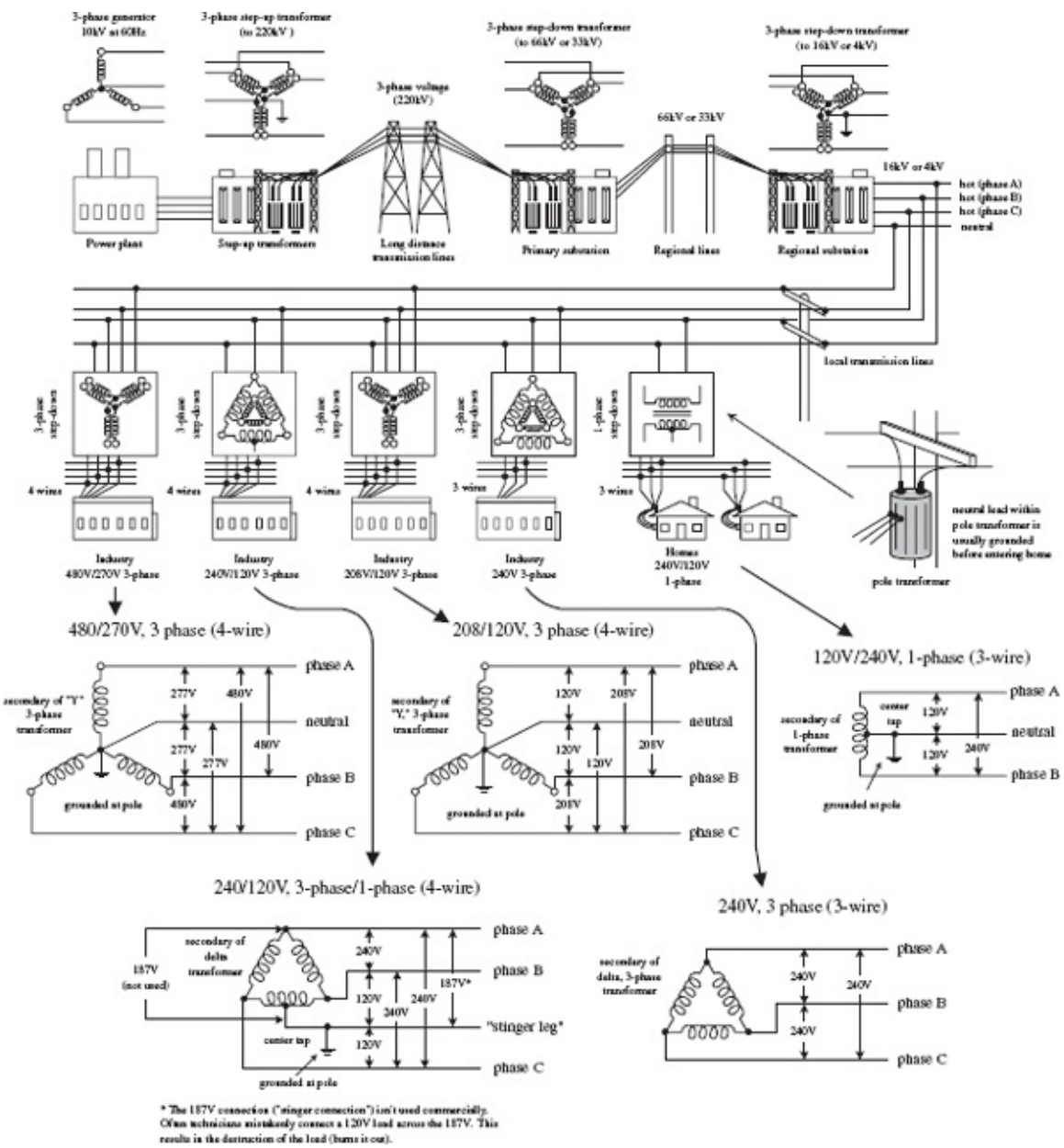


FIGURE A.1

As a note, ac is used in electrical distribution instead of dc because ac can be stepped up or down

easily by using a transformer. Also, over long distances, it is more efficient to send electricity via high-voltage/low-current transmission lines. By reducing the current, less power is lost to resistive heating during transmission (lowering I within $P = I^2R$ lowers P). Once electricity reaches a substation, it must be stepped down to a safe level before entering homes and businesses.

Notice that industry typically uses three-phase electricity. The natural sequencing of the three phases is particularly useful for devices that perform rhythmic tasks. For example, three-phase electric motors (found in grinders, lathes, welders, air conditioners, and other high-power devices) often turn in near synchrony with the rising and falling voltages of the phases. Also, with three-phase electric power, there is never a time when all three phases are at the same voltage. With single-phase power, whenever the two phases have the same voltage, there is temporarily no electric power available. This is why single-phase electric devices must store energy to carry them over these dry spells. In three-phase power, a device can always obtain power from at least one pair of phases.

A.2 A Closer Look at Three-Phase Electricity

The simple generator shown in Fig. A.2 can be used to generate single-phase voltage. As the magnet is rotated by a mechanical force, a voltage is induced within the two coils (spaced 180° apart) that yields a single sinusoidal voltage. The output voltage is usually expressed as an rms voltage ($V_{rms} = 1/\sqrt{2}\ V_0$).

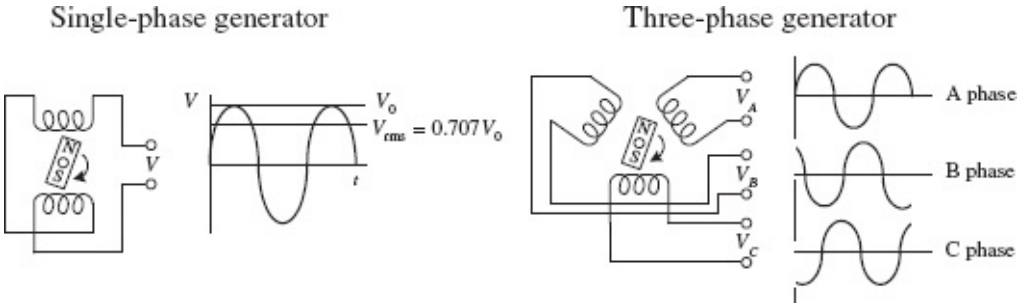
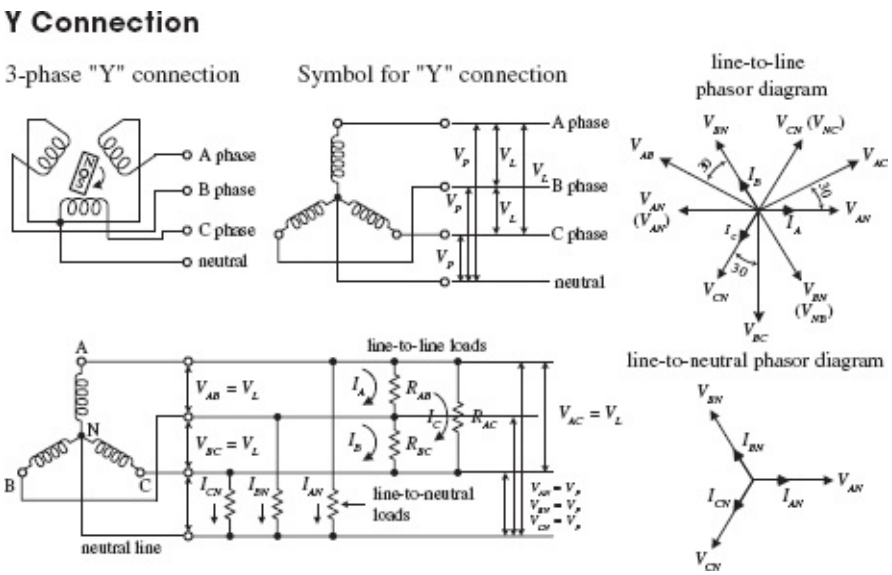


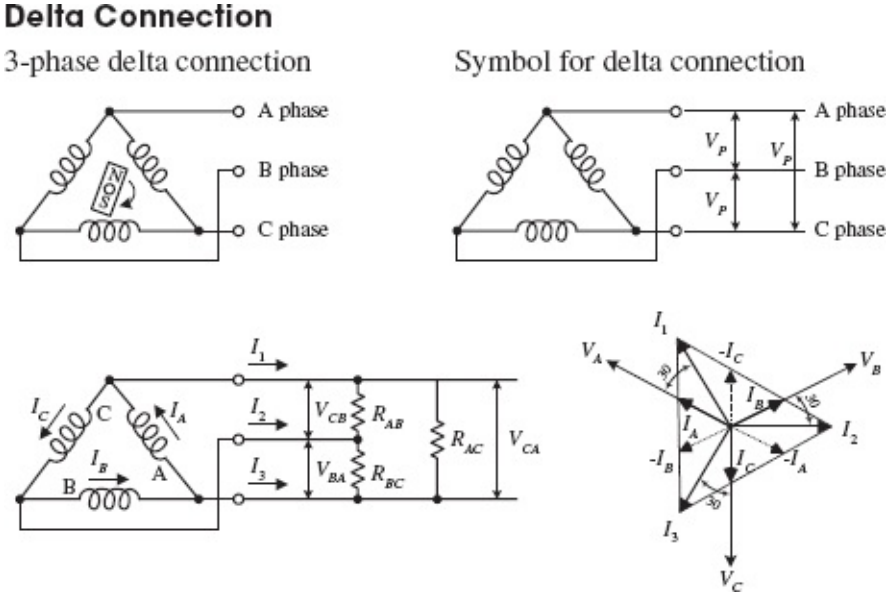
FIGURE A.2

In the three-phase generator, three separate voltages are generated by using three different coils spaced 120° apart. As the magnet rotates, a voltage is induced across each of the generator's coils. All the coil voltages are equal in magnitude but are 120° out of phase with each other. With this generator, you could power three separate loads of equal resistance, or you could drive a three-phase motor with a similar coil configuration; however, this requires using six separate wires. To reduce the number of wires, there are two tricks that can be used. The first trick involves rearranging the three-phase coil connections into what is called a *delta connection*—which yields three wires. The second trick involves rearranging the coil connection into what is called a *Y connection*—which yields four wires. Here are the details.



A Y configuration is made by connecting one end of each of the generator's coils together to form what is called the *neutral* lead. The remaining three coil ends are brought out separately and are considered the "hot" leads. The voltage between the neutral and any one of the hot lines is called the *phase voltage* V_p. The total voltage, or *line voltage* V_L, is the voltage across any two hot leads. The line voltage is the vector sum of the individual phase voltages. In a Y-loaded circuit, each load has two phases in series. This means that the current and voltage through and across a load must be determined by superimposing the phase currents and phase voltages. One way to do this is to make a phasor diagram, as shown in the figure. For practical purposes, what is important to note is the line voltage is about $\sqrt{3}$ times the phase voltage. Also, the line currents are 30° out of phase with the line voltages. When a neutral is used, the line currents and line voltages are equal to the individual phases from the generator. No current flows in the neutral when the loads are equal across each phase (balanced load).

FIGURE A.3



A delta connection is made by placing the three-phase generator's coils end to end, as shown in the figure. Because the delta configuration has no neutral lead, the phase voltages are equal to the line voltages. Again, like the Y configuration, all line voltages are 120° out of phase with each other. However, unlike the Y configuration, the line currents (I_1, I_2, I_3) are equal to the vector sum of the phase currents (I_A, I_B, I_C). When each of the phases are equally loaded, the line currents are all equal but 120° apart. The line currents are 30° out of phase with the line voltages, and they are $\sqrt{3}$ times the phase currents.

FIGURE A.4

A.3 Home Wiring

In the United States, three wires run from the pole transformer (or "green box" transformer) to the main service panel at one's home. One wire is the A-phase wire (black in color), another is the B-phase wire (black in color), and the third is the neutral wire (white in color). (Figure A.5 shows where these three

wires originate within the pole/ green box transformer.) The voltage between the A-phase and B-phase wires, or the "hot to hot" voltage, is 240 V, while the voltage between the neutral wire and either A-phase or B-phase wire, or the "neutral to hot" voltage, is 120 V. (These voltages are nominal and may vary from region to region.)

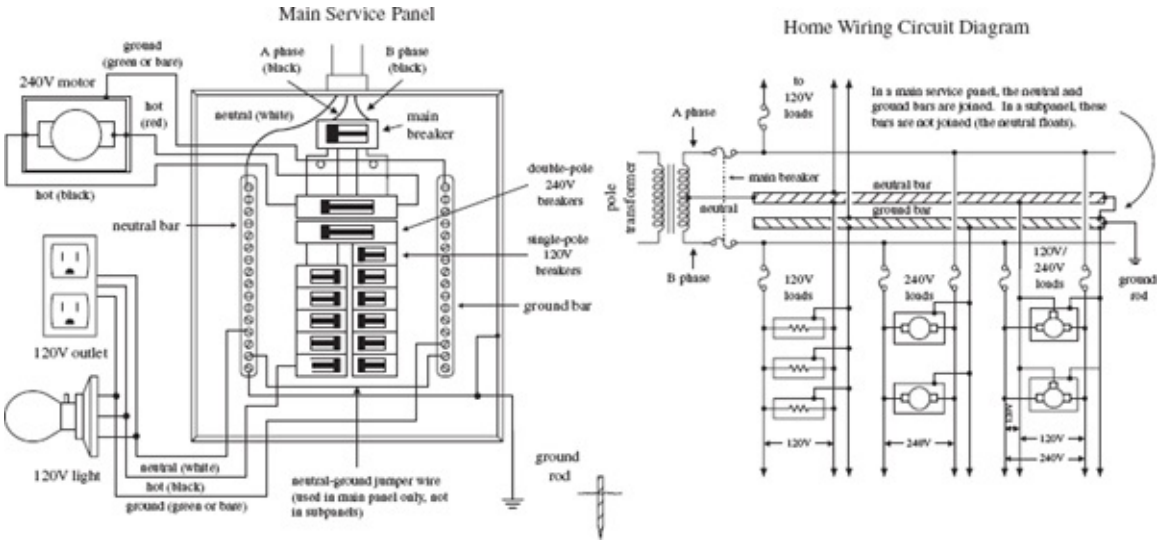


FIGURE A.5

At the home, the three wires from the pole/green box transformer are connected through a wattmeter and then enter a main service panel that is grounded to a long copper rod driven into the ground or to the steel in a home's foundation. The A-phase and B-phase wires that enter the main panel are connected through a main disconnect breaker, while the neutral wire is connected to a terminal referred to as the *neutral bar* or *neutral bus*. A *ground bar* also may be present within the main service panel. The ground bar is connected to the grounding rod or to the foundation's steel supports.

Within main service panels, the neutral bar and the ground bar are connected together (they act as one). However, within subpanels (service panels that get their power from the main service panel but which are located some distance from the main service panel), the neutral and ground bars are not joined together. Instead, the subpanel's ground bar receives a ground wire from the main service panel. Often the metal conduit that is used to transport the wires from the main service panel to the subpanel is used as the "ground wire." However, for certain critical applications (e.g., computer and life-support systems), the ground wire probably will be included within the conduit. Also, if a subpanel is not located in the same building as the main panel, a new ground rod typically is used to ground the subpanel. Note that different regions within the United States may use different wiring protocols. Therefore, do not assume that what I am telling you is standard practice where you live. Contact your local electrical inspector.

Within the main service panel, there are typically two bus bars into which circuit breaker modules are inserted. One of these bus bars is connected to the A-phase wire; the other bus bar is connected to the B-phase wire. To power a group of 120-V loads (e.g., upstairs lights and 120-V outlets), you throw the main breaker to the off position and then insert a single-pole breaker into one of the bus bars. (You can choose either the A-phase bus bar or the B-phase bus bar. The choice of which bus bar you use only becomes important when it comes to balancing the overall load—more on that in a second.) Next, you take a 120-V three-wire cable and connect the cable's black (hot) wire to the breaker, connect the cable's white (neutral) wire to the neutral bar, and connect the cable's ground wire (green or bare) to the ground bar. You then run the cable to where the 120-V loads are located, connect the hot and neutral wires across the load, and fasten the ground wire to the case of the load (typically a ground screw is supplied on an outlet mounting or light figure for this purpose). To power other 120-V loads that use their own breakers, you basically do the same thing you did in the last setup. However, to maximize the capacity of the main panel

(or subpanel) to supply as much current as possible without overloading the main circuit breaker in the process, it is important to balance the number of loads connected to the A-phase breakers with the number of loads connected to the B-phase breakers. This is referred to as *balancing the load*.

Now, if you want to supply power to 240-V appliances (e.g., ovens, washers, etc.), you insert a double-pole breaker between the A-phase and B-phase bus bars in the main panel (or subpanel). Next, you take a 240-V three-wire cable and attach one of its hot wires to the A-phase terminal of the breaker and attach its other hot wire to the B-phase terminal of the breaker. The ground wire (green or bare) is connected to the ground bar. You then run the cable to where the 240-V loads are located and attach the wires to the corresponding terminals of the load (typically within a 240-V outlet). Also, 120-V/240-V appliances are wired in a similar manner, except you use a fourwire cable that contains an additional neutral (white) wire that is joined at the neutral bar within the main panel (or subpanel). (As a practical note, you could use a fourwire 120-V/240-V cable instead of a 240-V three-wire cable for 240-V applications—you would just leave the neutral wire alone in this case.)

As a note of caution, do not attempt home wiring unless you are sure of your abilities. If you feel that you are capable, just make sure to flip the main breaker off before you start work within the main service panel. When working on light fixtures, switches, and outlets that are connected to an individual breaker, tag that breaker with tape so that you do not mistakenly flip the wrong breaker when you go back to test your connections.

A.4 Electricity in Other Countries

In the United States, homes receive a 60-Hz, 120-V single-phase voltage, whereas industry typically receives a 60-Hz, 208-V/120-V three-phase voltage. Most other countries, on the other hand, work with a 50-Hz, 230-V single-phase voltage and a 415-V three-phase voltage. Now, if you were to take a U.S.-built 120-V, 60-Hz device over to Norway—where 230 V is used—and plug that device directly into the outlet (you would need an adapter to do this—their outlets look different), you run a risk of damaging the device. Some devices may not “care” about the voltage and frequency differences, but others will. You could use a converter (transformer plug-in device) to step down the voltage from the outlet, but you would still be stuck with 50 Hz. The 10-Hz difference will not affect most devices, but other devices, such as TVs and VCRs, may not function properly.

Here’s a listing of single-phase voltages found in some countries. Note the plug types.

Country	Voltage V	Frequency Hz	Plug Type
Australia	240	50	I
Belgium	230	50	C, E
Brazil	110/220	60	A, B, C, D, G
Canada	120	60	A, B
Chile	220	50	C, L
China	220	50	I
Congo	230	50	C, E
Costa Rica	120	60	A, B
Egypt	220	60	C
France	230	50	C, E, F
Germany	230	50	F
Hong Kong	230	50	D, G
India	230	50	C, D
Iraq	220	50	C, D, G
Italy	127/220	50	F, L
Japan	100	50/60	A, B
Korea	110/220	60	A, B, D, G, I, K
Mexico	127	60	A
Netherlands	230	50	C, E
Norway	230	50	C, F
Philippines	110/220	60	A, B, C, E, F, I
Russia & former Soviet Republics	220	50	C, F
Spain	127/220	50	C, E
Switzerland	220	50	C, E, J
Taiwan	110	60	A, B, I
US	120	60	A, B
United Kingdom	230	50	G

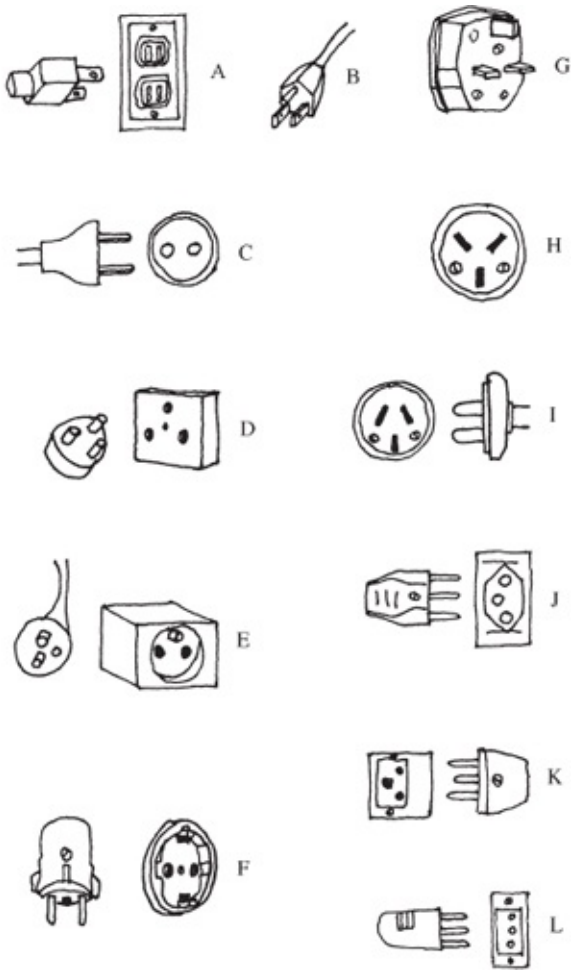


FIGURE A.6

APPENDIX B

Error Analysis

Reliability estimates of measurements greatly enhance their value. For example, saying the resistance of a resistor is 1000 Ω ± 50 Ω tells you much more than simply stating the resistance is 1000 Ω.

The term *error* is basically interchangeable with *uncertainty*, but does not have the same meaning as *mistake*. Mistakes, such as errors in calculations, should be corrected before estimating the experimental error. In estimating the reliability of a single quantity (such as resistance), you should recognize three different kinds of sources of error:

- Actual variations of the quantity being measured, for example, resistance changes due to temperature variations. In electronics, for accurate measurements, you should consider the temperature ratings and the corresponding errors specified in the data sheets.
- Test equipment in error. The errors introduced by test equipment. Make sure that all equipment is calibrated, and be certain to take into account input impedance characteristics, such as the input impedance of a multimeter and oscilloscope.
- Human error. With digital equipment displays this is less of a problem. A scope with a graphical display will not give you high accuracy (only around 5 percent).

B.1 Absolute Error, Relative Error, and Percent Error

If Δx is the *absolute error* (or *uncertainty* with a ± in front) in a measurement whose value is x, then $\Delta x/x$ is called the *relative error* (or *fractional uncertainty*). If we multiply the relative error by 100 percent, we get the *percent error*, or $100\% \cdot (\Delta x/x)$. The term *tolerance* is interchangeable with both absolute error and percent error. For example, tolerances in length measurements are typically given in terms of absolute error values, while tolerances in resistances are typically given in percent error.

Example 1: What's the relative error and percent error of 0.125A ± 0.01A?

Answer:

$$\text{Relative error} = \frac{\Delta x}{x} = \frac{0.01\ \text{A}}{0.125\ \text{A}} = 0.08$$

$$\text{Percent error} = 100\% \frac{\Delta x}{x} = 100\% \times \text{Relative error} = 100 \times 0.08 = 8\%$$

Example 2: What are the relative error and the absolute error or uncertainty of a 3300-Ω resistor with a tolerance of 5 percent? What's the guaranteed range of resistance?

Answer: Here, tolerance represents percent error, so

$$\text{Relative error} = \frac{\text{Percent error}}{100\%} = \frac{\text{Tollerance}}{100\%} = \frac{5\%}{100\%} = 0.05$$

The absolute error or uncertainty is:

$$\Delta x = x \text{ (relative error)} = (3300\ \Omega)(0.05) = \pm 165\ \Omega$$

The resistor is guaranteed for 3300 Ω ± 165 Ω or guaranteed to be between 3135 Ω and 3465 Ω.

B.2 Uncertainty Estimates

When dealing with equations with many independent variables, like the following RC charge response equation

$$I = \frac{V_C}{R} e^{-t/RC}$$

the uncertainty or error in the final result (e.g., current) will depend on the individual uncertainties (e.g., uncertainties in resistance, capacitance, voltage, and time). The propagation of errors can be explained by first analyzing simple arithmetic cases:

1. If the desired result is the *sum* or *difference* of two measurements, the absolute uncertainties add:

   Let Δx and Δy be the errors in x and y, respectively. For the *sum*, we have:

   $$z = x + \Delta x + y + \Delta y$$

   and a relative error of

   $$(\Delta x + \Delta y)/(x + y)$$

   Since the signs of Δx and Δy can be opposite, adding the absolute values gives a pessimistic estimate of the uncertainty. If errors have a Gaussian distribution and are independent, they combine in quadrature (square root of the sum of the squares):

   $$\Delta z = \sqrt{\Delta x^2 + \Delta y^2}$$

   For the *difference* of two measurements, we obtain a relative error of:

   $$(\Delta x + \Delta y)/(x - y)$$

   which becomes very large if x is nearly equal to y. This is an important point to note; you must avoid designing experiments where two large quantities are measured and their difference obtained.

2. If the desired result involves *multiplying* or *dividing* measured quantities, then the relative uncertainty of the result is the sum of the relative errors in each of the measured quantities. The most pessimistic case corresponds to adding the absolute value of each term, since the Δx_i and Δx_i can be of either sign:

   $$\Delta z = z \times \left[\sum_i \left| \frac{\Delta x_i}{x_i} \right| + \sum_i \left| \frac{\Delta y_i}{y_i} \right| \right]$$

   Again, if the measurement errors are independent and have a Gaussian distribution, the relative errors

will add in quadrature:

$$\Delta z = z \times \sqrt{\sum_i \left(\left| \frac{\Delta x_i}{x_i} \right| \right)^2 + \sum_i \left(\left| \frac{\Delta y_i}{y_i} \right| \right)^2}$$

3. If the desired result is a *power* of the measured quantity, the relative error in the result is the relative error in the measured quantity multiplied by the power. For example, the uncertainty in

$$\boldsymbol{z = x^n}$$

is:

$$\Delta z = z \times n \left(\frac{\Delta x}{x} \right)$$

4. For anything more complex, the following equation gives the general method for finding uncertainty in measurements. For example, if $R = f(x, y, z)$ is the functional relationship between three measurements x, y, z, then

$$dR = \frac{\partial f}{\partial x} dx = \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz$$

gives the uncertainty in R when the uncertainties dx, dy, and dz are known.

Usually, you don't have to go to this extreme—you can usually get by with the rules for adding, subtracting, multiplying, dividing, and rising to the power. To make life easy on you, a cheat sheet is provided here.

FORMULAS TO CALCULATE UNCORRELATED ERRORS

If $A = \bar{A} \pm a$, $B = \bar{B} \pm b$, and $C = \bar{C} \pm c$, where $\bar{A}$, $\bar{B}$, $\bar{C}$ are the measured values of the quantities A, B, C, and a, b, c are the respective errors, then the calculated values with error (assuming independent variables and Gaussian distribution) give:

1. $A + B = \bar{A} + \bar{B} \pm \sqrt{a^2 + b^2}$
2. $A + B + C = \bar{A} + \bar{B} + \bar{C} \pm \sqrt{a^2 + b^2 + c^2}$
3. $A - B = \bar{A} - \bar{B} \pm \sqrt{a^2 + b^2}$
4. $A \times B = \bar{A} \times \bar{B} \pm (\bar{A} \times \bar{B}) \sqrt{(a/\bar{A})^2 + (b/\bar{B})^2}$
5. $A \times B \times C = \bar{A} \times \bar{B} \times \bar{C} \pm (\bar{A} \times \bar{B} \times \bar{C}) \sqrt{(a/\bar{A})^2 + (b/\bar{B})^2 + (c/\bar{C})^2}$
6. $A/B = \bar{A}/\bar{B}) \pm (\bar{A}/\bar{B}) \sqrt{(a/\bar{A})^2 + (b/\bar{B})^2}$
7. $A^B = (\bar{A})^{\bar{B}} \pm (\bar{A})^{\bar{B}} \times \bar{B}(b/\bar{B})$

Example 1: The voltages across two series resistors are measured using two different voltmeters. One digital meter reads 6.24 V ± 0.01 V across the first resistor; a less precise analog meter measures 14.3 V ± 0.2 V. What is the total voltage across the pair, including uncertainty in the result?

Answer: We simply add voltages and use Eq. 1 to determine the resultant uncertainty:

$$V_1 + V_2 = 6.24\ \text{V} \pm 14.3\ \text{V} \pm \sqrt{(0.01\ \text{V})^2 + (0.2\ \text{V})^2} = 20.5\ \text{V} \pm 0.2\ \text{V}$$

Example 2: The current through a 180-Ω, 5 percent resistor is 1.256 A ± 0.005 A. Determine the voltage across the resistor with uncertainty included.

Answer: First convert the tolerance to absolute error (or uncertainty):

$$\Delta R = \frac{\text{Tolerance}}{100\%} R = \frac{5\%}{100\%}(180\ \Omega) = \pm 9\Omega$$

Since the equation for voltage is Ohm's law, $V = I \times R$, we use Eq. 4:

$$V = I \times R = 1.256\ \text{A} \times 180\ \Omega \pm (1.256\ \text{A} \times 180\ \Omega) \sqrt{\left(\frac{0.005\ \text{A}}{1.256\ \text{A}}\right)^2 + \left(\frac{9\ \Omega}{180\ \Omega}\right)^2} = 226 \pm 11\ \text{V}$$

Useful Facts and Formulas

C.1 Greek Alphabet

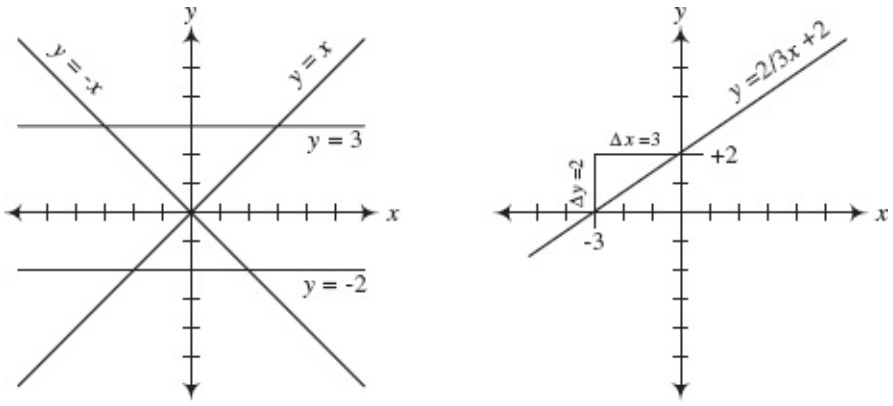
Alpha	A	α
Beta	B	β
Gamma	Γ	γ
Delta	Δ	δ
Epsilon	E	ε
Zeta	Z	ζ
Eta	E	η
Theta	Θ	θ
Iota	I	ι
Kappa	K	κ
Lambda	Λ	Λ
Mu	M	μ
Nu	N	ν
Xi	Ξ	ξ
Omicron	Π	o
Pi	Π	π
Rho	P	P
Sigma	Σ	σ
Tau	T	τ
Upsilon	Y	υ
Phi	Φ	φ
Chi	X	χ
Psi	X	ψ
Omega	Ω	ω

C.2 Powers of 10 Unit Prefixes

PREFIX	SYMBOL	MULTIPLYING FACTOR
tera	T	$\times 10^{12}$

giga	G	$\times 10^{9}$
mega	M	$\times 10^{6}$
kilo	K	$\times 10^{3}$
centi	c	$\times 10^{-2}$
milli	m	$\times 10^{-3}$
micro	μ	$\times 10^{-6}$
nano	n	$\times 10^{-9}$
pico	p	$\times 10^{-12}$

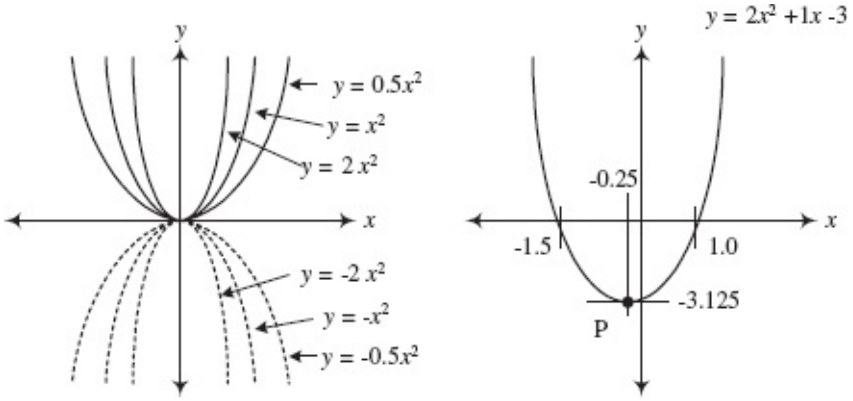
C.3 Linear Functions ($y = mx + b$)



The equation $y = mx + b$ represents the equation of a line. The slope of the line ($\Delta y/\Delta x$) is equal to m, while the vertical shift, or point where the line crosses the y axis, is equal to b.

FIGURE C.1

C.4 Quadratic Equation ($y = ax^2 + bx + c$)

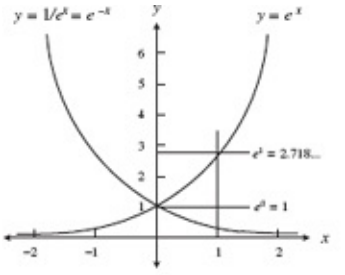


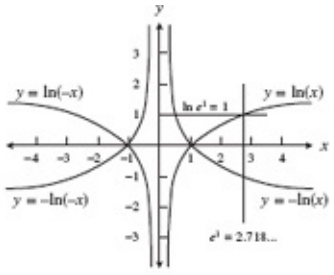
The equation $y = ax^2 + bx + c$ traces out a parabola in the xy plane. The narrowness of the parabola is influenced by a, the horizontal shift is given by $-b/2a$, and the vertical shift is given by $-b^2/a + c$. To determine the roots of the equation (points where the parabola crosses the x axis), use

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

FIGURE C.2

C.5 Exponents and Logarithms





$x^0 = 1$

$1/x^n = x^{-n}$

$x^{1/n} = \sqrt[n]{x}$

$x^m \cdot x^n = x^{m+n}$

$(xy)^n = x^n \cdot y^n$

$(x^n)^m = x^{n \cdot m}$

Base 10: if $10^n = x$, then $\log_{10} x = n$

Base e: if $e^m = y$, then $\ln y = m$

($\log_{10} 100 = 2$, since $10^2 = 100$,

$\ln e = 1$, since $e^1 = e = 2.718\ldots$)

Properties of any logarithm to the base b:

$\log_b 1 = 0$

$\log_b b = 1$

$\log_b 0 = \begin{cases} +\infty & b < 1 \\ -\infty & b > 1 \end{cases}$

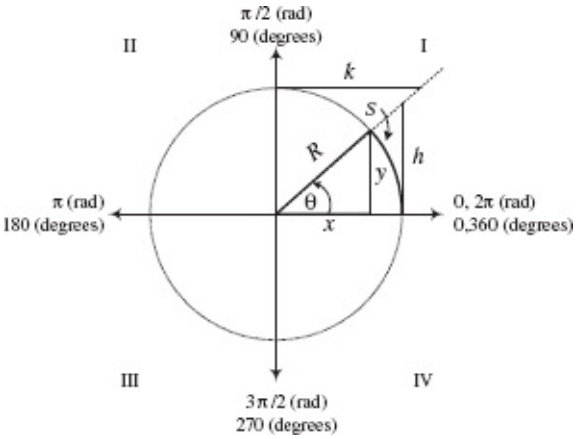
$\log_b (x \cdot y) = \log_b x - \log_b y$

$\log_b (x/y) = \log_b x - \log_b y$

$\log_b (x^y) = y \log_b x$

FIGURE C.3

C.6 Trigonometry



The angle θ subtended by an arc *S* of a circle of radius *R* is equal to the ratio θ = *S/R*, where θ is in radians. 1 radian = 180°/π = 57.296°, while 1° = π/180° = 0.17453 radians. If *R* is rotated counterclockwise from the positive *x* axis, θ is positive in sign. If *R* is rotated clockwise from the positive *x* axis, θ is negative in sign. The trigonometric functions of the angle θ are defined as specific ratios between the sides of the triangles shown in the figure and are expressed as

$$\sin\theta = \frac{y}{R} \qquad \text{if } R = 1 \rightarrow y = \sin\theta$$

$$\cos\theta = \frac{x}{R} \qquad \text{if } R = 1 \rightarrow x = \cos\theta$$

$$\tan\theta = \frac{y}{x} \qquad \text{if } R = 1 \rightarrow h = \tan\theta$$

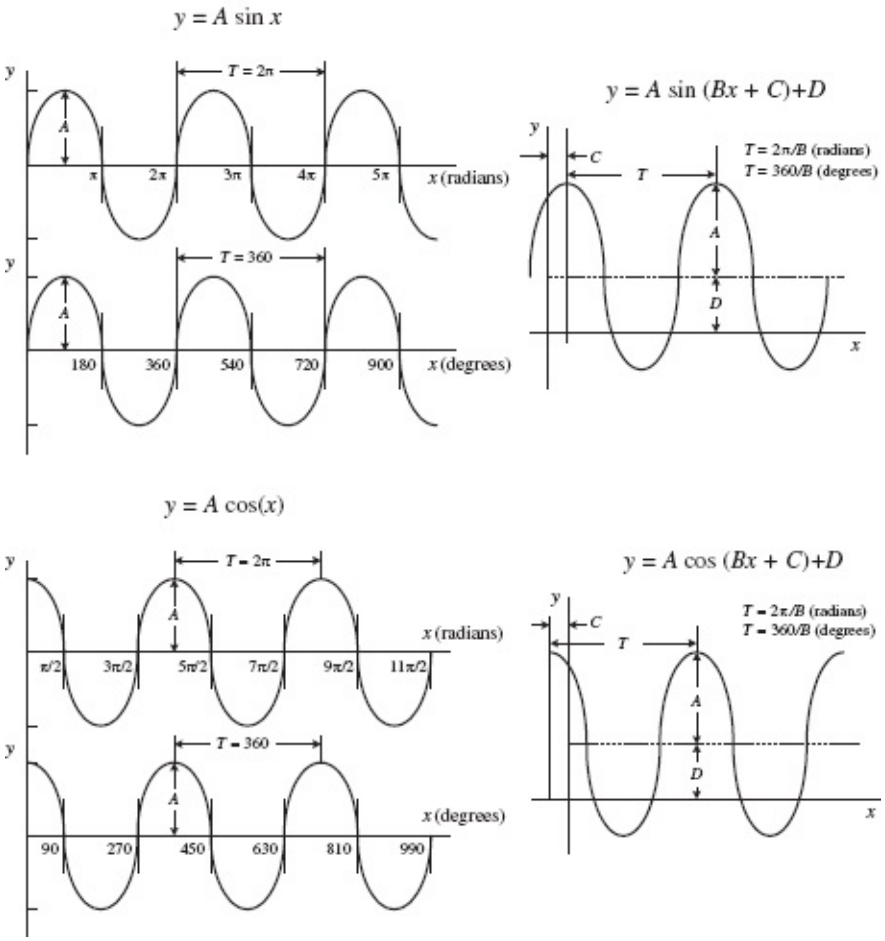
$$\cot\theta = \frac{x}{y} = \frac{1}{\tan\theta} \qquad \text{if } R = 1 \rightarrow k = \cot\theta$$

$$\sec\theta = \frac{R}{x} = \frac{1}{\cos\theta} \qquad \text{if } R = 1 \rightarrow \frac{1}{x} = \sec\theta$$

$$\csc\theta = \frac{R}{y} = \frac{1}{\sin\theta} \qquad \text{if } R = 1 \rightarrow \frac{1}{y} = \csc\theta$$

FIGURE C.4

Sine and Cosine Functions



The graph of $y = A \sin \theta$ is shown to the far left. To alter the vertical, horizontal, period, and phase of this function, alter the equation to the form $y = A \sin (Bx + C) + D$. A is amplitude, $2\pi/B$ is the period (T), C is the phase shift, and D is the vertical shift. In electronics, a voltage can be expressed as

$$V(t) = V_0 \sin(\omega t + \phi) + V_{dc}$$

V_0 is the peak voltage, V_{dc} is the dc offset, Φ is the phase shift, and ω is the angular frequency (rad/s), which is related to the conventional frequency (cycles/s) by

$$f = \frac{1}{T} = \frac{\omega}{2\pi}$$

The graph of $y = A \cos x$ is shifted in phase by $\pi/2$ radians (or 90°) with respect to the graph $y = A \sin x$. The following relations show how the sine and cosine functions are related:

$$\sin\left(\frac{\pi}{2} \pm x\right) = +\cos x \quad \text{or} \quad \sin(90° \pm x) = +\cos x$$

$$\sin\left(\frac{3\pi}{2} \pm x\right) = -\cos x \quad \text{or} \quad \sin(270° \pm x) = +\cos x$$

$$\cos\left(\frac{\pi}{2} \pm x\right) = \pm\sin x \quad \text{or} \quad \cos(90° \pm x) = \pm\sin x$$

$$\cos\left(\frac{3\pi}{2} \pm x\right) = \pm\sin x \quad \text{or} \quad \cos(270° \pm x) = \pm\sin x$$

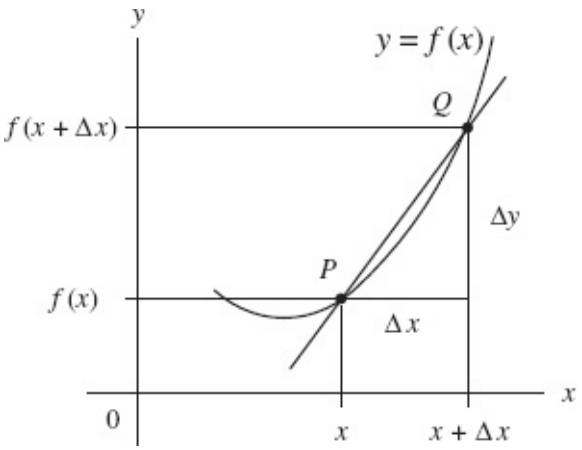
FIGURE C.5

C.7 Complex Numbers

Complex numbers are covered in detail in Chap. 2.

C.8 Differential Calculus

Say you have a function $f(x)$. This function may represent a line, parabola, exponential curve, trigonometric curve, etc. Now pretend that you take a point and move it along the curve of $f(x)$. At the same time, you envision a tangent line touching the curve at the point. As the point moves along the curve, the slope of the tangent line changes ("teeter-totters")—provided the curve is not a line. Now, the slope of the tangent line has great significance in real-life situations. For example, if you graph a curve of the position of an object versus time, the slope at a particular time along the curve represents the instantaneous velocity of the object. Likewise, if you have an electrical charge versus time graph, the slope at time t represents the instantaneous current-flow. Now, the trick to finding the slope of a tangent line for any point along a curve involves using differential calculus. What differential calculus does is this: If you have a function, say, $y = x^2$—through the tricks of differential calculus—you can find another function, which is called the *derivative of y* (usually expressed as y' or dy/dx), that tells you the slope at every point along the curve of y. For $y = x^2$, the derivative is $dy/dx = 2x$. If you are interested in the slope of y at $x = 2$, you plug 2 into dy/dx to give you a slope of 4. But how do you find the derivative of $y = x^2$? Better yet, how do you find the derivative of any given function? The following provides the basic theory.



To find the derivative of a function $y = f(x)$, you let $P(x,y)$ be a point of the graph of $y = f(x)$ and let $Q(x + \Delta x, y + \Delta y)$ be another point of the graph. The slope of the line between P and Q is then simply

$$\frac{f(x+\Delta x)-f(x)}{\Delta x}$$

Now you substitute the function into the preceding equation. For example, if $f(x) = x^2$, then $f(x + \Delta x) = (x + \Delta x)^2$, and the whole expression equals $[(x + \Delta x)^2 - x^2]/\Delta x$. Next, you hold x fixed and let Δx approach zero. If the slope approaches a value that depends only on x, you call this the slope of the curve at point P. The slope of the curve at P is itself a function of x, defined at every value of x at which the limit exists. You denote the slope by $f'(x)$ ("f prime"), or dy/dx ("$dydx$"), or df/dx ("$dfdx$"), and call any one of these terms (your choice) the derivative of $f(x)$:

$$f'(x)=\frac{dy}{dx}=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}$$

For the function $f(x) = x^2$, after carrying out the limit, you would get $f'(x) = dy/dx = 2x$ as the derivative.

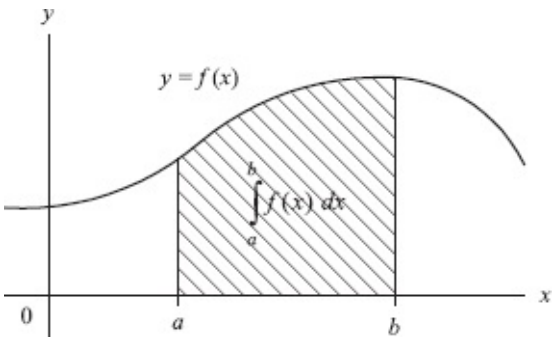
FIGURE C.6

Now, in practice, if you need to find the derivative of a function, you do not bother using the preceding equation. To do so would be very time-consuming and might require a number of nasty mathematical tricks, especially if you are trying to find the derivative of a complex function such as $2e^x \sin(3x + 2)$. Instead, what you do is memorize a few simple rules and memorize a few simple derivatives. The table below shows some of the rules and simple derivatives that will come in handy for many applications. In the table, a and n are constants, while u and v are functions.

DERIVATIVE	EXAMPLES
$\frac{d}{dx}a = 0$	$\frac{d}{dx}4 = 0$
$\frac{d}{dx}x^n = nx^{n-1}$ (note: $\frac{1}{x^n} = x^{-n}$)	$\frac{d}{dx}x = 1, \frac{d}{dx}x^2 = 2x, \frac{d}{dx}x^5 = 5x^4, \frac{d}{dx}x^{-1/2} = -\frac{1}{2}x^{-3/2}$
$\frac{d}{dx}e^x = e^x$	
$\frac{d}{dx}\ln x = \frac{1}{x}$	
$\frac{d}{dx}\sin x = \cos x$	
$\frac{d}{dx}\cos x = -\sin x$	
$\frac{d}{dx}au(x) = a\frac{d}{dx}u(x)$	$\frac{d}{dx}3x^2 = 3\frac{d}{dx}x^2 = 6x, \frac{d}{dx}3e^x = 3e^x, \frac{d}{dx}7\sin x = 7\cos x$
$\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$	$\frac{d}{dx}(2x + x^2) = \frac{d}{dx}(2x) + \frac{d}{dx}(x^2) = 2 + 2x$
$\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{v\,du/dx - u\,dv/dx}{v^2}$	$\frac{d}{dx}\left(\frac{x^2+1}{x^2-1}\right) = \frac{(x^2-1)\cdot 2x - (x^2+1)\cdot 2x}{(x^2-1)^2} = \frac{-4x}{(x^2-1)^2}$
Chain rule: If u is a function of y, and y is in turn a function of x, then	
$\frac{d}{dx}\{u[v(x)]\} = \frac{du}{dx}\cdot\frac{dv}{dx}$	$\frac{d}{dx}\sin(ax) = a\cos(ax), \frac{d}{dx}e^{2x} = 2e^{2x}$

C.9 Integral Calculus

In differential calculus, our goal was to find the derivative of a function. In integral calculus, our goal is to find the *functions of a derivative*. In reality, calling one thing a derivative and another thing a function is somewhat confusing because a derivative is usually a function in itself. To avoid possible complications, we simply call anything that's written as $y = f(x)$ a function, call anything that's written as $dy/dx = df(x)/dx$ a derivative, and call anything that's written as $\int dy = \int f(x)dx$ an antiderivative or integral. In the last case, the "$\int$" is called the integral sign, the function $f(x)$ is called the integrand, and dx is called the variable of integration. To integrate a function is to find all the functions that have it as a derivative—to find all of the given function's antiderivatives. But, the term *integration* has a second, less technical meaning—to give the sum total of. In this meaning, integration represents a mathematical process that enables us to calculate the area of a region that has curved boundaries (see graph in Fig. C.7).



If we're given an equation $dy/dx = f(x)$ and wish to find y, we must integrate (antidifferentiate). In this case, we first rearrange the equation into the form $dy = f(x)dx$. Next, we integrate both sides:

$$\int dy = \int (x)dx$$

$$y \pm C = \int f(x)\,dx$$

$$y = \int f(x)dx + C$$

Again, $\int$ is an integral sign, the function $f(x)$ is the integrand, and C is the constant of integration. We guessed that $\int dy = y \pm C$, since if we were to take the derivative of any function of the form $y \pm C$ ($y + 2$, $y - 54$, etc.) we'd get y as the answer—we worked backward. Since C could be positive or negative, it's arbitrary to give it a sign. This way, we simply add C to the left. In this form, we call the integral an indefinite integral.

Example: Given $dy/dx = 2$, find y.

$$dy = 2dx$$
$$\int dy = \int 2\,dx$$
$$y = 2x + C$$

In real-life situations, it's typically undesirable to have a constant in our answers, since we're looking for a definite result. To get rid of the constant, we apply boundary conditions. For example, if we take our last example $dy/dx = 2$, we say we're only interested in values of dy/dx from, say, 1 to 5—whatever range is appropriate for the situation. In that case, we take the definite integral

$$y = \int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a)$$

Without getting too technical, F represents the definite integral without the constant sign, while the (a) and (b) terms mean to place the boundaries points into the x term of F. If we consider our example $dy/dx = 2$ with boundary condition 1 to 5, we'd get:

$$y = \int_1^5 2\,dx = 2x\Big|_1^5 = 2(5) - 2(1) = 8$$

We call this a definite integral. At this point, it's worth introducing a graphical approach to visualizing integration. If we consider $dy/dx = 2$, and graph dy/dx versus x, we get a straight horizontal line at $dy/dx = 2$ for all values of x. By summing up the area under the curve, we get the integral. So incorporating the boundary conditions the area is $(5 - 1) \times 2 = 8$.

Now, if we attempt to integrate more complex functions, we'll have a hard time guessing what the functions of the derivative are. Instead, we must apply some special tricks along the way. These tricks stem from a basic idea called the *fundamental theorem of calculus*, and they get fairly involved. Presenting the theory in such a short space isn't going to do. However, in practice, you typically never use the fundamental theorem of calculus to find integrals of functions. Instead, you memorize a few fundamental integrals and learn a few tricks. The following list highlights some of the most common integrals and solutions. A mathematical handbook will provide a more extensive list. In the list, a u represents a function of x, while v represents another function of x.

FIGURE C.7

$$\int dx = x + C$$

$$\int a\ f(x)dx = a\int f(x)dx$$

$$\int (du(x) \pm dv(x)) = \int du(x) \pm \int dv(x)$$

$$\int u\,dv = uv - \int u\,dv \text{ (integration by parts)}$$

$$\int u^n\,du = \frac{u^{n+1}}{u+1} + C(n \neq 1)$$

$$\int \frac{1}{u}du = \ln u + C$$

$$\int e^u\,dx = e^u + C$$

$$\int \ln x\,dx = x\ln x - x + C$$

$$\int \sin x\,dx = -\cos x + C$$

$$\int \cos x\,dx = \sin x + C$$

$$\int x\sin x\,dx = \sin x - x\cos x + C$$

$$\int x\cos x\,dx = \cos x + x\sin x + C$$

$$\int \sin x\cos x = \tfrac{1}{2}\sin^2 x + C$$

INDEX

1-wire interface
    interfacing with microcontrollers
    voltage level conversion in microcontrollers
1.8 V, in microcontrollers
2’s complement representation
-3-dB frequency
3.3 V, in microcontrollers
4-bit transfer, on LCDs
+5-V power supply
8-bit parallel-to-serial interface, shift registers
8-bit serial-to-parallel converter with simultaneous data transfer, shift registers
10 percent rule, connecting resistors in series and
32-bit microcontrollers
74-BiCMOS logic family
74-low-voltage logic family
74F280 9-bit parity generator/checker
74HC74 switch debouncer
74HC85 comparator
74LS147 decimal-to-BCD encoder
74LS279A switch debouncer
117-volt plug/socket
120-volt power, mains power and
240-volt power, mains power and
315 MHz modules, RF modules
433 MHz modules, RF modules
555 timers
    applications of
    astable operation
    monostable operation
    overview of
    specifications and considerations regarding
    speed control of DC motors
556 timers, 691 558 timers
4017 decade counter/divider IC
7490 4-bit ripple counter, with MOD-2 and MOD-5 sections
7491A 8-bit serial-in/serial-out shift register
7492 divide-by-12 ripple counter, with MOD-2 and MOD-6 section
7493 4-bit ripple counter, with MOD-2 and MOD-8 sections
74148 octal-to-binary encoder
74160/63 presettable 4-Bit (MOD-16) synchronous up/down counters
74164 8-bit serial-in/serial-out shift register

74165 8-bit serial-in/serial-out shift register
74184 BCD-to-binary encoder
74185 binary-to-BCD encoder
74190/91 presettable decade (BCD or MOD-10) synchronous up/down counter
74192 presettable decade (BCD or MOD-10) synchronous up/down counter
74193 presettable 4-bit (MOD-16) synchronous up/down counter
74194 universal shift register
74299 universal shift/storage register with three-state interface shift register

A

A (amperes)
    measuring current
    measuring current through a conductor
A-phase wire, of mains power
absolute error
AC-actuated switch, in illustration of relay circuits
AC (alternating current)
    ac/dc resistance ratio as function of frequency
    in capacitors
    circuits. *See* circuits, AC
    inductors
    measuring RMS voltage and currents
    motor controller
    power distribution systems and
    relays
    resistors, RMS voltage, and current
    Thevenin’s theorem in AC form
    using oscilloscope as voltmeter
ac-coupling, capacitor functions
AC waveforms
    amplitude of
    frequency and period of
    overview of
    phase of
ACC (automatic current control), laser diode drive circuits
accelerometers
    illustration of
    measuring acceleration
    measuring tilt
accuracy principal, sensors and
active crossover networks
active filters
    bandpass filter design
    high-pass filter design
    low-pass filter design
    notch filter design
    overview of

RC audio filter
active mode/region, bipolar transistors
active probes, oscilloscopes
active rectifiers, op amp applications
active region, MOSFETs
adapters
illustration of
in laboratory
ADC (analog-to-digital conversion)
comparators used in
converters
DSP (digital signal processing) and
overview of
parallel-encoded (Flash conversion)
successive approximation
addition
binary adder
binary adder/subtractor
binary arithmetic
address labels, PBASIC2 language
adjustable inductors
AF100 filter IC, from National Semiconductor
Ah (ampere-hours), battery capacity in
air, as insulator or conductor
air coil inductors
air core capacitors
air core inductors
air core transformers
alkaline batteries
comparing primary batteries
internal voltage drop
selecting
transformers and
alligator connectors
alphanumeric LCD displays
alphanumeric LED displays
illustration of
overview of
smart display
alternating current. *See* AC (alternating current)
alternators
generating AC current
magnetic induction and
Altium Designer, electronic CAD programs for use in laboratory
aluminum boxes, enclosing the circuit boards in
aluminum electrolytic capacitor
ALUs (arithmetic logic units)

binary adder
binary adder/subtractor
overview of
AM (amplitude modulated), signal detection
American Standard Code for Information Interchange (ASCII)
nonprinting characters
printing characters
American Wire Gauge (AWG)
breadboard leads and
standards for wiring
ammeters
analog
measurement errors
measuring resistance
measuring voltage, current, and resistance with
amperage, safety hazards and
Ampere, Andre-Marie
ampere-hours (Ah), battery capacity in
amperes (A)
measuring current
measuring current through a conductor
amplifiers
bipolar transistor as complementary-symmetry amplifier
bipolar transistor as differential amplifier
digital
expressing gain in decibels
impedance matching and
inverting amplifier
JFET operation in
LM386 and LM383
MOSFET applications
noninverting amplifier
operational. *See* op amps (operational amplifiers)
optocoupler applications
overview of
power amp module
preamplifiers
ramp generator with
reducing hum in
in sound sensors
amplitude
oscillators generating
of sinusoidal waveform
amplitude modulated (AM), signal detection
analog
grounds
I/O in PBASIC2 language

ICs (integrated circuits) in analog devices
illustration of analog signal
interfacing analog input with microcontrollers
interfacing with microcontrollers
multiple switches to one analog input
oscilloscope selection
analog digital, common (floating) return
analog/digital interfaces
ADC (analog-to-digital) and DAC (digital-to-analog)
analog multiplexer/demultiplexer
analog switches
analog-to-digital converters
binary-weighted DAC
DAC ICs
DAC8083A device
digital codes used by DACs
logic used to drive external loads
overview of
R/2R ladder DAC
triggering simple logic responses from analog signals
analog multimeters
analog multiplexer/demultiplexer
analog sensors
analog switches
analog thermometer ICs
analog-to-digital conversion. *See* ADC (analog-to-digital conversion)
AND logic gate
diodes logic gates
as enable gate
overview of
performing basic addition
transistor logic gates
AND-OR-Inverter (AOI) logic gates
anemometer, for measuring wind speed
anode gate, SCSs (silicon-controlled switches)
anode terminal
of battery cell
of diode
antenna rods ferrite core, comparing inductor types
antistatic mats/sprays
AOI (AND-OR-Inverter) logic gates
APC (automatic power control), laser diode drive circuits
apparent power (VA), in AC circuits
arc suppression, capacitors and
Arduino
board models
features of Arduino boards

high-power digital outputs
IDE
libraries
microcontroller evaluation boards
modules
OneWire library
open source hardware and
popularity of Arduino boards
programming ATiny85 microcontrollers with Arduino library
robot project in
shields
taking offboard
arithmetic
with binary numbers
rules for complex numbers
arithmetic logic units. *See* ALUs (arithmetic logic units)
ASCII (American Standard Code for Information Interchange)
nonprinting characters
printing characters
associate law of addition, Boolean algebra
astable mode, of 555 timers
asynchronous (ripple) counters
7490 4-bit ripple counter, with MOD-2 and MOD-5 sections
7492 Divide-by-12 ripple counter, with MOD-2 and MOD-6 section
7493 4-bit ripple counter, with MOD-2 and MOD-8 sections
overview of
ATiny85 microcontrollers
accessing ATiny options from Atmel website
minimization of external components
overview of
programming with Arduino
programming with AVR studio
ATmega328
Atmel
32-bit microcontrollers
ATiny85 microcontrollers. *See* ATiny85 microcontrollers
AVR studio development environment
microcontroller evaluation boards
attenuation
attenuators
adjustable
attenuators
overview of
audible-signal devices
audio electronics
amplifiers
applications of MOSFETs

audible-signal devices
audio modules
comparing transformer types
crossover networks
example circuits
ICs (integrated circuits) for driving speakers
ICs (integrated circuits) for modular development
illustration of audio connectors
impedance matching and
microphones
mixers
overview of
preamplifiers
sound theory
speakers
automatic current control (ACC), laser diode drive circuits
automatic power control (APC), laser diode drive circuits
automatic power-up clear (reset) circuits
autotransformers
avalanche diodes
AVR studio, programming ATiny85 microcontrollers
AWG (American Wire Gauge)
breadboard leads and
standards for wiring
axial lead resistors

B

B-phase wire, of mains power
balanced coaxial cable
Balun chokes
banana connectors
banana plug/socket
band-reject (notch) filters. *See* notch (band-reject) filters
bandpass filters
active design
illustration of
inductor applications
inductor as reactive element in
overview of
passive design
bandwidth, resonant circuits and
bare perforated construction boards, PCB options
barrier strip, spade lug and
base (gate), SCSs (silicon-controlled switches)
BASIC Stamps
BASIC Stamp I. *See* BSI (BASIC Stamp I)
BASIC Stamp II. *See* BSII (BASIC Stamp II)

batteries
backup battery
C ratings
capacity of
cells in
combining in series for greater voltage
comparing primary batteries
comparing secondary batteries
comparing with supercapacitors
discharge curves for primary batteries
discharge curves for secondary batteries
engaging/disconnecting
flashlight voltage and
ground reference
internal voltage drop
mechanisms of voltage and
overview of
primary batteries
RAM (rechargeable alkaline-manganese) batteries
recharging with solar cells
reverse-polarity protection
secondary batteries
selecting
supercapacitors
for troubleshooting use in laboratory
types of primary batteries
types of secondary batteries
voltage regulation for car batteries
battery recharger, voltage regulation and
BCD (binary-coded decimal) numbers
74190/91 presettable decade (BCD or MOD-10) synchronous up/down counter
74192 presettable decade (BCD or MOD-10) synchronous up/down counter
BCD decoder driving LCDs
BCD-to-binary encoder
conversion table
decimal-to-BCD encoder
overview of
bellows, pressure sensors
bels
Bessel filters
designing
response curve of
bias
biasing methods
in bipolar transistors
forward and reversed biasing
bidirectional motors

bilateral switches
binary adder
binary adder/subtractor
binary-coded decimal numbers. *See* BCD (binary-coded decimal) numbers
binary coded switches
binary numbers
    arithmetic with
    BCD-to-binary encoder
    binary-to-BCD encoder
    conversion table
    overview of
binary operators, PBASIC2 language
binary-weighted DAC
bipolar steppers
    decoding leads of
    H-bridge used with
    overview of
bipolar transistors
    555 timers and
    applications of
    biasing methods
    common-collector amplifier
    common-emitter amplifier
    common-emitter configuration
    Darlington pair
    emitter follower and
    high-power digital outputs on microcontrollers
    how they work
    important information regarding
    mode of operation
    overview of
    phototransistors
    pinouts for
    transistor switch
    types of
    use in logic devices
    voltage regulator
    water analogy applied to
bits, binary
blade fuses
blank copper-plated boards, PCB options
bleeder current
bleeder resistance
    bleeder resistors discharging high-voltage
    overview of
blinking LEDs
blocking, capacitors

Bluetooth modules
BNC connectors
    crimpers for
    illustration of
    oscilloscopes and
    overview of
body colors, resistor labeling schemes
body terminal lead, MOSFETs
Boolean algebra
Boolean expressions (truth tables), using with Karnaugh maps
Boolean logic
boosting
    boost inductor
    transformers
boron, in p-type silicon
bouncing, switches and
bourdon tube, pressure sensors
braided wire
branching, in PBASIC2 language
breadboards. *See* solderless breadboards
breakdown voltage
    JFETs
    MOSFETs
    thyristors
breakout boards, in modular electronics
breakover frequency, frequency compensation in op amps
bridge rectifiers
    full-wave bridge rectifier
    rectifier packages used in power supplies
    types of diode packages
bridging, impedance matching and
broad hilltop, in resonant circuits
broadband transmission line transformers
BSI (BASIC Stamp I)
    PBASIC2 language and
    programming PIC16Cx microcontrollers
BSII (BASIC Stamp II)
    connecting to host PC
    EEPROM settings
    making a robot using
    PBASIC2 language and
    power supply
    programming PIC16Cx microcontrollers
    reset circuits
bubble pushing, De Morgan’s theorem and
buck-boost converter
bucking, buck inductor

buffer (unity gain amplifier), negative feedback examples
buffers, three-state octal
bulk-metal foil resistors
Butterworth filters
    active filter low-pass value
    designing
    response curve
button batteries
buzzers, audio circuits
bypassing capacitors
    capacitor functions
    overview of
    power supply decoupling
    selecting and placing decoupling capacitors
    as transient suppressor

C

C (capacity) ratings, batteries
    Ah and mAh measures of
    charge and discharge currents in batteries
C language
    Arduino C library
    programming ATiny in
cables. *See also* wire
    crimpers for
    finding characteristic impedance of
    laboratory supplies
    matching impedance
    overview of
    VSWR (voltage standing-wave ratio) and
    weird behavior in
CAD (computer-aided design)
    advantages in circuit design
    use in laboratory
calculus
    differential
    integral
calibration, of sensors
CAM (computer-aided manufacturing)
cameras, types of sensors
CANCER (Computer Analysis of Non-Linear Circuits Excluding Radiation)
capacitance
    by cable types
    defined
    determining
    impact on switch behavior
    overview of

stray capacitance
substitution boxes
variable capacitance diode
capacitive divider, in capacitors
capacitive effect, in cables
capacitive reactance
in capacitors
in circuits with sinusoidal sources
capacitive sensors
capacitors
alternating current in
applications of
arc suppression
blocking and coupling
bypassing
capacitive divider
capacitive reactance
charge-based model of current through
commercial
comparing types of
decoupling
determining capacitance
energy in
energy loss in
expressing voltage in complex form
labeling schemes
markings on
Maxwell’s displacement current and
overview of
in parallel
quality factor (Q) and
RC Ripple filter
RC time constant
real capacitors
in series
specifications
stray capacitance
supercapacitor applications
timing and sample-and-hold circuits and
in transient circuits
types of
voltage ratings and dielectric breakdown of
water analogy for understanding
carbon-composition resistors
comparing resistor types
power ratings
tolerance ratings

carbon-film resistors
carbon zinc batteries
    comparing primary batteries
    as primary battery type
    selecting
cartridge fuses
CAT5 cable
catalogs, setting up laboratory
cathode gate, SCSs (silicon-controlled switches)
cathode-ray tubes
    aiming the beam
    in oscilloscopes
    safety hazards
cathode terminal
    of battery cell
    of diode
CD players
CD4543B CMOS BCD-to-seven-segment latch/decoder/driver
cells
    batteries composed of
    how they work
cement resistors
center frequency, filter response and
center-tap transformers
central processing unit (CPU)
ceramic core inductors
ceramic fuses
ceramic (multilayer) capacitors
ceramic (single-layer) capacitors
character entry mode, on LCDs
characteristic impedance
charge-based model, of current through capacitors
Charlieplexed LED displays
chassis (frame) ground, symbol for
Chebyshev filters
    designing
    response curve
chemical sensors
    gas detectors
    humidity measurement
    overview of
    smoke detectors
chemistry
    battery cells converting chemical energy into electrical energy
    chemical items in laboratory
    chemical reactions as voltage source in batteries
    impacting resistance

chip resistor arrays
circuit breakers. *See also* fuses/circuit breakers
    illustration of
    protecting circuits with
CircuitMaker simulator
circuits
    analyzing complex with Norton’s theorem
    analyzing linear with superposition
    audio
    automatic power-up clear (reset) circuits
    examples of relay circuits
    introduction to electronics and
    Karnaugh maps for simplifying
    modeling complex
    open and short circuits
    simulating
    types of
circuits, AC
    AC waveforms
    amplitude of AC waveforms
    combining or separating DC currents
    combining sinusoidal sources
    frequency and period of AC
    waveforms
    generating sinusoidal waveforms
    importance of
    phase of AC waveforms
    water analogy for understanding
circuits, constructing
    creating PCBs with laser printer toner transfer
    creating prototypes
    creating schematic or diagram
    enclosing the circuit
    etch-resistant pen used for creating PCB layout
    hardware used in
    list of workbench items
    overview of
    PCB options
    photo-etching PCBs with OHP transparency film
    simulator programs
    soldering/desoldering
    troubleshooting
    using PCB service
circuits, transient
    determining current flow and voltage when switch is opened
    modeling with Kirchhoff’s voltage equation
    overview of

resistors, capacitors, and inductors and
series RLC circuit and
in switch example
circuits, with periodic nonsinusoidal sources
Fourier series and
overview of
squarewave RC circuit example
circuits, with sinusoidal sources
analyzing sinusoidal circuits with complex impedances
complex impedances
inductive reactance
odd phenomena in reactive circuits
overview of
phasor diagrams
resistance, capacitive reactance, and inductive reactance
sinusoidal voltage source in complex notation
clamp circuits, overvoltage protection
Clapp LC oscillator
class-D amplifiers
clock generators
60-Hz, 10-Hz, and 1-Hz clock-pulse generator
astable multivibrators (clocks)
monostable multivibrators (one-shot)
one-shot/continuous-clock generator
clock timing
digital applications of oscilloscopes
for digital circuits
closed pair, optoisolators
CMOS
555 timers and
4017 decade counter/divider IC
analog switches
combining different logic families
family of ICs
handling unused inputs
I/O voltages and noise margins
illustration of CMOS subfamilies
logic gate ICs
speed control of DC motors
CMRR (common-mode rejection ratio)
in differential amplifiers
op am specification
CNC router, cutting PCB with
coaxial cable
characteristic impedance
electrical and magnetic fields
finding characteristic impedance of

illustration of
keeping stock of
types of
code converters
coils, relay
cold-junction temperature
collector (anode gate), SCSs (silicon-controlled switches)
Colpits oscillator
combinational devices
ALUs (arithmetic logic units)
binary adder/subtractor
binary adders
comparators and magnitude comparator ICs
demultiplexers and decoders
encoders and code converters
logic gates
microcontroller control
multiplexers and bilateral switches
parity generator/checker
combinational logic gates
commercial capacitors
commercial power supplies
common-collector amplifier
common-emitter
amplifier
bipolar transistor operation in
phototransistors and
common (floating) return, grounds and
common-mode chokes, comparing inductor types
common-mode rejection ratio (CMRR)
in differential amplifiers
op am specification
commutative law of addition, in Boolean algebra
comparators
analog/digital interfacing
applications of
combining with triangle-wave generator
with hysteresis
ICs (integrated circuits) for
overview of
supply-side comparators
voltage-level indicator
window comparators
compensated attenuators
complementary-symmetry (push-pull follower) amplifier, bipolar transistor as
complex impedances, in circuit with sinusoidal source
complex notation, sinusoidal voltage source in

complex numbers
arithmetic rules for
expressing graphically
real and imaginary part of
trigonometric form of
computer-aided design (CAD)
advantages in circuit design
use in laboratory
computer-aided manufacturing (CAM)
Computer Analysis of Non-Linear Circuits Excluding Radiation (CANCER)
computers
connecting BSII to PC
DRAM (dynamic RAM)
multifunction PC instruments
PC-based oscilloscopes
RAM (random access memory)
setting up laboratory
condenser microphones
conduction
applying voltage across a conductor
inverse relationship of conductivity to resistivity
overview of
of thermal heat
conductors
applying voltage across
energy-band structure for
resistance impacted by shape of
resistivity of
semiconductors compared with
water, air, and vacuums as
conformal coated (dipped) inductors
connectors
illustration of
items in laboratory
overview of
symbols
constants, PBASIC2 language
contact noise, resistors and
contact resistance, potentiometers
control panel, oscilloscopes
horizontal mode controls
overview of
trigger mode controls
vertical mode controls
conversion tables, number systems
copper
conduction and

current carrying capacity of copper wire
specifications of copper wire
cores
inductor
magnetic
transformer
cosine functions, trigonometry
coulomb
in equation for determining volts
measuring current in coulombs per second
counter ICs
asynchronous (ripple counter)
clock-pulse generator
with displays
overview of
synchronous
coupling
capacitor
due to spikes, lightning, and other pulses
inductor as reactive element in coupling filter
magnetic
CPU (central processing unit)
Cramer’s rule, in analyzing circuits
crimp connectors
crimpers
crossover networks
filter networks
speakers and
crowbar circuit, overvoltage protection
crystal oscillators
current. *See* electric current
current dividers
ac current divider
equations
current mirror, applications of bipolar transistors
current noise, resistors and
current probes, oscilloscopes
current ratings, logic families of ICs
current ratio, transformers
current sense inductors
current source, applications of MOSFETs
current-to-voltage converter, op amp applications
current transformers
custom-etched boards, PCB options
cutoff region, bipolar transistors
cutoff voltage
JFETs

MOSFETs

D

D-Arsonval galvanometer
D-connectors
D-Sub crimping tool
D-type flip-flops
external asynchronous control signal
overview of
quad and octal D flip-flops
stop-go indicator circuits
three-state octal latches
triggering simple logic responses from analog signals
DA (dielectric absorption)
capacitor specification
nonideal capacitor parameters
in timing and sample-and-hold circuits
DAC (digital-to-analog)
binary-weighted DAC
DAC0808 8-bit DAC
DAC8083A device
digital codes used by
DSP (digital signal processing) and
ICs (integrated circuits) for
overview of
R/2R ladder DAC
damped harmonic oscillators
dark current
Darlington pair
directional control of DC motors
photodarlington
of transistors
types of bipolar transistors
data distributors (demultiplexers)
analog
overview of
data selector, multiplexers as. *See also* multiplexers
DC-actuated switch
dc-blocking, capacitors and
DC (direct current)
ac/dc resistance ratio as function of frequency
analyzing linear dc circuits
combining sinusoidal sources (ac and dc)
converting line voltage to
oscilloscope as voltmeter
pulsating
relays

restoration with diode voltage clamp
voltage sources
dc motors
continuous
controlling direction of
controlling speed of
dc power supplies
setting up laboratory
step-down transformer for
dc working voltage (dcwv)
DCR (direct current resistance)
dcwv (dc working voltage)
DDR3 memory
DDRAMM II (double-data-rate SRAM)
De Morgan’s theorem, applying to NAND and NOR logic gates
debouncing, switches
DEBUG command, PBASIC2 language
decibels
alternative representations of
bels and
expressing gain of amplifier in
overview of use of
ripple reduction in
decimal numbers
conversion table
converting to/from binary
converting to/from sign-magnitude or 2’s complement numbers
decimal-to-BCD encoder
decoders
driving LCDs
overview of
using in combination with encoder
decoupling capacitors
capacitor functions
power supply decoupling
selecting and placing
delay distortion, filter design and
delta connection, in three-phase electricity
demultiplexers (data distributors)
analog
overview of
depletion-type JFETs
depletion-type MOSFETs
derivatives
in differential calculus
finding functions of
design labs, prototype PCBs

DesignSoft TINA Lab II
desoldering braid (solder wick)
desoldering circuits
desoldering irons
desoldering pump (solder sucker)
Deumilanove, Arduino boards
DF (dissipation factor), capacitor specification
DFRobotShopRover, Arduino variants
diacs
    overview of
    types of thyristors
diagonal wire cutters (nippers)
diagrams, for circuit design
dielectric absorption. *See* DA (dielectric absorption)
dielectric breakdown, capacitor ratings
dielectric constant
dielectric withstanding voltage (dwv)
difference amplifier, negative feedback examples
differential amplifier. *See also* op amps (operational amplifiers)
    bipolar transistor as
    how op amps work
differential calculus
differential input impedance, op am specification
differential-input voltage range, op am specification
differential probes, oscilloscopes
differentiator circuits, negative feedback examples
Digi International, XBee modules from
digital
    amplifiers
    analog/digital interfacing. *See* analog/digital interfaces
    ASCII (American Standard Code for Information Interchange)
    character standard (ASCII)
    clock generators
    clock timing
    clocks (astable multivibrators)
    combinational devices. *See* combinational devices
    counter ICs. *See* counter ICs
    digital-to-analog converters
    displays. *See* displays, digital
    grounds
    high-power outputs on microcontrollers
    ICs (integrated circuits)
    illustration of digital signal
    interfacing with microcontrollers
    logic families. *See* logic families
    logic gates. *See* logic gates
    logic states

measurement with oscilloscopes
memory devices. *See* memory devices
monostable multivibrators (one-shot)
multimeters
number systems. *See* number systems
one-shot/continuous-clock generator
oscillator using Schmitt trigger inverter
oscilloscopes
overview of
pots
powering and testing logic ICs
sensors
sequential logic. *See* sequential logic
serial vs. parallel representation
shift registers. *See* shift registers
thermometer ICs
three-state octal buffers
three-state octal latches and flip-flops
transceivers
digital signal processing (DSP)
DIMM memory modules
DIN connectors
diodes
AM (amplitude modulated) signal detection
battery backup and
characteristics of
comparing types of
for creating simple ROM
diode clamps
diode switch
half-wave and full-wave doublers
half-wave and full-wave rectifiers
laser diodes. *See* laser diodes
LEDs (light emitting diodes). *See* LEDs (light emitting diodes)
logic gates
microwave diodes
motor inductive kickback protection
overview of
PIN diodes
pn-junction (rectifier) diodes
pros/cons of rectifier circuits
for reverse-polarity protection
ROM (read-only memory) and
Schottky diode termination
specifications to consider in selecting
transient suppressions with fly-back
types of

varactor diodes
for voltage regulation
water analogy for understanding
zener diodes
DIP (dual-inline package) switches
dips, voltage regulation and
direct current. *See* DC (direct current)
direct current resistance (DCR)
directional control, of dc motors
directional microphones
discrete active devices
discrete active/passive circuits
discrete passive circuits
displays
counter ICs with
modules
safety and
displays, digital
4-bit transfer on LCDs
addressing LCD module
character entry mode on LCDs
controlling LCD module
dot-matrix LCD modules
driving LCDs
driving numeric LED displays
LCD displays
LCD pins
LED displays
multiplexed LCD displays
multiplexed LED displays
overview of
sifting LCD display
user-defined graphics on LCDs
dissipation factor (DF), capacitor specification
distributive law, Boolean algebra
diverter switches
DLC (double layer capacitor). *See* supercapacitors
DMM (digital multimeters)
doping process, silicon
dot-matrix LCD modules
dot-matrix LED displays
double-data-rate SRAM (DDRAMM II)
double layer capacitor (DLC). *See* supercapacitors
double pole double throw. *See* DPDT (double pole double throw)
double pole single throw. *See* DPST (double pole single throw)
doublers, full-wave and half-wave
doubling power, decibels and

DPDT (double pole double throw)
  directional control of DC motors
  relay
  switch
DPST (double pole single throw)
  relay
  switch
drain-current
  JFETs
  MOSFETs
DRAM (dynamic RAM)
  overview of
  RDRAM (Rambus DRAM)
  synchronous
  use in computer memory
drift velocity, applying voltage across a conductor
drive circuits, laser diodes
DS18B20 temperature sensor
DSP (digital signal processing)
dual coaxial cable
dual complementary rectifier
dual-inline package (DIP) switches
dual-location switches
dual polarity power supplies
dual rectifier, triac application
DVD players
dynamic drives, multiplexed LCDs
dynamic microphones
dynamic speakers

E
Eagle PCB
EagleCAD
  advantages of CAD tools
  electronic CAD programs for laboratory
  schematic of offboard Arduino project
earth ground
  overview of
  symbol for
EDO (extended data out), types of DRAM
EEPROM
  microcontrollers
  in PBASIC2
  language
  PIC16Cx microcontroller and
  ROM devices
  settings

types of memory in ATiny microcontrollers
EI transformer
ELC (Emitter-Coupled Logic) logic family
electret microphones
electric current
circulating current in parallel resonant circuit
combining batteries in series for greater capacity
combining or separating DC currents
comparing use by devices
current dividers
current mirrors
current sense inductors
current-to-voltage converter
current transformers
determining current flow and voltage when switch is opened
IC current ratings
measurement errors
measuring
measuring AC current
measuring RMS voltage and currents
measuring with multimeters
measuring with oscilloscope
noise
offset compensation in op amps
oscilloscope probes for measuring
overview of
pots use in adjusting load current
sensors
sources
voltage regulation and
electric field meter, JFET use in
Electric Sheep
electricity
global differences in
home wiring
power distribution systems
three-phase electricity
electrocution, safety hazards
electrolytic capacitors
electromagnetic compatibility (EMC), inductor design and
electromagnetic interference. *See* EMI (electromagnetic interference)
electromagnetism
electromagnetic induction
electromagnetic spectrum
influence of magnetic fields
overview of
electromotive force. *See* EMF (electromotive force)

electron gun, in cathode-ray tubes
electronic introduction
electronic safety tips
electronics catalogs, in laboratory
electronics laboratory. *See* laboratory
Electronics Workbench simulator
electrons
    capacitors and
    conduction and
    EMF (electromotive force) pushing free electrons to create current
    in measuring flow of current
    mechanisms of voltage and
    qualities of conductors, insulators, and semiconductors
electrostatic discharge. *See* ESD (electrostatic discharge)
Elektor PCB Service
EMC (electromagnetic compatibility), inductor design and
EMF (electromotive force)
    in electron flows
    magnetic induction and
    mechanisms of voltage and
    reverse or back EMFs and forward EMFs
    self-induced
EMI (electromagnetic interference)
    design tips for inductors
    inductor specification
    preventing
emitter (cathode gate), SCSs (silicon-controlled switches)
Emitter-Coupled Logic (ELC) logic family
emitter follower
    amplifiers
    network
    phototransistors and
enclosed power supplies
encoders
    BCD-to-binary
    binary-to-BCD
    decimal-to-BCD encoder
    octal-to-binary encoder
    overview of
energy
    in capacitors
    conservation of (1st Law of Thermodynamics)
    in inductors
    storing
enhancement-type MOSFETs
EPROM
    OTP EPROM

PIC16Cx microcontroller and
types of microcontrollers
types of ROM devices
error analysis/detection
absolute error, relative error, and percent error
multimeters and
overview of
parity generator/checker and
uncertainty estimates
error voltage, voltage regulation and
ESD (electrostatic discharge)
component susceptibility to
grounds preventing
precautions in handling components
as transient
work area protections
ESL (equivalent series inductance)
capacitor specification
nonideal capacitor parameters
ESR (equivalent series resistance)
capacitor characteristics in timing and sample-and-hold circuits
nonideal capacitor parameters
selecting coupling capacitors and
etch-resistant pen, creating PCB layout
Ethernet
Arduino boards
Arduino shields
EtherTen
Euler
evaluation/development boards. *See also* Arduino
Exclusive NOR (XNOR) logic gate
overview of
performing basic addition
working with
Exclusive OR (XOR) logic gate
overview of
performing basic addition
working with
exponents
ExpressPCB
extended data out (EDO), types of DRAM

F

F-type connectors
fanout family of ICs
fans, for dissipating excess heat
Faraday shields, in isolation transformers

Faraday’s law of induction
fast-action fuses
fast recovery diodes
ferrite beads (ferrite choke)
ferrite core inductors
ferrite toroidal transformers
FETs (field-effect transistors)
  junction field-effect transistors. *See* JFETs (junction field-effect transistors)
  metal oxide semiconductor field-effect transistors. *See* MOSFETs (metal oxide semiconductor field-effect transistors)
  mode of operation
  overview of
  unijunction FETs. *See* UJTs (unijunction transistors)
field emission, electrons in conduction and
filament resistors
filters
  active bandpass filter design
  active filter design
  active high-pass filter design
  active low-pass filter design
  active notch filter design
  bandpass
  basic
  crossover networks
  EMI/EMC design tips
  high-pass
  ICs (integrated circuits) for
  inductors for filter circuits
  line filters
  low-pass
  notch
  overview of
  passive bandpass filter design
  passive high-pass filter design
  passive low-pass filter design
  passive notch filter design
  preparation for designing
  RL high-pass
  with sharper falloff and flatter passband response
  types of
Fio, Arduino boards
flash memory
  PIC16Cx microcontroller and
  types of memory in ATiny microcontrollers
  types of ROM devices
flip-flops
  applications of bipolar transistors

D-type flip-flops
examples of SR flip-flop ICs
JK flip-flops
quad and octal D flip-flops
SR flip-flop
three-state octal latches and
timing
triggering simple logic responses from analog signals
floating (common) return, grounds and
floating (ungrounded) dc power supply
flooded lead-acid batteries
flow measurement
fluid flow sensors
fluorescent lamps
flux, soldering and
fly-back diode, transient suppressions with
FM (frequency modulation)
foil resistors
comparing resistor types
frequency response and
tolerance ratings
force-sensitive resistors
illustration of
for sensing force
force sensors. *See* movement, force, and
pressure sensors FORTRAN
forward biasing
in diodes
in pn-junction (rectifier) diodes
water analogy for understanding diodes
four-layer diodes
overview of
types of thyristors
Fourier series
in circuits with periodic nonsinusoidal sources
nonperiodic sources and
overtones and
fractional uncertainty (relative) error
frame (chassis) ground, symbol for
Franklin, Benjamin
free-electron model
free electrons
adding energy to
EMF (electromotive force) pushing
Freescale microcontroller evaluation board
frequency
ac/dc resistance ratio as function of

compensation in op amps
input impedance and
oscillators generating
output impedance and
resonant frequency
scaling with filters
of sinusoidal waveform
in sound theory
frequency counter, in laboratory setup
frequency-division relationships, in oscilloscopes
frequency modulation (FM)
frequency response
active high-pass filter design
active low-pass filters
bandpass filters
Butterworth filters
crystal oscillators and
microphone specifications
normalization of filters and
passive high-pass filter design
resistors
selecting coupling capacitors and
speakers and
full-wave doublers
full-wave rectifiers
bridge rectifier
center-tap rectifier
use in power supplies
full-wave transformers
bridge transformer
center tap transformer
function generators
general-purpose
in multifunction PC instrument
NE566 function generator
fundamental theorem of calculus
fuse resistors
fuses/circuit breakers
overview of
protecting circuits with fuses
safety tips
types of

G

gain
defined
expressing amplifier gain in decibels

gamma rays
ganged pots
gas detectors
gas-discharge lamps
gas, heat transfer within
gauge number, wires
Gauss’s law
Geiger-Müller tube, in radiation detection
General Packet Radio Service (GPRS)
general-purpose function generator
general-purpose input/output (GPIO) pins, capacitive sensing and
general-purpose resistors
generalized power law
generators
    comparing single-phase and three-phase
    generating AC current
    magnetic induction and
Gerber files, PCB design files
germanium diodes
GFCI outlets, safety tips and
GHz (gigahertz), measuring waveform frequency in
glass fuses, comparing fuse types
global positioning systems (GPSs)
Global System for Mobile Communication (GSM)
glow discharges, damaging switch contacts and generating noise
GPIO (general-purpose input/output) pins, capacitive sensing and
GPRS (General Packet Radio Service)
GPSs (global positioning systems)
Greek alphabet
ground bar, in home service panel
ground bus
ground loops, noise resulting from
ground plane
ground reference, batteries and
grounds
    analog and digital
    earth ground
    grounding error
    mains power and
    noise due to ground issues
    overview of
    shock hazard
    symbols
GSM (Global System for Mobile Communication)
Gunn diodes

H

H-bridge
    directional control of DC motors
    driving stepper motors
H (henry), as unit of inductance
half-wave bridge transformers
half-wave doublers
half-wave rectifiers
Hall effect sensor, magnetic sensors
halogen lamps
hand tools, in laboratory
hardware
    for circuit construction
    items in laboratory
    open source
harmonic analyzer
harmonics, in sound theory
Hartley oscillators
heat
    importance of heat production
    power loss due to
    thermal heat conduction and thermal resistance
heat-shrinkable tubing, laboratory supplies
heat sinks
    for dissipating excess heat
    hardware needed for circuit construction
    protecting laser diode drive circuits
    voltage regulators and
henry (H), as unit of inductance
Hertz, Heinrich
hertz (Hz), measuring waveform frequency in
hexadecimal numbers
    conversion table
    overview of
high-current chokes, comparing inductor types
high-current screw-in fuses
high frequency JFETs
high frequency (RF), types of bipolar transistors
high-pass filters
    active design
    illustration of
    inductor applications
    inductor as reactive element in
    overview of
    passive design
    RC active audio filter
    transforming low-pass filter into
high-power digital outputs

high-power LEDs (HPLEDs)
high power MOSFETs
high speed diodes
high state, digital logic states
high voltage
    shock hazard and
    voltage states in digital electronics
holes
    LEDs and
    in p-type silicon
    silicon atom and
home wiring
    overview of
    power distribution and
hop-on/hop-off resistance, potentiometers
horizontal mode controls, oscilloscopes
hotwire sensor, for flow measurement
HPLEDs (high-power LEDs)
humidity
    measuring
    sensors
hysteresis
    comparators and
    inductors and
    voltage
Hz (hertz), measuring waveform frequency in

I

I/O (input/output)
    devices
    digital applications of oscilloscopes
    in PBASIC2 language
    voltages and noise margins
I/O interfacing gold-plated fingers
I²C (TWI)
    microcontroller-to-microcontroller communication
    voltage level conversion in microcontrollers
I²R loss, power loss due to heat
IC sockets, hardware needed for circuit construction
ICL7106, for driving LCDs
ICs (integrated circuits)
    555 timer IC. *See* 555 timers
    4017 decade counter/divider IC
    analog thermometer ICs
    asynchronous counter ICs
    in circuit design
    CMOS family of

    comparators and magnitude comparator ICs
    for crystal oscillator design
    DAC (digital-to-analog) converter
    digital logic gate ICs
    digital thermometer ICs
    for driving speakers
    fanout family of
    filter ICs
    for GPS modules
    humidity sensors
    integrated optoisolators
    introduction to electronics and
    magnetic field detection
    in modular electronics
    overview of
    packages
    PCB layout
    powering and testing logic ICs
    shift register ICs
    SR flip-flop ICs
    synchronous counter ICs
    TTL family of
    voltage regulators
ICSP (in-circuit system programming)
IDC connectors
ideal capacitors
ideal voltage sources
IDEs (integrated development environments)
    Arduino
    AVR studio
imaginary part, of complex numbers
imaging devices, types of sensors
IMPATT diodes, microwave diodes
impedance
    analyzing sinusoidal circuits with complex impedances
    cables and
    characteristic impedance
    complex impedances in circuit with sinusoidal source
    input impedance
    measuring with oscilloscope
    microphone specifications
    modeling impedance of cables
    output impedance
    speakers and
impedance matching
    audio electronics and
    networks

overview of
scaling impedance with filters
techniques
in transformers
impedance ratio, transformers
impedance transformers
in-circuit system programming (ICSP)
incandescent lamps
incremental current, inductor specification
inductance (L)
by cable types
defined
impact of signals
impact on switch behavior
inductive voltage spikes
inductor specification
inductors and
measuring in henrys
mutual inductance and magnetic coupling
straight wire inductance
substitution boxes
induction
electromagnetic
magnetic
self-induction
inductive dividers
inductive reactance
in circuits with sinusoidal sources
inductors and
inductive tolerance, inductors
inductors
alternating current and
applications of
constructing
cores
deenergizing LR circuits
design tips to avoid EMI/EMC problems
electromagnetism and
energizing LR circuits
energy in
equations
expressing voltage in complex form
for filters circuits
how they work
inductance and
inductive dividers
inductive reactance

inductive voltage spikes
influence of magnetic fields
labeling schemes
nonideal inductor model
for oscillators
overview of
problems with magnetic cores
quality factor (Q) and
for radio circuits
real inductors
RL time constant and
selecting by application
self-induction
in series and in parallel
specifications
straight wire inductance
for switching regulators
symbols for
in transient circuits
types of
water analogy for understanding
inertial devices, types of sensors
infrared
LEDs (light emitting diodes)
optical sensors
photons
PIR (passive infrared) detectors
thermometers
input bias current, op am specification
input devices
input impedance
attenuators and
load resistance and
overview of
input offset current, op am specification
input offset voltage, op am specification
input/output (I/O). *See* I/O (input/output)
input resistance
insulation resistance. *See* IR (insulation resistance)
insulators
energy-band structure for
resistivity of
semiconductors compared with
water, air, and vacuums as
integral calculus
integrated circuits. *See* ICs (integrated circuits)
integrated development environments. *See* IDEs (integrated development environments)

integrator circuits, negative feedback examples
intensity (loudness), in sound theory
interconnects, wiring
interfaces, microcontroller
    1-wire serial bus
    analog inputs
    debouncing switches
    high-power digital outputs
    I2C (TWI)
    keyboard interfaces
    LED display interfaces
    level conversion
    multiple switch interfaces
    overview of
    serial interfaces
    serial peripheral interface
    single switch interfaces
    sound interfaces
intermittent (pulsating) dc
interpreters, programming PIC microcontrollers and
interrupt service routine, structure of microcontrollers and
interrupter switches
INVERT (NOT) logic gate
inverting amplifiers
    negative feedback examples
    in negative feedback op amp
    overview of
inverting comparators
    with hysteresis
    overview of
inverting voltage regulators
ionizing radiation detection
IR (insulation resistance)
    capacitor specification
    in timing and sample-and-hold circuits
iron core inductors
iron core transformers
isolation transformers
Itlead Studio, PCB service from

J
jacks, connectors and
JFETs (junction field-effect transistors)
    amplifiers
    applications of
    basic current source and basic amplifier
    how they work

light dimmer example
mode of operation
overview of
sample problems
source follower
susceptibility to ESD
technicalities of
types of JFET packages
voltage-controlled resistor
water analogy for understanding
JK flip-flops
overview of
ripple (asynchronous) counters
synchronous counters
Johnson noise resistors
Johnson shift counters
Joule heating, power loss due to heat
joules, in equation for determining volts
jumper wires, in laboratory supplies
junction field-effect transistors. *See* JFETs (junction field-effect transistors)
junction rule. *See* Kirchhoff’s current law
junction threshold voltage, diodes

K
Karnaugh maps, for simplifying circuits
keypads, interfacing with microcontrollers
kHz (kilohertz), measuring waveform frequency in
KiCad, electronic CAD program
kickback protection, with Schottky diode
kilohertz (kHz), measuring waveform frequency in
Kirchhoff’s current law
connecting resistors in parallel
constructing voltage divider
illustration of
overview of
Kirchhoff’s voltage law
connecting resistors in series
deenergizing RL circuits
determining resistance of a resistor network
energizing RL circuits
illustration of
overview of
resistors and

L
labeling schemes
capacitors

inductors
potentiometers
resistors
laboratory setup
active probes
building workbench
CAD programs
chemical items
computers
current probes
DC power supplies
design labs or project boards
differential probes
electronics catalogs
frequency counter
general-purpose function generator
hand tools
hardware items
isolation transformers
multifunction PC instruments
multimeters
oscilloscope probes
oscilloscopes
overview of
parts stockpile
passive probes
prototype PCBs
solder and flux
solder-removing tools
soldering irons
solderless breadboards
substitution boxes
test cables, connectors, and adapters
test equipment
variac (variable transformer)
wire and cable
work area in
lamps
holders
incandescent lamps
overview of
technicalities of
types of
landscape lighting, low-voltage transformer for
LASCRs (light activated SCRs)
laser diodes
drive circuits

laser pointers
mode of operatio
optical fiber and
overview of
specifications
laser pointers
laser printer toner tra
laser sensors, for flo
latches, three-state o
latching relays
latching switch
latchup protection, o
LC circuits
filtering
parallel
parallel impedanc
power in LC para
power in LC serie
resonant circuits i
resonant circuits i
series impedance
LC oscillators
crystal oscillators
overview of
LCDs (liquid-crystal
4-bit transfer on
addressing LCD r
Arduino shields
character entry m
controlling LCD r
digital electronics
dot-matrix LCD n
driving
LCD pins
multiplexed displ
overview of
sifting display
user-defined grap
LDRs (light-depende
lead-acid batteries
lead free solders
leakage current, capa
leakage flux, transfor
leakage, nonideal ca
leakage reactance, tr
lease significant bit (
LED displays

N

n-channel JFETs
  configuration options for JFETs
  how they work
  in light dimmer
  as relay driver
  specification table for
  water analogy for understanding
n-channel MOSFETs
n-type semiconductors
  how UJTs work
  overview of
  in photodiodes
  silicon and
NAND logic gate
  cross-NAND SR flip flop
  De Morgan’s theorem simplifying
  overview of
  universal capability of
Nano, Arduino boards
nanoamps, measuring current in
narrow-band design
  active bandpass filters
  notch filters
  passive bandpass filters
NC (normally closed)
  relays
  switches
NE566 function generator
negative feedback
  buffer (unity gain amplifier) example
  integrator/differentiator examples
  inverting/noninverting amplifier examples
  overview of
  summing/difference amplifier examples
  types of
negative temperature coefficient (NTC) thermistors
negative voltage regulator
.NET Gadgeteer
  modules
  plug-and-play development in
Netduino modules
neutral bar, in home service panel
neutral wire, of mains power
NI MultiSim, CAD program for laboratory use
NICad (nickel-cadmium) batteries
comparing secondary batteries

internal voltage drop
selecting
nickel alkaline batteries
NiFe (nickel-iron) batteries
NIMH (nickel metal hydride) batteries
comparing secondary batteries
internal voltage drop
selecting
nine’s complement
nippers (diagonal wire cutters)
NiZn (nickel-zinc) batteries
comparing secondary batteries
selecting
NO (normally open)
relays
switches
noise (electrical)
checking potentiometers for static noise
due to ground issues
glow discharges causing
resistors
suppression of
noise margins, I/O voltages and
noninverting amplifier
negative feedback examples
noninverting clipper amplifier
types of amplifiers
noninverting comparators
with hysteresis
overview of
nonmetals, heat transfer within
nonohmic materials
nonperiodic sources
nonsinusoidal nonrepetitive sources
nonsinusoidal periodic circuits. *See* circuits, with periodic nonsinusoidal sources
nonvolatile RAM (NOVRAM)
NOR logic gate
cross-NOR SR flip flop
De Morgan’s theorem simplifying
overview of
universal capability of
normalization
creating normalized filter
of filter response curve
normally closed (NC)
relays
switches

normally open (NO)
relays
switches
Norton’s theorem, in analyzing complex circuits
NOT (INVERT) logic gate
notch (band-reject) filters
active design
illustration of
inductor as reactive element in
overview of
passive design
NOVRAM (nonvolatile RAM)
npn bipolar transistors
directional control of DC motors
making simple current source
overview of
rules governing
SCRs (silicon-controlled rectifiers) and
as transistor switch
use as relay driver
water analogy for understanding
NTC (negative temperature coefficient) thermistors
null frequency, filter response and
number systems
2’s complement representation
arithmetic with binary numbers
BCD (binary-coded decimal) numbers
binary numbers
conversion tables
octal and hexadecimal numbers
sign-magnitude representation
numeric LED displays
numeric values, in PBASIC2 language

O

observer effect, sensors and
OC (open-collector) outputs
driving relay with OC gates
logic gates with
octal numbers
conversion table
overview of
three-state octal buffers
three-state octal latches
transceivers
octal-to-binary encoder
Octopart website, for finding electronic parts

Ohm, Georg Simon
Ohmic heating
Ohmic materials
Ohmic region
  JFETs
  MOSFETs
ohmmeters
  analog multimeter functions
  decoding stepper leads
  measurement errors
  measuring voltage, current, and resistance with
Ohm’s law
  connecting resistors in series
  constructing voltage divider
  deenergizing RL circuits
  determining resistance of a resistor network
  electron flow and
  energizing RL circuits
  modeling transient circuit with
  overview of
  resistors and
  in terms of resistivity and conductivity
ohms (Ω)
OHP transparency film, photo-etching with
oil-filled capacitors
  comparing capacitor types
  use in high-current supplies
omnidirectional microphones
one-shot/continuous-clock generator, for timing flip-flops
Op Amp oscillator
op amps (operational amplifiers)
  analog/digital interfacing
  applications of
  buffer (unity gain amplifier) example of negative feedback
  comparators generally
  comparators with hysteresis
  constructing relaxation oscillators
  frequency compensation
  how they work
  integrator/differentiator examples of negative feedback
  inverting/noninverting amplifier examples of negative feedback
  negative feedback
  overview of
  positive feedback
  powering
  practical issues with
  specifications

summing/difference amplifier examples of negative feedback
supply-side comparators
theory of
types of
voltage and current offset compensation
voltage-level indicator
water analogy for understanding
window comparators
open circuits, circuit faults
open-collector (OC) outputs
driving relay with OC gates
logic gates with
open frame power supplies
open source hardware
operational amplifiers. *See* op amps (operational amplifiers)
operational feedback
optical fiber
optical reflective sensors
optical sensors
distance-measuring devices
flow measurement
overview of
slotted sensors
smoke detectors
optocouplers
optoelectronics
lamps
laser diodes. *See* laser diodes
LEDs (light emitting diodes). *See* LEDs (light emitting diodes)
optical fiber
optoisolators
overview of
photodiodes
photons
photoresistors
photothyristors
phototransistors
solar cells
optoisolators
applications of
overview of
OR logic gate
diode logic gates
as enable gate
overview of
transistor logic gates
oscillation prevention

oscillators
crystal oscillators
illustration of simple relaxation oscillator
inductors for
LC (sinusoidal) oscillators
microcontroller oscillators
overview of
RC relaxation oscillators
UJTs (unijunction transistors) in relaxation oscillator
varactor (varicap) diode and
voltage-controlled oscillators
Wien-bridge and twin-T oscillators
oscilloscope probes
active probes
current probes
differential probes
passive probes
selecting for laboratory use
suggestions for working with
types of
oscilloscopes
aiming the beam
applications of
cathode-ray tube in
checking potentiometers for static noise
comparing phase relationship between two voltage signals
control panel layout
digital applications
horizontal mode controls
how they work
initial settings
interior circuitry of
for laboratory use
measuring current with
measuring impedances
measuring sinusoidal voltage signal with
in multifunction PC instrument
overview of
trigger mode controls
uses of
vertical mode controls
OTP (one-time programmable)
OTP EPROM
types of microcontrollers
outlets, electric
output. *See also* I/O (input/output)
devices

op amp and comparator drivers
output impedance
attenuators and
load resistance and
overview of
output resistance
output voltage swing
overvoltage protection
op amp applications
voltage regulators
oxidation-reduction reactions, battery cells converting chemical energy into electrical energy

P

p-channel JFETS
configuration options for JFETs
how they work
light dimmer
specification table for
water analogy for understanding
p-channel MOSFETs
p-n junction (rectifier) diodes
how they work
LEDs and
mode of operation
p-type semiconductors
how UJTs work
overview of
in photodiodes
silicon in
paired cable
parallel
batteries in
capacitors in
diodes in
impedance in
inductors in
resistors in
resonance
transmission
parallel circuits
combining with series circuit
LC resonant circuit
overview of
parallel RLC circuits
power in LC parallel circuit
RLC resonant circuit
parallel-encoded analog-to-digital conversion (Flash conversion)

parallel impedance, LC circuits
parallel-in/serial-out shift register
parallel shift registers
    8-bit parallel-to-serial interface
    8-bit serial-to-parallel converter with simultaneous data transfer
    16-bit serial-to-parallel converter
    serial-in/parallel-out
parallel wires, characteristic impedance
parity bits
parity generator/checker
part stockpile, for laboratory
passband, filter response and
passband ripple, filter design and
passive crossover networks, speakers and
passive filters
    bandpass filter design
    high-pass filter design
    low-pass filter design
    notch filter design
    overview of
passive infrared (PIR) detectors
passive probes, oscilloscopes
Pauli exclusion principle, in quantum physics
PBASIC2 language
    debugging
    instructions
    overview of
PCBs (printed circuit boards)
    creating
    custom-etched boards
    design tips for inductors
    etch-resistant pen used for creating PCB layout
    layout for offboard Arduino project
    PCB kits
    perforated board
    photo-etching with OHP transparency film
    pre-etched perforated boards
    using PCB service
    wire-wrap board
PCs (personal computer). *See* computers
peak current rating
    diode selection and
    diodes
peak detectors, op amp applications
peak inverse voltage (PIV) rating. *See* PIV (peak inverse voltage) rating
peak-to-peak ripple voltage
percent error, error analysis

perforated PCB boards
  overview of
  with pads and buses
period, of sinusoidal waveform
periodic nonsinusoidal circuits. *See* circuits, with periodic nonsinusoidal sources
permanent-magnet steppers
personal computer (PCs). *See* computers
PF (power factor), in AC circuits
phase
  comparing phase relationship between two voltage signals
  of sinusoidal waveform
  transformers and
phasor diagrams
  of circuit with sinusoidal source
  expressing voltage in complex form
Phillips SAA1027, stepper controller chip
phone connectors
phono (RCA) connectors
phosphorus, in n-type silicon
photo diodes
photo-etching with OHP transparency film, creating PCB layout
photoconductivity
photodarlington
photodiodes
  how they work
  overview of
  in solar cells
  specifications
  types of
photoelectric emission, electrons in conduction and
photoFETs
photons
photoresistors
  applications of
  how they work and technicalities of
  overview of
photosensors, using in conjunction with supercapacitor
photothyristors
phototransistors
  applications of
  bipolar transistors
  how they work
  overview of
  triggering logic responses
  types of
photovoltaics
  photodiodes in

as voltage source
PIC16Cx microcontrollers
making a robot using BSII
overview of
PBASIC2 language and
programming
programming with BASIC Stamps
piezoelectric effects
crystal oscillators and
detecting shock and vibration
speakers and
as voltage source
PIN diodes
pinouts, for bipolar transistors
PIR (passive infrared) detectors
PIV (peak inverse voltage) rating
diode selection and
diodes and
power supplies
PL-259 (UHF) connectors
plastic boxes, enclosing the circuit boards in
plug-and-play (PnP) prototyping
plugs, connector male ends
pnp bipolar transistors
in current mirror application
directional control of DC motors
multiple current sources and
overview of
rules governing
SCRs (silicon-controlled rectifiers) and
as transistor switch
water analogy for understanding
PnP (plug-and-play) prototyping
point voltage, measuring voltage and
polar coordinate form, of complex numbers
poles, switch
polyester film capacitors
polypropylene film capacitors
polystyrene capacitors
comparing capacitor types
use in timing and sample-and-hold circuits
PolySwitches
POS (product of sums), AND-OR-Inverter gates and
position measuring sensors
positive feedback op amps
positive temperature coefficient (PTC) thermistors
positive voltage regulator

pot cores, comparing inductor types
potential/potential difference. *See also* voltage
potentiometers (pots)
    characteristics of
    checking for static noise
    illustration of
    noise related to
    pros/cons and applications of
    rotation measurement with
    variable (potentiometers) resistors
    voltage and current offset compensation
    voltage-level indicator
powdered core inductors
powdered-iron toroidal transformer
power
    defining volts in terms of
    distribution systems
    measuring consumption of
    power control modules
    resistor power ratings
    solar cells as power source
    transformer power ratio
power booster, op amp applications
power factor (PF), in AC circuits
power-film resistors
power, in AC circuits
    apparent power, real power, and reactive power
    in LC parallel circuit
    in LC series circuit
    power factor (PF)
    reactive factor (RF)
power law, finding power dissipated as heat by resistor
power of, unit prefixes
power resistors
power stepping sequence
power supply
    BSII (BASIC Stamp II) and
    commercial packages
    constructing
    DC power supplies for use in laboratory
    decoupling
    decoupling capacitors and
    dual polarity (± 12V)
    dual polarity with variable output
    EMI/EMC design tips
    engaging/disconnecting battery power
    ICs (integrated circuits) for modular development

introduction to electronics and
laser diodes
for op amps
PCB layout and
rectifier packages
regulated +5-V supply
regulating. *See* voltage regulators
ripple reduction
selecting transformer for
step-down transformer for DC power supply
switch mode
switchers (switching power supplies)
transforming. *See* transformers
power transformers
power transistors
power-up clear (reset) circuits
pre-etched perforated PCB boards
preamplifiers
precision current source, op amp applications
precision op amps
precision principal, sensors and
precision resistors
presets (trimmers)
capacitors
variable resistors in
pressure sensors. *See* movement, force, and pressure sensors
pretinned solid bus wire
primary batteries
comparing
defined
discharge curves for
overview of
types of
probes, oscilloscope
active probes
current probes
differential probes
passive probes
selecting for laboratory use
suggestions for working with
probes, safety tips for
product of sums (POS), AND-OR-Inverter gates and
programmable-gain, op amp applications
programmable op amps
programmable ROM
programmable unijunction transistors (PUTs). *See* PUTs (programmable unijunction transistors)
programming PIC16Cx microcontrollers

project PCB boards
PROM devices
propagation delays
    digital applications of oscilloscopes
    logic families of ICs
prototypes
    creating for circuit design
    PCBs
proximity and touch sensors
    capacitive sensors
    optical distance-measuring devices
    overview of
    summary of
    touch screens
    types of sensors
    ultrasonic distance-measuring devices
Pspice simulator
PTC (positive temperature coefficient) thermistors
pullup/pulldown resistors
pulse measurement, with oscilloscope
pulse transformers
pulse-width modulation (PWM). *See* PWM (pulse-width modulation)
pulses, unwanted coupling due
push-pull follower (complementary-symmetry) amplifier, bipolar transistor as
pushbutton switches
PUTs (programmable unijunction transistors)
    RC relaxation oscillator and
    types of UJTs
PWM (pulse-width modulation)
    amplifiers
    Arduino capabilities
    digital outputs and
    remote control of DC motors and
pyrometers

Q

Q-point (quiescent point), bipolar transistors
Q (quality factor). *See* quality factor (Q)
quadratic equations
quadrature encoder
    pros/cons and applications of
    rotation measurement with
quality factor (Q)
    capacitors
    filter response and
    inductors
    resonant circuits

quantum mechanics, velocity-distribution laws
quantum physics, Pauli exclusion principle
quarter-wave section, for impedance matching
quartz crystals, crystal oscillators and
quiescent point (Q-point), bipolar transistors

R
R/2R ladder DAC
R (resistance). *See* resistance (R)
radiation detection
radio circuits, inductors for
radio frequency. *See* RF (radio frequency)
radio frequency interference (RFI)
radioisotopes, use in smoke detectors
RAM (random access memory)
    computer memory and
    DIMMs
    DRAM (dynamic RAM)
    overview of
    SRAM (static RAM)
RAM (rechargeable alkaline-manganese) batteries
    comparing secondary batteries
    selecting
Rambus DRAM (RDRAM)
ramp generator
ramp waveforms
random access memory. *See* RAM (random access memory)
RC relaxation oscillators
    overview of
    PUTs (programmable) and
    UJTs (unijunction transistors) in
RC (remote control) servos
RC (resistor-capacitor) circuits
    RC networks in arc suppression
    squarewave RC circuit
    substitution boxes
RC Ripple filters
RC time constant, capacitors and
RCA (phono) connectors
RDRAM (Rambus DRAM)
reactance
    capacitive reactance
    in circuits with sinusoidal sources
    inductive reactance
    quality factor (Q) and
reactive circuits
reactive factor (RF), in AC circuits

reactive power, in AC circuits
read-only memory. *See* ROM (read-only memory)
read/write memory (RWM). *See* RAM (random access memory)
real capacitors
real inductors
real part, of complex numbers
real power, in AC circuits
real transformers
real voltage sources
receiver circuit, applications of phototransistors
receivers, in BSII robot
rechargeable alkaline-manganese (RAM) batteries. *See* RAM (rechargeable alkaline-manganese) batteries
recirculating memory registers
rectifiers
  active rectifiers
  diode/rectifier applications
  dual rectifier
  half-wave and full-wave rectifiers
  p-n junction. *See* p-n junction (rectifier) diodes
  packages used in power supplies
  pros/cons of rectifier circuits
  silicon-controlled. *See* SCRs (silicon-controlled rectifiers)
  types of diodes
reed relays
  comparing relay types
  illustration of
reference voltage, in voltage
regulators
reflective pair, optoisolators
regulators. *See* voltage regulators
relative (fractional uncertainty) error
relaxation oscillators. *See* RC relaxation oscillators
relays
  555 timer as relay driver
  Arduino shields
  bipolar transistor as relay driver
  controlling from digital output
  examples of relay circuits
  JFET as relay driver
  light-activated
  light-activated relay
  MOSFETs as relay driver
  overview of
  solid-state
  spike protection
  types of

UJT as relay driver
using logic to drive external loads
reset circuits, BSII (BASIC Stamp II)
reset condition. *See also* SR (set-reset) flip-flops
resistance (R)
ac/dc resistance ratio as function of frequency
chemistry impacting
in circuit with sinusoidal source
conductor shape impacting
determining for a resistor network
internal voltage drop in batteries due to
measurement errors
measuring
measuring with multimeters
in Ohm's Law
Ohm's law and
quality factor (Q) and
substitution boxes
thermal heat resistance
voltage coefficient of resistance
resistance taper, potentiometer characteristics
resistive temperature detectors (RTDs)
resistive touch screens
resistivity
conductivity and
defined
of insulators, conductors, and semiconductors
thermal resistivity
resistors
in 555 timers
AC (alternating current) and
bleeder resistors
body colors
characteristics of
determining resistance of a resistor network
expressing voltage in complex form
frequency response
labeling schemes
noise and
overview of
in parallel
PCB layout
potentiometers and
power ratings
pullup/pulldown resistors
resistance and Ohm's law
in series

stability and reliability of
symbols for
TCP (temperature coefficient of resistance)
temperature ratings
thermistors
tolerance
in transient circuits
types of
variable (potentiometers) resistors
voltage and current offset compensation
voltage coefficient of resistance
voltage-controlled
voltage ratings
resolution, potentiometer characteristics
resolution principal, sensors
resonant circuits. *See also* LC circuits
bandwidth and
capacitor losses
in LC parallel circuit
overview of
in parallel RLC circuits
quality factor (Q) and
quality factor (Q) of loaded circuits
in RLC circuit
voltage drop across components in RCL resonant circuit
resonant filters
resonant frequency, inductive reactance and
reverse polarity protection
diode/rectifier applications
op amps (operational amplifiers)
reversed biasing
in diodes
in pn-junction (rectifier) diodes
water analogy for understanding diodes
RF (radio frequency)
433 and 315 MHz modules
Bluetooth modules
chokes
grounds
GSM/GPRS modules
high frequency transistor
oscillator/transmitter
PIN diodes as RF switch
XBee modules
RFI (radio frequency interference)
RG-11AU cable, characteristic impedance of
RG-58/U cable, characteristic impedance of

RGB color, in LED displays
rheostats, variable resistors in
ribbon and plane cable
ribbon cable
ring counter (shift register sequencer)
ripple counters. *See* asynchronous (ripple) counters
ripple factor
    RC Ripple filter
    reducing ripple in power supplies
rise-time, resistor
RL circuits
    deenergizing
    energizing
    inductors and
    series impedance
RL high-pass filters
RLC circuits
    bandwidth and
    circulating current in parallel resonant circuit
    critically damped
    crystal oscillators and
    modeling complex circuits
    overdamped
    parallel RLC circuits
    quality factor (Q) and
    resonant circuits in
    series RLC circuit and
    underdamped
    voltage drop across components in
RMS (root mean square) voltage
    measuring
    overview of
robot project
    with Arduino
    with BASIC stamps
ROM (read-only memory)
    devices
    diode applications
    diodes for creating simple ROM
    overview of
    programmable ROM
    size and organization
    types of microcontrollers
rotary encoders
    illustration of
    rotation measurement with
rotary switches

rotation measurement
RTDs (resistive temperature detectors)
RWM (read/write memory). *See* RAM (random access memory)

S
SAA1027 (Phillips), stepper controller chip
safety goggles
safety issues
    electronic safety tips
    ESD (electrostatic discharge) and
    overview of
    precautions in handling components
sample-and-hold circuits
    capacitors in
    op amp applications
sample voltage, in voltage regulators
saturation current, inductor specification
saturation region
    bipolar transistors
    JFETs
sawtooth waveform, RC relaxation oscillators
schematics
    for circuit design
    of offboard Arduino project
Schmitt trigger inverter
Schmitt-triggered gates
Schottky diodes
    mode of operation
    motor kickback protection with
    overview of
    termination
scopes. *See* oscilloscopes
SCRs (silicon-controlled rectifiers)
    how they work
    overview of
    overvoltage protection and
    specification table for
    speed control of DC motors
    types of
    types of thyristors
SCSs (silicon-controlled switches)
    overview of
    types of thyristors
SDRAM (synchronous DRAM)
sealed lead acid batteries. *See* SLA (sealed lead acid) batteries
secondary batteries
    comparing

defined
discharge curves for
overview of
types of
secondary emission, electrons in conduction and
security alarm, switch applications in
Seebeck effect, thermocouples and
self-inductance, by inductors
self-resonant frequency (SRF)
semiconductors
diodes. *See* diodes
energy-band structure for
integrated circuits. *See* ICs (integrated circuits)
overview of
resistivity of
silicon. *See* silicon
technology of
thyristors. *See* thyristors
transient voltage suppressors. *See* transient voltage suppressors
transistors. *See* transistors
semiprecision resistors
sensitivity, microphone specifications
sensors
calibration of
chemical sensors
GPS
interfacing with microcontrollers
light, radiation, magnetism, and sound
modules
movement, force, and pressure. *See* movement, force, and pressure sensors
observer effect
optical sensors. *See* optical sensors
overview of
photosensors
precision, accuracy, and resolution
proximity and touch. *See* proximity and touch sensors
temperature. *See* temperature sensors
types of
sequential logic
automatic power-up clear (reset) circuits
D-type flip-flops
examples of SR flip-flop ICs
JK flip-flops
overview of
pullup and pulldown resistors
quad and octal D flip-flops
SR flip-flops

switch debouncers
timing flip-flops
serial access memory
serial I/O, in PBASIC2 language
serial-in/parallel-out shift register
serial-in/serial-out shift register
serial interfaces
1-wire serial bus
8-bit parallel-to-serial interface
8-bit serial-to-parallel converter with simultaneous data transfer
I²C (TWI)
microcontrollers and
old serial interface standard
serial peripheral interface (SPI)
TTL serial
USI (Universal Serial Interface)
serial peripheral interface (SPI)
serial port adapters, in microcontrollers
serial shift registers
8-bit serial-to-parallel converter with simultaneous data transfer
16-bit serial-to-parallel converter
7491A 8-Bit serial-in/serial-out
74164 8-Bit serial-in/serial-out
74165 8-Bit serial-in/serial-out
parallel-in/serial-out
serial transmission
series
batteries in
capacitors in
combining batteries for greater voltage
diodes in
impedance in
inductors in
resistors in
series circuits
combining with parallel
LC resonant circuits
overview of
power in LC series circuit
RLC circuits
RLC resonant circuits
series impedance
in LC circuits
in RL circuits
series resonance, crystals and
service panels
connecting mains to

in home wiring
servo motors
in BSII robot
controlling
RC (remote control)
set-reset (SR) flip-flops. *See* SR (set-reset) flip-flops
set-reset (SR) latches
shape, of waveforms
Sharp GP2YOA21YK optical sensor
sharp, resonant circuits
shielded inductors
shielded twin lead cable
shielding transformers
shields, Arduino
shift register ICs
7491A 8-Bit serial-in/serial-out
74164 8-Bit serial-in/serial-out
74165 8-Bit serial-in/serial-out
74194 universal
74299 universal shift/storage register with three-state interface
shift register sequencer (ring counter)
shift registers
8-bit parallel-to-serial interface
8-bit serial-to-parallel converter with simultaneous data transfer
16-bit serial-to-parallel converter
Johnson shift counter
overview of
parallel-in/serial-out
recirculating memory registers
ring counter (shift register sequencer)
serial-in/parallel-out
serial-in/serial-out
shock hazard
short circuits
Short Message Service (SMS)
short-wave receiver
shot noise, resistor
sign-magnitude representation
signal analyzer, in multifunction PC instrument
signal diodes
silicon. *See also* microcontrollers
applications of
clarifying confusion regarding
doping process
overview of
silicon chips, in pressure sensors
silicon-controlled rectifiers. *See* SCRs (silicon-controlled rectifiers)

silver mica capacitors
silver oxide batteries
  comparing primary batteries
  internal voltage drop
  selecting
SIMM memory modules
simulation
  example, with SPICE computer program
  MultiSim circuit simulator
  programs
  programs for circuit design
sine
  trigonometry functions
  wave generator
  waveforms
single pole double throw (SPDT)
single pole multiple throw (SP(n)T)
single pole single throw. *See* SPST (single pole single throw)
single-pulse generators
  monostable multivibrators (one-shot)
  one-shot/continuous-clock generator
single switch interfaces, microcontrollers
sinusoidal voltage signal
sinusoidal waveforms. *See also* circuits, AC
  circuits with. *See* circuits, with sinusoidal sources
  combining sinusoidal sources
  overview of
skin effect
  energy loss in inductors and
  wires and cables and
SLA (sealed lead acid) batteries
  comparing secondary batteries
  internal voltage drop
  selecting
slew rate, op am specification
slotted coupler/interrupter, optoisolators
SMA connectors
  illustration of
  oscilloscopes and
small signal JFETs
small signal transformers
small signal transistors
small switching JFETs
small switching transistor
SMD (surface mounted devices)
smoke detectors
SMS (Short Message Service)

snap (microswitch) switches
snubber networks
    in arc suppression
    overview of
socket-and-bus strips, for prototype PCBs
SODIMM memory modules
solar cells
solar energy, storing without using battery
soldering
    circuits
    safety tips
    solder and flux supplies in laboratory
    tools for removing solder
soldering irons
solderless breadboards
solid core wire
solid-state relays
SOP (sum of products), AND-OR-Inverter gates and
sound
    interfacing with microcontrollers
    in PBASIC2 language
    theory
sound-activated switches
sound sensors
source follower, JFETs (junction field-effect transistors)
spade lug, barrier strip and
spade terminals
SparkFun PIR sensor
SPDT (single pole double throw)
speaker wire
speakers
    in BSII robot
    crossover networks
    ICs (integrated circuits) for driving
    impedance matching and
    overview of
spectrum analyzers, displaying Fourier transformation with
speed control, for dc motors
SPI (serial peripheral interface)
SPICE computer program
    how it works
    limitations of
    overview of
    simulation example with
spikes
    inductive voltage spikes
    protection diodes

transient suppressors for eliminating
unwanted coupling due to spikes, lightning, and other pulses
voltage regulation and
spiral coil inductor
SP(n)T (single pole multiple throw)
SPST (single pole single throw)
relays
switch debouncers
switches
square waveforms
squarewave generators
microcontroller oscillator as
VCOs (voltage-controlled oscillators)
squarewave oscillators
squarewave oscillators (clocks)
squarewave RC circuit
SR (set-reset) flip-flops
edge triggered
examples of SR flip-flop ICs
latched temperature of light alarm
level-triggered
overview of
pulse-triggered
rules for deciphering
switch debouncers
SR (set-reset) latches
SRAM (static RAM)
DDRAMM II (double-data-rate SRAM)
nonvolatile
overview of
simple example
types of memory in ATiny microcontrollers
SRF (self resonant frequency)
stall current ratings, motors
Stamps
BSI. *See* BSI (BASIC Stamp I)
BSII. *See* BSII (BASIC Stamp II)
overview of
standard twisted nematic display, LCDs (liquid-crystal displays)
standing wave pattern. *See also* VSWR (voltage standing-wave ratio)
static electricity. *See also* ESD (electrostatic discharge)
checking potentiometers for static noise
sources of voltage
static RAM (SRAM). *See* SRAM (static RAM)
steepness factor, normalized stop frequency for filters
Stefan-Boltzmann law
Steinhart-Hart equation

step-down regulators
step-down transformers
  converting line voltage to dc
  for DC power supply
  overview of
step-up regulators
stepper motors
  controlling
  driving
  identifying characteristics of
  overview of
  permanent-magnet
  translator circuit controlling drivers
  types of
  variable-reluctance
stop-band frequency, filter response and
strain gauges, for sensing force
strained hookup wire, in laboratory supplies
stranded wire
  overview of
  reducing resistance due to skin effect
stray capacitance
  capacitors and
  characteristics of real transformers
strip line cable
stubs, for impedance matching
subminiature relays
substitution boxes, setting up laboratory
substitution method, for analyzing circuits
subtraction
  binary adder/subtractor
  in binary arithmetic
successive approximation ADCs, in analog-to-digital converters
sum of products (SOP), AND-OR-Inverter gates and
summing amplifier, negative feedback examples
supercapacitors
  applications of
  comparing capacitor types
  overview of
  selecting batteries by application
superposition theorem
  analyzing linear circuits
  analyzing sinusoidal circuits with complex impedances
supertwist nematic display, LCDs (liquid-crystal displays)
supply current, op am specification
supply-side comparators
supply-side op amps

surface mounted devices (SMD)
SURFBOARDS, PCB options
surgectors
    devices used as transient suppressor
    overview of
switch debouncers
switched-capacitor filters
switchers (switching power supplies)
    overview of
    step-up, step-down, and inverting
    switch mode power supplies
    switch-mode variable power supplies
switches
    analog switches
    applications of
    bilateral switches
    debouncing
    determining current flow and voltage when switch is opened
    diode switch
    electrically activated. *See* relays example of use of transient circuits
    glow discharges damaging contacts
    how they work
    inductance and capacitance influencing behavior of
    interfacing with microcontrollers
    overview of
    sound-activated switches
    symbols for
    transistor switch
    triac application
    types of
switching
    MOSFETs
    UJTs
switching power supplies. *See* switchers (switching power supplies)
switching regulators, inductors for
symbols
    in circuit design
    ground
    inductor
    JFETs (junction field-effect transistors)
    LEDs (light emitting diodes)
    resistor
    switch
    wire and connector
synchronous counter ICs
    74160-63 presettable 4-bit (MOD-16) synchronous up/down counter
    74190/91 presettable decade (BCD or MOD-10) synchronous up/down counter

74192 presettable decade (BCD or MOD-10) synchronous up/down counter
74193 presettable 4-bit (MOD-16) synchronous up/down counter
applications of
overview of
synchronous DRAM (SDRAM)
synthesizing sound

T
T-connectors
tachometer, applications of phototransistors
tangent delta (tanδ), capacitor specification
tantalum electrolytic capacitors
TC (temperature coefficient)
capacitor specification
inductor specification
resistors
TCP (temperature coefficient of resistance), ceramic (single-layer) capacitors
Teensyduino
temperature
protecting laser diode drive circuits
relationship of resistivity and conductivity to
resistor ratings
temperature coefficient. *See* TC (temperature coefficient)
temperature coefficient of resistance (TCP), ceramic (single-layer) capacitors
temperature coefficient of resistivity (a)
temperature sensors
analog/digital interfacing
analog thermometer ICs
digital thermometer ICs
DS18B20. *See* DS18B20 temperature sensor
infrared thermometers/pyrometers
interfacing with microcontrollers
overview of
RTDs (resistive temperature detectors)
summary of
thermistors
thermocouples
types of sensors
ten’s complement
test equipment
introduction to electronics and
miscellaneous
overview of
thermal heat
conductivity
overview of
resistivity

thermal (white) noise, in resistors
thermionic emission, electrons in conduction and
thermistors
    illustration of
    overview of
    pros/cons and applications of
thermocouples
    illustration of
    pros/cons and applications of
    in temperature sensors
thermodynamics, conservation of energy and
thermoelectric effects, sources of voltage
thermometers
    analog thermometer ICs
    digital thermometer ICs
    infrared thermometers/ pyrometers
    using thermistor as
Thevenin’s theorem
    in AC form
    analyzing complex circuits
    examples applying
thick film chip resistors
    comparing resistor types
    illustration of
    power ratings
thin film chip resistors
    comparing resistor types
    illustration of
Thomson, Joseph
three-phase electricity
three-state octal buffers
three-state octal latches and flip-flops
throws, switch
thyristors
    four-layer diodes and diacs
    overview of
    SCRs (silicon-controlled rectifiers)
    solid-state relays
    triacs
    types of
tilt measurement
timbre (overtones), in sound theory
time, in PBASIC2 language
time-lag (slow-blow) fuses
timers
    555 timers
    applications of 555 timer

applications of UJTs
astable operation of 555 timer
ICs (integrated circuits) for modular development
monostable operation of 555 timer
specifications and considerations regarding 555 timer
timing circuits, capacitors in
TINA Lab II, from DesignSoft
TINAPro simulator program
tinning, soldering irons
tip connectors
TL783 voltage regulators
TMP36
analog thermometer ICs
illustration of
interfacing with microcontrollers
toggle switches
tokens, BSII (BASIC Stamp II)
tolerances
error and
resistor ratings
tone generator, audio circuits
toroidal core transformer
toroidal inductor
touch screens. *See also* proximity and touch sensors
transceivers, connecting devices that share common bus
transconductance
JFETs
MOSFETs
transducers, measurements with
transformer utilization factor, bridge rectifiers and
transformers
applications of
autotransformers
basic operations
boosting and bucking configurations
center-tap transformers
characteristics of real transformers
constructing power supplies and
converting line voltage to dc
core shapes
current ratio
gear analogy
in home wiring
for impedance matching
impedance matching and
impedance ratio
increasing/decreasing voltage with

isolation transformers
line filter and
power ratio
precautions
selecting for power supply
shielding
split supply from single transformer winding
standard and specialized types
switch mode power supplies and
variable
windings
transient voltage suppressor diodes. *See* TVSs (transient voltage suppressor diodes)
transient voltage suppressors
avalanche diodes
devices used as
for eliminating spikes
fly-back diode for
MOVs (metal oxide varistors)
overview of
PolySwitches
surgector
switching regulator supplies
TVSs (transient voltage suppressor diodes)
voltage regulators and
transistor arrays
transistor relays
transistor sockets, hardware needed for circuit construction
transistor switches
transistor-transistor logic (TTL). *See* TTL (transistor-transistor logic)
transistors
bipolar. *See* bipolar transistors
constructing relaxation oscillators
junction field-effect transistors. *See* JFETs (junction field-effect transistors)
metal oxide semiconductor field-effect transistors. *See* MOSFETs (metal oxide semiconductor field-effect transistors)
overview of
solid-state relays
unijunction transistors. *See* UJTs (unijunction transistors)
used in logic devices
translator circuits, driving stepper motors
transmitters, in BSII robot
transparent latches. *See* SR (set-reset) flip-flops
triacs
applications of
how they work
optoisolated
overview of

thyristors and
types of
triangle-wave generator
combining with comparator
VCOs (voltage-controlled oscillators)
triangle waveforms
trigger mode controls, oscilloscopes
trigonometry
complex numbers and
overview of
trimmers (presets)
capacitors
variable resistors in
truth tables (Boolean expressions), using with Karnaugh maps
TTL (transistor-transistor logic)
74194 universal shift register
combining different logic families
handling unused inputs
I/O voltages and noise margins
illustration of TTL subfamilies
logic families
logic gate ICs
TTL serial
voltage level conversion in microcontrollers
tunnel diodes, microwave diodes
TVs, safety tips
TVSs (transient voltage suppressor diodes)
applications of
devices used as transient suppressor
overview of
specifications
tweeters, frequency response of speakers and
TWI (Two-Wire Interface)
microcontroller-to-microcontroller communication
voltage level conversion in microcontrollers
twin lead cable
twin-T oscillators
twisted pair cable

U

UART (universal asynchronous receiver transmitter)
UJTs (unijunction transistors)
applications of
constructing relaxation oscillators
how they work
mode of operation
overview of

RC relaxation oscillators
speed control of DC motors
types of
use in relaxation oscillator
Ultiboard PCB Layout
ultrasonic distance-measuring devices
ultrasonic sensors, for flow measurement
ultraviolet photons
unary operators, PBASIC2 language
unbalanced coaxial cable
uncertainty
error and
uncertainty estimates in error analysis
Understanding Digital Signal Processing (Lyons)
unijunction transistors. *See* UJTs (unijunction transistors)
unipolar steppers
unity gain amplifier (buffer), negative feedback examples
unity-gain frequency, frequency compensation in op amps
universal asynchronous receiver transmitter (UART)
universal gates
Universal Serial Interface (USI), in ATiny microcontrollers
universal steppers
universal synchronous/asynchronous receiver transmitter (USART)
Uno, Arduino boards
up/down counters
74160-63 presettable 4-bit (MOD-16) synchronous up/down counter
74190/91 presettable decade (BCD or MOD-10) synchronous up/down counter
74192 presettable decade (BCD or MOD-10) synchronous up/down counter
74193 presettable 4-bit (MOD-16) synchronous up/down counter
USART (universal synchronous/asynchronous receiver transmitter)
USB boards, Arduino
USBDroid, Arduino
user-defined graphics, on LCDs
USI (Universal Serial Interface), in ATiny microcontrollers

V

V (volts)
formal definition of
in Ohm’s Law
VA (apparent power), in AC circuits
vacuum, as insulator or conductor
vacuum capacitors
valve-regulated lead acid (VRLA)
batteries
varactor (varicap) diodes
mode of operation
variable capacitance diode

variable-reluctance steppers
variable resistors. *See* potentiometers
variable transformers
Variac (variable transformer)
    overview of
    setting up laboratory
varicap (varactor) diodes
    mode of operation
    variable capacitance diode
varistors
    metal oxide. *See* MOVs (metal oxide varistors)
    multilayer. *See* MLTV (multilayer varistor)
VCOs (voltage-controlled oscillators)
Venus638FLPx GPS module
vertical mode controls, oscilloscopes
VI-322-DP, driving LCDs
vibration sensors
    illustration of
    overview of
visible light LEDs
Visual Studio, .NET Gadgeteer programmed in
volt-ohm-milliammeter. *See* VOM (volt-ohm-milliammeter)
voltage
    alternating current. *See* AC (alternating current)
    applying across a conductor
    capacitor ratings
    combining batteries in series to increase
    comparing phase relationship between two voltage signals
    constructing voltage divider
    determining current flow and voltage when switch is opened
    diodes in regulation of
    direct current. *See* DC (direct current)
    equation for
    formal definition of volt
    generating. *See* generators
    global differences in electricity supply
    half-wave voltage doubler
    in home wiring
    inductive voltage spikes
    internal voltage drop in batteries due to resistance
    level conversion in microcontrollers
    line voltage as safety hazard
    measurement errors
    measuring generally
    measuring RMS voltage and currents
    measuring sinusoidal voltage signal with oscilloscope
    measuring with multimeters

measuring with oscilloscope
mechanisms of
offset compensation in op amps
overview of
overvoltage protection
pots use in adjusting voltage levels
power distribution systems
regulating. *See* voltage regulators
resistor ratings
sensors
shifter
sinusoidal voltage source in complex notation
sources
suppressors
transforming. *See* transformers
trippler and quadrupler
water analogy for understanding
voltage coefficient of resistance
voltage-controlled oscillators (VCOs)
voltage-controlled resistors
voltage dividers
ac voltage
capacitive
constructing
equations
inductive
light-sensitive
monitoring output voltage
potentiometer as
resistors and
variable resistors as
voltage dropper, diode/rectifier applications
voltage feedback, types of negative feedback
voltage gain, op am specification
voltage-level indicator, op amps (operational amplifiers)
voltage regulators
applications of
bleeder resistors and transient
suppressors
booster
diode/rectifier applications
ICs (integrated circuits) for
overview of
overvoltage protection
ripple reduction features in
specifications
step-up, step-down, and inverting

switching power supplies
zener diodes
voltage standing-wave ratio (VSWR)
voltage states
voltage-to-current converter, op amp applications
voltmeters
analog multimeter functions
LED voltmeter
measurement errors
measuring voltage, current, and resistance with
oscilloscope as
volts (V)
formal definition of
in Ohm’s Law
VOM (volt-ohm-milliammeter). *See also* multimeters
analog
digital
overview of
VRLA (valve-regulated lead acid) batteries
VSWR (voltage standing-wave ratio)

W

wall plug-in power supplies
warbler siren, audio circuits
wattage
increasing wattage rating of zener diode
watts (W) as unit of power
waveforms
adjustable waveform clipper
enable/disable logic gates in generation of
microcontroller oscillator generating
modifier and limiter
oscillators generating shape, frequency, and amplitude of
VCOs (voltage-controlled oscillators) generating
waveforms, AC
amplitude
frequency and period of
overview of
phase of
wavelengths, LEDs (light emitting diodes)
white light, perception of photons
wide-band design
active bandpass filters
passive bandpass filters
Wien-bridge oscillators
windings
split supply from single transformer winding

transformers and
window, between reference voltages
window comparators
wire
ac/dc resistance ratio as function of frequency
applying voltage across a conductor
cables. *See* cables
connectors
gauges
laboratory supplies
overview of
skin effect and
straight wire inductance
symbols
types of
wire and plane cable
wire cutters
wire strippers
wire wound resistors
frequency response and
illustration of
power ratings
power wirewound
precision wirewound
wire-wrap PCB board
wireband chokes
woofers, frequency response of speakers and
work area
in laboratory
safety tips
work function, of metals
workbench
building
creating work area in laboratory
list of items on
wrist straps, for ESD protection

X
X-rays
XBee XRF module
xenon flash lamps
XNOR (exclusive NOR) logic gate. *See* Exclusive OR (XOR) logic gate
XOR (exclusive OR) logic gate. *See*
Exclusive NOR (XNOR) logic gate xy graphics mode, oscilloscopes

Y
Y-connection, in three-phase electricity

Z

zener diodes
    applications of
    biasing methods
    devices used as transient suppressor
    mode of operation
    overview of
    overvoltage protection and
    selecting
    voltage regulator
zero-Ohm resistors
zinc-air batteries
    comparing primary batteries
    internal voltage drop
    selecting
zinc carbon batteries
zinc-chloride batteries
    comparing primary batteries
    selecting
zinc-mercuric oxide batteries