

# BUILDING SCIENTIFIC APPARATUS

Fourth Edition

---

Unrivalled in its coverage and unique in its hands-on approach, this guide to the design and construction of scientific apparatus is essential reading for every scientist and student of engineering, and physical, chemical, and biological sciences.

Covering the physical principles governing the operation of the mechanical, optical and electronic parts of an instrument, new sections on detectors, low-temperature measurements, high-pressure apparatus, and updated engineering specifications, as well as 400 figures and tables, have been added to this edition. Data on the properties of materials and components used by manufacturers are included. Mechanical, optical, and electronic construction techniques carried out in the lab, as well as those let out to specialized shops, are also described. Step-by-step instruction supported by many detailed figures, is given for laboratory skills such as soldering electrical components, glassblowing, brazing, and polishing.

This fourth edition contains new sections on detectors, low-temperature measurements, high-pressure apparatus, updated engineering specifications for all those who bought the previous editions, and over 400 hundred figures and tables to permit specification of the components of apparatus.

JOHN H. MOORE is Professor Emeritus at the University of Maryland. He is a Fellow of the American Physical Society and the American Association for the Advancement of Science. His research has included plasma chemistry, high-energy electron scattering, and the design and fabrication of instruments for use in the laboratory and on spacecraft.

CHRISTOPHER C. DAVIS is Professor of Electrical and Computer Engineering at the University of Maryland. He is a Fellow of the Institute of Physics, and a Fellow of the Institute of Electrical and Electronics Engineers. Currently his research deals with free space optical and directional RF communication systems, plasmonics, near-field scanning optical microscopy, chemical and biological sensors, interferometry, optical systems, bioelectromagnetics, and RF dosimetry.

MICHAEL A. COPLAN is Professor and Director of the Chemical Physics Program at the University of Maryland. He is a Fellow of the American Physical Society and has research programs in space science, electron scattering, and neutron detection.

SANDRA C. GREER is Professor of Chemistry and Biochemistry and Professor of Chemical and Biomolecular Engineering at University of Maryland. She is a Fellow of the American Physical Society, and recipient of the American Chemical Society Francis P. Garvan–John M. Olin Medal. Her research deals with experimental thermodynamics and statistical mechanics of fluids and fluid mixtures, living polymers, biopolymers, and polymer solutions.



# BUILDING SCIENTIFIC APPARATUS

---

A Practical Guide to Design and Construction

**Fourth Edition**

John H. Moore ♦ Christopher C. Davis ♦ Michael A. Coplan,  
with a chapter by Sandra C. Greer

 **CAMBRIDGE**  
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,  
São Paulo, Delhi

Cambridge University Press  
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by  
Cambridge University Press, New York

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9780521878586](http://www.cambridge.org/9780521878586)

© J. Moore, C. Davis, M. Coplan, and S. Greer 2009

This publication is in copyright. Subject to statutory exception and  
to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without  
the written permission of Cambridge University Press.

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

*A catalog record for this publication is available from the British Library.*

*Library of Congress Cataloging in Publication Data*

ISBN 978-0-521-87858-6 hardback

Cambridge University Press has no responsibility for the persistence or  
accuracy of URLs for external or third-party Internet Web sites referred to  
in this publication, and does not guarantee that any content on such  
Web sites is, or will remain, accurate or appropriate.

To our families



# CONTENTS

Preface xiii

## **1 MECHANICAL DESIGN AND FABRICATION** 1

### **1.1 Tools and Shop Processes** 2

- 1.1.1 Hand Tools 2
- 1.1.2 Machines for Making Holes 2
- 1.1.3 The Lathe 4
- 1.1.4 Milling Machines 7
- 1.1.5 Electrical Discharge Machining (EDM) 9
- 1.1.6 Grinders 9
- 1.1.7 Tools for Working Sheet Metal 10
- 1.1.8 Casting 11
- 1.1.9 Tolerance and Surface Quality for Shop Processes 12

### **1.2 Properties of Materials** 13

- 1.2.1 Parameters to Specify Properties of Materials 13
- 1.2.2 Heat Treating and Cold Working 15
- 1.2.3 Effect of Stress Concentration 15

### **1.3 Materials** 18

- 1.3.1 Iron and Steel 18
- 1.3.2 Nickel Alloys 20
- 1.3.3 Copper and Copper Alloys 21
- 1.3.4 Aluminum Alloys 22
- 1.3.5 Other Metals 22
- 1.3.6 Plastics 23
- 1.3.7 Glasses and Ceramics 24

### **1.4 Joining Materials** 25

- 1.4.1 Threaded Fasteners 25
- 1.4.2 Rivets 28
- 1.4.3 Pins 29

- 1.4.4 Retaining Rings 29
- 1.4.5 Soldering 30
- 1.4.6 Brazing 31
- 1.4.7 Welding 33
- 1.4.8 Adhesives 34
- 1.4.9 Design of Joints 34
- 1.4.10 Joints in Piping and Pressure Vessels 38

### **1.5 Mechanical Drawing** 39

- 1.5.1 Drawing Tools 40
- 1.5.2 Basic Principles of Mechanical Drawing 41
- 1.5.3 Dimensions 44
- 1.5.4 Tolerances 47
- 1.5.5 From Design to Working Drawings 49

### **1.6 Physical Principles of Mechanical Design** 50

- 1.6.1 Bending of a Beam or Shaft 50
- 1.6.2 Twisting of a Shaft 53
- 1.6.3 Internal Pressure 53
- 1.6.4 Vibration of Beams and Shafts 54
- 1.6.5 Shaft Whirl and Vibration 56

### **1.7 Constrained Motion** 57

- 1.7.1 Kinematic Design 57
- 1.7.2 Plain Bearings 59
- 1.7.3 Ball Bearings 60
- 1.7.4 Linear-Motion Bearings 62
- 1.7.5 Springs 62
- 1.7.6 Flexures 64

### **Cited References** 67

### **General References** 67

### **Chapter 1 Appendix** 69

## **2** WORKING WITH GLASS 77

---

### **2.1 Properties of Glasses** 77

- 2.1.1 Chemical Composition and Chemical Properties of Some Laboratory Glasses 77
- 2.1.2 Thermal Properties of Laboratory Glasses 78
- 2.1.3 Optical Properties of Laboratory Glassware 79
- 2.1.4 Mechanical Properties of Glass 79

### **2.2 Laboratory Components Available in Glass** 79

- 2.2.1 Tubing and Rod 79
- 2.2.2 Demountable Joints 80
- 2.2.3 Valves and Stopcocks 81
- 2.2.4 Graded Glass Seals and Glass-to-Metal Seals 82

### **2.3 Laboratory Glassblowing Skills** 82

- 2.3.1 The Glassblower's Tools 82
- 2.3.2 Cutting Glass Tubing 83
- 2.3.3 Pulling Points 84
- 2.3.4 Sealing Off a Tube: The Test-Tube End 85
- 2.3.5 Making a T-Seal 86
- 2.3.6 Making a Straight Seal 88
- 2.3.7 Making a Ring Seal 88
- 2.3.8 Bending Glass Tubing 89
- 2.3.9 Annealing 89
- 2.3.10 Sealing Glass to Metal 90
- 2.3.11 Grinding and Drilling Glass 92

**Cited References** 93

**General References** 93

## **3** VACUUM TECHNOLOGY 94

---

### **3.1 Gases** 94

- 3.1.1 The Nature of the Residual Gases in a Vacuum System 94
- 3.1.2 Gas Kinetic Theory 94
- 3.1.3 Surface Collisions 96
- 3.1.4 Bulk Behavior versus Molecular Behaviour 96

### **3.2 Gas Flow** 96

- 3.2.1 Parameters for Specifying Gas Flow 96
- 3.2.2 Network Equations 97
- 3.2.3 The Master Equation 97
- 3.2.4 Conductance Formulae 98
- 3.2.5 Pumpdown Time 99

- 3.2.6 Outgassing 99

### **3.3 Pressure and Flow Measurement** 99

- 3.3.1 Mechanical Gauges 99
- 3.3.2 Thermal-Conductivity Gauges 101
- 3.3.3 Viscous-Drag Gauges 102
- 3.3.4 Ionization Gauges 102
- 3.3.5 Mass Spectrometers 104
- 3.3.6 Flowmeters 104

### **3.4 Vacuum Pumps** 105

- 3.4.1 Mechanical Pumps 106
- 3.4.2 Vapor Diffusion Pumps 110
- 3.4.3 Entrainment Pumps 113

### **3.5 Vacuum Hardware** 116

- 3.5.1 Materials 116
- 3.5.2 Demountable Vacuum Connections 119
- 3.5.3 Valves 122
- 3.5.4 Mechanical Motion in the Vacuum System 124
- 3.5.5 Traps and Baffles 126
- 3.5.6 Molecular Beams and Gas Jets 128
- 3.5.7 Electronics and Electricity *in Vacuo* 131

### **3.6 Vacuum-System Design and Construction** 133

- 3.6.1 Some Typical Vacuum Systems 134
- 3.6.2 Differential Pumping 139
- 3.6.3 The Construction of Metal Vacuum Apparatus 141
- 3.6.4 Surface Preparation 144
- 3.6.5 Leak Detection 145
- 3.6.6 Ultrahigh Vacuum 145

**Cited References** 146

**General References** 147

**Manufacturers and Suppliers** 147

## **4** OPTICAL SYSTEMS 150

---

### **4.1 Optical Terminology** 150

### **4.2 Characterization and Analysis of Optical Systems** 152

- 4.2.1 Simple Reflection and Refraction Analysis 153
- 4.2.2 Paraxial-Ray Analysis 154
- 4.2.3 Nonimaging Light Collectors 165
- 4.2.4 Imaging Systems 165
- 4.2.5 Exact Ray Tracing and Aberrations 168
- 4.2.6 The Use of Impedances in Optics 176
- 4.2.7 Gaussian Beams 182



**4.3 Optical Components** 185

- 4.3.1 Mirrors 185
- 4.3.2 Windows 190
- 4.3.3 Lenses and Lens Systems 190
- 4.3.4 Prisms 199
- 4.3.5 Diffraction Gratings 204
- 4.3.6 Polarizers 207
- 4.3.7 Optical Isolators 213
- 4.3.8 Filters 214
- 4.3.9 Fiber Optics 219
- 4.3.10 Precision Mechanical Movement Systems 222
- 4.3.11 Devices for Positional and Orientational Adjustment of Optical Components 224
- 4.3.12 Optical Tables and Vibration Isolation 229
- 4.3.13 Alignment of Optical Systems 232
- 4.3.14 Mounting Optical Components 232
- 4.3.15 Cleaning Optical Components 234

**4.4 Optical Materials** 238

- 4.4.1 Materials for Windows, Lenses, and Prisms 239
- 4.4.2 Materials for Mirrors and Diffraction Gratings 248

**4.5 Optical Sources** 251

- 4.5.1 Coherence 251
- 4.5.2 Radiometry: Units and Definitions 252
- 4.5.3 Photometry 253
- 4.5.4 Line Sources 254
- 4.5.5 Continuum Sources 255

**4.6 Lasers** 265

- 4.6.1 General Principles of Laser Operation 270
- 4.6.2 General Features of Laser Design 271
- 4.6.3 Specific Laser Systems 273
- 4.6.4 Laser Radiation 286
- 4.6.5 Coupling Light from a Source to an Aperture 287
- 4.6.6 Optical Modulators 289
- 4.6.7 How to Work Safely with Light Sources 291

**4.7 Optical Dispersing Instruments** 293

- 4.7.1 Comparison of Prism and Grating Spectrometers 296
- 4.7.2 Design of Spectrometers and Spectrographs 298
- 4.7.3 Calibration of Spectrometers and Spectrographs 302
- 4.7.4 Fabry–Perot Interferometers and Etalons 302
- 4.7.5 Design Considerations for Fabry–Perot Systems 310
- 4.7.6 Double-Beam Interferometers 311

**Endnotes** 316**Cited References** 316**General References** 321**5****CHARGED-PARTICLE OPTICS** 327**5.1 Basic Concepts of Charged-Particle Optics** 327

- 5.1.1 Brightness 327
- 5.1.2 Snell's Law 328
- 5.1.3 The Helmholtz–Lagrange Law 328
- 5.1.4 Vignetting 329

**5.2 Electrostatic Lenses** 329

- 5.2.1 Geometrical Optics of Thick Lenses 329
- 5.2.2 Cylinder Lenses 331
- 5.2.3 Aperture Lenses 334
- 5.2.4 Matrix Methods 335
- 5.2.5 Aberrations 336
- 5.2.6 Lens Design Example 339
- 5.2.7 Computer Simulations 340

**5.3 Charged-Particle Sources** 341

- 5.3.1 Electron Guns 342
- 5.3.2 Electron-Gun Design Example 343
- 5.3.3 Ion Sources 345

**5.4 Energy Analyzers** 347

- 5.4.1 Parallel-Plate Analyzers 348
- 5.4.2 Cylindrical Analyzers 349
- 5.4.3 Spherical Analyzers 351
- 5.4.4 Preretardation 352
- 5.4.5 The Energy-Add Lens 353
- 5.4.6 Fringing-Field Correction 354
- 5.4.7 Magnetic Energy Analyzers 355

**5.5 Mass Analyzers** 356

- 5.5.1 Magnetic Sector Mass Analyzers 356
- 5.5.2 Wien Filter 356
- 5.5.3 Dynamic Mass Spectrometers 357

**5.6 Electron- and Ion-Beam Devices: Construction** 357

- 5.6.1 Vacuum Requirements 357
- 5.6.2 Materials 358
- 5.6.3 Lens and Lens-Mount Design 359
- 5.6.4 Charged-Particle Detection 360
- 5.6.5 Magnetic-Field Control 361

**Cited References** 363

**6 ELECTRONICS** 365**6.1 Preliminaries** 365

- 6.1.1 Circuit Theory 365
- 6.1.2 Circuit Analysis 368
- 6.1.3 High-Pass and Low Pass Circuits 370
- 6.1.4 Resonant Circuits 374
- 6.1.5 The Laplace-Transform Method 376
- 6.1.6 RLC Circuits 379
- 6.1.7 Transient Response of Resonant Circuits 380
- 6.1.8 Transformers and Mutual Inductance 382
- 6.1.9 Compensation 382
- 6.1.10 Filters 383
- 6.1.11 Computer-Aided Circuit Analysis 384

**6.2 Passive Components** 384

- 6.2.1 Fixed Resistors and Capacitors 385
- 6.2.2 Variable Resistors 387
- 6.2.3 Transmission Lines 387
- 6.2.4 Coaxial Connectors 398
- 6.2.5 Relays 402

**6.3 Active Components** 403

- 6.3.1 Diodes 403
- 6.3.2 Transistors 406
- 6.3.3 Silicon-Controlled Rectifiers 420
- 6.3.4 Unijunction Transistors 421
- 6.3.5 Thyratrons 422

**6.4 Amplifiers and Pulse Electronics** 423

- 6.4.1 Definition of Terms 423
- 6.4.2 General Transistor-Amplifier Operating Principles 426
- 6.4.3 Operational-Amplifier Circuit Analysis 430
- 6.4.4 Instrumentation and Isolation Amplifiers 432
- 6.4.5 Stability and Oscillators 435
- 6.4.6 Detecting and Processing Pulses 436

**6.5 Power Supplies** 443

- 6.5.1 Power-Supply Specifications 443
- 6.5.2 Regulator Circuits and Programmable Power Supplies 445
- 6.5.3 Bridges 448

**6.6 Digital Electronics** 449

- 6.6.1 Binary Counting 449
- 6.6.2 Elementary Functions 449
- 6.6.3 Boolean Algebra 449

- 6.6.4 Arithmetic Units 451
- 6.6.5 Data Units 452
- 6.6.6 Dynamic Systems 452
- 6.6.7 Digital-to-Analog Conversion 452
- 6.6.8 Memories 460
- 6.6.9 Logic and Function 462
- 6.6.10 Implementing Logic Functions 465

**6.7 Data Acquisition** 467

- 6.7.1 Data Rates 467
- 6.7.2 Voltage Levels and Timing 471
- 6.7.3 Format 471
- 6.7.4 System Overhead 472
- 6.7.5 Analog Input Signals 474
- 6.7.6 Multiple Signal Sources: Data Loggers 476
- 6.7.7 Standardized Data-Acquisition Systems 476
- 6.7.8 Control Systems 479
- 6.7.9 Personal Computer (PC) Control of Experiments 484

**6.8 Extraction of Signal from Noise** 494

- 6.8.1 Signal-to-Noise Ratio 494
- 6.8.2 Optimizing the Signal-to-Noise Ratio 495
- 6.8.3 The Lock-In Amplifier and Gated Integrator or Boxcar 496
- 6.8.4 Signal Averaging 497
- 6.8.5 Waveform Recovery 497
- 6.8.6 Coincidence and Time-Correlation Techniques 499

**6.9 Grounds and Grounding** 503

- 6.9.1 Electrical Grounds and Safety 503
- 6.9.2 Electrical Pickup: Capacitive Effects 506
- 6.9.3 Electrical Pickup: Inductive Effects 507
- 6.9.4 Electromagnetic Interference and r.f.i 508
- 6.9.5 Power-Line-Coupled Noise 508
- 6.9.6 Ground Loops 509

**6.10 Hardware and Construction** 510

- 6.10.1 Circuit Diagrams 511
- 6.10.2 Component Selection and Construction Techniques 511
- 6.10.3 Printed Circuit Boards 518
- 6.10.4 Wire Wrap™ Boards 524
- 6.10.5 Wires and Cables 526
- 6.10.6 Connectors 531

**6.11 Troubleshooting** 537

- 6.11.1 General Procedures 537
- 6.11.2 Identifying Parts 540

**Cited References** 540**General References** 541**Chapter 6 Appendix** 544

## 7 DETECTORS 550

- 7.1 Optical Detectors** 550
- 7.2 Noise in the Optical Detection Process** 551
  - 7.2.1 Shot Noise 551
  - 7.2.2 Johnson Noise 552
  - 7.2.3 Generation-Recombination (gr) Noise 552
  - 7.2.4 1/f Noise 552
- 7.3 Figures of Merit for Detectors** 553
  - 7.3.1 Noise-Equivalent Power 553
  - 7.3.2 Delectivity 553
  - 7.3.3 Responsivity 555
  - 7.3.4 Quantum Efficiency 555
  - 7.3.5 Frequency Response and Time Constant 555
  - 7.3.6 Signal-to-Noise Ratio 555
- 7.4 Photoemissive Detectors** 557
  - 7.4.1 Vacuum Photodiodes 558
  - 7.4.2 Photomultipliers 558
  - 7.4.3 Photocathode and Dynode Materials 562
  - 7.4.4 Practical Operating Considerations for Photomultiplier Tubes 564
- 7.5 Photoconductive Detectors** 570
- 7.6 Photovoltaic Detectors (Photodiodes)** 575
  - 7.6.1 Avalanche Photodiodes 578
  - 7.6.2 Geiger Mode Avalanche Photodetectors 580
- 7.7 Detector Arrays** 580
  - 7.7.1 Reticons 580
  - 7.7.2 Quadrant Detectors 582
  - 7.7.3 Lateral Effect Photodetectors 582
  - 7.7.4 Imaging Arrays 583
  - 7.7.5 Image Intensifiers 585
- 7.8 Signal-to-Noise Ratio Calculations** 585
  - 7.8.1 Photomultipliers 585
  - 7.8.2 Direct Detection with  $p-i-n$  Photodiodes 587
  - 7.8.3 Direct Detection with APDs 588
  - 7.8.4 Photon Counting 589

## 7.9 Particle and Ionizing Radiation Detectors 589

- 7.9.1 Solid-State Detectors 593
- 7.9.2 Scintillation Counters 596
- 7.9.3 X-Ray Detectors 596

## 7.10 Thermal Detectors 596

- 7.10.1 Thermopiles 597
- 7.10.2 Pyroelectric Detectors 598
- 7.10.3 Bolometers 599
- 7.10.4 The Goly Cell 600

## 7.11 Electronics to be Used With Detectors 600

## 7.12 Detector Calibration 601

**Endnotes** 602

**Cited References** 602

**General References** 604

## 8 MEASUREMENT AND CONTROL OF TEMPERATURE 605

### 8.1 The Measurement of Temperature 605

- 8.1.1 Expansion Thermometers 606
- 8.1.2 Thermocouples 607
- 8.1.3 Resistance Thermometers 610
- 8.1.4 Semiconductor Thermometers 614
- 8.1.5 Temperatures Very low: Cryogenic Thermometry 615
- 8.1.6 Temperatures Very High 617
- 8.1.7 New, Evolving, and Specialized Thermometry 618
- 8.1.8 Comparison of Main Categories of Thermometers 618
- 8.1.9 Thermometer Calibration 618

### 8.2 The Control of Temperature 619

- 8.2.1 Temperature Control at Fixed Temperatures 619
- 8.2.2 Temperature Control at Variable Temperatures 619

**Cited References** 626

**General References** 629



## PREFACE

---

*Building Scientific Apparatus* provides an overview of the physical principles that one must grasp to make useful and creative decisions in the design of scientific apparatus. We also describe skills, such as mechanical drawing, circuit analysis, and optical ray-tracing and matrix methods that are required to design an instrument. A large part of the text is devoted to components. For each class of components – electrical, optical, thermal and so on – the parameters used by manufacturers to specify their products are defined and discussed. Useful materials and components such as infrared detectors, metal alloys, optical materials, and operational amplifiers are discussed, and examples and performance specifications are given. Of course, having designed an apparatus and chosen the necessary components, one must *build* it. We deal in considerable detail with basic laboratory skills: soldering electrical components, glassblowing, brazing, polishing, and so on. Described in lesser detail are operations such as lathe turning, milling, casting, laser cutting, and printed-circuit production, which one might let out to an outside shop. Understanding the capabilities and limitations of shop processes is necessary to fully exploit them in designing and building an instrument. Overall, we recognize that there are many engineering and technical texts that cover every aspect of instrument design; our goal in *Building Scientific Apparatus* has been to winnow the available information down to the essentials required for practical work by the designer and builder of scientific instruments.

In the 25 years since *Building Scientific Apparatus* first appeared, there have been profound changes in the way in which unique apparatus is assembled, as well as in the

scientists who require unique apparatus. Twenty-five years ago the technology discussed in the first edition – optics, electronics, charged-particle beams, and vacuum systems – was generally the purview of engineers. Today sophisticated apparatus is conceived of and built by chemists, physicists, and biologists, as well as medical and social scientists. In the past, microprocessors were a component that needed to be programmed and wired into an apparatus. Now computer controls are an integral part of devices such as power supplies, pressure gauges, and machine tools. Similarly, lasers were once assembled in the lab from individual optical components; today a complete laser is itself a component, less expensive than the elements from which it is assembled. The fourth edition of *Building Scientific Apparatus* recognizes this evolution in the complexity of available components and the concomitant implications for the instrument designer.

We recognize the importance that the World Wide Web has acquired as a resource for research. Throughout the text we mention particular suppliers of materials and components. There are certainly many that have escaped our attention. The suppliers of equipment, devices, components, and materials along with specifications, availability, and cost can be readily located using a suitable search engine.

This book was written with the goal of contributing to the quality and functionality of apparatus designed and built for research in disciplines from the engineering and physical sciences to the life sciences. We welcome comments; many suggestions from the past have been incorporated in the current text.



# MECHANICAL DESIGN AND FABRICATION

---

Every scientific apparatus requires a mechanical structure, even a device that is fundamentally electronic or optical in nature. The design of this structure determines to a large extent the usefulness of the apparatus. It follows that a successful scientist must acquire many of the skills of the mechanical engineer in order to proceed rapidly with an experimental investigation.

The designer of research apparatus must strike a balance between the makeshift and the permanent. Too little initial consideration of the expected performance of a machine may frustrate all attempts to get data. Too much time spent planning can also be an error, since the performance of a research apparatus is not entirely predictable. A new machine must be built and operated before all the shortcomings in its design are apparent.

The function of a machine should be specified in some detail before design work begins. One must be realistic in specifying the job of a particular device. The introduction of too much flexibility can hamper a machine in the performance of its primary function. On the other hand, it may be useful to allow space in an initial design for anticipated modifications. Problems of assembly and disassembly should be considered at the outset, since research equipment rarely functions properly at first and often must be taken apart and reassembled repeatedly.

Make a habit of studying the design and operation of machines. Learn to visualize in three dimensions the size and positions of the parts of an instrument in relation to one another.

Before beginning a design, learn what has been done before. It is a good idea to build and maintain a library of commercial catalogs in order to be familiar with what is

available from outside sources. Too many scientific designers waste time and money on the reinvention of the wheel and the screw. Use nonstandard parts only when their advantages justify the great cost of one-off construction in comparison with mass production. Consider modifications of a design that will permit the use of standardized parts. An evening spent leafing through the catalog of one of the major tool and hardware suppliers can be remarkably educational – catalogs from McMaster-Carr or W. M. Berg, for example, each list over 200 000 standard fasteners, bearings, gears, mechanical and electrical parts, tools etc.

Become aware of the available range of commercial services. In most big cities, specialty job shops perform such operations as casting, plating, and heat-treating inexpensively. In many cases it is cheaper to have others provide these services rather than attempt them for oneself. Some of the thousands of suppliers of useful services, as well as manufacturers of useful materials, are noted throughout the text.

In the following sections we discuss the properties of materials and the means of joining materials to create a machine. The physical principles of mechanical design are presented. These deal primarily with controlling the motion of one part of a machine with respect to another, both where motion is desirable and where it is not. There are also sections on machine tools and on mechanical drawing. The former is mainly intended to provide enough information to enable the scientist to make intelligent use of the services of a machine shop. The latter is presented in sufficient detail to allow effective communication with people in the shop.

## 1.1 TOOLS AND SHOP PROCESSES

A scientist must be able to make proper use of hand tools to assemble and modify research apparatus. A successful experimentalist should be able to perform elementary operations safely with a drill press, lathe, and milling machine in order to make or modify simple components. Even when a scientist works with instruments that are fabricated and maintained by research technicians and machinists, an elementary knowledge of machine-tool operations will allow the design of apparatus that can be constructed with efficiency and at reasonable cost. The following is intended to familiarize the reader with the capabilities of various tools. Skill with machine tools is best acquired under the supervision of a competent machinist.

### 1.1.1 Hand Tools

A selection of hand tools for the laboratory is given in Table 1.1. A research scientist in physics or chemistry will have use for most of these tools, and if possible should have the entire set in the lab. The tool set outlined in Table 1.1 is not too expensive for any scientist to have on hand.

A laboratory scientist should adopt a craftsman-like attitude toward tools. Far less time is required to find and use the proper tool for a job than will be required to repair the damage resulting from using the wrong one.

### 1.1.2 Machines for Making Holes

Holes up to about 25 mm (1 in.) diameter are made using a *twist drill* (Figure 1.1) in a drill press. A *boring bar*

---

**Table 1.1 Tool Set for Laboratory Use**

---

#### Screwdrivers:

- No. 1, 2, and 3 drivers for slotted-head screws
- No. 1, 2, and 3 Phillips screwdrivers
- Allen (hex) drivers for socket-head screws, both a fractional set (1/16–1/4 in.) and a metric set (1.5–10 mm)
- Nut drivers for hex-head screws and nuts, both fractional (1/8–1/2-in.) and metric (4–11 mm)
- Set of jeweller's screwdrivers

#### Wrenches:

- Combination box and open-end wrenches, both a fractional set (3/8–1 in.) and a metric set (10–19 mm)
- 3/8 in. square-drive ratcheting socket driver with both fractional (3/8–7/8 in.) and metric (10–19 mm) socket sets
- Adjustable wrenches (small, medium, and large)
- Pipe wrench

#### Pliers:

- Slip-joint pliers
- Channel-locking pliers
- Large and small needle-nose pliers
- Large and small diagonal cutters
- Small flush cutters
- Hemostats

#### Hammers:

- Small and medium ball-peen hammers
- Soft-faced hammer with plastic or rubber inserts

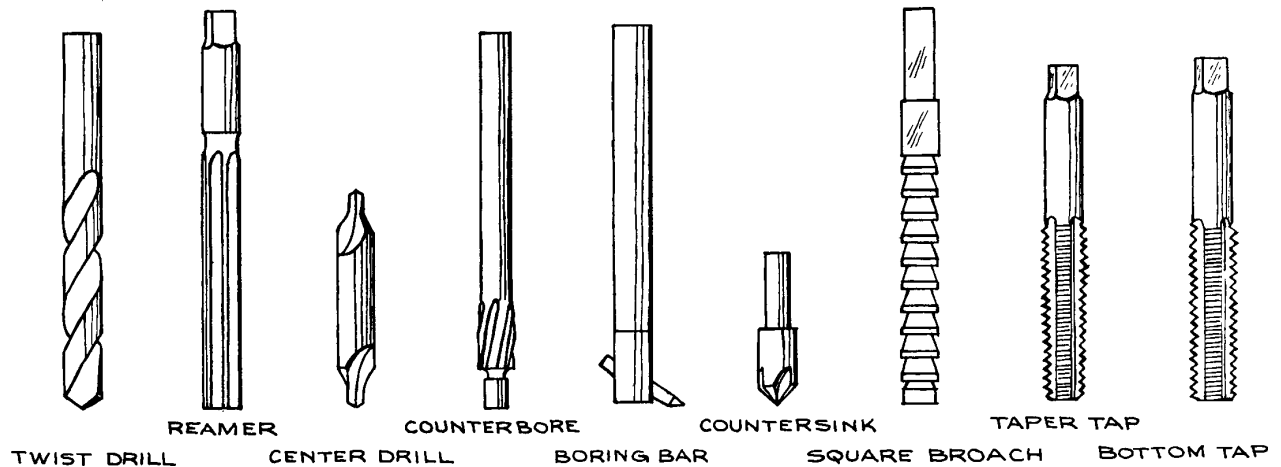
#### Files:

- Second-cut, flat, half-round, and round files with handles
- Smooth-cut, flat, half-round, and round files with handles
- Six-piece Swiss-Pattern file set

#### Miscellaneous:

- Dial caliper or micrometer
  - Forceps
  - Sheet-metal shears
  - Hacksaw
  - Tubing cutter
  - Center punch
  - Scriber
  - Small machinist's square
  - Steel scale
  - Divider
  - Tapered hand reamer
  - Tap wrench with:
    - 4-40 to 1/4-20 UNC and 3 × 0.6 to 12 × 1.75 metric tap sets
    - 1/8 to 1/2 NPT or 6 mm to 15 mm BPST (metric) pipe thread tap sets
  - Electric hand drill motor or small drill press
  - Drills, 1/16–1/2 in. in 1/32 in. increments, in drill index
  - Drills, Nos. 1–60, in drill index
  - Small bench vise
-





**Figure 1.1** Tools for making and shaping holes.

(Figure 1.1) is used in a lathe or vertical milling machine to bore out a drilled hole to make a large hole. Of course, a hole can be drilled with a twist drill in a handheld drill motor; this method, although convenient, is not very accurate and should only be employed when it is not possible to mount the work on the drill press table.

Twist drills are available in fractional inch sizes and metric sizes as well as in number and letter series of sizes at intervals of only a few thousandths of an inch. Sizes designated by common fractions are available in 1/64 in. increments in diameters from 1/64 to 1 3/4 in., in 1/32 in. increments in diameters from 1 3/4 to 2 1/4 in., and in 1/16 in. increments in diameters from 2 1/4 to 3 1/2 in.; metric sizes are available in 0.05 mm increments in diameters from 1.00 mm to 2.50 mm, in 0.10 mm increments from 2.50 mm to 10.00 mm, and in 0.50 mm increments from 10.00 mm to 17.50 mm. Number drill sizes are given in Appendix 1.1. The included angle at the point of a drill is 118°. A designer should always choose a hole size that can be drilled with a standard-size drill, and the shape of the bottom of a blind hole should be taken to be that left by a standard drill unless another shape is absolutely necessary.

If many holes of the same size are to be drilled, it may be worthwhile to alter the drill point to provide the best performance in the material that is being drilled. In very hard materials the included angle of the point should be increased to as much as 140°. For soft materials such as plastic or fiber

it should be decreased to about 90°. Many shops maintain a set of drills with points specially ground for drilling in brass. The included angle of such a drill is 118°, but the cutting edge is ground so that its face is parallel to the axis of the drill in order to prevent the drill from digging in.

A drilled hole can be located to within about 0.3 mm (0.01 in.) by scribing two intersecting lines and making a punch mark at the intersection. The indentation made by the punch holds the drill point in place until the cutting edges first engage the material to be drilled. With care, locational accuracy of 0.03 mm (.001 in.) can be achieved in a milling machine or jig borer. Locational error is primarily a result of the drill's flexing as it first enters the material being drilled. This causes the point of the drill to wander off the center of rotation of the machine driving the drill; the hole should be started with a *center drill* (Figure 1.1) that is short and stiff. Once the hole is started, drilling is completed with the chosen twist drill.

A drill tends to produce a hole that is out-of-round and oversize by as much as 0.2 mm (.005 in.). Also, a drill point tends to deviate from a straight line as it moves through the material being drilled. This run-out can amount to 0.2 mm (.008 in.) for a 6 mm (1/4 in.) drill making a 25 mm (1 in.) deep hole; more for a smaller diameter drill. It is particularly difficult to make a round hole when drilling material that is so thin that the drill point breaks out on the under side before the shoulder

enters the upper side. Clamping the work to a backup block of similar material alleviates the problem. When roundness and diameter tolerances are important, it is good practice to drill a hole slightly undersize and finish up with the correct size drill; better yet, the undersized hole can be accurately sized using a *reamer*.

Before drilling in a drill press, the location of the hole should be center-punched and the work should be securely clamped to the drill-press table. The drill should enter perpendicular to the work surface. When drilling curved or canted surfaces, it is best to mill a flat, perpendicular to the hole axis at the location of the hole.

The speed at which the drill turns is determined by the maximum allowable surface speed at the outer edge of the bit as well as the rate at which the drill is fed into the work. The rate at which a tool cuts is typically specified as meters per minute (m/min) or surface feet per minute (sfpm). Suggested tool speeds are given in Table 1.2. A drill (or any cutting tool) should be cooled and lubricated by flooding with soluble cutting oil, kerosene, or other cutting fluid. Brass or aluminum can be drilled without cutting oil if necessary.

A drilled hole that must be round and straight to close tolerances is drilled slightly undersize and then reamed using a tool such as is shown in Figure 1.1. Reamers with a round shank are meant to be grasped in the collet chuck of a milling machine; the reamer inserted after the drill is removed from the chuck without moving the work-piece on the bed of the milling machine. A reamer with a square shank is to be grasped in a tap handle for use by hand. A hand reamer has a slight initial taper to facilitate starting the cut. The diameter tolerance on a reamed hole can be 0.03 mm (.001 in.) or

better. The chamfer (taper) tolerance can be kept to 0.0002 millimeter per millimeter (or 8 microinch per inch) or better. Tapered drill and reamer sets are available for preparing the tapered holes for standard taper pins used to secure one part to another with great and repeatable precision.

A drilled hole can be threaded with a *tap* (shown in Figure 1.1). Cutting threads with a tap is usually carried out by hand. A tap has a square shank that is clamped in a tap handle. The tap is inserted in the hole and slowly turned, cutting as it goes. The tool should be lubricated and should be backed at least part way out of the hole after each full turn of cutting in order to clear metal chips from the tool. Taps are chamfered (tapered) on the end so that the first few teeth do not cut full depth. This makes for smoother cutting and better alignment. For more precise tapping the tap can be placed in a drill press with the work piece held underneath. The drill chuck can be rotated by hand to start the tap off correctly, parallel to the hole. Some drill presses come with a foot-operated reversing mechanism so that with the drill operating at a slow speed the correct action of tap–reverse–tap can be carried out. The chamfer on a tap extends for nearly 10 teeth in a *taper tap*, 3 to 5 teeth on a *plug tap*, and 1 to 2 teeth on a *bottoming tap*. The first two are intended for threading through a hole, the latter for finishing the threads in a blind hole. The hole to be threaded is drilled with a tap drill with a diameter specified to allow the tap to cut threads to about 75% of full depth. Appendix 1.1 gives tap drill sizes for American National and metric threads.

The head of a bolt can be recessed by enlarging the entrance of the bolt hole with a *counterbore* (shown in Figure 1.1).

A keyway slot can be added to a drilled hole or a drilled hole can be made square or hexagonal by shaping the hole with a *broach* (Figure 1.1). A broach is a cutting tool with a series of teeth of the desired shape, each successive cutting edge slightly larger than the one preceding. The broach can be driven through the hole by a hand-driven or hydraulic press. In some broaching machines the tool is pulled through the work. A broach can, at some expense, be ground to a nonstandard shape. The expense is probably only justified if many holes are to be broached.

### 1.1.3 The Lathe

A lathe (Figure 1.2) is used to produce a surface of revolution such as a cylindrical or conical surface. The work to

**Table 1.2 Tool Speeds for High-speed-steel Tools**  
(Speeds can be increased  $2 \times$  with carbide-tipped tools)

Material	m/min (sfpm)		
	Drill	Lathe	Mill
Aluminum	60 (200)	100 (300)	120 (400)
Brass	60 (200)	50 (150)	60 (200)
Cast iron	30 (100)	15 (50)	15 (50)
Carbon steel	25 (80)	30 (100)	20 (60)
Stainless steel	10 (30)	30 (100)	20 (60)
Copper	60 (200)	100 (300)	30 (100)
Plastics	30 (100)	60 (200)	60 (200)

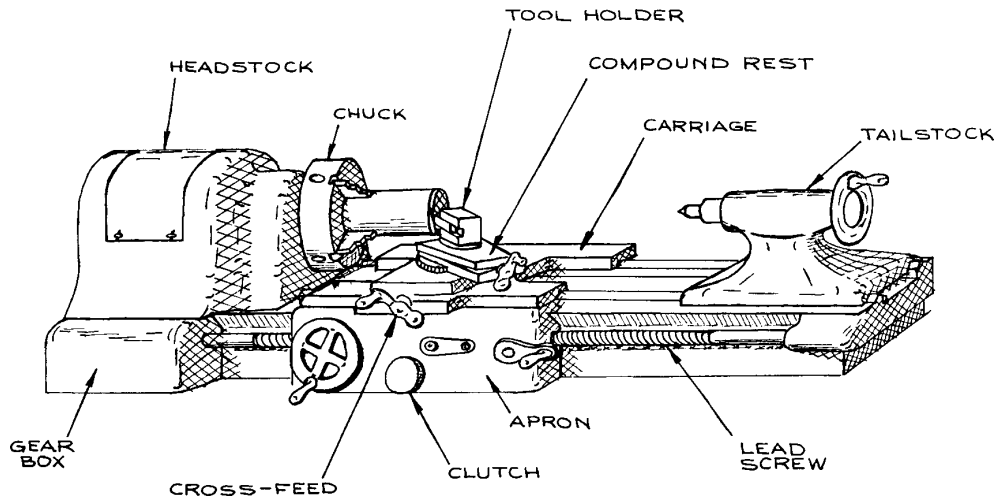


Figure 1.2 A lathe.

be turned is grasped by a *chuck* that is rotated by the driving mechanism within the lathe *headstock*. Long pieces are supported at the free end by a center mounted in the *tailstock*. A cutting tool held atop the lathe carriage is brought against the work as it turns. As shown in Figure 1.2, the *tool holder* is clamped to the *compound rest* mounted to a rotatable table atop the *cross-feed* that in turn rests on the *carriage*. The carriage can be moved parallel to the axis of rotation along slides or ways on the lathe bed. A cylindrical surface is produced by moving the carriage up or down the ways, as in the first cut illustrated in Figure 1.3. Driving the cross-feed produces a face perpendicular to the axis of rotation; driving the tool with the compound-rest screw produces a conical surface.

Most lathes have a *lead screw* along the side of the lathe bed. This screw is driven in synchronization with the rotating chuck by the motor drive of the lathe. A groove running the length of the lead screw can be engaged by a clutch in the carriage *apron* to provide power to drive either the carriage or the cross-feed in order to produce a long uniform cut. When cutting threads the lead screw can be engaged by a split nut in the apron to provide uniform motion of the carriage.

A variety of attachments are available for securing work to the spindle in the headstock. Most convenient is the three-jaw chuck. All three jaws are moved inward and

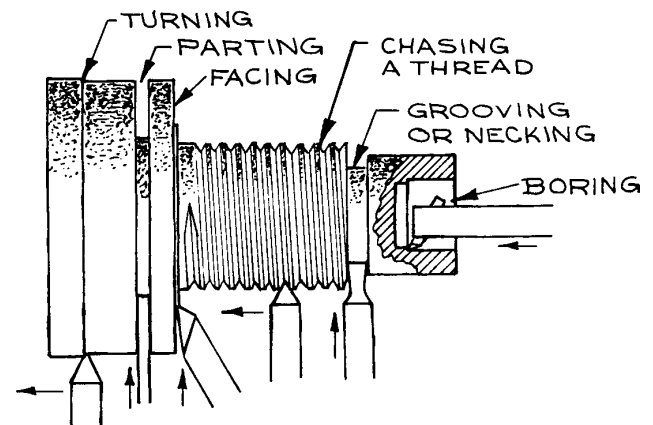
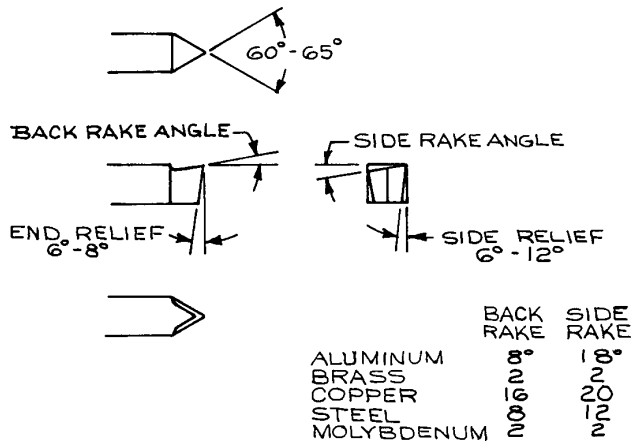


Figure 1.3 Cuts made on a lathe.

outward by a single control so that a cylinder placed in the chuck is automatically centered. A four-jaw chuck with independently controlled jaws is used to grasp a workpiece that is not cylindrical or to hold a cylindrical piece off center. Large irregular work can be bolted to a face plate that is attached to the lathe spindle. Small round pieces can be grasped in a *collet chuck*. A collet is a slotted tube with an inner diameter of the same size as the work and a slightly tapered outer surface. The work is clamped



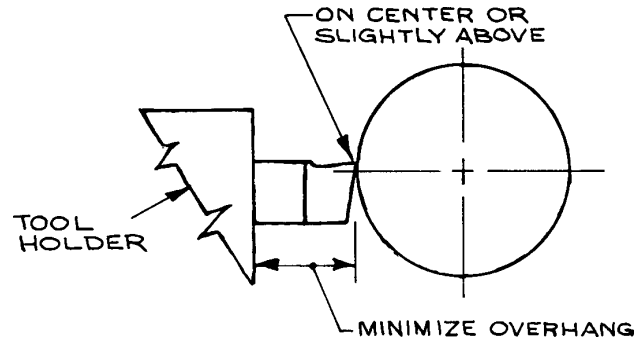
**Figure 1.4** Tool angles for a right-cutting round-nose tool. A right-cutting tool has its cutting edge on the right when viewed from the point end.

in the collet by a mechanism that draws the collet into a sleeve mounted in the lathe spindle.

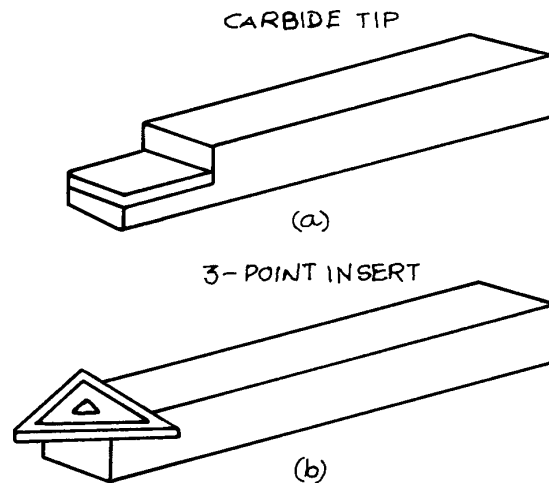
The cutting tool largely determines the quality of work produced in a lathe. The efficiency of the tool bit used in a lathe depends upon the shape of the cutting edge and the placement of the tool with respect to the work-piece. A cutting tool must be shaped to provide a good compromise between sharpness and strength. The sharpness of the cutting edge is determined by the *rake angles* indicated in Figure 1.4. The indicated *relief angles* are required to prevent the noncutting edges and surfaces of the tool from interfering with the work. Placement of the tool in relation to the work-piece is illustrated in Figure 1.5.

In the past, a machinist was obliged to grind tool steel stock to the required shape to make a tool bit. Now virtually all work is done with prepared tool bits. These may be simply ground-to-shape tool bits. Especially sharp, robust tools are available with sintered tungsten carbide (so-called “carbide”) or diamond tips. Many tools are made with replaceable carbide inserts. The insert, which is clamped to the end of the tool, is triangular or square to provide three or four cutting edges by rotating the insert in its holder. Examples are illustrated in Figure 1.6.

As in drilling, the cutting speed for turning in a lathe depends upon the material being machined. Cutting speeds for high-speed steel tools are given in Table 1.2. Modern



**Figure 1.5** Placement of tool with respect to the work in a lathe.



**Figure 1.6** (a) Single-point carbide-tipped lathe tool. (b) Lathe tool with replaceable carbide insert. The insert has three cutting points.

carbide- and ceramic-tipped cutters are much faster than tool-steel bits; they also produce a cleaner, more precise cut. Typically, a cut should be 0.1 to 0.3 mm (.003–.010 in.) deep, although much deeper cuts are permissible for rough work if the lathe and work-piece can withstand the stress.

Holes in the center of a work-piece may be drilled by placing a twist drill in the tailstock and driving the drill into the rotating work with the hand-wheel drive of the tailstock. The hole should first be located with a *center drill* (Figure 1.1), or the drill point will wander off center.

A drill, a lathe tool, or a milling cutter is usually bathed with a cutting fluid to cool the tool and to produce a smooth cut. Cutting fluids include soluble oils, mineral oils, and base oils. *Soluble oils* form emulsions when mixed with water and are used for cutting both ferrous and nonferrous metals, when cooling is most important. *Mineral oils* are petroleum products including paraffin oils and kerosene. They are typically used for light, high-speed cutting. *Base oils* are inorganic or fatty oils with or sulfur-containing additives. Base oils are called for when making heavy cuts in ferrous materials.

Tolerances of 0.1 mm (.004 in.) can be maintained with ease when machining parts in a lathe. Diameters accurate to  $\pm 0.01$  mm ( $\pm .0004$  in.) can be obtained by a skilled operator at the expense of considerable time. Any modern lathe will maintain a straightness tolerance of 0.4 mm/m (0.005 in/ft) provided the work-piece is stiff enough not to spring away from the cutting tool.

### 1.1.4 Milling Machines

Milling, as a machine-tool operation, is the converse of lathe turning. In milling the work-piece is brought into contact with a rotating cutter. Typical milling cutters are illustrated in Figure 1.7. A *plain milling cutter* has teeth only on the periphery and is used for milling flat surfaces.

A *side milling cutter* has cutting edges on the periphery and either one or both ends so that it can be used to mill a channel or groove. An *end mill* is rotated about its long axis and has cutters on both the end and sides. There are also a number of specially shaped cutters for milling dovetail slots, T-slots, and Woodruff key-slots. A *fly cutter* is another useful milling tool. It consists of a cylinder with a single, movable cutting edge and is used for cutting round holes and milling large flat surfaces. Radial saws are also used in milling machines for cutting narrow grooves and for parting off. Ordinary milling cutters are made of tough, hard steel known as high-speed steel. Other alloys are used as well, frequently with coatings of titanium nitride (TiN) or titanium carbonitride (TiCN) for a harder surface and improved lubricity. Both carbide-tipped cutters and tools with replaceable carbide inserts are available. Solid carbide end mills can be had for about twice the cost of high-speed steel mills.

There are two basic types of milling machine. The *plain miller* has a horizontal shaft, or *arbor*, on which a cutter is mounted. The work is attached to a movable bed below the cutter. The plain miller is typically used to produce a flat surface or a groove or channel. It is not much used in the fabrication of instrument components. The *vertical mill* has a vertical spindle located over the bed. Milling cutters can be mounted to an arbor in the spindle of the vertical

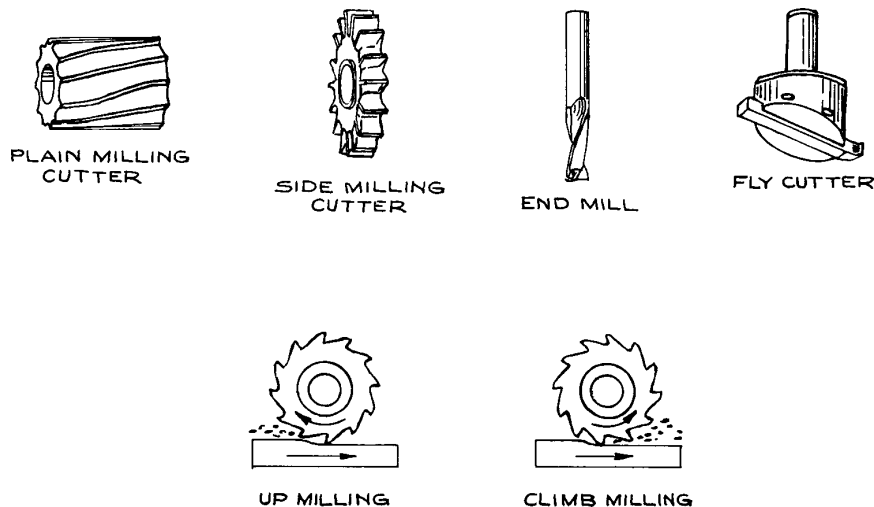


Figure 1.7 Milling cutters.

mill, or a collet chuck for grasping an end mill can replace the arbor. Motion of the mill bed in three dimensions is controlled by hand-wheels. Big machines may have power-driven beds. An essential accessory for a vertical mill is a rotating table so that the work-piece can be rotated under the cutter for cutting circular grooves and for milling a radius at the intersection of two surfaces.

The two possible cutting operations are illustrated in Figure 1.7. *Climb milling*, in which the cutting edge enters the work from above, has the advantage of producing a cleaner cut. Also, climb milling tends to hold the work flat and deposits chips behind the direction of the cut. There is however a danger of pulling the work into the cutter and damaging both the work and the tool. *Up milling* is preferred when the work cannot be securely mounted and when using older, less rigid machines. Cutter speeds are given in Table 1.2.

Dimensional accuracy of  $\pm 0.1$  mm ( $\pm .004$  in.) is easily achieved in a milling operation; flatness and squareness of much higher precision are easily maintained. Both the mill operator and the designer specifying a milled surface should be aware that milled parts tend to curl after they are unclamped from the mill bed. This problem is particularly acute with thin pieces of metal. It can be alleviated somewhat if cuts are taken alternately on one side and then the other, finishing up with a light cut on each side.

The vertical milling machine is the workhorse of the model shop where instruments are fabricated. A scientist contemplating the design of a new instrument should become familiar with the milling machine's capability and, if possible, gain at least rudimentary skill in its operation.

Two electronic innovations have significantly increased the utility and ease of operation of the milling machine: the electronic digital position readout and computer control of the motion of the mill arbor and the mill bed.

All machine tools suffer from backlash in the mechanical controls. In a milling machine the position of the mill bed is read off vernier scales on the hand-wheels that drive the screws that position the bed. In addition to the inconvenience caused by this sort of readout, the operator must realize that reversing the direction of rotation of the hand-wheel does not instantly reverse the direction of travel of the bed, owing to inevitable clearances between the threads of the drive screw and the nut that it engages. This backlash must be accounted for in even the roughest

work. The problem is entirely obviated by the fitting of electronic position sensors on the bed that read out in inches or millimeters on displays mounted to the machine. This simple innovation significantly increases the speed of operation for a skilled operator, while at the same time reducing the number of errors. These displays invariably improve the quality of work of a relatively unskilled operator.

Even modest shops now use milling machines in which the bed and arbor are driven by electric motors under computer control – CNC (computer numerical control) machines. Most modern machinists, working from mechanical drawings provided by the designer, can efficiently program these machines. The time required is frequently offset by time saved by the machine operating under computer control, so in many instances it is quite reasonable to use a CNC machine for one-off production. This is particularly true when a complex sequence of bed and arbor motions is required to turn out a part. Conversely, when a CNC machine is available, the designer can contemplate many more complex shapes in a design than would be economically feasible to produce with manually controlled machines. An additional advantage for the instrument designer is that complex parts can initially be turned out in an inexpensive material, such as polyethylene, to check shape and fit before the final part is machined from some expensive material.

The ultimate application of computer-controlled machines is for them to be operated directly by programs produced by the software the designer employs in the process of preparing the engineering drawings. This is the integration of computer-aided design (CAD) with computer-aided manufacturing (CAM) in a so-called CAD/CAM system. At present, however, this mode of fabrication is not usually practical for the scientist-designer. The time involved in learning to use sophisticated CAD/CAM software, as well as its cost, cannot be justified. In addition, few model shops have machines that can be operated by the output of high-level CAD programs.

Mills and lathes are both relatively powerful machines. One cannot expect the machine to stop should rotating parts of the machine get caught on loose clothing, such as a shirt cuff or necktie, or, worse, on a limb or digit. Initial operation of these machines should be under the guidance of a competent instructor. One must always be

certain that the machine is in proper operating condition and that the work in the machine is securely clamped to the bed (of the milling machine) or by the jaws (of the lathe chuck); a loose work-piece can become a projectile. Metal chips may come flying off the cutting tool; eye protection is mandatory.

### 1.1.5 Electrical Discharge Machining (EDM)

A spark between an electrode and a work-piece will remove material from the work as a consequence of highly localized heating and various electron- and ion-impact phenomena. As improbable as it may seem, the process of spark erosion has been developed into an efficient and very precise method for machining virtually any material that conducts electricity. In electrical-discharge machining (EDM), the electrode and the work-piece are immersed in a dielectric oil or de-ionized water, a pulsed electrical potential is applied between the electrode and the work, and the two are brought into close proximity until sparking occurs. The dielectric fluid is continuously circulated to remove debris and to cool the work. The position of the work-piece is controlled by a computer servo that maintains the required gap and moves the work-piece into the electrode to obtain the desired cut.

There are two types of electrical discharge machines: the *plunge* or *die-sinking EDM* and the *wire EDM*. The plunge EDM is used to make a hole or a well. The electrode is the male counterpart to the female concavity produced in the work. The electrode is machined from graphite, tungsten, or copper. Making the electrode in the required shape is a significant portion of the entire cost of the process. The electrode for wire EDM is typically a vertically traveling copper or copper-alloy wire 0.05 to 0.4 mm (.002 to .012 in.) in diameter. The wire is eroded as cutting progresses. It comes off a spool to be fed through the work and is discarded after a single pass. With two-dimensional horizontal (XY) control of the work-piece, a cutting action analogous to that of a bandsaw is obtained. XYZ-control permits the worktable to be tilted up to 20° so that conical surfaces can be generated.

In an EDM process, the cavity or *kerf* is always larger than the electrode. This *overcut* is highly predictable and may be as small as 0.03 mm (.001 in.) for wire EDM. As a

consequence, an accuracy of  $\pm 0.03$  mm ( $\pm .001$  in.) is routine and an accuracy of  $\pm .003$  mm ( $\pm .0001$  in.) is possible. Furthermore, EDM produces a very high-quality finish; a surface roughness less than 0.0003 mm (12 micro-inches) RMS is routinely obtained. Part of the reason for the great accuracy obtained in EDM is that there is no force applied to the work and hence no possibility of the work-piece being deflected from the cutting tool. There is no work-hardening and no residual stress in the material being cut. The efficacy of EDM is independent of the hardness of the material of the work. Tool steel, conductive ceramics such as graphite and carbide, and refractory metals such as tungsten are cut with the same ease as soft aluminum. A particular advantage is that the material can be heat treated before fabrication since very little heat is generated in the cutting operation.

A precision CNC electrical-discharge machine may cost in excess of \$100 000. As a consequence these machines are seldom found in the typical university model shop; the EDM machine has become, however, the workhorse of the tool-and-die industry. Owing to the cost of the machines and the fact that, once programmed, they can run virtually 24 hours a day, job shops are usually anxious to have outside work to keep the machines running full time and usually welcome scientists with one-off jobs.

### 1.1.6 Grinders

Grinders are used for the most accurate work and to produce the smoothest surface attainable in most machine shops. A grinding machine is similar to a plain milling machine except that a grinding wheel rather than a milling cutter is mounted on the rotating arbor. In most machines the work is clamped magnetically to the table and the table is raised until the work touches the grinding wheel. The table is automatically moved back and forth under the wheel at a fairly rapid rate. Many lathes incorporate, as an accessory, a grinder that can be mounted to the compound rest in place of the tool holder, so that cylindrical and conical surfaces can be finished by grinding.

Even the hardest steel can be ground, but grinding is seldom used to remove more than a small fraction of a millimeter (a few thousandths of an inch) of metal. For complicated pieces to be made of hardened bronze or steel, it is advantageous to soften the stock material by annealing,

machine it slightly oversize with ordinary cutting tools, re-harden the material, and then grind the critical surfaces.

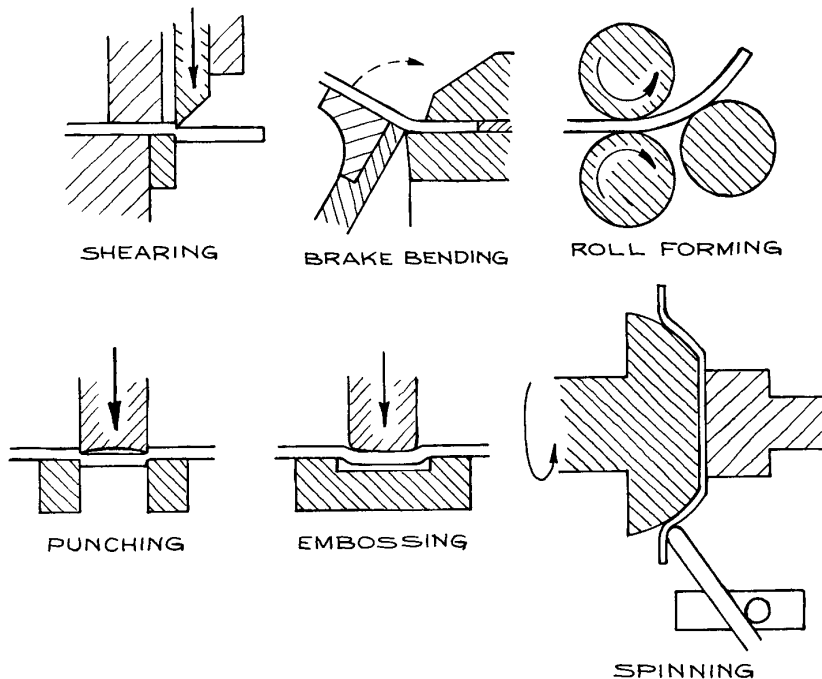
A flatness tolerance of  $\pm 0.003$  mm ( $\pm 0.0001$  in.) can be maintained in grinding operations. The average variation of a ground surface should not exceed 0.001 mm (50 microinches); a surface roughness less than 0.0003 mm (10 microinches) RMS is possible.

Safety is a primary concern in any grinding operation. Grinding wheels are typically held together with an inorganic ceramic cement. They are brittle and can fail catastrophically. Significant forces come into play when grinding: the centrifugal force on the spinning wheel; the force generated between the wheel and the work; and especially the shock produced when the wheel first contacts the work. A guard should enclose the wheel. The exposed portion of the wheel should be no more than necessary to carry out the operation at hand. The bearings and pilot shaft supporting the wheel must be in good condition; check for balance before spinning up; unbalanced forces can be destructive. The flanges clamping the grinding wheel to the

drive shaft must be correct for the wheel in use. Wheel speed should not exceed that specified for the wheel. The wheel speed is usually specified as meters per minute (m/min) or surface feet per minute (sfpm), thus requiring the operator to calculate the rotational speed required to attain a specified linear speed at the outer circumference of the wheel. In general, the surface speed at the outer edge of an inorganic-bonded wheel should not exceed about 2000 m/min (6000 sfpm). In a dry grinding operation an exhaust system should be in place to carry away the considerable dust and metal residue that is generated. ANSI standards for safe grinding operations are given in condensed form in *Machinery's Handbook Pocket Companion*.<sup>1</sup>

### 1.1.7 Tools for Working Sheet Metal

Most machine shops are equipped with the tools necessary for making panels, brackets, and rectangular and cylindrical boxes of sheet metal. The basic sheet-metal processes are illustrated in Figure 1.8.



**Figure 1.8** Sheet-metal shop processes.



Sheet metal is cut in a guillotine *shear*. Shears are designed for making long straight cuts or for cutting out inside corners. A typical shear can make a cut a meter in length in sheet metal of up to 1.5 mm (1/16 in.) in thickness.

A *sheet-metal brake* is used to bend sheet stock. A typical instrument-shop brake can accommodate sheet stock at least a meter wide and up to 1.5 mm (1/16 in.) thick. The minimum bend radius is equal to the thickness of the sheet metal. Dimensional tolerances of 1 mm (.04 in.) can be maintained.

Sheet can be formed into a simple curved surface on a *sheet-metal roll*. The roll consists of three long parallel rollers, one above and two below. Sheet is passed between the upper roller and the two lower rollers. The upper roller is driven. The distance between the upper roller and the lower rollers is adjustable and determines the radius of the curve that is formed.

Holes can be punched in sheet metal. A sheet-metal punch consists of a *punch*, a *guide bushing*, and a *die*. The punch is the male part. The cross section of the punch determines the shape and size of the hole. The punch is a close fit into the die, so that sheet metal placed between the two is sheared by the edge of the punch as it is driven into the die. Round and square punches are available in standard sizes. A punch-and-die set to make a nonstandard hole can be fabricated, but the cost may be justified only if a large number of identical holes are required. On the other hand, a very precisely shaped hole can be made in a punch-and-die operation since the tools can be made with great precision.

Sheet metal can be embossed with a *stamp and die*, similar to a punch and die except that the die is somewhat larger than the stamp, so that the metal is formed into the die rather than sheared off at the edge.

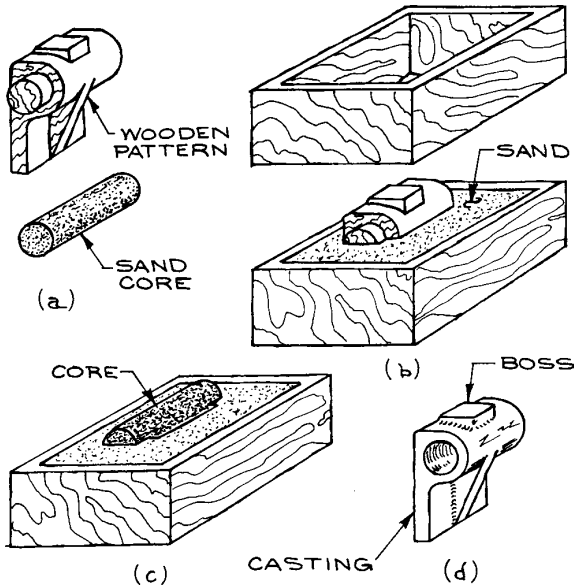
Sheet metal can be formed into surfaces that are figures of revolution by *spinning*. The desired shape is first turned in hard wood in a lathe. A circular sheet-metal blank is then clamped against the wooden form by a rubber-faced rotating center mounted in the lathe tailstock. Then as the wooden form is rotated the sheet metal is gradually formed over the surface of the wood by pressing against the sheet with a blunt wooden or brass tool. Spinning requires few special tools and is economical for one-off production.

### 1.1.8 Casting

Sand casting is the most common process used for the production of a small number of cast parts. Although few instrument shops are equipped to do casting, most competent machinists can make the required wooden patterns that can then be sent to a foundry for casting in iron, brass, or aluminum alloy. There are both mechanical and economical advantages to producing some complicated parts by casting rather than by building them up from machined pieces. Very complicated shapes can be produced economically because the patternmaker works in wood rather than metal. Castings can be made very rigid by the inclusion of appropriate gussets and flanges that can only be produced and attached with difficulty in built-up work. Most parts on an instrument require only a few accurately located surfaces. In this case the part can be cast and then only critical surfaces machined.

A designer must understand the sand-casting process in order to design parts that can be produced by this method. A sand-casting mold is usually made in two parts, as shown in Figure 1.9. The lower mold box, called the *drag*, is filled with sand, and the wooden pattern is pressed into the sand. The sand is leveled and dusted with dry parting sand. Then the upper box, or *cope*, is positioned over the lower and packed with sand. The two boxes are separated, the pattern is removed, and a filling hole, or *sprue*, is cut in the upper mold. A deep hole can be cast by placing a sand *core* in the impression, as shown in Figure 1.9(c). Note that the pattern has lugs, called *core prints*, which create a depression to support the core. The two halves of the mold are then clamped together and filled with molten metal.

It is obvious that a part to be cast must include no involuted surfaces above or below the parting plane, so that the pattern can be withdrawn from the mold without damaging the impression. It is also good practice to taper all protrusions that are perpendicular to the parting plane to facilitate withdrawal of the pattern from the mold. A taper or *draft* of 1–2° is adequate. Surfaces to be machined should be raised to allow for the removal of metal. That is the purpose of the *boss* shown in Figure 1.9(d). When possible, all parts of the casting should be of the same thickness to avoid stresses that may build up as the molten metal solidifies. All corners and edges should be



**Figure 1.9** Sand casting: (a) the pattern and core; (b) making an impression in the lower mold box; (c) inserting the sand core in the impression; (d) the finished casting.

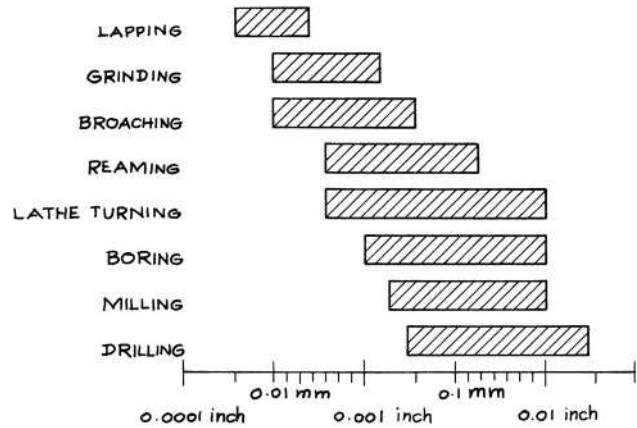
radiused. With care, a tolerance of 1 mm (.03 in.) can be maintained.

### 1.1.9 Tolerance and Surface Quality for Shop Processes

The designer must decide on the precision and accuracy required in making each part of an apparatus. Also a decision must be made about the quality of the surfaces of each part to assure proper fit and function. It follows that the capabilities of the various shop processes must be taken into consideration, as well as the cost of production, to meet desired tolerances and surface quality (note Figure 1.45).

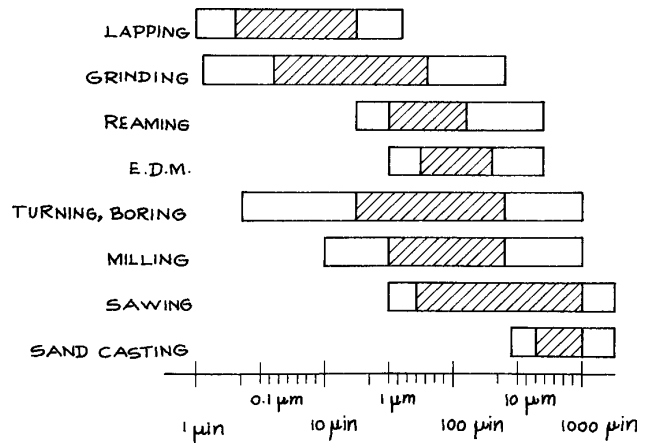
Precision and accuracy are specified as the tolerance (+/−) on each dimension of a part. The surface roughness is specified as the RMS variation in the height of a surface.

Figure 1.10 gives an approximate indication of the tolerance normally obtainable for the shop processes discussed in this section. Figure 1.11 specifies the surface quality normally obtained in various shop process as well in industrial processes used in the manufacturer of materials from which parts are machined. An important point to



**Figure 1.10** Tolerance that can be reasonably maintained in shop processes for a work piece dimension of 1 to 10 cm. (Adapted from ASME, ANSI Std. B4.1-1967)

be made here is that it is often possible to use materials as they come from the supplier without further machining of the surfaces. It should be noted as well that obtainable tolerances and surface quality depend upon the size of the part being made. A tolerance of  $\pm 0.05\text{mm}$  ( $\pm .002$  in.) is easily obtained in a milling operation on a part that is only a few centimeters across. In a part that is 100 times



**Figure 1.11** Surface quality as RMS variation in surface height for various machining and production processes. (Adapted from ASME, ANSI Std. B46.1-1961)

larger, a tolerance of  $\pm 0.5$  mm ( $\pm 0.02$  in.) is the best that can be obtained in a milling operation without the expenditure of considerable effort.

## 1.2 PROPERTIES OF MATERIALS

The materials employed in the construction of an apparatus – metals, plastics, glass, ceramic, even wood – are characterized by their strength, flexibility, hardness, toughness, machinability, electrical and thermal conductivity, and so on. In order to wisely choose a material for a part of an apparatus, it is essential to understand the ways in which these properties are specified, how these properties can be modified by heating and working, and how to design to best exploit these properties.

### 1.2.1 Parameters to Specify Properties of Materials

The strength and elasticity of metal are best understood in terms of a stress–strain curve such as is shown in Figure 1.12. *Stress* is the force applied to the material per unit of cross-sectional area [ $\text{MN m}^{-2}$ , psi (pounds per square inch)]. This may be a stretching, compressing, shearing, or twisting force. *Direct strain* is the change in length per unit length that occurs in response to an in-line stress [strain is unitless (cm/cm, in/in) or expressible as a percent]. *Shear strain* is the sideward displacement per unit length in response to a transverse stress (again without units). A torsional stress induces a shear strain. Most metals deform in a similar way under compression or elongation. When a stress is applied to a metal, the initial strain is *elastic* and the metal will return to its original dimensions if the stress is removed. Beyond a certain stress however, *plastic* strain occurs and the metal is permanently deformed.

The *tensile strength* or ultimate strength of a metal is the stress applied at the maximum of the stress–strain curve. The metal is very much deformed at this point, so that in most cases it is impractical to work at such a high load.

A more important parameter for design work is the *yield strength*. This is the stress required to produce a stated, small, plastic strain in the metal – usually 0.2% permanent deformation. For some materials the *elastic limit* is specified.

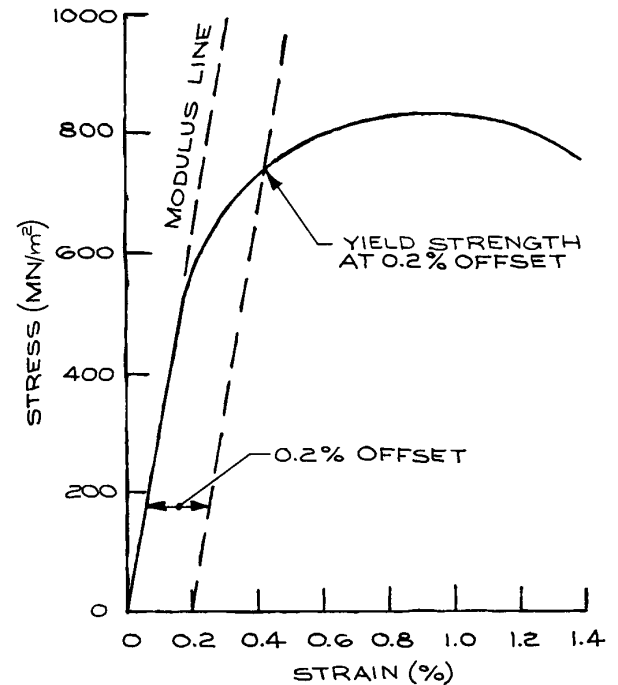


Figure 1.12 Stress–strain curve for a metal.

This is the maximum stress that the material can withstand without permanent deformation.

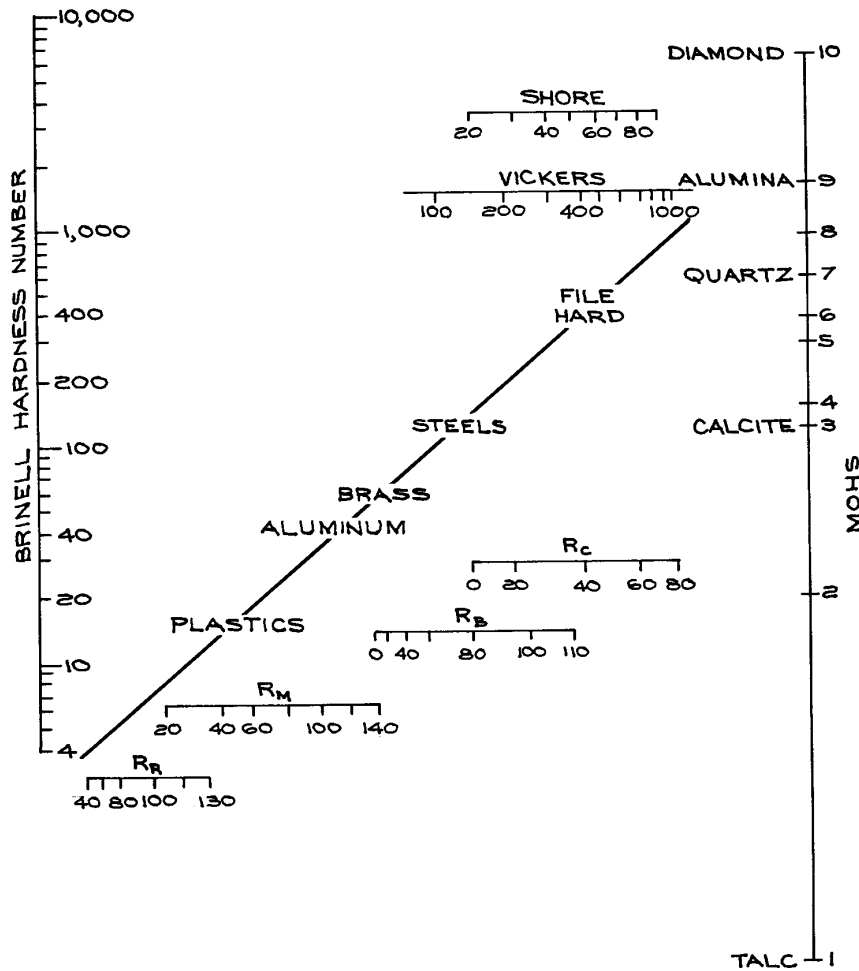
The slope of the straight-line portion of the stress–strain curve is a measure of the stiffness of the material; this is the *modulus of elasticity*,  $E$ , sometimes referred to as “Young’s modulus.” It is worthwhile to note that  $E$  is about the same for all grades of steel (about  $210 \text{ GN/m}^2$  or  $30 \times 10^6$  psi) and about the same for all aluminum alloys (about  $70 \text{ GN/m}^2$  or  $10 \times 10^6$  psi), regardless of the strength or hardness of the alloy. The effect of elastic deformation is specified by *Poisson’s ratio*,  $\mu$ , the ratio of transverse contraction per unit dimension of a bar of unit cross-section to its elongation per unit length, when subjected to a tensile stress. For most metals,  $\mu = 0.3$ . A modulus of elasticity in shear or *shear modulus*,  $G$ , is similarly defined for a torsional stress. The shear modulus is typically about one third of Young’s modulus for a metal.

The *hardness* of a material is a measure of its resistance to indentation and is usually determined from the force required to drive a standard indenter into the surface of

the material or from the depth of penetration of an indenter under a standardized force. The common hardness scales used for metals are the Brinell hardness number (BHN), the Vickers scale (VHN), and the Rockwell C scale. Type 304 stainless steel is about 150 BHN, 160 Vickers, or 60 Rockwell C. A file is 600 BHN, 650 Vickers, or 60 Rockwell C. The Shore hardness scale is determined by the height of rebound of a steel ball dropped from a specified distance above the surface of a material under test. Mineralogists and ceramic engineers use the

Mohs hardness scale. It is based upon standard materials, each of which will scratch all materials below it on the scale. Knoop hardness is used as a measure for very brittle materials or thin sheets, where only a microindentation with a pyramidal diamond point can be made. Figure 1.13 shows the approximate relation of the various hardness scales.

A designer must consider the machinability of a material before specifying its use in the fabrication of a part that must be lathe-turned or milled. In general, the harder and



**Figure 1.13** Approximate relation of Brinell (BHN), Rockwell ( $R_R$ ,  $R_M$ ,  $R_B$ ,  $R_C$ ), Vickers (VHN), Shore, and Mohs hardness scales.

stronger a material, the more difficult it is to machine. On the other hand, soft metals, such as copper and some nearly pure aluminum alloys, are also difficult to machine because the metal tends to adhere to the cutting tool and produce a ragged cut. Some metals are alloyed with other elements to improve their machinability. Free-machining steels and brass contain a small percentage of lead or sulfur. These additives do not usually affect the mechanical properties of the metal, but since they have a relatively high vapor pressure, their out-gassing at high temperature can pose a problem in some applications.

### 1.2.2 Heat Treating and Cold Working

The properties of many metals and metal alloys can be considerably changed by heat treating or cold working to modify the chemical or mechanical nature of the granular structure of the metal.<sup>2</sup> The effect of heat treatment depends upon the temperature to which the material is heated relative to the temperatures at which phase transitions occur in the metal, and the rate at which the metal is cooled. If the metal is heated above a transition temperature and quickly cooled or *quenched*, the chemical and physical structure of the high-temperature phase may be frozen in, or a transition to a new metastable phase may occur. Quenching is accomplished by plunging the heated part into water or oil. This process is usually carried out to harden the metal. Hardened metals can be softened by *annealing*, wherein the metal is heated above the transition temperature and then slowly cooled. It is frequently desirable to anneal hardened metals before machining and re-harden after working, although hardening and annealing can result in distortion, owing to differences in density of the high and low temperature phases. *Tempering* is an intermediate heat treatment wherein previously hardened metal is reheated to a temperature below the transition point in order to relieve stresses and then cooled at a rate that preserves the desired properties of the hardened material.

Repeated plastic deformation will reduce the size of the grains or crystallites within the metal. This is *cold working* that accompanies bending, rolling, drawing, hammering, and, to a lesser extent, cutting operations. Not all metals benefit from cold working, but for some the strength is greatly increased. Because of the annealing effect, the strength and hardness derived from cold working begin to

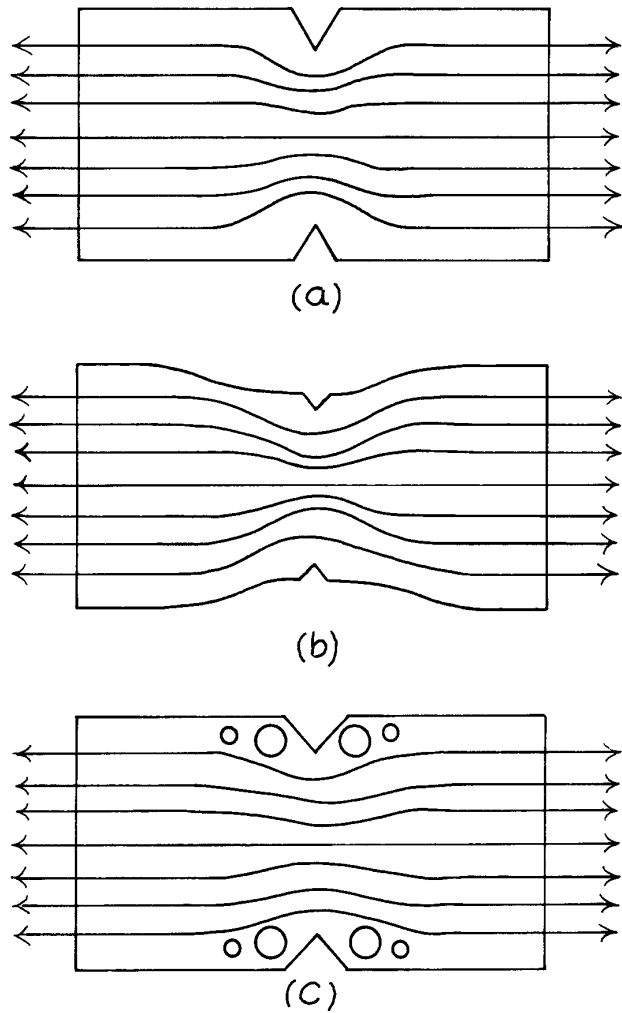
disappear as a metal is heated. This occurs above 250 °C for steel and above 125 °C for aluminum. In rolling or spinning operations, some metals will work-harden to such an extent that the material must be periodically annealed during fabrication to retain its workability. Cold working reduces the toughness of metal. The surface of a metal part can be work-hardened, without modifying the internal structure, by peening or shot blasting and by some rolling operations. The strength of metal stock may depend upon the method of manufacture. For example, sheet metal and metal wire are usually much stronger than the bulk metal because of the work hardening produced by rolling or drawing.

The surface of a metal part may be chemically modified and then heat-treated to increase its hardness while retaining the toughness of the bulk of the material beneath. This process is called *case hardening*. The many methods of case hardening are usually named for the chemical that is added to the surface. *Carburizing* (C), *cyaniding* (C and N), and *nitriding* (N) are common processes for case-hardening steel. Of these, carburizing of low-carbon steel is most conveniently performed in the lab or shop. The part to be hardened is packed in bone charcoal or Kasenit in an open metal box and heated in a furnace to 900 °C (salmon-red heat). The time in the furnace, from 15 minutes to several hours, determines the depth of the case. Steel will absorb carbon by this process to a maximum depth of about 0.4 mm (0.015 in.) The part is removed from the furnace and promptly quenched in water. The quenching hardens the case, but the core remains tough and ductile.

### 1.2.3 Effect of Stress Concentration

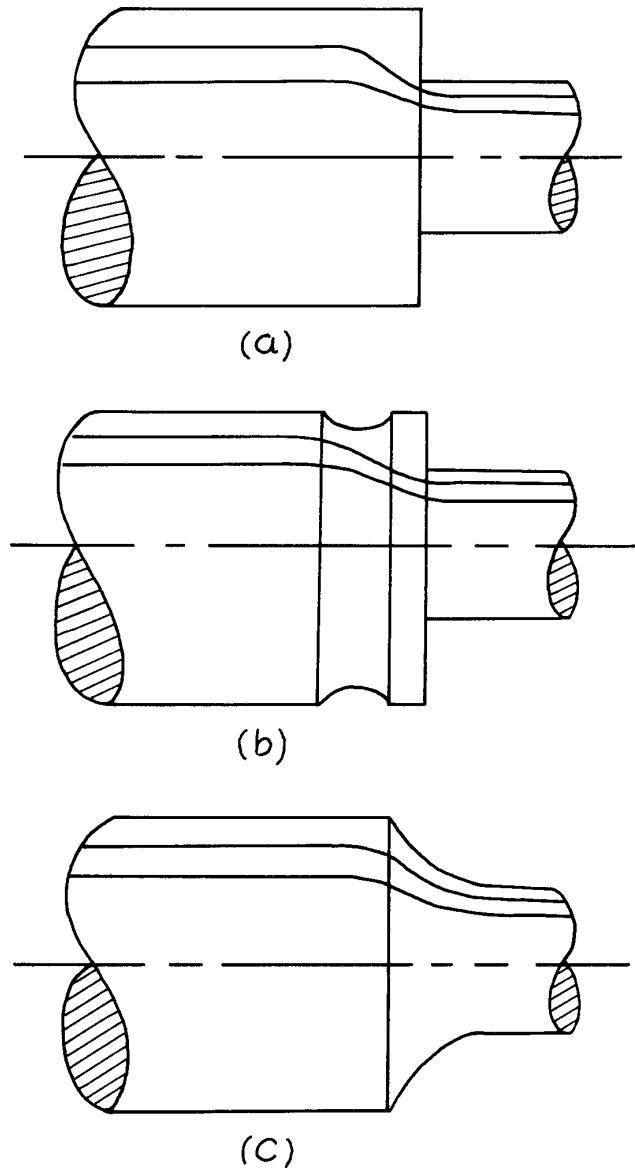
In choosing a material for a particular job, one must consider not only the stress the part is expected to bear, but also the way in which the stress is distributed. Careless design can lead to failure in very strong materials under relatively light loads if the stress of the load is concentrated in a small volume.

An abrupt change in the cross-section of a loaded beam or shaft produces a concentration of stresses, as shown in Figures 1.14(a) and 1.15(a). For the element in tension [Figure 1.14(a)] the lines indicate the direction of tensile force; this is a so-called “force flow depiction.” For an element under torsional stress, Figure 1.15(a), the lines indicate the cross-section of concentric cylindrical shells



**Figure 1.14** Force flow lines in a beam under stress: (a) A notch in a beam under tension concentrates stress in the vicinity of the notch. (b) and (c) suggest means to mitigate the stress concentration.

each under the same torque. Stress increases when the force flow lines or equitorque surfaces become sharply curved and crowded together. Stress concentrations are located at steps, grooves, keyways, holes, dents, and scratches. Failure occurs as a result of shear forces in the region of highest

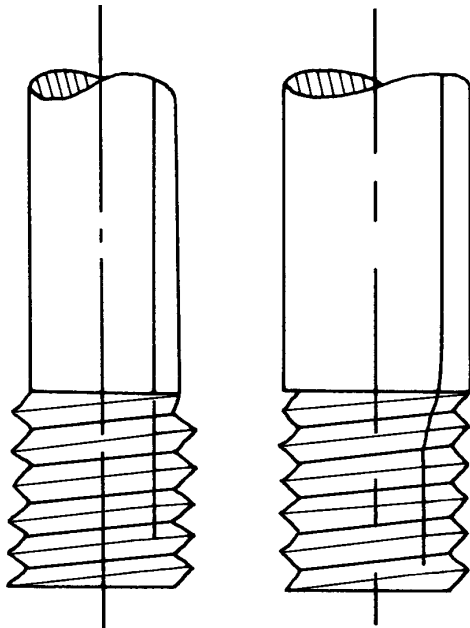


**Figure 1.15** A shaft under torsional stress. Lines indicate concentric equitorque surfaces: (a) A step in a shaft under a torsional load concentrates stress in the vicinity of the step. (b) and (c) suggest means to mitigate the stress concentration.

stress concentration. Anyone who has ever broken a bolt is familiar with this effect; a bolt invariably fails at the step where the head joins the shaft. The effects of stress concentration may be especially important in fatigue failure resulting from cyclic applications of stress.

Stress is often concentrated in the area of the smallest radius of curvature of a part, as suggested in Figures 1.14(a) and 1.15(a). This observation explains the manner in which a crack propagates once a part has begun to fail: the radius of curvature is small at the end of a crack, thus stress is concentrated at the tip of the crack and further failure ensues. Increasing the radius of curvature of a stress riser can mitigate stress concentration. It is often more practical to introduce features that move the force flow lines away from the stress riser when it is not possible to modify the offending aspect. Examples are given in Figures 1.14 and 1.15.

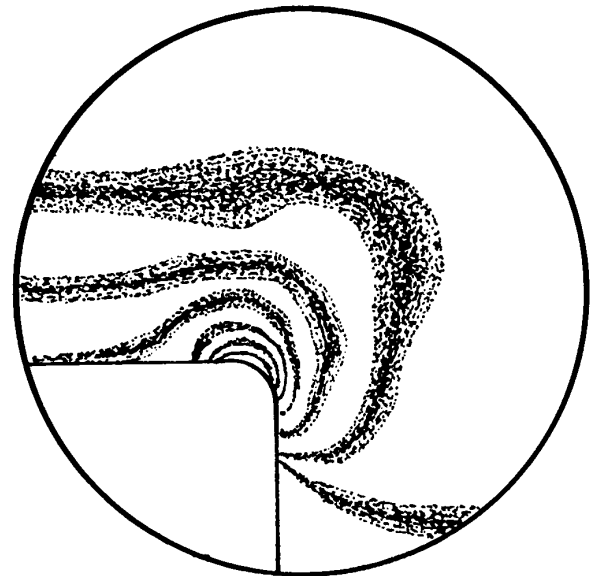
Threaded parts are often a problem. The “V” bottom of the thread groove is an insipient stress riser. As suggested in Figure 1.16, reducing the diameter of the adjoining unthreaded portion of a shaft has the effect of moving the equitorque surfaces away from the threads.



**Figure 1.16** Threads on a shaft may produce a concentration of stress. Reducing the diameter of the adjoining shaft moves the equitorque surfaces away from the threads.

There are a number of means of detecting stress in a part. One of the most useful is the *photoelastic method*.<sup>3</sup> In this technique a transparent plastic model is stressed in the same way as the element of interest without necessarily duplicating the magnitude of the stress. The model is illuminated with polarized, monochromatic light and viewed through a polarizer. As shown in Figure 1.17, the stress distribution appears as fringes in the image; closely spaced fringes characterize stress concentrations. Most glassblowing shops are equipped with a polarizing device of this type for detecting residual stresses in worked glass (see Figure 2.16, next chapter).

In a similar fashion, fabricated parts can be nondestructively tested by means of holography. A hologram of the object is made; then the part is stressed and a second hologram is made. The two holograms are then superposed and placed in the holographic projector. The resulting three-dimensional image consists of patterns of fringes that are densest in the areas of greatest distortion in the stressed object. Holographic nondestructive testing (HNDT) is carried out as a commercial service.



**Figure 1.17** Stress distribution in a transparent plastic model as observed when the part is illuminated with polarized light and observed with a crossed polarizer.

A particularly simple and inexpensive means of locating stress concentrations in a working part has been developed by Magnaflux. The part to be tested is sprayed with a substance that forms a hard, brittle layer over the surface. The part is returned to service for a time, and any flexing of the part causes the surface coating to crack. The part is then removed and dipped in a dye that permeates the cracks to reveal the areas where the strain under load was greatest.

## 1.3 MATERIALS

We shall discuss in a qualitative manner the properties of useful metals, plastics and ceramics. Quantitative data are gathered together in Table 1.3.

### 1.3.1 Iron and Steel

For instrument work, iron is used primarily as a casting material. The advantages of cast iron are hardness and a high degree of internal damping. Cast iron is harder than most steels and is used for sliding parts, where resistance to wear is important. The damping capacity of iron, that is, the ability to dissipate the energy of vibration, is about 10 times that of steel and for this reason the frame of a delicate instrument is often made of cast iron.

Cast iron contains a few % carbon. The carbon is in a free state in gray cast iron, while it is chemically incorporated into the structure of white cast iron. Gray iron is inexpensive and easy to machine. Class-30 gray iron has a Brinell hardness of about 200 BHN. It is, however, rather brittle and is only about half as strong as steel. White iron is stronger and harder than gray iron, and thus it is very wear-resistant. White iron is difficult to machine and is usually shaped by grinding. The machinability and ductility of cast irons can be considerably improved by various heat treatments.

Steel is an iron alloy. There are more than 10 000 different types of steel. Steels are classed as cast or wrought steels depending on the method of manufacture. *Wrought steels* are produced by rolling and are almost the only kind used for one-off machine work. Steels are also classified as either carbon steels or alloy steels. *Carbon steels* are specified by a four-digit number. Plain carbon steels are specified as 10xx, where xx is a two-digit number that

indicates the carbon content in hundredths of %. In general the strength and hardness of carbon steel increases with carbon content, but these properties also depend upon the nature of the heat treatment and cold working that the metal has received. Alloy steels containing elements in addition to carbon have special properties especially useful for the manufacture of tools and dies. These include, for example, manganese steels (designated 13xx for tool steel with 1.8% Mn), nickel alloy steel (23xx, 3.5% Ni; 25xx, 5% Ni), molybdenum steel (40xx, 0.20 % Mo), and so on.

Steels with a carbon content in the 0.10–0.50% range are referred to as *low* and *medium-carbon steels* or mild steels. These steels are supplied rolled or drawn and are most suitable for machine work.

*High-carbon steels* are difficult to machine and weld, although their machinability is improved if they are first annealed by heating to 750 °C (orange heat) and cooling slowly in air. They may be returned to their hard state by reheating to 750 °C and quenching in water; however, this process inevitably produces some distortion.

The most common *alloy steels* are the types known as *stainless steels*. The main constituents of stainless steels are iron, chromium, and nickel. They are classified as ferritic, martensitic, or austenitic, depending on the nickel content. The high-nickel austenitic alloys comprise the 200 and 300 series of stainless steels. 300-series stainless steels are most useful for instrument construction because of their superior toughness, ductility, and corrosion resistance. The austenitic stainless steels are sensitive to heat treatment and cold working, although the heat-treatment processes are rather more complicated than for plain carbon steels. For example, quenching from 1000 °C leaves these steels soft. Cold working, however, can more than double the strength of the annealed alloy. The heat associated with welding of austenitic stainless steels can result in carbide precipitation at the grain boundaries in the vicinity of a weld, leaving the weld subject to attack by corrosive agents. The carbides can be returned to solution by annealing the steel after welding. There are also some special stainless steels that do not suffer from carbide precipitation. In general, stainless steels are more expensive than plain carbon steels, but their superior qualities often offset the extra cost.

Type 303 is the most machinable of the 300-series alloys. Type 316 has the greatest resistance to heat and corrosion.



Table 1.3 Properties of Research Materials

<i>Material</i>	<i>Density</i> ( $\text{kg m}^{-3}$ )	<i>Yield Strength</i> ( $\text{MN/m}^2$ )	<i>Tensile Strength</i> ( $\text{MN/m}^2$ )	<i>Modulus of Elasticity</i> ( $\text{GN/m}^2$ )	<i>Shear Modulus</i> ( $\text{GN/m}^2$ )	<i>Hardness<sup>a</sup></i>	<i>Coefficient of Thermal Expansion</i> ( $10^{-6}\text{K}^{-1}$ )	<i>Comments</i>
<b>Ferrous Metals</b>								
Cast gray iron (ASTM 30)	6990	170	210	90	36	200 BHN	11	
1015 steel (hot-finished)	7850	190	340	200	76	100 BHN	15	Low carbon (0.15% C)
1030 steel (hot-finished)	7850	250	470	200	76	137 BHN	15	Medium carbon (0.30% C)
1050 steel (hot-finished)	7850	340	620	200	76	180 BHN	15	High carbon (0.50% C)
Type 304 stainless steel (annealed)	7910	240	590	190	69	150 BHN	17	Austenitic
Type 304 stainless steel (cold-worked)	7910	520	760	190	69	240 BHN	17	Austenitic
Type 316 stainless steel (cold-worked)	7910	410	620	190	69	190 BHN	16	Austenitic
<b>Nickel Alloys</b>								
Monel 400: 25 °C	8820	170	480				14	
500 °C		150	310	180		110 BHN	16	Slightly magnetic
Monel 500: 25 °C	8460	280–1000	600–1200				14	
500 °C		620	65	180		140 BHN	16	Nonmagnetic
Inconel 600: 25 °C	8400	250	620	210			12	
650 °C		180	450	140		120 BHN	16	Strong, resists high-temperature oxidation
Invar (0–100 °C)	8130	160	480	140		140 BHN	1.3	small temperature coefficient
<b>Aluminum Alloys</b>								
1100-0	2760	30	90	70	30	23 BHN	23	99% Al
2024-T4	2710	320	470	70	30	120 BHN	23	3.8% Cu, 1.2% Mg, 0.3% Mn
2024-T4 (200 °C)	2710	80	120					
6061-T6	2760	280	310	70	30	95 BHN	23	0.15% Cu, 0.8% Mg, 0.4% Si
7075-T6	2790	500	570	70	30	150 BHN	23	5.1% Zn, 2.1% Mg, 1.2% Cu
<b>Copper Alloys</b>								
Yellow brass (annealed)	8460	410	510	100	40		20	65% Cu, 35% Zn
Yellow brass (cold-worked)	8460	410	510	100	40	150 BHN	20	65% Cu, 35% Zn
Cartridge brass (1/2 hard)	8510	360	480	110	40	145 BHN	20	70% Cu, 30% Zn
Beryllium copper (precipitation-hardened)	8210	960	1200	130	48	380 BHN	17	98% Cu, 2% Be

Table 1.3. (contd.)

Table 1.3. (contd.)								
<b>Unalloyed Metals</b>								
Copper	8930	70	220	110	45	44 BHN	17	
Molybdenum	10200	560	650	320	120	190 VHN	5.4	Refractory, nonmagnetic
Tantalum	16590	330	460	180	70	80 VHN	6.5	Refractory, somewhat ductile
Tungsten	19270	1520	1520	410	150	350 VHN	5	Refractory, very dense
<b>Plastics</b>								
Phenolics	1350		50	7		125 R <sub>M</sub>	81	Bakelite, Formica
Polyethylene (low density)	910		10	0.2		10 R <sub>R</sub>	180	
Polyethylene (high density)	940		30	0.8		40 R <sub>R</sub>	216	
Polyamide	1110	80		2.8		118 R <sub>R</sub>	90	Nylon
Polymethylmethacrylate	1190		50	2.9		90 R <sub>M</sub>	72	Lucite, Plexiglas
Polytetrafluoroethylene	2130		20	0.4		60 R <sub>R</sub>	99	Teflon
Polychlorotrifluoroethylene	2100		40	1.7		110 R <sub>R</sub>	70	Kel-F
Poly(amide-imide)	1440		150	5.8		80 R <sub>E</sub>	28	Torlon < -200 °C to 220 °C
Polyimide	1440		80	2.6		45 R <sub>R</sub>	54	Vespel -270 °C to 290 °C
Polycarbonate	1190		60	2.3		70 R <sub>M</sub>	70	Lexan
<b>Ceramics</b>								
Alumina (polycrystalline)	3900		240	330		9 Mohs	8	99% alumina
Macor	2520		30	60	25	400 VHN	13	Corning machinable ceramic
Steatite	3590		60	100			10.6	
<b>Wood</b>								
Douglas fir (air-dried)	300		65,3 <sup>b</sup>	10			6,35 <sup>b</sup>	Typical of softwoods
Oregon white oak (air-dried)	700		70,6 <sup>b</sup>	16			5,55 <sup>b</sup>	Typical of hardwoods

<sup>a</sup> BHN = Brinell hardness number; R = Rockwell hardness; VHN = Vickers hardness.

<sup>b</sup> Parallel to fiber, across fiber.

All of these alloys are fairly nonmagnetic; types 304 and 316 are the least magnetic. Machining and cold working tends to increase the magnetism of these alloys. Residual magnetism can be relieved by heating to 1100 °C [safely above the temperature range (450–900°) where carbide precipitation occurs] and quenching in water. Thin pieces that might distort excessively may be quenched in an air blast.

### 1.3.2 Nickel Alloys

Nickel alloys are typically tough and strong and resistant to corrosion and oxidation. The machining characteristics of the wrought alloys are similar to those of steel, and they are

readily brazed and welded. Nickel alloys mostly do not go by a standard specification like steel, but rather are referred to by trade names. These names often allude to an alloy or class of alloys that have been designed to optimize a particular property. Nickel alloys are expensive compared to alloy steels, but not so much so that they shouldn't be used when an instrument can exploit their special properties.

The Monel alloys are nickel-copper alloys whose properties make them a better choice than austenitic stainless steels in some applications. The most familiar is Monel 400. It is resistant to attack by both acids and bases and is sufficiently ductile for parts to be fabricated by spinning and stamping. Monel 404 is similar, but it is quite

nonmagnetic, and its magnetic properties are not affected by cold working. Monel 500 is similarly nonmagnetic and retains most of its strength up to 500 °C.

Inconel alloys are notable for their strength at high temperatures. The Inconel 600 series of alloys retain their strength to 700 °C and are very resistant to oxidation at high temperatures. Inconel 600 is nonmagnetic. The Inconel 700 series are so-called “superalloys” with tensile strengths on the order of 1000 MN/m<sup>2</sup> (150 000 psi). They are used in highly-stressed parts that will maintain their strength to 1000 °C. They must be heat-treated to obtain the maximum high-temperature strength. The Inconel 700 alloys, especially Inconel 702 and Inconel X-750, are nonmagnetic.

The Hastelloy alloys, of which Hastelloy B is most common, contain primarily nickel and molybdenum and are nearly as strong as the Inconel 700 alloys and somewhat more ductile. These alloys are notable for their resistance to corrosion.

In certain instances the dimensions of a component of an instrument must be very stable in the face of variations of temperature. Iron–nickel alloys containing about 36% nickel have been found to meet this requirement. Invar (Carpenter Technology Corp.) is one such alloy. This material has a coefficient of expansion of  $1.3 \times 10^{-6}/\text{K}$  in the range 0 to 100 °C and about twice this at temperatures up to 200 °C. Better yet, the alloy Super-Invar (from Eagle Alloys or High Temp Metals) has a coefficient of  $0.3 \times 10^{-6}/\text{K}$  in the range 0 to 100 °C. By comparison, the coefficient of expansion of stainless steel is about  $1.5 \times 10^{-5}/\text{K}$ . One disadvantage of Invar in certain applications is that it is magnetic below 277 °C.

After being machined, Invar may require stress relieving to achieve maximum stability. Rosebury<sup>4</sup> suggests the following heat treatment: heat to about 350 °C for one hour and allow to cool in air; heat to a temperature slightly above the operating temperature and cool slowly to a temperature below the operating temperature; repeat step 2. To achieve the minimum coefficient of thermal expansion for Super-Invar, Eagle Alloys recommend the following heat treatment: hold at 830 °C for 10 minutes, quench rapidly in water or with an air blast; hold at 315 °C for 60 minutes and then air cool; hold at 100 °C for 24 hours and then air cool.

Kovar is an iron–nickel–cobalt alloy designed for making metal-to-glass seals. Its thermal expansion characteristics match that of several hard glasses including Corning

7052 and 7056. Unfortunately, Kovar is magnetic and easily oxidized.

### 1.3.3 Copper and Copper Alloys

Copper is a soft, malleable metal. Of all metals (short of the noble metals) it is the best electrical and thermal conductor. For instrument applications, oxygen-free high-conductivity (OFHC) copper is preferred because of its high purity and excellent conductivity and because it is not subject to hydrogen embrittlement. Ordinary copper becomes brittle and porous when exposed to hydrogen at elevated temperatures such as may occur in welding or brazing.

Alloys of copper are classified as either brasses or bronzes. *Brass* is basically an alloy of copper and zinc. Generally, brass is much easier to machine than steel but not quite so strong, and its electrical and thermal coefficients are about half those of pure copper. Its properties vary considerably with the proportions of copper to zinc, and also with the addition of small amounts of other elements. Soldering or brazing readily joins brass parts. Brass is expensive, but, because it is easy to work, it is well suited to instrument construction, where the cost of fabrication far outweighs the cost of materials.

For general machine work, yellow brass (35% Zn) is most suitable. Free-machining brass is similar except that it contains 0.5–3% lead. Cartridge brass (30% Zn) is very ductile and is suitable for the manufacture of parts by drawing and spinning, and for rivets. It work-hardens during cold forming and may require periodic annealing to 600 °C. Stresses that have built up in a cold-formed part during manufacture can be relieved by heating to 250 °C for an hour.

The name bronze originally meant an alloy of copper and tin, but the term now also refers to alloys of copper and many other elements, such as aluminum, silicon, beryllium, or phosphorus. Bronzes, generally harder and stronger than brasses, often are of special value to the designer of scientific apparatus. Beryllium copper or beryllium bronze is a versatile material whose properties can be considerably modified by heat treatment. In its soft form it has excellent formability and resistance to fatigue failure. Careful heat treating will produce a material that is harder and stronger than most steels. It is useful for the manufacture of springs and parts that must be corrosion-resistant.

Also it is nearly impossible to strike a spark on beryllium bronze; this makes it useful for parts to be used in an explosive atmosphere. Beryllium bronze hand tools are available for use around explosive materials. Phosphor bronze cannot be heat-treated, but it is very formable and has great resistance to fatigue. It is used for bellows and springs. Aluminum bronze is highly resistant to corrosion and is strong and tough. It is used in marine applications. Bronze stock can be manufactured by sintering. Compacting and heating a bronze powder at a temperature below the melting point produces sintered bronze. It is quite porous in this form and can be used as a filter or can be impregnated with lubricant for use as a bearing.

### 1.3.4 Aluminum Alloys

Aluminum alloys are valued for their light weight, good electrical and thermal properties, and excellent machinability. Aluminum is about one third as dense as steel. Its electrical conductivity is 60% that of copper, but it is a better conductor per unit weight.

Aluminum is resistant to most corrosive agents except strong alkalis, owing to an oxide film that forms over the surface. This layer is hard and tenacious and forms nearly instantaneously on a freshly exposed surface. Because the oxide layer is an insulator, electrical connections to aluminum parts are problematic. In some applications aluminum parts can accumulate a surface charge. This problem can be overcome by copper- or gold-plating critical surfaces. The oxide layer makes plating difficult, but reliable plating can be carried out by specialty plating shops. Aluminum parts cannot be easily soldered or brazed (see, however, Handy & Harman AL802 flux-cored aluminum solder). Heliarc or TIG (tungsten inert gas) welding is quite practical with aluminum (see, for example, the plethora of welded-aluminum-frame bicycles on the market). Welds in aluminum tend to be porous, and welding weakens the metal in the vicinity of the weld because of the rapid conduction of heat into the surrounding metal during the operation.

Aluminum stock is produced in cast or wrought form. Wrought aluminum, produced by rolling, is used for most machine work. A four-digit number that indicates the composition of the alloy, followed by a suffix, beginning with T or H that specifies the state of heat treatment or work hardness, specifies aluminum alloys. The heat-treatment scale

extends from T0 through T10. T0 indicates that the material is dead soft, and T10 that it is fully tempered. The 1000-series alloys are more than 99% aluminum. The 2000-series alloys are aluminum alloyed with copper. The 6000-series alloys contain magnesium and silicon. The 7000-series are alloys with zinc. Readily available stock includes 1100-T0, a soft, formable alloy that is nearly pure aluminum and is used for extrusions. 2024-T4 or -T6 is as strong as annealed mild steel and is used for general machine work. 6061-T6 is not quite so strong as 2024-T4, but is more easily cold-worked. Its machinability is excellent. The strongest readily available aluminum alloy is 7075-T6. It is easily machined; its corrosion resistance, however, is inferior to that of other alloys.

Aluminum is available as plate, bar, and round and square tube stock. Aluminum jig plate is useful for instrument construction. This is rolled plate that has been fly-cut by the manufacturer to a high degree of flatness.

Aluminum may be anodized to give it a pleasing appearance and to improve corrosion resistance. Anodizing of aluminum is carried out inexpensively by most commercial plating services. A number of proprietary formulas, such as Birchwood-Casey AlumaBlack, are available to blacken aluminum to reduce its reflectivity. Beware, however, of difficulties in making a reliable electrical connection to an anodized aluminum surface.

### 1.3.5 Other Metals

Magnesium alloys are useful because their density is only 0.23 times that of steel. Magnesium is very stiff per unit weight, and is completely nonmagnetic. It is very easy to machine, although there is some fire hazard. Magnesium chips can be ignited, but not the bulk material. The strength of magnesium alloys is considerably reduced above 100 °C. The cost of these alloys is at least 10 times that of steel.

Titanium is about half as dense as steel, while titanium alloys are stronger than steel. Titanium is very resistant to corrosion and completely nonmagnetic. It is readily machined, if one proceeds slowly with sharp carbide-tipped cutters, and can be welded under an inert gas atmosphere. Titanium and titanium alloys (99%) are useful for extreme-temperature service, since they maintain their mechanical properties from -250 °C to well above 400 °C. A titanium-manganese alloy (Ti-8Mn; 6.5–9% Mn) retains

a tensile strength greater than 100 000 psi to temperatures in excess of 400 °C. Its strength and hardness (up to 36 R<sub>c</sub>) can be varied considerably by heat treatment. Titanium–aluminum–vanadium alloy (Ti-6Al-4V; 5.5–6.7% Al; 3.5–4.5% V) is a similarly strong, versatile material. These alloys are more easily machined than unalloyed titanium. Titanium fasteners are available.

Molybdenum is a refractory metal. It exhibits a very low, uniform, surface potential and is used for fabricating electrodes and electron-optical elements. Molybdenum is brittle and must be machined slowly using very sharp tools. Molybdenum is produced from a melt or from a sintered powder. The sintered stock is very difficult to machine.

Tungsten is the most dense of the readily available elements. Its density is 2.5 times that of steel. Compact counterweights in scientific mechanisms are often made of this material. Tungsten is also used for extreme high-temperature service, such as torch nozzles and plasma electrodes. Suppliers of refractory metals such as tungsten and molybdenum include A. D. Mackay and Philips Elmet.

### 1.3.6 Plastics

Plastics are classified as either thermoplastic or thermosetting. Within limits, *thermoplastics* can be softened by heating and will return to their original state upon cooling. *Thermosetting plastics* undergo a chemical change when heated during manufacture, and they cannot be re-softened.

Phenol-formaldehyde plastics, or phenolics, are the most widely used thermosetting plastics. Phenolics are hard, light in weight, and resistant to heat and chemical attack. They make excellent electrical insulators. Phenolics are usually molded to shape in manufacture and typically incorporate a paper or cloth filler. Thus reinforced, the phenolic plastics are easily sawn and drilled. Bakelite is a linen-reinforced phenolic plastic.

Representative thermoplastics include polyethylene, nylon, Delrin, Plexiglas, Teflon, and Kel-F.

Polyethylene is inexpensive, machinable, and very resistant to attack by most chemicals. It is soft and not very strong. Its physical properties vary with the molecular weight and the extent of chain branching within the polymer. It is produced in high-density and low-density forms that are whitish and translucent, respectively. It is also manufactured in a blackened form that is not degraded by ultraviolet radi-

ation, as is the white form. Polyethylene softens at 90 °C and melts near 110 °C. It can be welded with a hot wire or a stream of hot air from a heat gun. It can be cast, although it shrinks a great deal upon solidifying. Care must be exercised when heating polyethylene because it will burn. Polyethylene is an excellent electrical insulator, with a dielectric strength of 40 000 V/mm [1000 V/mil (0.001 in.)].

Nylon, a polyamide, is strong, tough, and very resistant to fatigue failure. It retains its mechanical strength up to about 120 °C. It does not cold-flow, and is self-lubricating and thus is useful for all types of bearing surfaces. Nylon is available as rod and sheet in both a white and a black form. Nylon balls and gears are also readily available. There is no fire hazard with nylon, since it is self-extinguishing. It is hygroscopic, and its volume increases and strength decreases somewhat when moisture is absorbed.

Delrin is a polyacetal resin with properties similar to hard grades of nylon in most respects, except that it is not hygroscopic.

All plastics are electrical insulators, although the hygroscopic ones are liable to surface leakage currents. DuPont produces a polyamide film known as Kapton that is widely used as an electrical insulator. This material has a high dielectric strength in thin films. It is strong and tough, resists chemicals and radiation, and is a better thermal conductor than most plastics. It is used in sheets as an insulator between layers of windings in magnets. Kapton-insulated wire is available. DuPont also supplies bulk polyamide in standard shapes under the trade-name Vespel. This material is relatively hard, strong and machinable. It is probably the best plastic material for use in vacuum because of its low vapor pressure and its high-temperature stability. Amoco produces a line of engineering plastics under the name Torlon that are based upon poly(amide-imide) copolymer. These thermoplastics are very strong and can be used up to 250 °C. Torlon is also an excellent choice when plastics must be used in vacuum.

Plexiglas, Lucite, and Perspex are all trade names for polymethylmethacrylate (PMMA). These materials, also known as acrylic plastics, are strong, hard, and transparent. Because they are machinable, heat-formable, and shatter-resistant, polymethylmethacrylates are used in many applications as replacements for glass.

Thermoforming of acrylic sheet into curved surfaces is a fairly simple operation. Begin with a wooden or metal

pattern of the desired shape. Rest the sheet on top of the pattern, warm the plastic to about 130 °C, and permit gravity to pull the softened sheet down over the form. The forming may be done in an oven, or the plastic may be heated with infrared lamps or with hot air from a heat gun. It is necessary to apply the heat slowly and uniformly.

Polymethylmethacrylate dissolves in a number of organic solvents. Parts can be cemented by soaking the edges to be joined in acetone or methylene chloride for about a minute and then clamping the softened edges together until the solvent has evaporated.

General Electric manufactures a polycarbonate plastic known as Lexan. It is transparent and machinable, it has excellent vacuum properties, and is so tough as to be nearly unbreakable. It is the preferred plastic replacement for window glass. Lexan is one of the most useful plastics for instrument construction.

Teflon is a fluorocarbon polymer. It is only slightly stronger and harder than polyethylene, but it is useful at sustained temperatures as high as 250 °C and as low as -200 °C. Teflon is resistant to all chemicals, and nothing will stick to it. In most applications it is not as good as nylon as a bearing surface, since it tends to cold-flow away from the area where pressure is applied. A thin film of Teflon, such as is found on modern cookware, provides dry lubrication and prevents dust and moisture adhesion. Specialty shops that serve the food industry can apply these films. Teflon is about ten times as expensive as nylon.

Kel-F is a fluorochlorocarbon polymer. Its chemical resistance is similar to Teflon, but it is much stronger and harder. Like Teflon, Kel-F does not absorb water.

A number of rubber-like polymeric materials, known as elastomers, find application in scientific apparatus. Materials of interest include Buna-N, Viton-A, and RTV. Buna-N is a synthetic rubber. It can be used at sustained temperatures up to 80 °C without suffering permanent deformation under compression. It is available as sheet or in block form. It is useful as a gasket material and for vibration isolation. Viton-A is a fluorocarbon elastomer similar in appearance and mechanical properties to Buna-N. Useful up to 250 °C, it tends to take a set at higher temperatures. Unlike Buna-N, Viton-A shows no tendency to absorb water or cleaning solvents. RTV, a self-curing silicone rubber produced by General Electric, comes in a semiliquid form and cures to a rubber by reaction with moisture in the air. It is useful as a

sealant, a potting compound for electronic apparatus, and an adhesive. It can be cast in a plastic mold.

### 1.3.7 Glasses and Ceramics

Borosilicate glasses such as Corning Pyrex 7740 or Kimble KG-33 are used extensively in the lab. These glasses are strong, hard, and chemically inert, and they retain these properties to 500 °C. As will be discussed in the succeeding chapter, they are conveniently worked with a natural-gas-oxygen flame.

Quartz or fused silica has a number of unique properties. It has the smallest known coefficient of thermal expansion of any pure material:  $1 \times 10^{-6}/\text{K}$ . By comparison, the thermal expansion coefficient of the borosilicate glasses is  $3 \times 10^{-6}/\text{K}$ , and that of steel is  $15 \times 10^{-6}/\text{K}$ . Quartz can be drawn into long, thin fibers. The internal damping in these fibers is very low, and their stress-strain relation for twisting is linear to the breaking point. Springs of quartz fibers provide the best possible realization of Hooke's law. Torsion springs and coil springs made of a quartz fiber are used in delicate laboratory balances. Quartz retains its excellent mechanical properties up to 800 °C. It softens at 1500 °C and hence can only be worked with a hydrogen-oxygen flame.

Alumina ( $\text{Al}_2\text{O}_3$ ) is the most durable and heat-resistant of the readily available ceramics. Only diamond, borazon, and a few exotic metal carbides are harder. The compressive strength of alumina exceeds 2000 MN/m<sup>2</sup> (300 000 psi), although its tensile strength and shear strength are somewhat less than those of steel. Alumina is brittle and, because of its hardness, it ordinarily is shaped or cut by grinding with a diamond-grit wheel. Alumina rod, tube, thermocouple wells, and electrical insulators are available. These shapes are produced by molding a slurry of alumina powder and various binders to the desired shape and then firing at high temperature to produce the hard ceramic. One particularly useful product is alumina rod that has been centerless-ground to a roundness tolerance of better than  $\pm 0.03$  mm ( $\pm 0.001$  in.) and a straightness of 2 mm/m (1/32 in./ft.) (see for example, McDanel Refractory Porcelain Co.). A transparent form of alumina is sapphire, which is a clear, hard material very valuable in optical applications (see Chapter 4). Polycrystalline, as well as single-crystal, (sapphire) balls are available that have been ground to a roundness tolerance of  $\pm 0.001$

mm ( $\pm .00003$  in.) (Industrial Tectonics, Inc.). They are surprisingly inexpensive: balls of 3 to 12 mm (0.125 to 0.500 in.) diameter cost only a few dollars. These balls are used as electrical insulators and bearings, and in valves and flowmeters. Alumina circuit board substrate is available as plate 0.3 to 1.5 mm (.010 to .060 in.) thick and 15 cm (6 in.) square. Owing to the scale of the circuit board industry, the facilities for custom fabrication by laser cutting of alumina plate are widely available and inexpensive.

Corning produces a machinable glass ceramic known as Macor, and Aremco produces machinable ceramics called Aremcolox. These materials are not nearly so strong or hard as alumina, but can be machined to intricate shapes that find wide application in instrument work. They can be drilled and tapped and can be turned and milled with conventional high-speed steel tools. They resist chipping and are insensitive to thermal shock. Machinable ceramics have a dielectric strength of 40 000–80 000 V/mm [1000–2000 V/mil (.001 in.)].

Aluminum silicate, or lava, is another useful, machinable ceramic. Blocks of grade A lava are available in the unfired condition. This material is readily machined into intricate shapes that are subsequently fired at 1050 °C in air for half an hour. The fired ceramic should be cooled at about 150 K/hour to prevent cracking. It grows about 2% upon firing.

The development of so-called technical ceramics has burgeoned over the past two decades. A remarkable range of pure oxide ceramics as well as composite materials are available with mechanical, electrical, magnetic, and thermal properties to fit all sorts of special applications. The ultra-low thermal expansion glass ceramic, Zerodur, produced by Schott Glass, is an example that has found use in the construction of highly stable mechanical elements and mirrors. A glass ceramic consists of a crystalline phase dispersed in a glass phase. The quartz-like crystalline phase in Zerodur has a negative thermal coefficient, while the glassy phase has a positive coefficient. In the appropriate proportion a material with a thermal expansion coefficient of the order of  $10^{-8}/\text{K}$  is obtained.

## 1.4 JOINING MATERIALS

Soldering, brazing, welding, or riveting may permanently join the parts of an apparatus, or demountable fasteners such as screws, rivets, pins, or retaining rings may join them.

### 1.4.1 Threaded Fasteners

Threaded fasteners are used to join parts that must be frequently disassembled. A thread on the outside of a cylinder, such as the thread on a bolt, is referred to as an external or male thread. The thread in a nut or a tapped hole is referred to as an internal or female thread. The terminology used to specify a screw thread is illustrated in Figure 1.18. The *pitch* is the distance between successive crests of the thread, equivalent to the axial distance traveled in one turn of a screw. The pitch of metric threads is specified in millimeters. In English units the pitch is specified as the number of turns or threads per inch (tpi). The *major diameter* is the largest diameter of either an external or an internal thread. The *minor diameter* is the smallest diameter. In the United States, Canada, and the United Kingdom, the American Standard specifies the form of a thread. SI metric threads conform to ISO and DIN standards. Both the *crest* and *root* of these threads are flat, as shown in Figure 1.18, or slightly rounded. The American Standard and SI *thread angle* is always 60°.

Two American Standard thread series are commonly used for instrument work. The *coarse-thread series*, designated UNC for Unified National Coarse, is for general use. Coarse threads provide maximum strength. The *fine-thread series*, designated UNF, is for use on parts subject to shock or vibration, since a tightened fine-thread nut and bolt are less likely to shake loose than a coarse-thread nut and bolt. The UNF thread is also used where fine adjustment is necessary. The pitch of standard metric fasteners is intermediate between UNC and UNF-series threads. There

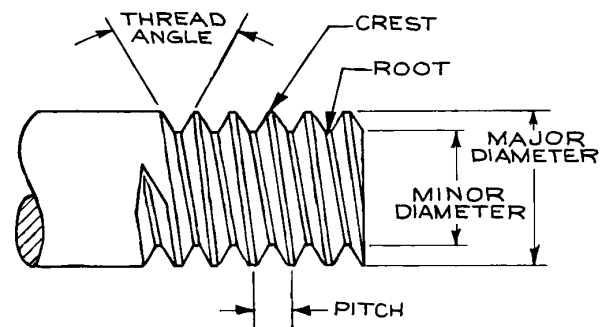


Figure 1.18 Screw-thread terminology.

is no interchangeability between American Standard and metric threaded fasteners. An obsolete thread type for small screws that can show up in some older equipment is the BA (British Association) thread type, which was very close to a metric thread.

The fit of American Standard threaded fasteners is specified by tolerances designated as 1A, 2A, 3A, and 5A for external threads and 1B, 2B, 3B, and 5B for internal threads. The fit of 2A and 2B threads is adequate for most applications, and such threads are usually provided if a tolerance is not specified. Many types of machine screws are only available with 2A threads. The 2A and 2B fits allow sufficient clearance for plating. 1A and 1B fits leave sufficient clearance that dirty and scratched parts can be easily assembled. 3A and 3B fits are for very precise work. 5A and 5B are interference fits, such as are used on studs that are to be installed semipermanently. The metric tolerance class 6g is approximately equivalent to American Standard tolerance class 2A. Class 4g6g is a more precise thread approximately equivalent to class 3A.

A left-hand thread may be specified to assure uniqueness or when it is thought that a disturbing rotation or vibration would tend to loosen a right-hand threaded fastener. Circumferential notches between the flats or an arrow on the face indicating the direction to be turned to tighten the nut identify a left-hand thread nut.

In the American Standard, a gauge number specifies major diameters less than 1/4 inch. The relation between gauge number,  $n$ , and diameter,  $d$ , in inches, is given by

$$d = (0.013n + 0.060) \text{ in.}$$

The specification of threaded parts on a mechanical drawing is illustrated by the following examples:

(1) An externally threaded part with a nominal major diameter of 3/8 in., a coarse thread of 16 threads per inch, and a 2A tolerance is

$$3/8-16 \text{ UNC}-2A$$

(2) An internally threaded part with a nominal major diameter of 5/8 in., a fine thread of 18 tpi, and a 2B tolerance is designated

$$5/8-18 \text{ UNF}-2B$$

(3) An American Standard, 8 gauge (0.164 in. diameter) coarse thread (32 tpi) is designated

$$8-32 \text{ UNC}$$

(4) A metric thread with a nominal major diameter of 5 mm and a pitch of 0.8 mm is designated

$$M5 \times 0.8$$

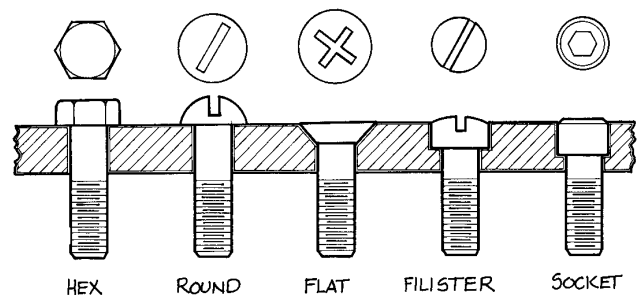
The letters LH following these designations specifies a left-hand thread.

Specifications for UNC and UNF thread forms are given in Appendix 1.2 and for metric threads in Appendix 1.3. The indicated tap drill size specifies the diameter of the hole to be drilled before cutting threads with a tap. For metric threads, the tap drill size is the diameter of the corresponding screw in millimeters minus the pitch in millimeters.

The common forms of machine screws are illustrated in Figure 1.19.

*Hex-head* cap screws are ordinarily available in sizes larger than 1/4 in. They have a large bearing surface and thus cause less damage to the surface under the head than other types of screws. A large torque can be applied to a hex-head screw, since it is tightened with a wrench.

*Slotted-head* screws and *Phillips* screws are driven with a screwdriver. They are available with *round*, *flat*, and *fillister* heads. The flat head is countersunk so that the top of the head is flush. Fillister screws are preferred to round-head screws, since the square



**Figure 1.19** Machine screws: hex head; round head (slotted); flat head (phillips); head; socket head.



shoulders of the head provide better support for the blade of a screwdriver.

*Socket-head* cap screws with a hexagonal recess are preferred for instrument work. These are also known as *Allen* screws. L-shaped, straight, and ball-pointed hex drivers are available that permit Allen screws to be installed in locations inaccessible to a wrench or screwdriver. An Allen screw can only be driven by a wrench of the correct size, so the socket does not wear as fast as the slot in a slotted-head screw. Socket-head screws have a relatively small bearing surface under the head and should be used with a washer where possible.

*Setscrews* are used to fix one part in relation to another. They are often used to secure a hub to a shaft. In this application it is wise to put a flat on the shaft where the setscrew is to bear; otherwise the screw may mar the shaft, making it impossible for it to be withdrawn from the hub. Setscrews should not be used to lock a hub to a hollow shaft, since the force exerted by the screw will deform the shaft. In general, setscrews are suitable only for the transmission of small torque, and their use should be avoided if possible.

Machine screws shorter than 5 cm (2 in.) are threaded their entire length. Longer screws are only threaded for part of their length. Screws are usually only available with class 2A coarse or fine threads.

The type of head on a screw is usually designated by an abbreviation such as “HEX HD CP SCR” or “FILL HD MACH SCR.” For example, a socket-head screw 1 1/2 in. long with an 8–32 thread is designated

8–32 UNC 1 1/2 SOC HD CAP SCR

The torque  $T$  required to produce a tension load  $F$  in a bolt of diameter  $D$  is

$$T = CDF,$$

where the coefficient  $C$  depends upon the state of lubrication of the threads. In general,  $C$  may be taken as 0.2. If the threads are oiled or coated with molybdenum disulfide, a value of 0.15 may be more accurate. If the threads are very clean,  $C$  may be as large as 0.4. The tension load on the bolt is equal to the compressive force exerted by the underside of the bolt head.

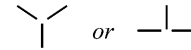
Steel bolts meeting SAE specifications are identified by markings on the head:

SAE grades 0, 1, 2: no mark

SAE grade 3:



SAE grade 5:



SAE grade 8:



The strength of the bolt increases with the SAE grade number. Most common steel bolts are SAE grade 2. In sizes up to 1 in., these bolts have a yield strength of 30 000 to 50 000 psi. The yield strength of SAE grade 5 bolts exceeds 90 000 psi while that of grade 8 bolts exceeds 130 000 psi.

When a bolt is to be tightened to a specified torque, the threads should be first seated by an initial tightening. It should then be loosened and retightened to the computed torque with a torque wrench. The torque corresponding to the yield strength should not be exceeded during this operation.

To obtain maximum load-carrying strength, a steel bolt engaging an internal thread in a steel part should enter the thread to a distance equal to at least one bolt diameter. For a steel bolt entering an internal thread in brass or aluminum, the length of engagement should be closer to two bolt diameters.

Machine screws are readily available in a remarkable range of materials and finishes. Run-of-the-mill steel screws are often zinc plated, but nickel- and chrome-plated screws are available for enhanced corrosion resistance and better appearance. In addition to steel screws, major suppliers provide screws and nuts in types 304 and 316 stainless steel, Monel, brass, silicon bronze, 2024-T4 aluminum, titanium, Nylon, Teflon, and fiberglass-reinforced plastic. The choice of the material for a threaded fastener depends on the desired strength in the joint to be made, the materials to be joined, and the environment of the finished assembly. The possibility of corrosion weakening a joint or preventing disassembly is often a major factor. Galvanic action between the materials of the fastener and the joined assembly is the culprit here. The material of the fastener and of the metals being joined should be close to one

another in the galvanic series. Metal screws are to be avoided when joining metals far apart in the galvanic series; brazing is preferable in this situation.

As a threaded assembly is tightened, the interlocking threaded surfaces are held together by the stress created by elastic deformation of the fastener and the clamped parts. Shock, vibration, and thermal cycling can relieve this tension and the assembly becomes loose. All manner of schemes have been developed to keep threaded parts locked together. For instrument assembly, lockwashers often suffice. These come in two forms: the spring lockwasher and the serrated washer. The former is simply a washer in the form of a single-turn spring that is placed under the head of a bolt or under a nut before it is screwed onto the bolt. The washer is flattened as the assembly is tightened. The spring force creates friction to resist the bolt turning. A toothed or serrated washer simultaneously engaging the surface of the joined part and the underside of a bolt or nut similarly resists a bolt's turning loose. As an aside it is usually noted that a lockwasher should not be used along with a flat washer since the lockwasher will engage the flat washer on one side, but the other side of the flat washer will slip. Self-locking screws are made with a nylon insert in the threads. The nylon deforms as the screw is installed to take up the clearance between the mating threaded surfaces. Nuts with nylon inserts, so-called "aircraft nuts," are readily available and quite effective in resisting loosening under vibration. A simple means of securing a nut and bolt assembly is to install a second nut, the locknut, after the first nut is tightened. With the first nut held securely with a wrench to prevent its turning, the locknut is tightened down on the first. The effect is to create opposing tension in the bolt that takes up clearance in the threads and increases friction between the mating surfaces. A threaded assembly can also be secured with adhesive applied to the threads to take up clearances and stick the surfaces together. Specialized adhesives for this purpose (Loctite) are available in a range of strengths for both removable and permanent installations.

Excessive wear may result when a steel bolt is screwed into a threaded (tapped) hole in a soft material such as plastic or aluminum. Threaded inserts (Heli-Coils) can be installed in the softer material to alleviate this problem. These inserts are tightly wound helices of stainless steel or phosphor bronze wire with a diamond-shaped cross

section. The insert is placed in a tapped oversize hole and the mating bolt is screwed into the insert. Special taps are required to prepare a hole for the insert, and a special tool is required to drive the insert into the hole.

### 1.4.2 Rivets

Rivets are used to join sheet-metal or sheet-plastic parts. They are frequently used when some degree of flexibility is desired in a joint, as when joining the ends of a belt to give a continuous loop. Common rivet shapes are shown in Figure 1.20. Rivets are made of soft copper, aluminum, carbon steel, and stainless steel, as well as plastic. As for

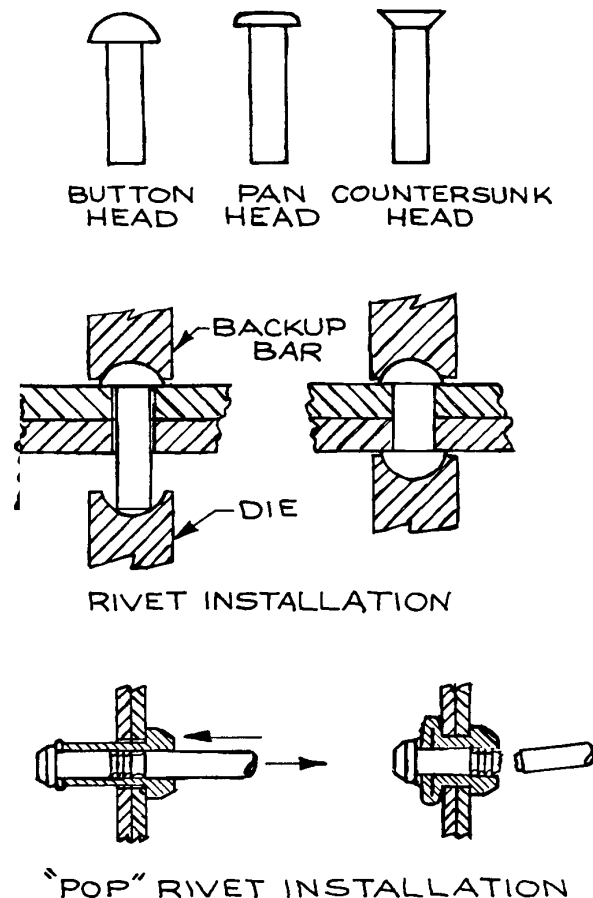


Figure 1.20 Rivets and riveted joints.

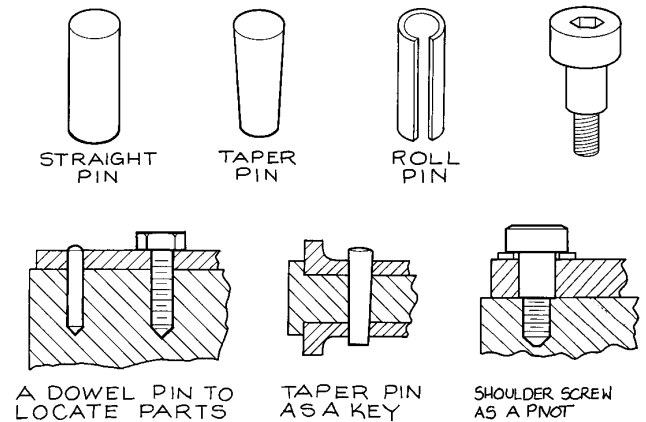
threaded fasteners, the strength required in a joint and the possibility of corrosion are considered in choosing the material for a rivet. To join two pieces, a hole, slightly larger than the body of the rivet, is drilled or punched in each piece. The holes are aligned and the rivet is inserted. The length of the rivet should be such that it protrudes a distance equal to one to one and half times its diameter. The head of the rivet is backed-up with a heavy tool while the plain end of the rivet is “upset” by hammering to draw the two pieces together. The hammering action swells the body of the rivet to fill the hole. Vise-like hand tools with appropriate die sets are available for setting rivets in a one-handed operation.

*Blind rivets* or “pop” rivets, illustrated in Figure 1.20, are useful in the lab. These can be installed without access to the back side of the joint. An inexpensive blind rivet tool is required. The tool grasps a mandrel that passes through the center of the rivet, the rivet is inserted into the hole, and the rivet tool pulls the mandrel back until it breaks. The head of the mandrel rolls the stem of the rivet over to form a head. Some pop rivets are designed so that the remaining broken portion of the mandrel seals the center hole of the rivet. The installation of pop rivets in carefully positioned, reamed holes is an excellent means of aligning two thin pieces with respect to one another. Pop riveted joints can be regarded as semipermanent. The rivet is easily removed by drilling into its center with an oversize drill until the head separates from the body. With care, the drill will never touch the parts joined by the rivet.

### 1.4.3 Pins

Pins such as are shown in Figure 1.21 are used to precisely locate one part with respect to another or to fix a point of rotation. To install a pin, the two pieces to be joined are clamped together and the hole for the pin is drilled and then reamed to size simultaneously in both pieces.

*Straight pins* are made of steel that has been ground to a diameter tolerance of  $\pm 0.003$  mm ( $\pm 0.0001$  in.). They require a hole that has been reamed to a close tolerance; straight pins are to be pressed into place. If a straight pin is to serve as a pivot or if it is to locate a part that is to be removable, the hole in the movable part is reamed slightly oversize.



**Figure 1.21** Pins and shoulder screws.

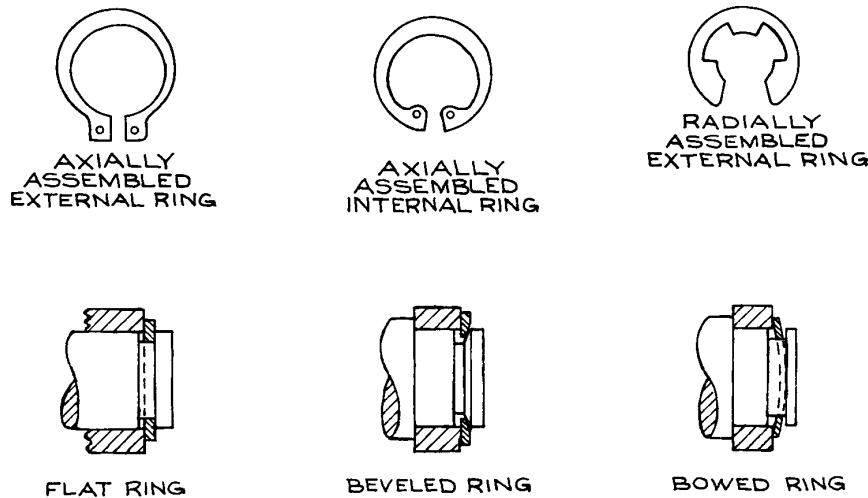
A *roll pin* has the advantage of ease of installation and removal and does not require a precision-reamed hole. The spring action of the wall of the pin holds it in place.

*Taper pins* are installed in holes that are shaped with a special reamer. The American taper is 0.250 in. per ft. The metric taper is 1:50. Plain taper pins are driven into place. Taper pins with the small end threaded are drawn into place with a nut that then secures the installed pin.

A *shoulder screw* is effectively a straight pin with a shoulder to provide location and a threaded end below for attachment. Shoulder screws are used as pivots. These screws are hardened, and the shoulder is ground to a diameter tolerance of  $\pm 0.0001$  in.

### 1.4.4 Retaining Rings

A retaining ring (Figure 1.22) serves as a removable shoulder to position an assembly of parts on a shaft or in a hole. An axially assembled external retaining ring is expanded slightly with a special pair of pliers, then slipped over the end of a shaft and allowed to spring shut in a groove on the shaft. An axially assembled internal ring is compressed, inserted in a hole and permitted to spring open into a groove. A radially assembled external retaining ring is forced onto a shaft from the side. It springs open as it slides over the diameter and then closes around the shaft. Some retaining rings have a beveled edge, as shown in the figure.



**Figure 1.22** Retaining rings and retaining ring installations.

Spring action of the ring on the edge of its groove produces an axial load to take up unwanted clearances between assembled parts. This scheme is frequently used to take up the end-play in a ball bearing. A bowed retaining ring provides another solution to end-play take-up. The variety and range of sizes of retaining rings are truly remarkable. The Truarc Division of Walde Kohinoor is a major supplier of retaining rings.

### 1.4.5 Soldering

In soldering, as well as brazing, two metal surfaces are joined by an alloy that is applied in the molten state to the joint. The melting point of the alloy is below that of the metal of the parts to be joined (the so-called *base metal*). Soldering and brazing differ in the composition of the fusible alloy. Solder is a tin alloy that melts at a relatively low temperature – in the 200–300 °C range. Brazing was originally done with a copper alloy that melts at about 900 °C. There are now many different brazing alloys with working temperatures from 400 °C to more than 2000 °C. Today the term “brazing” usually implies “silver brazing” with silver-containing alloys that flow in the 450–800 °C range.

The molten soldering or brazing alloy must wet the surfaces of the metal parts being joined. When the alloy

wets the surfaces of closely spaced parts, capillary action draws the alloy into the joint to provide the largest possible contact area and correspondingly the strongest possible joint. A dirty or oxidized metal surface will not be wetted. A chemically active flux is applied as the base metal is heated to remove oxides and protect the surface thus cleansed. The flux employed is, to a certain extent, specific to the base metal and the alloy as well as the temperature required for melting the solder or braze alloy.

For routine soldering of copper or brass parts, 50–50 tin–lead solder (50% tin, 50% lead) has traditionally been the alloy of choice. Lead-containing solder is no longer used for piping in contact with potable water. Tin–lead solder has been replaced by 95–5 tin–antimony for joining copper to copper. The 96–4 tin–silver eutectic (96.5 % tin, 3.5% silver) has melting and flow properties similar to tin–lead solder and, with the use of the proper flux, will wet copper, brass, steel, and stainless steel. 60–40 tin–lead solder is used to join electronic components. Soldered joints are appropriate when great strength and cleanliness are not required and when it is most convenient to work at low temperatures. The tensile strength of a copper sleeve joint with tin–lead solder is about 40 MN/m<sup>2</sup> (6000 psi). The shear strength is about 35 MN/m<sup>2</sup> (5000 psi). The strength of a tin–lead joint is considerably reduced at temperatures above 100 °C.

Steel and stainless steel, as well as brass and copper, can be soldered with tin–silver alloy solders. The 96–4 tin–silver eutectic alloy is almost universally applicable in the laboratory for low-temperature soldering. It melts at 221 °C and can produce a joint with a tensile strength of 100 MN/m<sup>2</sup> (15 000 psi) and a shear strength of 75 MN/m<sup>2</sup> (11 000 psi). The manufacturers of soldering and brazing alloys (including Engelhard, Handy & Harman/Lucas–Milhaupt, Harris Products, Wesgo Metals, CWB Ltd.) provide tin–silver alloy solders in wire form, in wire with a rosin flux core for electronic applications, and as a paste or slurry of alloy powder in a liquid flux. The latter is simply painted on the surface where desired and heated indirectly. The absence of flux on adjoining areas keeps solder from flowing where it is not wanted.

Soldering alloys containing lead, zinc, antimony, or cadmium should be avoided. These components are volatile. Zinc is sufficiently volatile at room temperature to be a significant contaminate in high-vacuum applications. Cadmium vapor at soldering temperatures is a serious health hazard.

The first step in soldering a joint is to clean the surfaces to be joined with fine emery paper or steel wool. The surfaces are then covered with soldering flux, assembled, and heated. A gap of 0.1–0.2 mm (.003–.006 in.) is desirable between surfaces to be soldered. In some cases it may be necessary to insert a piece of brass shim stock to maintain this clearance. When heated, the flux removes oxides to cleanse the surfaces so the molten solder will wet them. Sal ammoniac (NH<sub>4</sub>Cl) flux, available from hardware suppliers, is used for lead–tin solders. The suppliers of lead-free solders provide proprietary fluxes. A natural-gas–air flame or propane–air flame is most convenient. The assembly may also be heated on an electric or gas hot plate. Flame heating should be done indirectly if possible. Play the flame on the metal near the joint, particularly on the heavier of the two pieces. Periodically test the temperature of the two pieces by touching each with the solder. When each piece is hot enough to melt the solder, remove the flame and apply solder to the joint. The solder should flow by capillary action into and through the joint. Sufficient solder should be applied to fill the joint and produce a meniscus or fillet of solder at the juncture of the two parts (Figure 1.23). The fillet radius should be 1/16–1/8 in. Residual flux must be removed. This is why it is important that the joint is filled

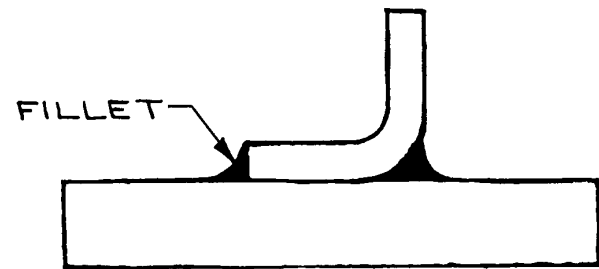


Figure 1.23 A soldered joint.

with molten solder. Flux trapped in the joint will lead to hidden corrosion. Inorganic flux can be dissolved in hot water. Organic fluxes will require an organic solvent.

### 1.4.6 Brazing

A brazed joint is strong and leak tight. The brazing process is amenable to the laboratory and the simplest shop. A brazed assembly can offer real advantages to the designer of an apparatus. Surprisingly often a complex part can be inexpensively made by brazing together an assembly of standard off-the-shelf parts. Brazing offers a relatively easy means of joining materials with very different properties: a hard metal can be joined to a flexible metal stem; a metal with desirable electrical and thermal properties can be incorporated in an assembly made of a metal with entirely different properties; a strong, metal-to-ceramic, brazed joint is possible.

There are hundreds of brazing alloys.<sup>5</sup> The choice of an alloy depends upon the metals to be joined and the application of the brazed part. The strength of the alloy in a joint is seldom an issue in a properly designed joint. In some instances the joint strength will exceed that of the base metal. The wetting and flow characteristics of the molten brazing alloy are often of primary importance. Of course the working temperature of the brazing alloy must be less than the melting point of the base metal, sometimes, much less, to avoid an undesirable phase transition in the base metal.

Alloys are characterized by a *solidus*, the temperature at which, upon heating, liquid mixed with solid first appears, and a *liquidus*, the temperature at which the alloy becomes completely liquid. The difference between the

two temperatures is the *melting range*. In brazing a joint, an alloy with a narrow melting range (0–20 °C) becomes fluid quickly and will tend to flow into the narrowest gap. For a joint with relatively large gap (greater than about 0.1 mm), a large melting range is desired so that upon heating, the alloy remains in the melting range and flows sluggishly to fill the joint. A eutectic alloy melts at a single temperature; there is no melting range.

For routine brazing of steel parts, borate-flux-coated, brass brazing rod is available from hardware suppliers as well as specialized suppliers of brazing and welding materials. The tensile strength of joints made with brass brazing alloy can be of the order of 50 000 psi. An oxyacetylene torch is required for brass brazing alloys.

Silver brazing alloys are preferred for most laboratory work. Silver alloys are lower-melting than brass alloys and can be used on brass as well as steel parts. The tensile strength, as well as the shear strength, of silver-brazed joints is in the 200–350 MN/m<sup>2</sup> (30 000–50 000 psi) range, depending upon the alloy and the joint design. The alloy of approximately 60% silver, 30% copper, and 10% tin is generally applicable. It melts at 600 °C (solidus) and flows at 720 °C (liquidus), and can be worked with the same gas-oxygen torch used for laboratory glassblowing. A fluoride flux is required. Parts made of copper, brass, steel, stainless steel, and even refractory metals can be joined with this brazing alloy.

When volatile metals cannot be tolerated, the silver-copper eutectic (72% Ag, 28% Cu) makes a suitable brazing alloy for copper, stainless steel, Kovar, and many refractory metals. This alloy melts at 780 °C. For torch brazing, an oxyacetylene flame is required. The addition of indium produces a lower-melting alloy with much the same properties as the silver-copper eutectic. Incusil-15 (62% Ag, 24% Cu, 15% In, Wesgo), for example, has a solidus of 630 °C and a liquidus of 705 °C. Silver-bearing brazing alloys containing cadmium or other volatile metals should be avoided. Cadmium vapor is toxic. Volatile components (i.e., zinc) of a brazing alloy can cause significant contamination in a high-vacuum system.

The manufacturers (including Handy & Harman/Lucas-Milhaupt, Engelhard, Harris Products, Wesgo, Aufhauser, CWB Ltd.) provide alloys for a remarkable range of applications along with the fluxes appropriate for each alloy-base metal combination. Brazing alloys are offered

in rod or wire form, as either flux-coated or flux-cored rod, and as a paste incorporating powdered alloy and flux.

Torch brazing is the simplest and most convenient brazing operation. An oxyacetylene torch is required for brass brazing alloys. A natural-gas-oxygen flame can be used for lower-melting alloys. The flame must be nonoxidizing. The flame should be adjusted to provide a well-defined, blue, inner cone 8–10 mm long. The oxyacetylene flame should have a greenish feather at the tip of the inner cone. A fireproof working surface is also required. Most fluxes contain sodium, which emits yellow light when heated. A pair of didymium eyeglasses, such as are used in glassblowing, will filter out the sodium yellow lines and make the work much more visible while brazing. Parts to be brazed should, if possible, be designed to be self-locating. A clearance between parts of 0.03–0.1 mm (.001–.003 in.) will assure maximum strength in the finished joint. Surfaces to be wetted by the brazing alloy must be clean.

The brazing process is not unlike the soldering process described above. Coat the joint surfaces with flux and assemble the parts. Clamp if necessary: parts move as a consequence of differential expansion as they are heated; the blast of gas from a torch can misalign an assembly. Preheat the base metals in the area of the joint nearly to the melting temperature of the brazing alloy. The flux should melt from its crystalline form to a glassy fluid as the brazing temperature is approached. Continue heating while testing both pieces to be joined by touching the base metal with the brazing rod. The rod should melt in contact with both pieces. Do not apply the flame directly to the brazing rod. When the brazing alloy flows, continue to heat the entire area of the joint. There are two dangers: insufficient heat and too localized heating. The molten alloy flows toward the hottest area of the joint; a void may be left in a cool area to weaken the joint. Intense localized heating will consume flux and leave the base metal exposed to oxidation that will prevent further wetting by the molten alloy, once again leaving a weak place in the joint. Overheating also mars the finish of the metal. In general, make sure there is adequate flux and that the quantity of alloy is sufficient to fill the joint and leave a fillet at the juncture of the parts being joined. It is sometimes convenient to pre-place snippets of brazing wire in or near the joint before heating, rather than applying the alloy after the work is heated. Alloy/flux pastes provide another effective means

for assuring that the alloy is effectively introduced in a joint. Finally it is essential that corrosive flux residues be removed from the finished joint. One of the easiest approaches is to quench the brazed part in water before it is completely cool. The thermal shock fractures the glassy residue which then dissolves in the hot water. Be careful here: the thermal shock may also distort thin sections in the part being brazed. A wire brush and emery cloth may be used to finish the job. A rinse in an acid solution may be necessary if the part has been overheated to the extent that the base metal has been oxidized.

Many metals, as well as brazing fluxes and alloys, contain toxic components that are volatile at brazing temperatures. Brazing should be carried out in a well-ventilated hood. Cadmium-containing brazing alloys should not be used in the laboratory. One should be aware that many standard parts are plated with cadmium or zinc. The plating must be removed before heating to brazing temperatures.

Copper, Kovar, Monel, nickel, steel, and stainless steel can be brazed without flux if the metal is heated in a hydrogen atmosphere in a furnace. The hydrogen serves as a flux by reducing surface oxides to produce a clean surface that will be wetted by the molten brazing alloy. Alloys containing chromium must be brazed in dry hydrogen. The surface oxide of chromium found on stainless steel is reduced by dry hydrogen at temperatures above 1050 °C. Furnace brazing provides for more uniform heating of the work so there is usually less distortion than in torch brazing. The work is also stress-relieved to some extent. Hydrogen-brazed parts come out clean and bright. Although it is possible to carry out hydrogen brazing in the lab, the job is better let out to a specialty shop. Commercial heat-treating shops operate dissociated-ammonia tunnel kilns. The work travels slowly through the kiln on a car. Ammonia introduced in the middle of the furnace dissociates to hydrogen and nitrogen to provide the flux.

Parts to be hydrogen-brazed should be chemically clean. Avoid fingerprints, since inorganic salts from the skin are not affected by hydrogen and will remain as stains on the finished work. Assemble the parts and place snippets of filler-metal wire at the junction of the parts to be brazed. For parts that are not securely self-aligned it may be worthwhile to pin the joints to maintain precise alignment. A rough calculation of the volume to be filled will determine

the amount of wire to be used. When the filler metal melts, it is drawn into the nearby joint by capillary action. If the joints to be brazed are not visible from outside the furnace, place a piece of brazing metal on top of the work where it can be observed by the operator, in order to determine when the brazing temperature has been reached. If the walls of the brazing apparatus are opaque, a thermocouple will be required for this purpose.

The filler metal used in hydrogen brazing should melt at a temperature higher than that required for the hydrogen to effectively clean the surface of the base metal. OFHC (oxygen-free high-conductivity) copper has a melting point of 1083 °C and is an excellent filler metal for brazing steel, stainless steel, Kovar, and nickel. If low-melting alloys, such as the silver-copper eutectic are used, a small amount of flux is required. Brazing alloys that contain volatile metals should never be used for hydrogen brazing. It is unwise to braze parts of ordinary electrolytic copper or to use ordinary copper as a filler in a hydrogen atmosphere. Oxides in the metal are reduced to water that is trapped within the crystalline structure of the metal. This water can cause out-gassing problems in vacuum applications, and, if the copper is heated, the expanding water vapor will produce internal cracks (hydrogen embrittlement). For instrument work, OFHC copper should be used for all copper parts that are to be brazed.

A brazed joint between ceramic pieces or between metal and ceramic is entirely possible although the operation must be carried out in a vacuum furnace or in a furnace under an inert atmosphere. Appropriate filler metal-flux pastes are used (for example, Lucanex from Lucas-Milhaupt). One must take into account the difference in thermal expansion of the parts to be joined. Refractory metals, such as molybdenum, tantalum, or tungsten, have coefficients of thermal expansion close to those of most ceramics. Ferrous and cuprous alloys can be brazed to ceramic materials if the metal part is thin in the region of the joint to allow for some give in the metal part as it cools down from the brazing temperature.

### 1.4.7 Welding

In welding, parts are fused together by heating above the melting temperature of the base metal. Sometimes a filler metal of the same type as the base metal is used. Fusion

temperatures are attained with a torch, with an electric arc, or by ohmic heating at a point of electrical contact. For most instrument work the preferred method is arc welding under an argon atmosphere using a nonconsumable electrode. This method is referred to as tungsten-inert-gas (TIG) welding or heliarc welding. All metals can be welded, although refractory metals must be welded under an atmosphere that is completely free of oxygen. Arc welding requires special skills and equipment, but most instruments shops are prepared to do TIG welding of steel, stainless steel, and aluminum on a routine basis.

The effect of the high temperatures involved must be considered when specifying a welded joint. Some distortion of welded parts is inevitable. Provision should be made for remachining critical surfaces after welding. The state of heat treatment of the base metal is affected by welding. Hardened steels will be softened in the vicinity of a weld. Stainless steels may corrode because of carbide precipitation at a welded joint.

Spot welding or resistance welding is a good technique for joining metal sheets or wires in the lab. The joint is heated by the brief passage of a large electric current through a spot contact between the pieces to be joined. Steel, nickel, and Kovar, as well as refractory metals such as tungsten and molybdenum, can be spot-welded. Copper and the precious metals are not easily spot-welded because their electrical resistance is so low that very little power can be dissipated at the point of contact. Small spot-welding units with hand-held electrodes are available from commercial sources (Ewald Instruments, Miyachi Unitek, PUK).

### 1.4.8 Adhesives

Epoxy resins are the most universally applicable adhesives for the laboratory. They are available in a variety of strengths and hardnesses. Epoxies that are either thermally or electrically conducting are also available. Epoxy adhesives consist of a resin and a hardener that are mixed just prior to use. The mixing proportions and curing schedule must be controlled to achieve the desired properties. A variety of epoxy adhesives is available prepackaged in small quantities in the correct proportions. Epoxies will adhere to metals, glass, and some plastics. Hard, smooth

surfaces should be roughened prior to the application of the adhesive. Sandblasting is a convenient method. Parts of a surface that are not to receive the adhesive can be masked with tape during this operation. To obtain maximum strength in an epoxied joint, the gap filled by the epoxy should be 0.05–0.2 mm (.002–.006 in.) wide. To maintain this gap between flat smooth surfaces, shims can be inserted to hold them apart.

Self-curing silicone rubber such as General Electric RTV (room temperature vulcanizing) can be used as an adhesive. RTV is chemically stable and will stick to most surfaces. The cured rubber is not mechanically strong; this can be an advantage when making a joint that may occasionally have to be broken.

Cyanoacrylate contact adhesives, so-called “super glues,” have found wide use in instrument construction. Eastman 910 and Techni-Tool Permabond are representative of this type of adhesive. These adhesives are monomers that polymerize rapidly when pressed into a thin film between two surfaces. They will adhere to most materials, including metal, rubber, and nylon. Cyanoacrylate adhesives are not void-filling and only work to bond surfaces where contours are well matched. An adhesive film of about 0.03 mm (0.001 in.) gives best results. A firm set is achieved in about a minute, and maximum strength is usually reached within a day. When joining metals and plastics, a shear strength in excess of 7 MN/m<sup>2</sup> (1000 psi) is possible.

Ceramic cements are available for a remarkable range of applications (Sauereisen, Dylon). These are inorganic materials (silica, zirconia, graphite) in powder, paste, or liquid form that are used to join ceramics, graphite, and metals. Some are intended to make electrically insulating coatings; others may be used for casting small ceramic parts. The tensile strength of these materials is generally not great, but the materials are serviceable up to 1100 °C and higher (much higher in the case of graphite cements).

### 1.4.9 Design of Joints

The design of the joints between parts of a machine deserves special attention. Experience indicates that failure of an assembly frequently occurs at or near a joint. A well-designed joint must suit the designer's purposes



while at the same time assuring that there will not be an undesirable concentration of stress in the vicinity of the joint.

The basic idea is to design the joint so that the assembly allows for a smooth flow of tensile, bending, and torsional forces through the joint. One may not intend to stress an assembly to the point that noticeable distortion occurs, but imagine that one has done just that: does the strain appear as a smooth bend or flex or does the strained part have a “kink” in the vicinity of a joint?

There are two types of joints, the *butt joint* and the *overlap joint*, as shown in Figure 1.24. A butt joint permits parts to be joined without an increase in thickness. The same advantage is realized with a hybrid butt/overlap joint, at the expense of additional machining and fitting. The woodworker’s or shipwright’s scarf joint is an example of a butt/overlap joint.

A joint should be as strong as the parts leading up to the joint in order to avoid an undesirable stress concentration. It follows that a butt joint is only amenable to welding. The contact area in a butt joint is usually too small for a brazed or soldered joint to provide the strength of the adjoining parts. For thin parts butted together, a full-penetration weld

should be specified. (Imagine two thin pieces of steel butt-welded with only partial weld penetration. If the assembly were flexed it would hinge and kink at the weld.)

An overlap joint can be soldered, brazed, or glued, provided the contact area is sufficient to attain the desired strength. The stronger the material being joined, the longer must be the overlap so that the joint strength is commensurate with the strength of the material.

The butt/overlap joint is amenable to joining with braze, solder, glue, or rivets. The advantage of a constant cross-section is lost if the joint is bolted, unless the material being joined is sufficiently thick that a recessed-head design is possible.

The designer must decide how permanent a joint is to be. A brazed, soldered, or welded joint is ordinarily quite permanent. Reheating, however, can easily disassemble a soldered electrical connection. Similarly, wires and thin metal parts that have been lightly spot-welded together can be broken apart at the weld without damaging the components. Spot-welding is a useful means for making electrical connections between refractory metal parts and parts to be used in a high-temperature environment. A glued joint may be permanent, but, if a soft adhesive such as self-curing silicone is used, the joint can be disassembled, although not conveniently. A riveted joint can be disassembled if required and, of course, a bolted joint can be assembled and disassembled repeatedly. A force fit permits parts to be assembled in precise alignment and without fasteners. A force fit usually involves inserting a round piece into a round hole. The inner piece is made larger than the hole so that the parts interfere upon assembly. The parts are permanently or semipermanently assembled depending on the extent of interference. The parts must be made with great precision to achieve the desired result. Mechanical design and drafting books provide tables of fits. A “light force fit” of parts that can be disassembled with a press or pin driver requires an interference of approximately 0.01 to 0.03 mm (.0005 to .001 in.) in parts up to 5 cm (2 in.) in diameter. Interference two to three times greater requires a heavy press or heating to expand the outer part (or cooling to contract the inner part) to be assembled. Immersion of a part into liquid nitrogen before it is placed into a hole can lead to a very strong force fit. Heavy force fits or “shrink fits” are effectively permanent.

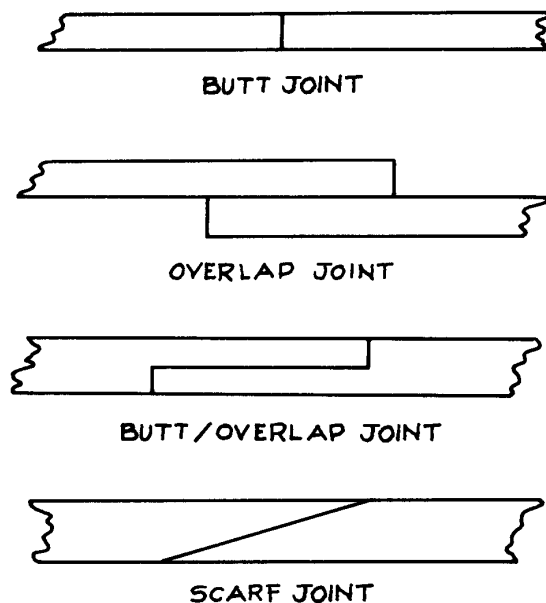
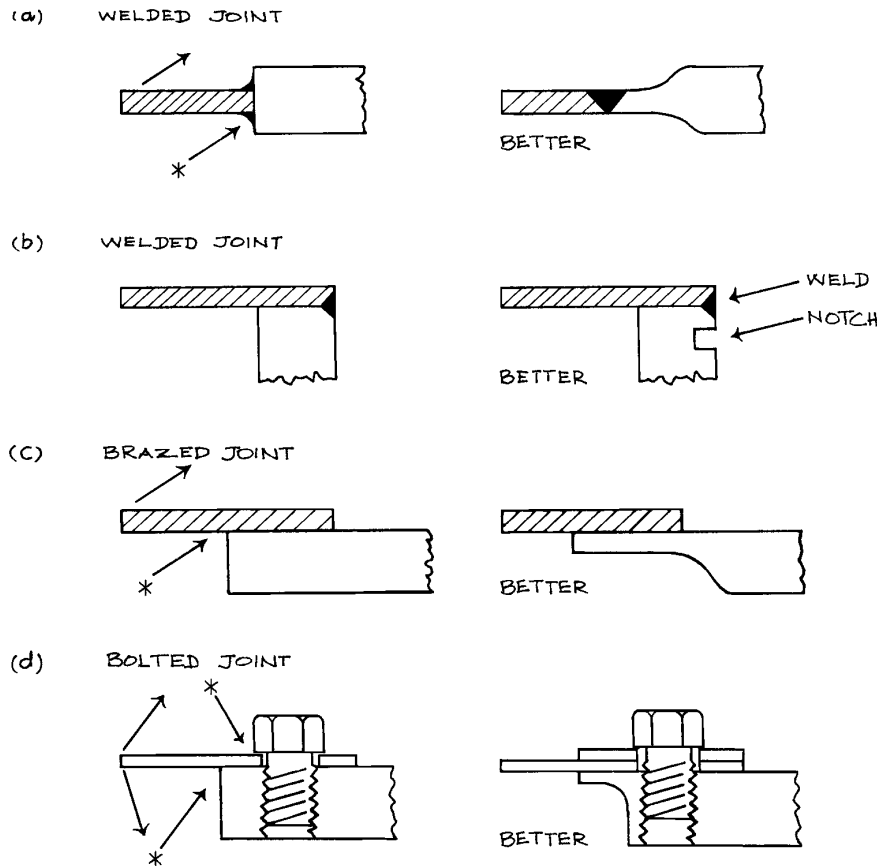


Figure 1.24 Joints.

It is sometimes desirable to allow for some “give” in a joint. For example, there is much to be said in favor of rivets being used to join soft materials such as plastics or fabric. A riveted lap joint in this case will allow for some movement between the parts when a stress is applied. This helps relieve stress concentration at the location of the rivets. As another example, consider a situation where it is anticipated that joined parts will change their relative position or shape in response to a sudden temperature change or differential. A springy fastener such as a bowed retaining ring may be called for.

As general rule, a joint should be designed so that the pieces being joined are of roughly equal strength at the point where they are joined, as suggested in Figure 1.25.

For parts of the same material, this implies that the parts have the same cross-section at the joint or the same thickness if the parts are of the same width. This is important in the design of welded and brazed joints for reasons of both manufacturer and application. Welding (as well as brazing) parts of different thicknesses risks a weak joint because of the differential in the rate of heat loss through the parts being joined. The thinner part is overheated or the thicker part does not achieve the temperature needed for fusion to the proper depth (or, in the case of brazing, to be properly wetted). The resulting state of heat treatment may be different on one side from the other. This will create a stress riser since one side may be more flexible than the other. In general, the dimension of weld penetration should be the



**Figure 1.25** A joint should be designed so that the joined parts are equally strong at the joint.

same as the thickness of the parts near the joint. When a thick piece of metal must be welded to a thin piece, it is common practice to cut a groove in the thick piece, near the joint, to limit the rate of heat loss into the heavier piece, to effectively equalize the dimensions of the weld and the parts at the location of the weld, and to avoid a concentration of stress at the joint [Figure 1.25(b)]. This situation arises when welding a thin tube into a heavy flange (Section 3.6.3 and Figure 3.39).

Joints with discrete fasteners such as bolts or rivets have unique problems. Flexing of the assembled parts may lead to a stress concentration at the corners of the head of the bolt or rivet, as suggested in Figure 1.25(d). The designer must also allow adequate clearance for manufacture and assembly. A tapped hole must leave the machinist clear access with a drill and tap. For assembly there must be sufficient clearance to insert a bolt or screw, as well as space for the wrench or screwdriver used to tighten the fastener. A rivet hole requires access for a drill and reamer as well as clearance for the tools employed to set the rivet. A threaded hole that does not pass clear through a part, that is, a “blind hole,” is sometimes necessary but should be avoided if possible. A blind hole is difficult to keep clean and free of debris that will bind the threaded parts when they are being assembled. There is also the possibility of damage should a bolt longer than specified be driven into the bottom of a blind hole. (The author once spent two weeks looking for a vacuum leak through a crack at the bottom of a blind hole created by an overlong bolt’s being driven into the bottom.)

The strength of a brazed or soldered overlap joint depends upon the length of the lap and the shear strength of the filler metal. An oft-cited rule of thumb for a brazed joint suggests that the length of the lap should be at least three times the thickness of the thinner of two pieces being joined. More precisely, *The Brazing Book*,<sup>6</sup> [also published online by Handy and Harman ([www.handyharmancanada.com](http://www.handyharmancanada.com))], specifies the length of overlap in a brazed joint as

$$\ell = \frac{S_u t}{K S_s},$$

where  $S_u$  is the tensile strength of the weaker of the two parts being joined,  $t$  is the thickness of that part,  $S_s$  is the shear strength of the filler metal, and  $K$  is a “joint integrity

factor” to account for how effectively the filler metal has penetrated the joint and wetted the surfaces being joined (a value of about 0.8 is typical). Note that the overlap must increase when joining high-strength materials.

The space between brazed and soldered parts has a significant effect upon the strength of the joint. A brazed joint between parts spaced 0.03 to 0.10 mm (.001 to .003 in.) apart can be stronger than the base metal. The clearance between “slip-fitted” machined parts is usually adequate, although, for parts of materials of very different coefficients of thermal expansion, one must consider the clearance at the brazing temperature. Shims may be necessary to keep polished surfaces apart. A blind joint is undesirable. The filler metal entering a joint must displace the flux. Residual flux trapped in a joint will lead to corrosion. Additionally, visual inspection is required to assure that the filler metal has flowed completely through a joint. If a blind joint is unavoidable, vent holes should be provided for flux escape and inspection into the deepest part of the joint.

In general, parts to be joined should be designed to be self-aligning when clamped together. If this is not possible, or if very precise alignment is required, parts can be pinned together. The parts are assembled and clamped in the proper position, small holes are drilled and reamed, and pins inserted to maintain alignment while the joints are brazed, welded or glued.

Special consideration must be given to a joint between parts made of dissimilar metals. Two important issues are the differential in the rates of thermal expansion of the joined parts and electrical currents produced at the point of contact. Welds between dissimilar metals may crack as the parts cool or develop cracks when exposed to large temperature swings. Brazed joints between dissimilar metals may be more tolerant of changes in temperature, but electrolytic corrosion can be a problem. If a brazed joint is contemplated, a brazing alloy more noble than the base metals should be used. If electrolytic effects are a problem, one solution may be to interpose an insulating layer between the parts being joined. A layer of paint will suffice provided that the parts are secured in such a way that relative movement between the parts does not abrade the insulating layer. In a wet environment, avoid threaded connections between dissimilar metals; corrosion will occur at the threads where they are hidden from inspection.

### 1.4.10 Joints in Piping and Pressure Vessels

Safety is the primary consideration in specifying or designing a joint in a pressurized conduit or vessel. At high pressures a sudden failure in a joint threatens life and limb. At lower pressures a ruptured joint can result in expensive damage if not personal injury. Of course, having considered safety in the design of a joint, the next requirement is that the joint not leak. Short of personal injury, there is perhaps no worse lab disaster than to spring a water leak late Friday after everyone has gone home for the weekend.

It is generally desirable that a pressure vessel be built up of cylindrical sections: the design and analysis of a cylindrical vessel is easiest and most reliable (see Section 1.6.3); joints between concentric cylindrical sections are most easily manufactured and assembled. Joints in the side of a cylindrical section are to be avoided: the design of such a joint is difficult at best, the joint is a stress riser, and the fabrication is inevitably difficult and expensive. In any event, a joint in a pressure vessel must be as strong as the joined parts.

A joint in a pressure vessel may be demountable or permanent depending on the particular requirements. A demountable joint requires a compressible element of rubber or metal – a gasket – that is trapped between more robust surfaces at the joint. This element is elastically deformed to provide the force required for sealing. A permanent, leak-proof joint can be welded, brazed, soldered, or glued depending on the pressure in the fluid being contained. A pipe-threaded joint may be demountable, but often inconveniently so.

Pipes and pipe fittings are threaded together. Pipe threads are tapered so that a seal is formed when an externally threaded pipe is screwed into an internally threaded fitting. The American Standard Pipe Thread, designated NPT for National Pipe Thread, has a taper of 1 in 16. The diameter of a pipe thread is specified by stating the nominal internal diameter of a pipe that will accept that thread on its outside. For example, a pipe with a nominal internal diameter of 1/4 in. and a standard thread of 18 tpi is designated

1/4–18 NPT

or simply

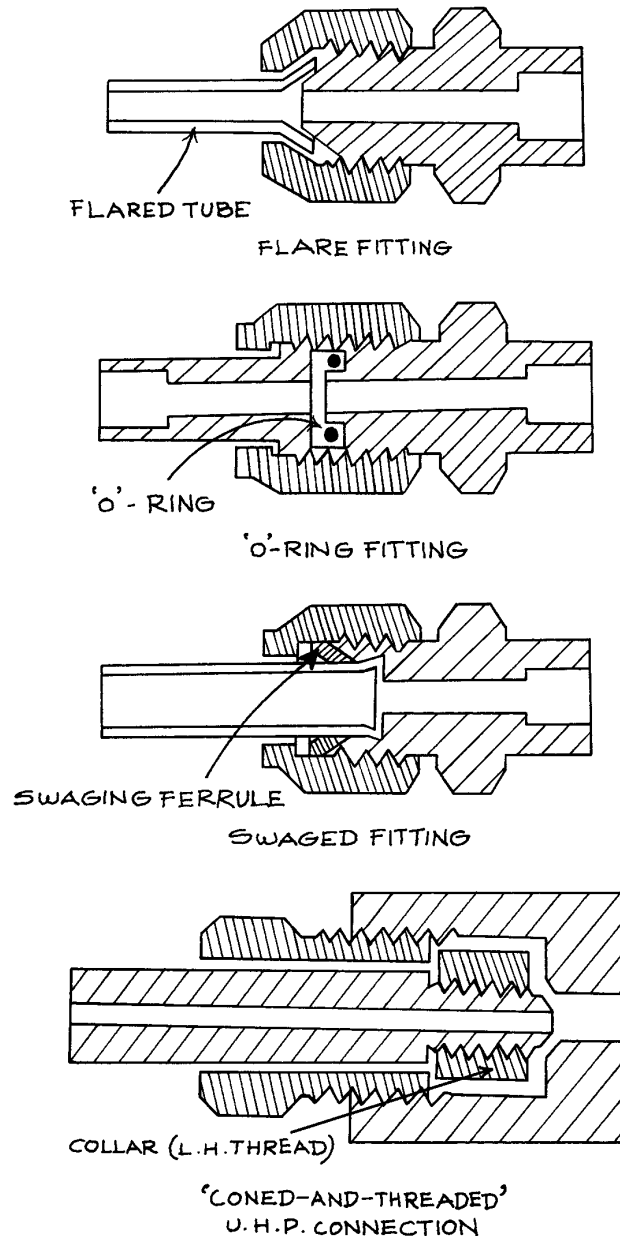
1/4–NPT.

The American Standard Pipe Thread specifications are listed in Appendix 1.4.

In Britain and Europe the usual tapered pipe thread specification is the British Standard Pipe-thread Tapered (BSPT, BS EN 10226 equivalent to ISO 7). The taper is 1 in 16 and the size is given as the nominal bore of a pipe with an external thread. The specifications for the British Standard Pipe-thread Tapered are given in Appendix 1.5. One should take care not to confuse tapered pipe threads (ISO 7) with the British Standard Pipe thread Parallel (ISO 228). The latter are not intended to make a leak-tight joint.

Demountable pressure fittings are illustrated in Figure 1.26. These include joints where the tubing itself is flared or swaged to make one element of the joint, joints where the seal is effected by a rubber O-ring or soft metal gasket, and ultrahigh-pressure joints, in which a conical surface machined on the end of thick-wall tubing is driven against a conical receptacle by a threaded assembly (a so-called “coned-and-threaded” joint). The latter are commonly found in the laboratory as connections to gas cylinders and pressure regulators. The configuration of cone and thread are specific to each gas. These connections have been standardized by the Compressed Gas Association and are identified by a “CGA” number. The mechanical configurations of 33 CGA fittings are given in Appendix 1.6. Note that some of these fittings have a left-hand thread to prevent mixing of fittings for incompatible gases. For all mechanical fittings it is important to follow the manufacturer’s instructions for assembly (this is especially important for the assembly of flare and swage fittings, where precise deformation of the tubing is essential to a leak-proof seal). Pay attention as well to the manufacturer’s specification of pressure limits, bearing in mind that the pressure limitation for each type of connection is likely to decrease with an increase in the diameter of the tubing.

The commercial market offers a remarkable number of inexpensive assemblies that employ standard demountable joints. These include tees, crosses, Ys, reducers, valves, and pumps in sizes from a few millimeters to at least 20 cm in brass, steel, and stainless steel. These assemblies may be intended for the dairy, automotive, or hydraulic equipment industries, but can serve well in the lab. An entire apparatus can be assembled from readily available



**Figure 1.26** Demountable joints for tubing: (a) a flare fitting, (b) a swaged fitting (Swagelok®), (c) an “O”-ring gasketed joint (Swagelok® VCO), (d) a “coned-and-threaded” ultrahigh-pressure connection.

standard assemblies at significant savings in money and time. An hour’s research on the web can save reinventing the wheel.

Permanent joints in piping and pressure vessels are typically soldered or brazed cylindrical overlap joints, or welded cylindrical butt joints.

Roughly speaking and assuming the ratio of diameter ( $D$ ) to wall thickness ( $t$ ) does not exceed about 10, a joint soldered with soft lead–tin solder can be used up to 1 MPa (150 psi) and tin–silver solder to perhaps 3 MPa; brazed joints are acceptable to at least 10 MPa and welded joints to 100 MPa (assuming an appropriately certified welding technician). Above 100 MPa and up to 1 GPa, a system should be built up of specialty ultrahigh-pressure assemblies employing coned-and-threaded ultrahigh-pressure fittings. For brazed or soldered, thin-wall vessels ( $D/t > 10$ ), one must be careful that hoop stress (Section 1.6.3) under pressure does not exceed the shear strength of the filler metal.

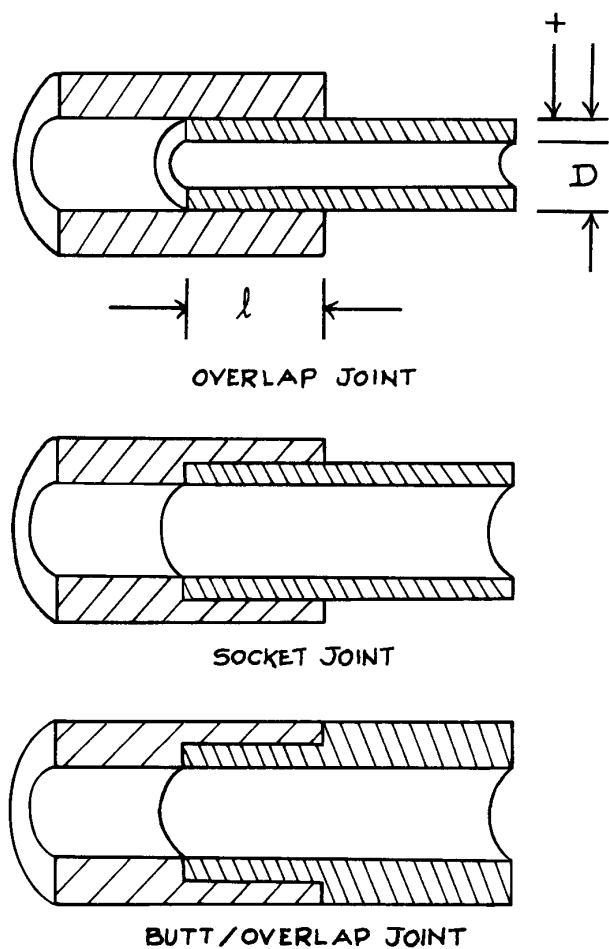
Brazing or soldering joints in cylindrical sections is most practical in the lab. Joint design is illustrated in Figure 1.27. The simple overlap joint requires no machining provided one finds stock tubing that nests. The butt/overlap joint makes for a constant wall thickness. The socket joint is used to install a fitting on a length of tubing (see Figure 1.26). Manufacturers of demountable fittings provide fittings with sockets for tubing of standard diameters. *The Brazing Book* gives the length of overlap for a brazed tubular joint as:

$$\ell = \frac{S_u t (D - t)}{K S_s D} \quad (1.1)$$

where  $S_u$  is the tensile strength of the weaker of the two tubes being joined,  $t$  is the thickness of that tube,  $S_s$  is the shear strength of the filler metal, and  $K$  is the joint integrity factor (see above).

## 1.5 MECHANICAL DRAWING

Mechanical drawing is the language of the instrument designer. Initial ideas for a design are expressed in terms of simple, full-scale drawings that develop in complexity and detail as the design matures. The construction of an



**Figure 1.27** Brazed joints in tubing: (a) an overlap joint, (b) a socket joint, (c) a butt/overlap joint.

apparatus is realized through communications to shop personnel in the form of working drawings of each part of the device. Successful design and construction relies on the designer's command of the language of mechanical drawing. Even in the event that a scientist does both the design and machine work, the completed instrument will benefit from the preparation of carefully executed mechanical drawings.

### 1.5.1 Drawing Tools

Earlier editions of *Building Scientific Apparatus* provided a list of the draftsman's tools – pencils, triangles, T-square,

and so on – along with instructions in their use. The traditional instruments are still appropriate for the preparation of one or two drawings, but, more and more, computer-aided design (CAD) programs have replaced them. For the scientist for whom instrument design is a small, occasional activity in the pursuit of broader scientific goals, a CAD program offers relatively simple, but important, advantages. Lines can be positioned precisely on a drawing and the line weight is consistently and accurately controlled. Scaling is done precisely so that dimensions can be determined directly from the drawing. Deletions are done at the stroke of a key – a significant time-saver compared to the difficult process of erasing pencil lines. Shapes can be duplicated and tracings can be made with ease. Files can be saved in an orderly fashion and easily retrieved for modification and updating. Lettering, once the laborious trademark of good mechanical drawing, is done at the keyboard.

In truth, CAD programs offer much more than can be effectively employed by the scientist-draftsman. With the remarkable power of modern CAD programs, “modeling” has evolved as a new method of design. The traditional draftsman-designer developed the plan of an instrument in two-dimensional orthographic projections. With a powerful CAD program it is possible to model an object in a three-dimensional representation and extract the required two-dimensional projections after the fact. This approach has considerable appeal since the fit between parts can be seen graphically and interferences easily avoided, however, from the point of view of this book, it is hard to justify the time and effort required to learn to use a three-dimensional CAD program, as well as the cost of many of these programs. Another apparent advantage of computer-assisted design with powerful drafting programs is that the program can produce computer files that will drive the lathes and milling machines in the shop, thus obviating the need for drawings on paper, however, the integration of computer-aided design and computer-aided machining (CAD/CAM) is rarely practical for small model shops building one-off instruments. The preparation of two-dimensional orthographic working drawings is, at least for the present, an essential step in building an instrument. The designer of scientific apparatus must learn the language of mechanical drawing.

The most modest drafting software (as opposed to “drawing” software) will ordinarily serve the requirements of the

builder of scientific apparatus – Claris Cad, MiniCAD, Graphite, or Vector Works for Mac and AutoCAD-LT or ProEngineer for PC are examples of entirely satisfactory programs. If a three-dimensional modeling program is available, one quickly discovers that the two-dimensional routines at the front end of these programs will fulfill most needs. It follows that only modest computing power will be required.

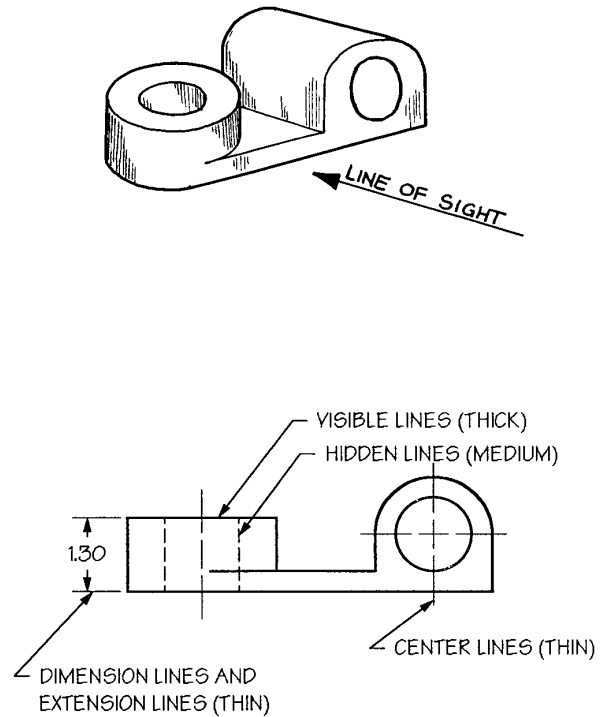
A drafting program incorporates hundreds or even thousands of commands. The user must become conversant with a significant subset of these commands to make progress without frequent reference to a manual. Many successful designers make a practice of learning at least one new command during every session at the computer.

In contrast to the modest CPU requirements for computer drafting, it is advantageous to have the largest possible monitor – a 21in. monitor is none too large. The conception of an instrument in the drawing phase is definitely enhanced if an entire, full-scale, drawing can be viewed at once.

A printer or a plotter is an essential element of a CAD system since one ordinarily must send paper copies of working drawings to the shop. Although reduced-scale drawings are appropriate for large industrial projects, the instructions for the fabrication of the parts of an instrument are best communicated in full-scale drawings. The letter-size format of an office laser printer is usually too small for mechanical drawings. Investment in a large-format laser or ink-jet printer or a pen plotter is frequently worthwhile. Ink-jet printers accepting paper a meter wide are now available. The large-format, E-size, pen plotter is the drafting-room standard.

### 1.5.2 Basic Principles of Mechanical Drawing

If mechanical drawing is the designer's language, then the line is the alphabet of this language. Some of the basic lines used in pencil drawing are illustrated in the plane view shown in Figure 1.28. Three line widths are employed: *thick lines* for visible outlines, short breaks, and cutting planes (explained below); *medium lines* for hidden outlines; *thin lines* for center, extension, dimension, and long break lines. A CAD program typically chooses a

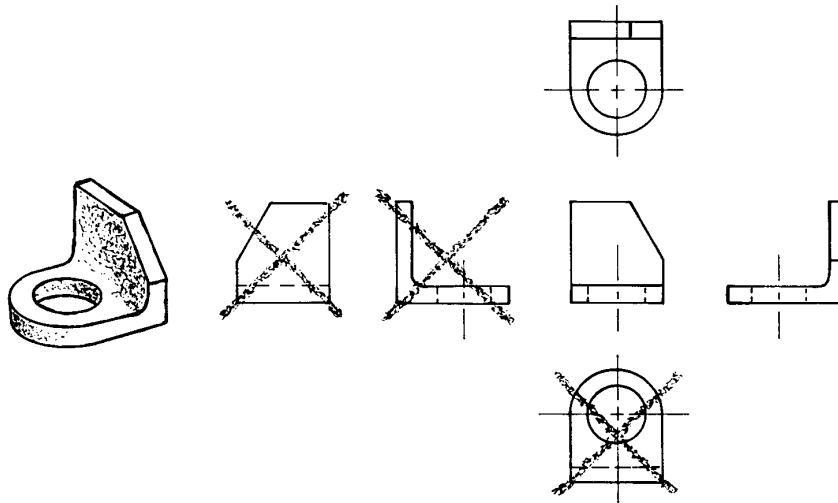


**Figure 1.28** Lines used in mechanical drawing exemplified in a CAD drawing of a side view of the object at the top.

line width of 0.5 mm for thick lines, 0.3 mm for medium lines, and 0.2 mm for thin lines. If only two line weights are available, medium lines are shown thin. For pencil-and-paper work, H-grade lead is used for heavy lines and lettering, 2H for medium and light lines, 4H for layout.

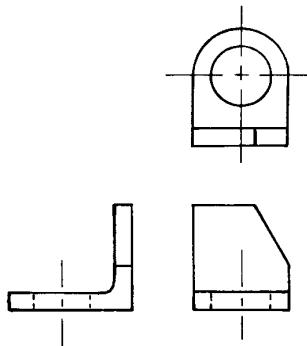
The shape of an object is described by *orthographic projection*. This is the view that one obtains when an object is far from the eye so that there is no apparent perspective. There are six principal views of an object, as illustrated in Figure 1.29. As indicated in the illustration, not all views are required to completely describe an object. The draftsman should present only those views that are necessary for a complete description.

In America the relative position of the views in orthographic projection must be as shown in Figure 1.29. Thus the view seen from the right side of an object is placed to the right of the front view; the view seen from the top is placed above the front view. This is the so-called



**Figure 1.29** The six principal views of the bracket shown on the left. This is a third-angle projection. This is the preferred projection in the New World. The views not required to describe the bracket in this particular case have been crossed out.

*third-angle projection.* Of course, for clarity or convenience, the draftsman is initially free to choose any side of the object as the “front.” In Europe and Asia the *first-angle projection* is employed (Figure 1.30). In this projection the view shown to the right of the front view represents what would be seen if the object in front view were rolled to the right; the view shown above the front



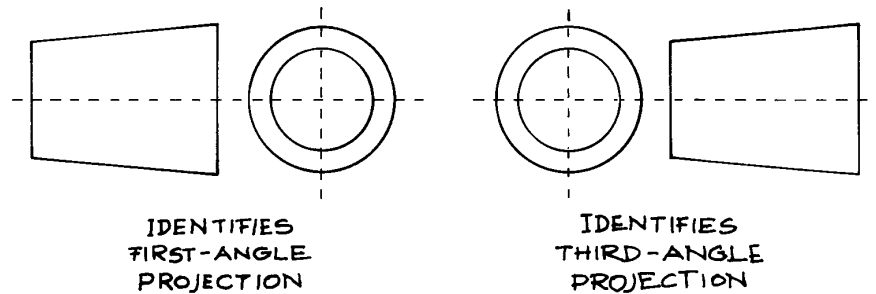
**Figure 1.30** A first-angle projection of the bracket shown in Figure 1.29. This is the projection preferred in Europe and Asia.

view represents what would be seen were the object in front view rolled  $90^\circ$  upwards. Of course it is very important that anyone who views a drawing understands the convention being used. Figure 1.31 shows orthographic projections of a “tapered stopper” that can be unambiguously identified as either a first-angle projection (on the left) or a third-angle projection (on the right). These “pictograms” are recommended for use as symbols on drawings. In this text, the third-angle projection is employed throughout.

In some cases an *auxiliary view* will simplify a presentation. For example, a view normal to an inclined surface may be easier to draw and more informative than one of the six principal views. Such a case is illustrated in Figure 1.32. The relationship of the auxiliary view to the front view must be as shown. That is, the auxiliary view of an inclined surface must be placed in the direction of the normal to the surface. Note the extension of the center line from the auxiliary view to the front view to show the direction of the auxiliary view.

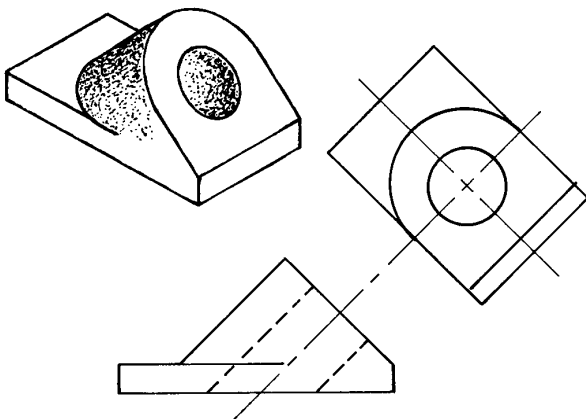
Often the internal structure of an object cannot be clearly indicated with hidden edges indicated by dashed





**Figure 1.31** One or the other of these pictograms that can be placed on a drawing to distinguish a “third-angle” projection from a “first-angle projection.”

lines. In this case it is useful to show the view that would be seen if the object were cut open. Such a view is called a *section*. Several examples of full sections are shown in Figure 1.33. The *cutting plane* in the section is represented by cross-hatching with fine lines to suggest a saw cut. The cross-hatch lines should always be at an oblique angle to the heavy lines that indicate the outline of the section. If the cutting plane extends through more than one part of an assembly, the cross-hatch lines on two adjacent parts should be in distinctly different directions. The cutting plane in a view perpendicular to the section is shown by a heavy *section line* of alternating long and two short dashes. Arrows on the section line indicate the direction of sight from which the section is viewed.



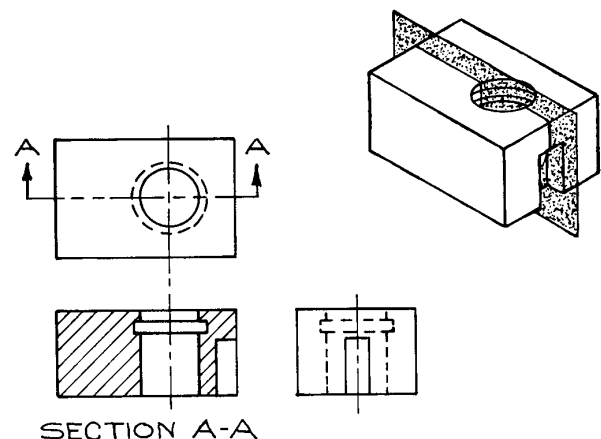
**Figure 1.32** Effective use of an auxiliary view.

Sections may be identified with letters at each end of the cutting-plane line.

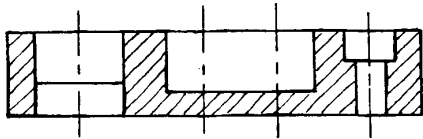
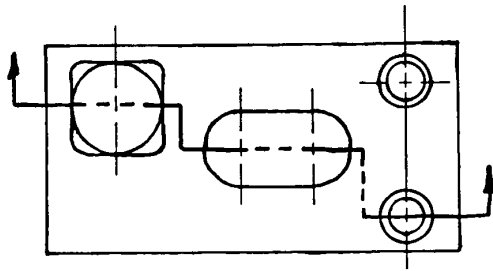
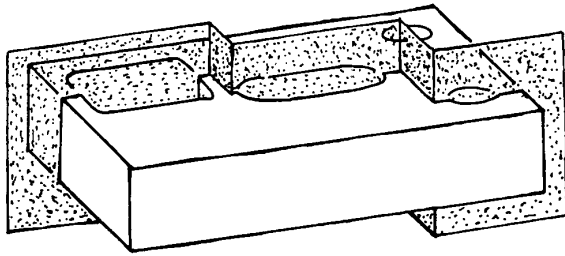
It is often useful to have the cutting plane change directions so that it passes through several features. Angled surfaces are then revolved into a common plane and offset surfaces are projected onto a common plane to give an *aligned section*, as illustrated in Figures 1.34 and 1.35, rather than a true projection of the cutting plane.

Other convenient sectioning techniques include the *half section* of a symmetrical object, as shown in Figure 1.36 and the *revolved section* of a long bar or spoke, as shown in Figure 1.37.

Many details, such as screw threads, rivets, springs, and welds, are so tedious to draw, and appear so often,



**Figure 1.33** Full section along the indicated plane.

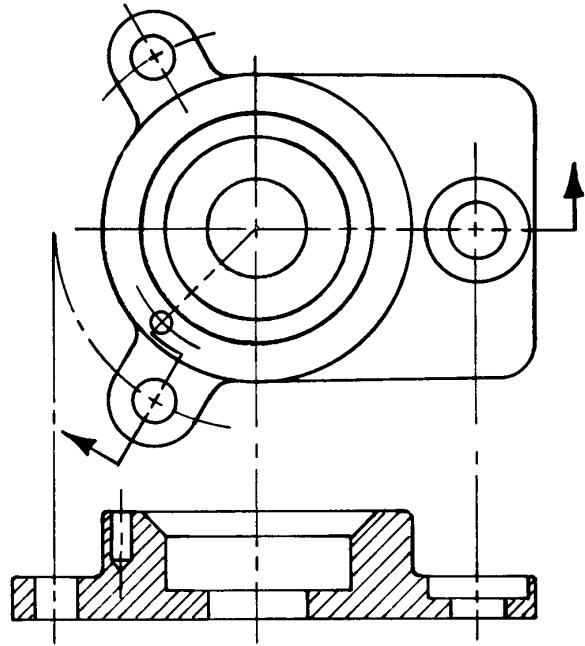


**Figure 1.34** An aligned view.

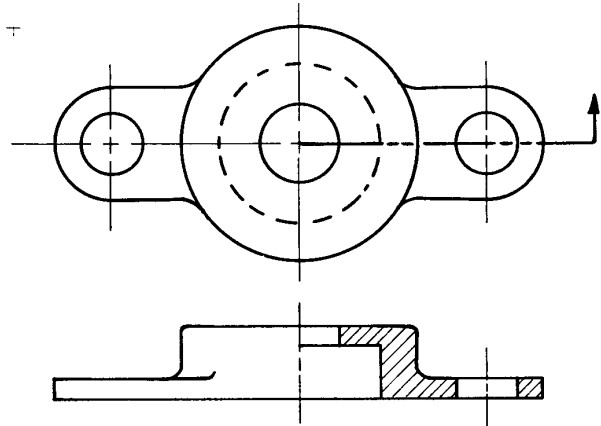
that they are designated by symbols rather than faithful representations. Symbols for threads are illustrated in Figure 1.38.

Drawings of parts too long to be conveniently fitted on a piece of drawing paper can be represented as though a piece from the middle of the part had been broken out and the two ends moved together. Three conventional breaks are illustrated in Figure 1.39.

Mechanical drawings include dimensions, tolerances, and special descriptions and instructions in written form. Lettering on mechanical drawings employs only uppercase letters in a *sans serif* font. Letters about 6 mm (1/4 in.) high are usually appropriate.



**Figure 1.35** An aligned section.



**Figure 1.36** A half section.

### 1.5.3 Dimensions

After the shape of an object is described by an orthographic projection, dimensions and notes on the drawing specify its size. The dimensions to be specified depend

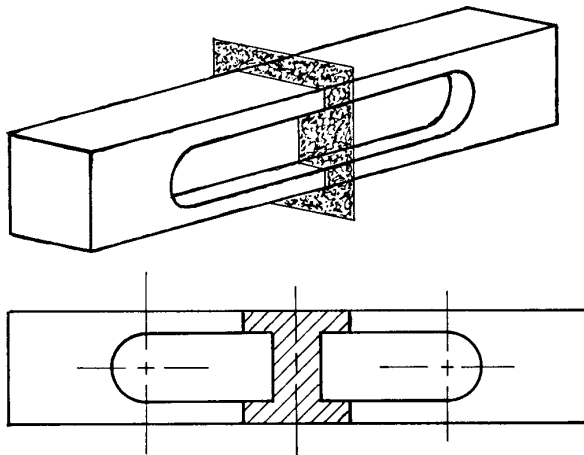


Figure 1.37 A revolved section.

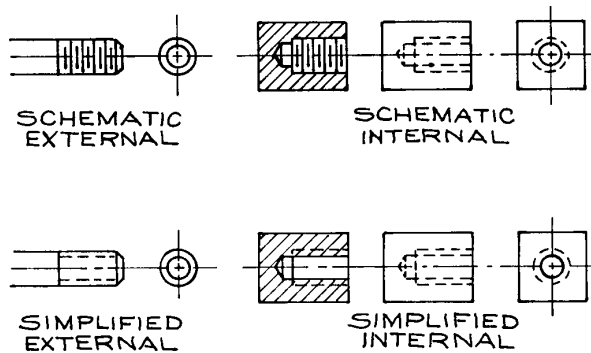


Figure 1.38 Thread symbols.

upon the function of the object and upon the machine operations to be performed by the workman when fabricating the object.

Either a dimension or a note specifies distance on a drawing. A *dimension* indicates the distance between points, edges, or surfaces (Figure 1.40). A *note* is a written instruction that gives information on the size and shape of a part (Figure 1.41). If SI units are intended, the linear unit of distance is the millimeter. In the New World the customary unit in engineering drawing is the decimal inch. The choice of unit is stated in a note: **UNLESS OTHERWISE SPECIFIED, ALL DIMENSIONS**

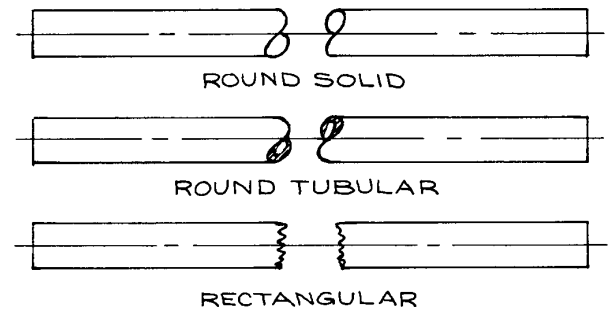


Figure 1.39 Breaks.

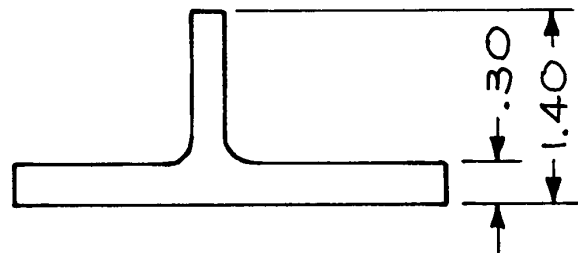
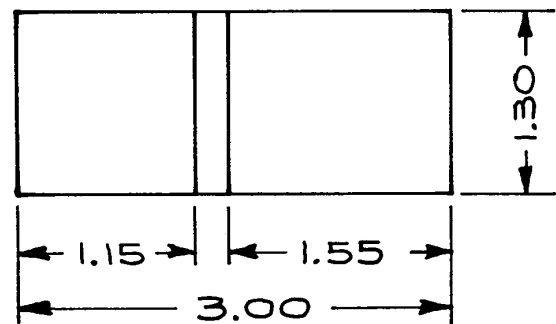


Figure 1.40 Dimensions.

ARE IN MILLIMETERS (or IN INCHES, as applicable). Apart from this note, the units of a dimension are not ordinarily specified on a drawing. Contrary to usual scientific practice, when dimensions are in inches, a decimal dimension less than unity is written without a zero preceding the decimal point: .256 rather than 0.256. On the other hand, when specifying millimeter dimensions, a zero precedes the decimal for a dimension less than one millimeter: 0.8 rather than .8.

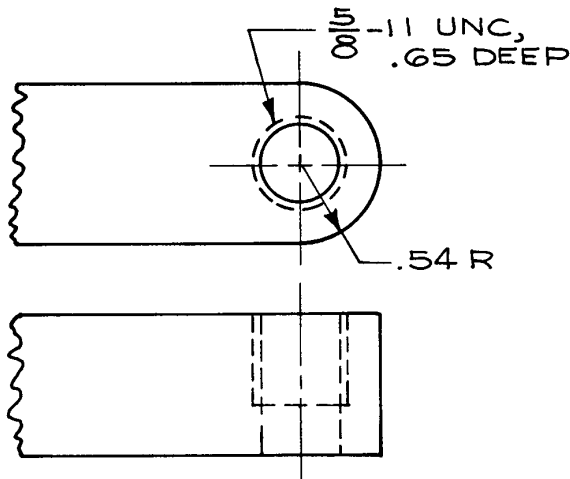


Figure 1.41 A note.

Dimensions may be given in *series* or *parallel* (Figure 1.42). If a combination of these two methods is used, care must be taken to ensure that the size of the object is not over-determined. An example of an over-dimensioned drawing is given in Figure 1.43. Dimensions should not be duplicated, and a drawing should include no more dimensions than are required. When several features of a drawing are obviously identical, it is only necessary to dimension one of them.

Dimensions are often specified with respect to a *datum*. This is a single feature whose location is assumed to be exact. When choosing a datum, consider the functional importance of locating other features of an object relative to the datum, as well as the ease with which the workman can locate the datum itself. (In Figure 1.46, the center of the .427R circular section serves as a datum point.)

If possible, all dimensions should be placed outside the view with *extension lines* leading out from the view. As shown in Figure 1.42, the shorter dimensions are nearer to the object line. Dimension lines never cross extension lines. Extension lines may cross extension lines if necessary. Dimensions within the view are permissible to eliminate long extension lines and if the dimension does not obscure some detail of the view. The dimension closest to the outside of a view should be about 12 mm (1/2 in.) from the edge. Successive dimensions should be placed at

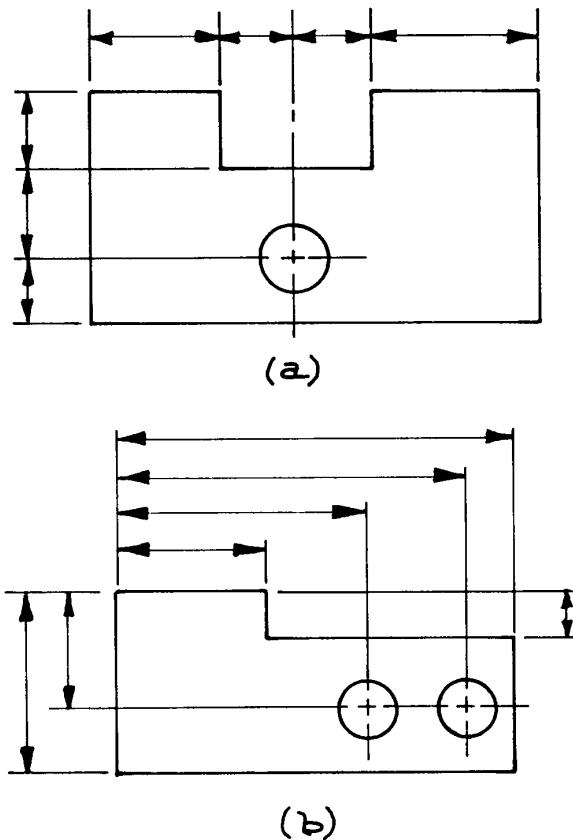
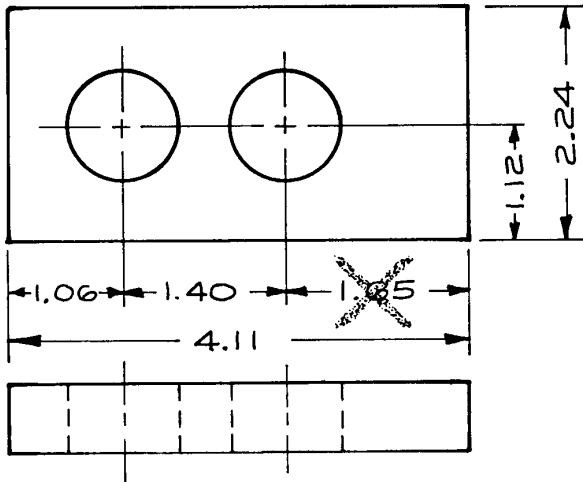


Figure 1.42 Dimensions: (a) series dimensions; (b) parallel dimensions.

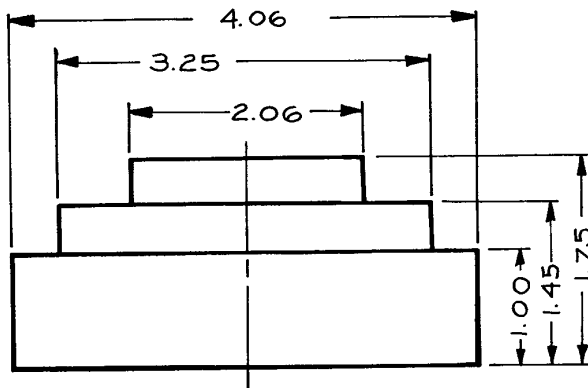
9 to 12 mm (3/8 to 1/2 in.) intervals. Dimension values should be staggered, as in Figure 1.44, rather than stacked vertically or horizontally. Dimension values are oriented so that they can be read from the bottom or the right side of the drawing. Notice that the extension lines do not touch the outline of the view, but rather stop about a millimeter (1/16 in.) short.

Giving the diameter or radius, and dimensions to locate the center of the radius of curvature, can specify a round hole or circular part. As in Figure 1.43 and elsewhere, the centerlines are extended to become extension lines.

Notes are used to specify the size of standard parts and parts too small to be dimensioned. Notes are also used to specify a feature that is derived from a standard operation



**Figure 1.43** An overdimensioned drawing. One of the dimensions is superfluous.



**Figure 1.44** Placement of dimensions and dimension values.

such as drilling, reaming, tapping, or spotfacing. Notes such as “TAP 10-32 UNF, .75 DEEP” or “1/2 DRILL, SPOTFACE .75D  $\times$  .10 DEEP,” that refer to a specific feature, are accompanied by a leader emanating from the beginning or end of the note and terminating in an arrowhead that indicates the location of the feature (Figure 1.41). General notes, such as “REMOVE BURRS” or “ALL FILLETS 1/4 R,” that refer to the whole drawing, do not require a leader.

Many abbreviations are used in writing dimensions and notes. Common abbreviations used on mechanical drawings are given in Table 1.4.

### 1.5.4 Tolerances

One of the most important parts of design work is the selection of manufacturing tolerances so that a designed part fulfills its function and yet can be fabricated with a minimum of effort. The designer must carefully and thoughtfully specify the function of each part of an apparatus in order to arrive at realistic tolerances on each dimension of a drawing. It is important to appreciate the capabilities of the machines that will be used in manufacture and the time and effort required to maintain a given tolerance. As indicated in Figure 1.45, the cost of production is approximately inversely related to the tolerances.

Usually only one or two tolerance specifications will apply to most of the dimensions on a drawing. It is then convenient to encode the tolerance specification in the dimension rather than affix a tolerance to each dimension. In instrument design most dimensions are written as decimals. The number of decimal places of the dimension can then indicate the tolerance on a dimension. Dimensions expressed to two decimal places have one tolerance specification, to three decimal places another, and so on. The specifications are given in a note. An example of this system is given in Figure 1.46.

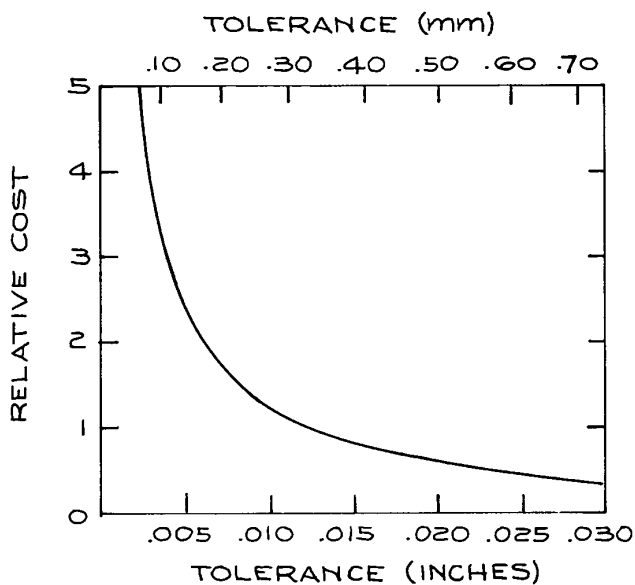
Fractions may be used in the American system to give the size of some features derived from standardized operations, such as drilling or threading; similarly a whole number millimeter specification may be used in the European system. A tolerance specification is unnecessary in these cases, since both the designer and workman should know what precision could be expected.

There are two methods of expressly stating the tolerance on a dimension. The allowable variation, plus and minus, may be stated after the dimension. Alternatively, the limits of the dimension can be stated without giving a tolerance at all. Examples of these two methods are given in Figure 1.47. The trend is toward limit dimensioning.

When the absolute size of two mating parts is not important, but the clearance between them is critical, specify the desired *fit*. Fits are classified as *sliding fits*,

Table 1.4 Abbreviations Used on Mechanical Drawings

Word	Abbr.	Word	Abbr.	Word	Abbr.
Bearing	BRG	Fillister	FIL	Right hand	RH
Bolt circle	BC	Grind	GRD	Rivet(ed)	RIV
Bracket	BRKT	Groove	CRV	Round	RD
Broach(ed)	BRO	Ground	GRD	Root mean square	RMS
Bushing	BUSH	Head	HD	Screw	SCR
Cap screw	CAP SCR	Inside diameter	ID	Socket	SOC
Center line	CL	Key	K	Space(d)	SP
Chamfer	CHAM	Keyway	KWY	Spot-face(d)	SF
Circle	CIR	Left hand	LH	Square	SQ
Circumference	CIRC	Long	LG	Stainless	STN
Concentric	CONC	Maximum	MAX	Steel	STL
Counterbore	CBORE	Minimum	MIN	Straight	STR
Counterdrill	CDRILL	Not to scale	NTS	Surface	SUR
Countersink	CSK	Number	NO	Taper(ed)	TPR
Cross section	XSECT	Opposite	OPP	Thread(ed)	THD
Diameter	DIA, D, Ø	Outside diameter	OD	Tolerance	TOL
Drawing	DWG	Pipe thread	PT	Typical	TYP
Drill(ed)	DR	Press	PRS	Vacuum	VAC
Each	EA	Punch	PCH	Washer	WASH
Equal(ly)	EQ	Radius	R	With	W/
Fillet	FIL	Reference line	REF	Without	W/O



**Figure 1.45** Approximate relation between tolerance and the cost of production.

*clearance locational fits, transition fits, interference fits, and force fits.* The uses of the various classes of fits and tables of the tolerances required to give such fits are given in standard texts on mechanical drawing.

Beware of an undesirable accumulation of tolerances. In series dimensioning, the tolerance on the distance between two points separated by two or more dimensions is equal to the sum of the tolerances for each of the intervening dimensions. In Figure 1.42(a) the tolerance on the lateral location of the hole with respect to the right side of the object is given by the sum of two tolerances. For parallel dimensions, the location of each feature relative to the datum depends on only one tolerance; however, the difference in the distance between two locations depends upon two tolerances. In Figure 1.42(b), the tolerance on the distance between the two holes is given by the sum of the tolerances on the dimensions that locate the holes.

As much as possible, dimensions and tolerances should be chosen so that materials can be used in their stock sizes and shapes. Do not call for unnecessary finishing. For

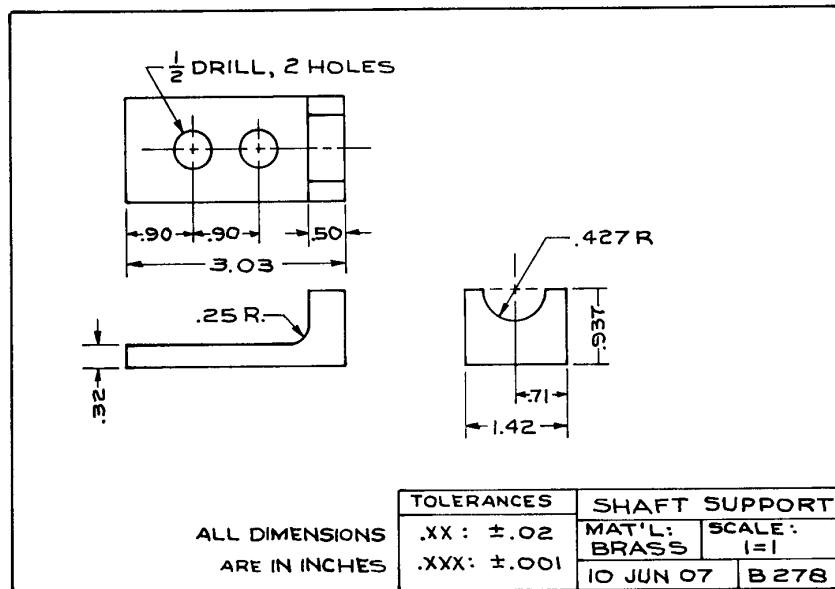


Figure 1.46 A decimal system for the specification of tolerances.

example, the area around a hole in a casting or other rough piece of metal can be spotfaced to provide a bearing surface for a bolt if the quality of the rest of the surface is not critical.

The dimensions and tolerances on a finished drawing should be carefully checked. Imagine yourself as the machinist who must use the drawing to make a part. Proceed mentally through each step of the fabrication to see that all necessary dimensions appear on the drawing and that the specified dimensions are easy to use. Also satisfy yourself that all desired tolerances are within the capabilities of the anticipated machine-tool operations.

### 1.5.5 From Design to Working Drawings

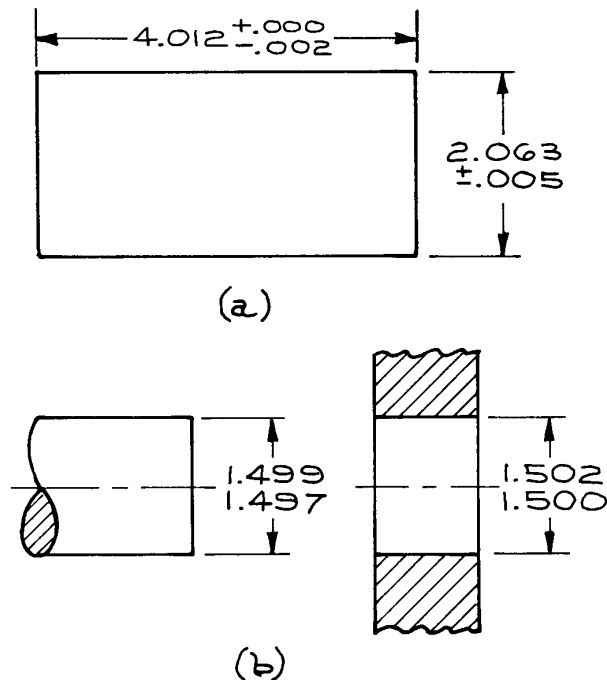
The first step in designing an instrument is to draw the entire assembly. Start with rough sketches in pencil and paper before turning to the computer. Then, with the CAD program, draw the outline of the most important parts of the instrument. Work to full scale if possible. Add subsidiary parts after the configuration of the essential elements has been established. Most dimensions, tolerances, and fine details can be omitted at this point.

Pay attention to the manner in which pieces fit together. Will the apparatus be convenient to assemble and disassemble? Is there sufficient clearance to insert bolts, pins and other fasteners? Are some parts much stronger or weaker than required? Does a strong part depend upon a weak one for location or support? Are all shapes as simple as possible? Are standard shapes of materials and standard bolts, couplings, bearings, and shafts used wherever possible?

Time spent at the drawing board mentally assembling and disassembling a device will save hours of frustration in the laboratory. Careful consideration of production techniques can save hundreds of dollars in the shop.

With the initial orthographic drawing of the apparatus in hand, it is wise to discuss it with the people who will do the work.

After the design has reached its final form, the designer must make a working drawing of each piece of the apparatus. A working drawing is a fully dimensioned and toleranced drawing to be used in the shop. Work to scale if possible. The CAD program is especially useful at this point. A new layer can be superimposed on the assembly drawing and the outline of a single part transferred to the new layer without fear of error. Alternatively the outline



**Figure 1.47** Two methods of stating the tolerance on a dimension: (a) as the allowable variation on a dimension; (b) in terms of the allowable limits of the dimension.

of an individual part can be copied and transferred to a new file. The draftsman should develop a systematic way of naming drawings so they can be stored and easily retrieved.

The critical test of a set of working drawings is their usefulness. All necessary communication between scientist and machinist should appear on the drawings. The machinist should not require oral instructions, and he should not be required to make decisions affecting the performance of an instrument.

In the above discussion we have assumed that the scientific designer has the services of a model shop. This is the case in industrial laboratories and in many university laboratories. The designer should follow the same procedure when he fabricates his own apparatus. All questions of sizes, tolerances, and fits must be answered before construction. When a scientist attempts to make these decisions as he proceeds with construction, the results are invariably poor.

## 1.6 PHYSICAL PRINCIPLES OF MECHANICAL DESIGN

No material is perfectly rigid. When any member of a machine is subjected to a force, no matter how small, it will bend or twist to some extent. Even in the absence of an external load, a mechanical element bends under its own weight (so-called “body forces”). A machine element subjected to a force that varies in time will vibrate. A designer must appreciate the extent of deflection of mechanical parts under load. The forces applied to or contained by an apparatus must be anticipated in specifying the material and dimensions of each element. The hallmark of a well-designed machine is that every element bears the same level of stress and that stress does not threaten the function of the machine or the capabilities of the materials of which the machine is constructed or the fasteners that hold the parts together.

### 1.6.1 Bending of a Beam or Shaft

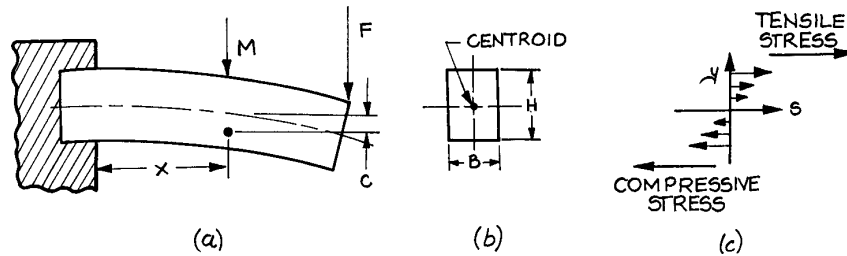
When a beam bends, one side of the beam experiences a tensile load and the other a compressive load. Consider a point in the flexed beam in Figure 1.48. The stress at this point depends upon its distance,  $c$ , from the centroid of the beam in the direction of the applied force. The *centroid* is the center of gravity of the cross section of the beam [see Figure 1.48(b)]. The stress is given by:

$$s = \frac{Mc}{I} \quad (1.2)$$

where  $M$  is the bending moment at the point of interest and  $I$  is the centroidal moment of inertia of the section of the beam containing the point of interest. Clearly the stress in a flexed beam is greatest at the outer surface of the beam.

The *centroidal moment of inertia* of the section (more correctly known as the second moment of the area of the cross-section) is the integral of the square of the distance from the centroid multiplied by the differential area at that distance. The integral is taken along a line through the centroid and parallel to the applied force. For a rectangular





**Figure 1.48** A flexed beam: (a) the beam bending under a load; (b) a cross section showing the centroid; (c) the distribution of shear forces along the long axis of the beam.

section of height  $H$  and width  $B$ , such as that shown in Figure 1.48(b), the centroidal moment of inertia is:

$$I = \int_{-B/2}^{B/2} \int_{-H/2}^{H/2} h^2 dhdb = \frac{BH^3}{12} \quad (1.3)$$

The centroidal moments of inertia of some common symmetrical sections are given in Figure 1.49.

The *bending moment* is proportional to the curvature produced by the applied force:

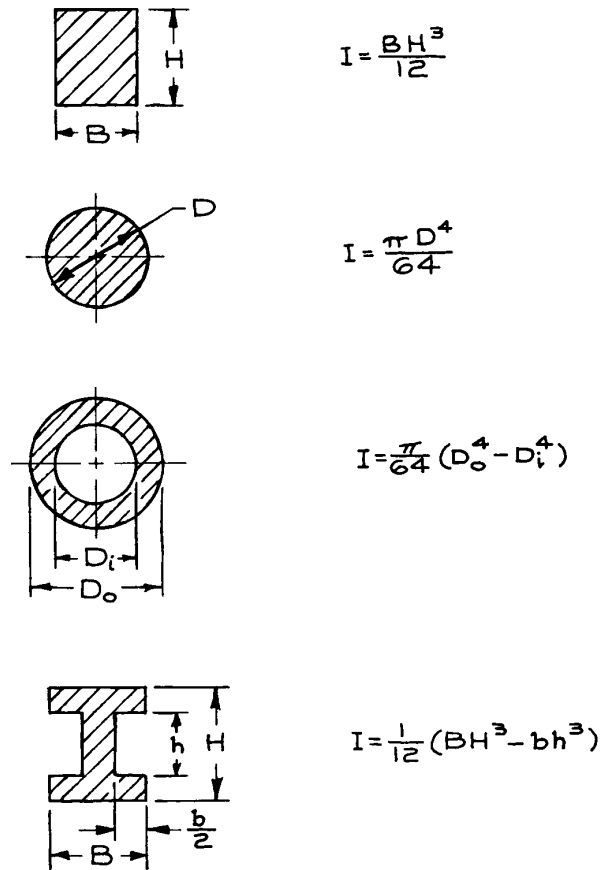
$$M = EI \frac{d^2y}{dx^2} \quad (1.4)$$

where  $y$  is the deflection produced by the force and  $E$  is the modulus of elasticity of the material of which the beam is composed. For the example given in Figure 1.48, assuming the weight of the shaft can be ignored, the bending moment is simply:

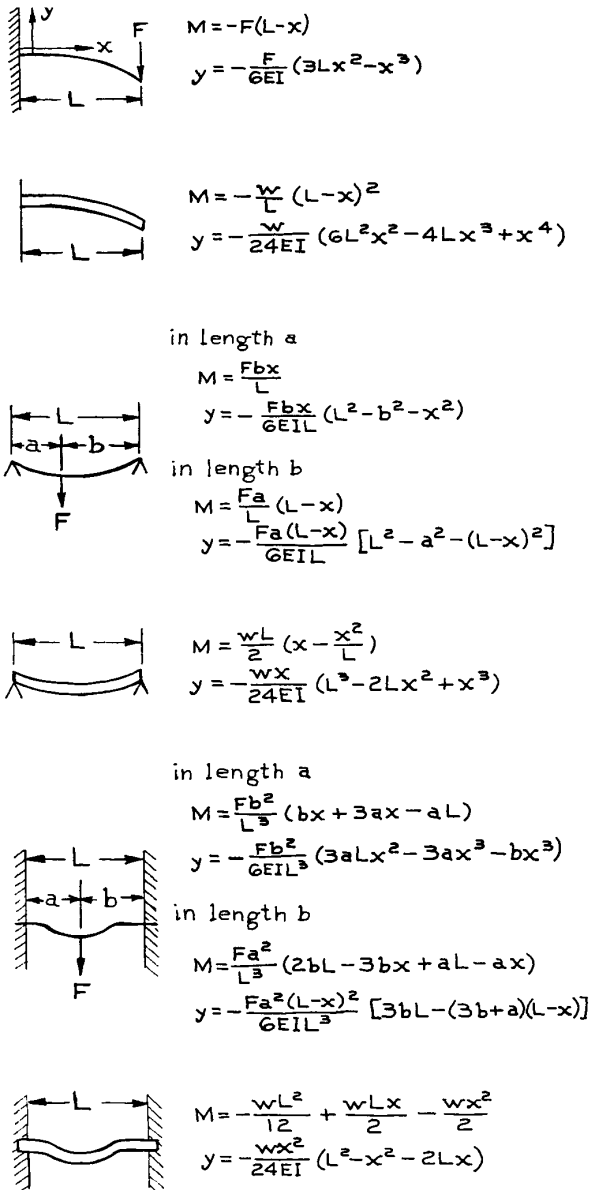
$$M = -F(L - x) \quad (1.5)$$

where  $L$  is the length of the beam. The bending moments for this and other simple systems are given in Figure 1.50.

The deflection of a stressed beam at a point a distance  $x$  from a supported end of the beam depends upon the modulus of elasticity of the material, the centroidal moment of the section of the beam, and the nature of the support. We consider the case where the beam is rigidly fixed to its support(s) (as in Figure 1.48) and the case where the beam is supported at each end by a simple fulcrum. Expressions for the deflection of both point-loaded and uniformly loaded beams are given in Figure 1.50.



**Figure 1.49** Centroidal moments of inertia.



**Figure 1.50** Bending formulae:  $M$  = bending moment;  $E$  = modulus of elasticity;  $I$  = centroidal moment of inertia;  $w$  = weight per unit length.

A sufficiently large bending moment will result in the permanent deformation of a beam. This is the yield point. A conservative relation between the bending moment at the yield point,  $M_y$ , and the yield strength of the material,  $s_y$ , is given by  $M_y = s_y I / c_{\max}$ . [For the rectangular section beam supported at one end (Figure 1.48),  $c_{\max} = H/2$  and  $M_y = s_y BH^2/6$ .] To determine the maximum allowable load on a beam or shaft one simply sets  $M_y$  equal to the expression for the maximum value of the bending moment on the beam and solves for the force. (For the rectangular section beam supported at one end,  $F_y = s_y BH^2/6L$ .) In reality the yield strength of a beam in bending is greater than for a beam in tension, since the inner fibers of the material must be loaded to their yield point to a considerable depth before the beam is permanently deformed.

Even in the absence of applied forces, the elements of a machine bend under their own weight or if the machine is otherwise accelerated or decelerated. An important conclusion of the discussion above is that this effect is reduced if the size of the machine is reduced; a mechanical structure becomes stiffer as it is scaled down in size. Consider a horizontal beam of rectangular cross-section cantilevered from a wall. From Figure 1.50, setting  $x = L$ , the weight of the beam causes a deflection at the free end:

$$y_{\max} = \frac{wL^4}{8EI} \quad (1.6)$$

where  $w$  is weight per unit length of the beam. Taking  $w = \rho BH$ , where  $\rho$  is the weight density of the material of the beam, and substituting for  $I$  gives:

$$y_{\max} = \frac{3\rho L^4}{2EH^2} \quad (1.7)$$

If the dimensions of the beam are scaled down by a factor  $\zeta$ :

$$y_{\max} = \frac{3\rho(\zeta L)^4}{2E(\zeta H)^2} = \zeta^2 \frac{3\rho L^4}{2EH^2} \quad (1.8)$$

The deflection decreases as the square of the scale factor; the relative deflection,  $y_{\max}/L$  decreases linearly with the scale factor.

It should be noted that all of the preceding discussion assumes the tensile strength of a material to be the same as the compressive strength. For materials such as cast iron where this is not true, the bending equations are much more complicated. The foregoing also ignores shear stresses in a loaded beam. For very short beams, shear stresses become important and the shear strength of the material must be considered.<sup>7</sup>

### 1.6.2 Twisting of a Shaft

The stress at a point in a hollow round shaft subjected to a torsional load is:

$$s = \frac{Tc}{J}, \quad J = \frac{\pi}{32}(D_o^4 - D_i^4) \quad (1.9)$$

where  $c$  is the distance of the point of interest from the center of the shaft and  $T$  is the applied torque.  $J$  is the *centroidal polar moment of inertia of the section* of the shaft, and  $D_o$ , and  $D_i$  are the outer and inner diameters of the shaft.

The total angle of twist in a shaft of length  $L$  is:

$$\theta = \frac{57LT}{GJ} \text{ deg} \quad (1.10)$$

where  $G$  is the modulus of elasticity in shear, or *shear modulus*, of the material of the shaft. For metals, the shear modulus is about 1/3 the elastic modulus.

### 1.6.3 Internal Pressure

In the scientific literature, “high pressure” implies pressures in excess of 1 GPa (150000 psi). Beyond 1 GPa most fluids condense to solid phases. At pressures in excess of 1 GPa, remarkable pressure-induced chemical changes are observed. Not coincidentally, 1 GPa corresponds approximately to the yield strength of the strongest materials. We shall not discuss here the special technologies, notably the diamond anvil, that are required above 1 GPa.

To an engineer, 0.1 to 1 GPa is “ultrahigh pressure.” In spite of what this term might imply, familiar technologies are appropriate. Ultrahigh pressure tubing and tube fittings,

demountable joints, valves and piston pumps are commercially available (for example, High Pressure Equipment Company, PO Box 8248, Erie, PA 16505, USA). As noted previously, scientific apparatus to be operated in this pressure range should be built up entirely of commercial components with coned and threaded joints; joints brazed or welded in the lab or the local shop are not safe or reliable at pressures above 0.1 GPa. Indeed, considerations of both safety and economics favor the use of commercial components when building apparatus to operate at elevated pressures.

A pressure vessel should incorporate only cylindrical sections if at all possible. The design and analysis of a cylinder under internal pressure is easiest and most reliable. Joints between cylindrical sections are most easily manufactured and assembled (see Section 1.4.10). Of course the stress exerted by an internal pressure is easily calculated for vessels with other simple geometric shapes (spheres, cones); the manufacture, however, can be daunting. In any event, it is recommended that one design for at least *five times* the anticipated pressure, unless a professional engineer is consulted.<sup>8</sup>

The maximum stress on a cylinder under internal pressure is a tensile stress tangent to the inner surface of the cylinder. For a cylinder with inner and outer radii,  $R_1$  and  $R_2$ , respectively, under pressure  $P$ , the maximum value of this tangential stress, the so-called “hoop stress,” is:

$$s_t = \frac{P(R_1^2 + R_2^2)}{R_2^2 - R_1^2} \quad (1.11)$$

Shearing forces are most often responsible for rupture in ductile materials. The maximum shear stress is at the inner surface where:

$$s_{\text{sheer}} = \frac{PR_2^2}{R_2^2 - R_1^2} \quad (1.12)$$

For a thin-wall cylinder the expression for the hoop stress reduces to:

$$s_t = P \frac{R}{t} \quad (1.13)$$

where  $R$  is the inner radius and  $t$  the wall thickness, and it is assumed  $t < 0.1R$ . The longitudinal (axial) stress and the shear stress in this approximation are then just half the hoop stress.

Expansion of a vessel under internal pressure may be an issue in some instances. In the thin-wall approximation the hoop strain is:

$$\varepsilon = \frac{PR}{2Et}(2 - \mu) \quad (1.14)$$

where  $E$  is the modulus of elasticity and  $\mu$  is Poisson's ratio for the material of which the cylinder is made. Recalling that *strain* is the increase in length per unit length, the dilation, or radial growth of a cylinder under internal pressure is:

$$\delta = \frac{PR^2}{2Et}(2 - \mu) \quad (1.15)$$

The longitudinal strain is:

$$\varepsilon = \frac{PR}{2Et}(1 - 2\mu) \quad (1.16)$$

The end closure of a cylindrical vessel may be spherical, torospherical (elliptical in cross-section) or flat. The first two are certainly more robust than a flat plate closure, but the manufacture will be difficult and expensive. The stress in a flat end plate depends upon how securely the end plate is attached to the cylindrical vessel. If there is the possibility of movement between the end plate and the cylinder to which it is attached, such as for a removable, gasketed and bolted, endplate, the maximum stress occurs at the center of the plate where, for a plate of thickness  $d$  and radius  $R$ :

$$S = \frac{3}{8} \left(\frac{R}{d}\right)^2 (1 + \mu)P \quad (1.17)$$

The deflection under pressure at the center of the plate is:

$$\delta = \frac{3PR^4(5 + \mu)(1 - \mu)}{16Ed^3} \quad (1.18)$$

For a permanently affixed flat closure, such as an end plate welded to a cylindrical vessel, the maximum stress is at the edge where the radial stress is:

$$S_r = \frac{3}{4} \left(\frac{R}{d}\right)^2 P \quad (1.19)$$

and the tangential stress is:

$$S_t = \frac{3}{8} \left(\frac{R}{d}\right)^2 (1 + \mu)P \quad (1.20)$$

Deflection under pressure at the center of a permanently fixed end plate is:

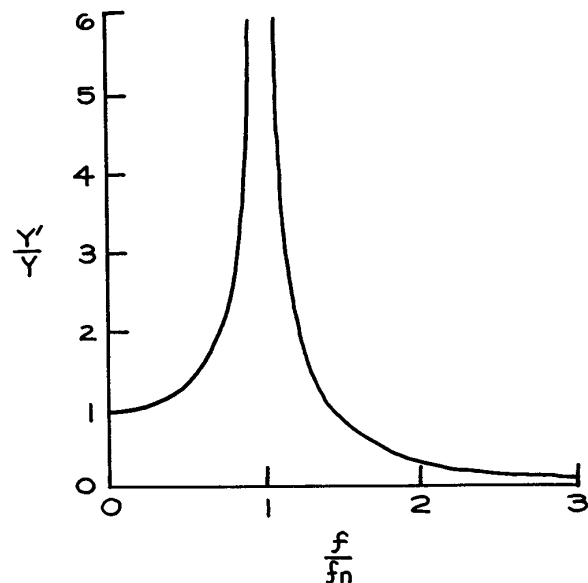
$$\delta = \frac{3PR^4(1 - \mu^2)}{16Ed^3} \quad (1.21)$$

Design for external pressure is discussed in Section 3.6.2 in the context of vacuum chambers.

## 1.6.4 Vibration of Beams and Shafts

In many instruments, vibration of a supporting beam or shaft can adversely affect operation. The designer must pay special attention to the natural frequencies of vibration of the parts of an instrument. A periodic disturbance with a frequency near that of one of the natural frequencies will induce a large and possibly destructive vibration. For the purpose of illustration, the response of an oscillator to the frequency of a driving force is shown in Figure 1.51.

There are two ways for the designer to prevent destructive vibrations. The critical element can be designed so



**Figure 1.51** Response of a machine element to a periodically varying force.  $Y'$  is the amplitude of vibration of the element and  $f_n$  is its natural frequency of vibration.  $Y$  and  $f$  are the amplitude and frequency of the driver.

that its natural frequency of vibration is far removed from the frequency of a disturbing force, or the disturbing force can be inhibited in the vicinity of the natural frequency of the critical element. The latter scheme is called damping. Very low-frequency vibrations can be damped by a hydraulic or friction shock absorber such as is employed on automobile suspensions. High frequencies can be damped by coupling the objectionable source of vibration to its surroundings through a member that has a very low natural frequency of vibration. Mounts made of rubber or cork are very effective for absorbing high-frequency vibrations and sudden shocks, since these materials possess a low modulus of elasticity.

The fundamental frequency of vibration (in Hz) of a shaft or beam with one or more concentrated masses attached is given by:

$$f_n^2 = \frac{\frac{a}{4\pi^2} \int_0^L \frac{M^2}{EI} dx}{\int_0^L \rho AY^2 dx + \sum_i F_i Y_i^2} \quad (1.22)$$

where:

$I$  = centroidal moment of inertia of the shaft,

$M = M(x)$  = maximum bending moment at  $x$ , i.e., the amplitude of the bending moment induced by the forces on the shaft at  $x$ ,

$a$  = inertial acceleration experienced by the shaft and its weights (for a nonrotating shaft,  $a = g$ ),

$E$  = modulus of elasticity of shaft material,

$\rho$  = weight density of the shaft ( $\rho/g$  is the mass density),

$A$  = cross-sectional area of the shaft,

$L$  = length of the shaft

$Y = y(x)$  = maximum deflection of the shaft at  $x$ , i.e., the amplitude of vibration at  $x$ ,

$F_i$  = force exerted by the  $i$ th mass (for a nonrotating shaft, this is the weight of the  $i$ th mass),

$Y_i$  = maximum deflection at the location of the  $i$ th mass.

For a system similar to one of the static deflection cases treated in Figure 1.50, one need only substitute the appropriate expressions for the bending moment and deflections and solve the resulting equation to determine the natural frequency. Consider for example a beam of negligible mass that is rigidly supported at one end and supports a mass of

weight  $W$  at its free end. This is the case illustrated at the top of Figure 1.50. The natural frequency of vibration is:

$$f_n^2 = \frac{\frac{g}{4\pi^2} \int_0^L \frac{W^2(L-x)^2}{EI} dx}{W \left\{ \frac{-W}{6EI} [3Lx^2 - x^3]_{x=L} \right\}^2},$$

$$f_n = \frac{1}{2\pi} \left( \frac{3gEI}{WL^3} \right)^{1/2} \text{ Hz.} \quad (1.23)$$

In general, the shape of the deflection curve is not known. However, the assumption of some reasonable deflection curve usually will provide a useful result. In most cases one can assume a sine curve for a shaft supported (but not clamped) at both ends, or a parabola for a shaft supported at only one end. More precise approaches to complex systems are given in many engineering texts on vibration.<sup>9</sup>

Consider once again the beam of negligible mass, rigidly supported at one end, with a weight  $W$  at the free end. A parabolic deflection curve with the correct limit properties is given by:

$$y = -\frac{y_0}{L^2}(L-x)^2 \quad (1.24)$$

where  $y_0$  is the maximum deflection of the free end. The bending moment is then:

$$M = EI \frac{d^2y}{dx^2}$$

$$= -EI \left( \frac{y_0}{L^2} \right). \quad (1.25)$$

Substituting in the equation for the natural frequency gives:

$$f_n^2 = \frac{\frac{g}{4\pi^2} \int_0^L \frac{[-2EI(y_0/L^2)]}{EI} dx}{W(-y_0)^2}$$

$$f_n = \frac{1}{\pi} \left( \frac{gEI}{WL^3} \right) \text{ Hz} \quad (1.26)$$

in reasonable agreement with the exact solution derived above.

For the special case of a heavy shaft or beam with a centrally placed load, a correct solution can be derived by assuming the shaft to be weightless and adding half the weight of the shaft to the concentrated load.

### 1.6.5 Shaft Whirl and Vibration

A shaft and rotor can never be perfectly balanced, and the axis of rotation can never be exactly located by the shaft bearings. As a shaft rotates, centrifugal force on the unbalanced mass deflects the shaft and causes it to *whirl* around its axis of rotation. This whirling motion appears as a vibration to a stationary observer, and can be treated as such. There is a *critical speed* at which the whirl becomes violent. In this unstable condition, the shaft or its bearings are likely to be damaged and intense vibration will be transmitted through the bearing mount to other parts of the machine. This critical condition occurs when the shaft speed (in rps) equals the natural frequency (in Hz) of the stationary shaft and rotor assembly. It follows that the mathematical derivations of natural frequencies of assemblies of beams and weights can be applied to the calculation of critical speeds for shafts and rotors. The actual centrifugal loads on the rotating shaft need not be determined. This is because the load  $F$  always appears as the ratio  $F/a$  and:

$$\frac{F}{a} = \frac{W}{g} = m \quad (1.27)$$

where  $W$  is the weight of the element that is exerting the load.

The critical speed for a rotor-shaft assembly depends upon how rigidly the shaft bearings support the shaft. Two extreme cases are considered: a thin bearing, such as a ball bearing, for which the angle of the shaft passing through the bearing is not fixed; and a thick bearing, such

as a plain sleeve bearing, that rigidly maintains the angular alignment of the shaft. The critical speed for a rotor on a weightless shaft supported on thin bearings is:

$$n_c = 0.276 \left( \frac{gEIL}{Wa^2b^2} \right)^{1/2} \text{ rps} \quad (1.28)$$

where  $W$  is the weight of the rotor,  $I$  is the centroidal moment of inertia of the shaft,  $E$  the modulus of elasticity of the shaft material,  $L$  the length of the shaft between the bearings, and  $a$  and  $b$  are the distances from the rotor to the bearings. If the ends of the shaft are rigidly supported in long bearings the critical speed is:

$$n_c = 0.276 \left( \frac{gEIL^3}{Wa^3b^3} \right)^{1/2} \text{ rps} \quad (1.29)$$

Notice that the critical speed can be increased by placing the rotor near one end of the shaft so that  $a$  or  $b$  is small. In this case the rotor acts as a gyro tending to stiffen the shaft.

For an unloaded shaft on thin bearings the critical speed is:

$$n_c = 1.57 \left( \frac{EIg}{\rho AL^4} \right)^{1/2} \text{ rps} \quad (1.30)$$

and for an unloaded shaft on long bearings:

$$n_c = 3.53 \left( \frac{EIg}{\rho AL^4} \right)^{1/2} \text{ rps} \quad (1.31)$$

where  $\rho$  is the weight density of the shaft material. The units of the individual quantities appearing in brackets in the four equations above must be such that overall the quantity in brackets has units of  $s^2$ . For the particular case of a solid, round, steel shaft the critical speeds are:

---


$$\begin{aligned} n_c &= 80\,000 \frac{D}{L^2} \text{ rps [thin bearings; } D = \text{diameter (in.); } L = \text{length (in.)]} \\ &= 3\,150 \frac{D}{L^2} \text{ rps [thin bearings; } D = \text{diameter (mm); } L = \text{length (mm)}] \\ n_c &= 180\,000 \frac{D}{L^2} \text{ rps [long bearings; } D = \text{diameter (in.); } L = \text{length (in.)]} \\ &= 7\,100 \frac{D}{L^2} \text{ rps [long bearings; } D = \text{diameter (mm); } L = \text{length (mm)}] \end{aligned} \quad (1.32)$$


---

The special case of a centrally placed rotor on a heavy shaft can be treated by adding half the weight of the shaft to the rotor and taking the shaft to be weightless. The critical speed for a shaft carrying several rotors can be approximated by:

$$n_c = \left( \frac{1}{n_s^2} + \sum_i \frac{1}{n_i^2} \right)^{-1/2} \quad (1.33)$$

where  $n_s$  is the critical speed of the shaft alone and  $n_i$  is the critical speed for the  $i$ th rotor alone on a weightless shaft.

The designer must manipulate the dimensions and material of a shaft and rotor, and the location of the rotor, so that its critical speed does not coincide with the design speed of an assembly. It is usually sufficient for these two speeds to differ by a factor of  $\sqrt{2}$ . The most desirable situation is for the critical speed to exceed the design speed. If this is not possible, the drive motor for the rotating assembly should be so powerful that the shaft accelerates very rapidly through the region of the critical speed. The drive motor will require an excess of power because considerable power is dissipated to vibration at the critical speed.

The formulae above give the lowest frequency of vibration and the lowest critical speed for a shaft and rotor assembly. A shaft with a number of weights has as many different fundamental modes of vibration as it has weights. In addition, vibration will occur at harmonics of the fundamentals.

For rotational speeds far away from a critical speed, a rotating assembly can be considered to be rigid. Nevertheless, imbalance can lead to inertial (centrifugal) forces that may damage the shaft or its bearings. If the unbalanced mass lies within a plane perpendicular to the axis of rotation, as for the case of a thin rotor on a shaft, the rotor can be statically balanced. The condition for *static balance* is that the axis of rotation passes through the center of gravity. In practice this can be accomplished by placing the shaft with its rotor across a pair of knife-edge rails. The shaft will roll or rock until coming to rest with the unbalanced inertial force pointed radially downwards. Weight is then added or removed along the vertical radius until balance is achieved.

If there are unbalanced masses acting at different points along the axis, then dynamic balancing is required. The conditions for *dynamic balance* are that the axis of rotation passes through the center of gravity and that the axis of rotation coincides with a principal axis of the inertial force. Dynamic balancing is difficult without specialized equipment. Balancing is done inexpensively and accurately as a commercial service.

## 1.7 CONSTRAINED MOTION

In addition to a rigid support structure, most machines include some moving parts. The motion of these parts must usually be constrained in some fashion to obtain the required function. Most often the desired motion is translation in a plane or along a line, or rotation about one or more axes. The design of the constraining mechanism for a moving part must meet dimensional, strength, and wear-resistance specifications.

To achieve constrained motion the designer may either create a support structure for the moving part that will give the desired alignment, or make use of a standard, commercially available machine element. In the former case, it is necessary to understand the principles and limitations of geometric design. In the latter, the designer must be familiar with the range and precision of commercial bearings, shafts, sliders, and so on. The economics of the choice should not be ignored.

### 1.7.1 Kinematic Design

A rigid body has six independent degrees of freedom of motion. These are usually taken to be the translational motions along three orthogonal axes through the center of gravity and the rotational motions about these axes. Any motion can be described as a linear combination of these six motions. Similarly, six coordinates define the position of a body: three position coordinates and three angle coordinates. Motion is constrained and the number of degrees of freedom is reduced if one of these coordinates is fixed or if some linear combination of these coordinates is fixed. For example, a rigid object will be constrained to move in a plane if one of its position coordinates is fixed. Kinematic design for constrained motion

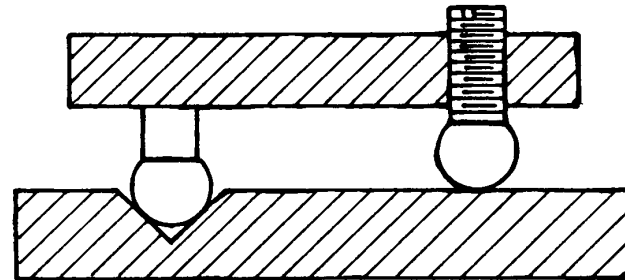
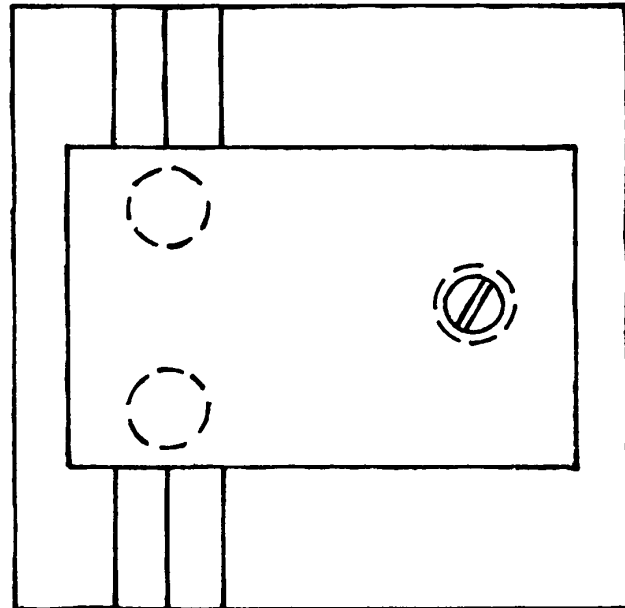
consists of fixing a point on a body for each degree of freedom to be constrained.<sup>10</sup>

In general, one degree of freedom can be removed by constraining a point on a body to remain in contact with a reference surface. The following examples illustrate this principle of kinematic design:

- (1) Consider a sphere constrained to maintain point contact with each of two reference planes: a ball resting in a V-groove. The ball is free to rotate about any axis, but it can only translate along a line parallel to the intersection of the planes. The ball has lost two degrees of freedom.
- (2) A three-legged “milking stool” that is constrained to maintain three points of contact with a plane surface has only three degrees of freedom: translation in two directions parallel to the plane and rotation about an axis perpendicular to the plane.
- (3) A body in contact at three points with a cylindrical surface has three degrees of freedom: translation in the direction of the cylinder axis, revolution about the cylinder axis, and rotation about an axis perpendicular to the plane defined by the three points of contact.

The object of kinematic design is to permit motion to be constrained without depending upon precision of manufacture. A complication exists if two contact points are degenerate in the sense that the function of the two points can be served by a single point of contact. Consider a four-legged table *vis-à-vis* a three-legged table. One of the four legs provides a degenerate point of contact with the floor. The fourth leg may provide some much-needed stability, but it also introduces uncertainty in the location of the table top unless all four legs are of precisely the same length. Of course, the sort of considerations that lead to the production of four-legged tables often dictates a *semikinematic* or degenerate kinematic design for other devices as well.

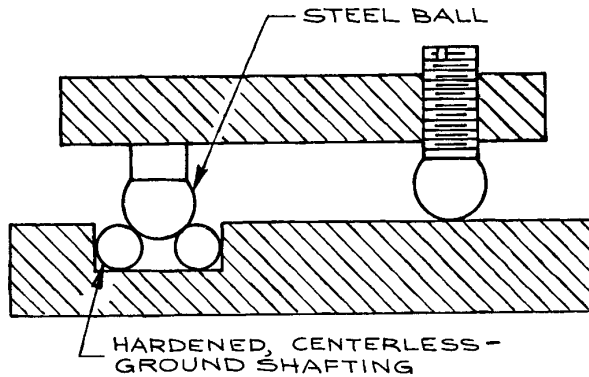
Kinematic designs are illustrated in Figures 1.52–1.54. The ball feet on the carriage shown in Figure 1.52 provide five points of contact between the carriage and its base. Thus the carriage is only free to slide in one direction. One problem with this design is that it is difficult to mill a V-groove with smooth surfaces. The surface quality of the groove can be improved by lapping with carborundum, using a V-shaped brass lap that fits in the groove.



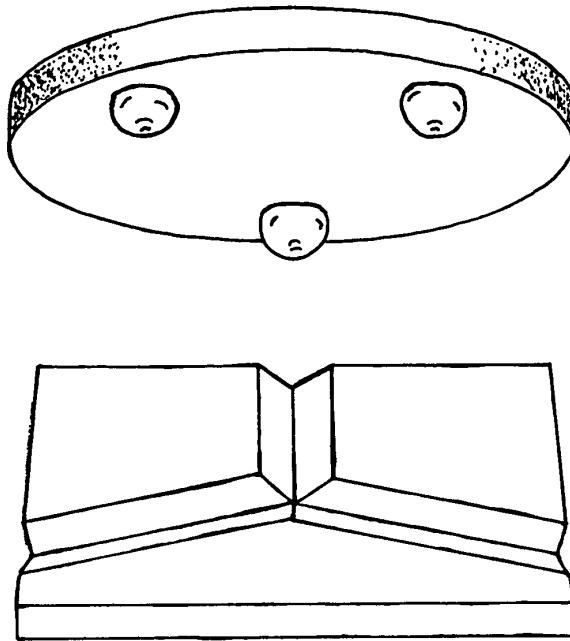
**Figure 1.52** Kinematic design that constrains a carriage to move in a straight line.

An improved version of the design in Figure 1.52 is shown in Figure 1.53. A pair of steel rods replaces the V-groove, and balls replace the feet of the carriage. This design is both more precise and more economical than the previous one. Stainless-steel shafting, hardened and centerless-ground to a diameter tolerance of  $\pm 0.001$  mm ( $\pm .00005$  in.), is available at a cost of a few euros a foot. Stainless-steel balls with a roundness tolerance of  $\pm 0.001$  mm are also inexpensive. The channel that contains the rods must be milled, but the surface quality in this channel





**Figure 1.53** An improved version of the design shown in Figure 1.52.



**Figure 1.54** Kinematic design that permits an accurately located part to be removed and replaced in the same position.

is not critical. In a milling operation, the sides of the channel can easily be kept straight and parallel to within 0.01 mm/m (.0001 in./ft.).

The three grooves in the platform shown in Figure 1.54 allow precise relocation of a three-legged table after it has

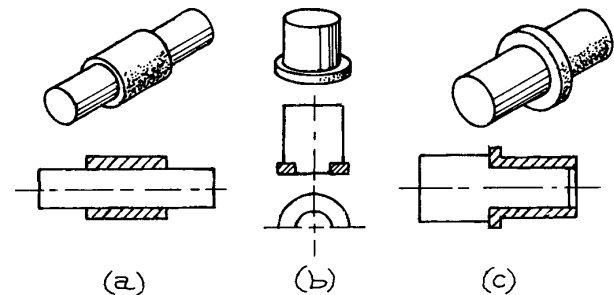
been removed. The ball feet of the table make six points of contact with the platform, so that there are no degrees of freedom.

In applying the principles of kinematic design it must be borne in mind that all materials are more or less elastic. Deformation or deflection of the elements of a kinematic structure under load leads to a deviation from true kinematic behavior. In the absence of careful analysis of such effects, it is often possible to obtain higher precision location or movement, as well as greater stability, through the use of standard, nonkinematic elements such as bearings and ball sliders. Smith and Chetwynd discuss the analysis and exploitation of elastic deformation on kinematic designs in their excellent book on the design of precision mechanism.<sup>11</sup>

### 1.7.2 Plain Bearings

A bearing is a stationary element that locates and carries the load of a moving part. Bearings can be divided into two categories depending upon whether there is sliding or rolling contact between the moving and stationary parts. Sliding-contact bearings are called *plain bearings*. Rolling bearings will be discussed in the next section.

Plain bearings may be designed to carry a radial load, an axial load, or both. Different types are illustrated in Figure 1.55. A radial bearing consists of a cylindrical shaft or *journal* rotating or sliding within a shell, which is the *bearing proper*. The entire assembly is referred to as a *journal bearing*. An axial bearing consists of a flat bearing surface, like a washer, against which the end of the shaft



**Figure 1.55** Plain bearings: (a) a journal bearing; (b) a thrust bearing; (c) a flanged journal bearing.

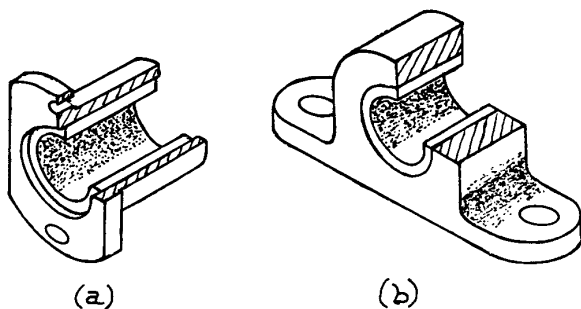
rests. These are called *thrust bearings*. A journal bearing may incorporate a *flanged journal*, in which case it will support a radial load as well as an axial load.

A journal is usually hardened steel or stainless steel. Precision-ground shafts and shafts with precision-ground journals are available in diameters from 1 to 25 mm (1/32 to 1 in.). Bearing shells of bronze or oil-impregnated bronze are available to fit. Nylon bearings for light loads and oil-free applications are also available. Commercial shafting and bearings are manufactured to provide a clearance of 0.005–0.02 mm (.0002–.0010 in.).

Many different methods of lubrication can be employed instead of oil impregnation. The inner surface of the bearing can be grooved, and oil or grease can be forced into the groove through a hole in the shell. If oil is objectionable, a groove on the inner surface of the bearing can be packed with molybdenum disulfide or other dry lubricant.

A plain bearing is installed by pressing it into a hole in the supporting structure. An interference of about 0.03 mm (.001 in.) is desirable for bearings up to an inch in diameter. That is, the outer diameter of the bearing shell should be about 0.03 mm (.001 in.) larger than the hole into which it is pressed. If the interference is too great, the inner diameter of the bearing may be significantly reduced.

A variety of *bearing housings* and *pillow blocks* (Figure 1.56) are available. These mountings are bored to accept standard bearings and in many instances the bearing is premounted. These mounts replace precision-bored bearing mounts.



**Figure 1.56** (a) A bearing housing; (b) a bearing mounted in a pillow block.

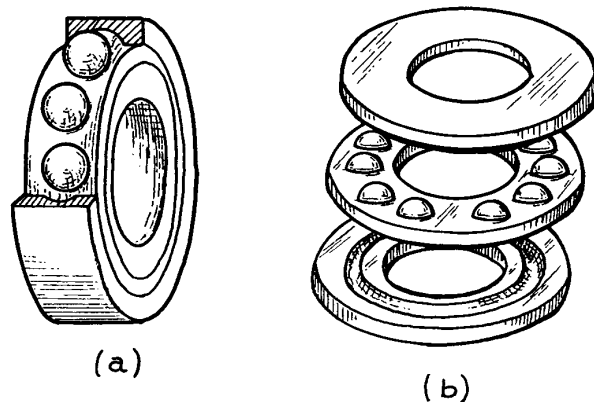
It is usually necessary to provide axial location for a shaft in a journal bearing. This can be accomplished with a retaining ring in a groove on the shaft or a collar secured by a setscrew.

Plain bearings run smoothly and quietly, and have a high load-carrying ability. Properly installed and lubricated, they have a very long life. Because of the close clearances between parts, they are not easily fouled by dirt in their environment. Plain bearings are limited to relatively low-speed operation. Speeds in excess of a few hundred rpm are not practical without forced lubrication. The primary disadvantage of plain bearings is their high starting friction, although, when properly installed and lubricated, their running friction can be very low.

### 1.7.3 Ball Bearings

The rolling element in a rolling contact bearing may be a ball, cylinder, or cone. *Ball bearings* are used for light loads and high speeds. *Roller bearings* employing cylindrical or conical rollers are suitable for very heavy loads, but are not often used for instrument work. We shall discuss only ball bearings.

As with plain bearings, there are both radial and thrust ball bearings (Figure 1.57). A *radial ball bearing* consists of an inner and an outer *race* with a row of balls between. The grooves in each race have a radius slightly larger than the radius of the balls so that there is only point contact



**Figure 1.57** Ball bearings: (a) a radial ball bearing; (b) a thrust ball bearing.

between the balls and the race. The balls are separated by a *retainer* that prevents the balls from rubbing against one another and keeps them uniformly spaced around the bearing. A radial ball bearing can tolerate a substantial thrust load, but for pure axial loads a thrust bearing should be used. A *thrust bearing* is similar to an axial bearing except that it has upper and lower races rather than inner and outer.

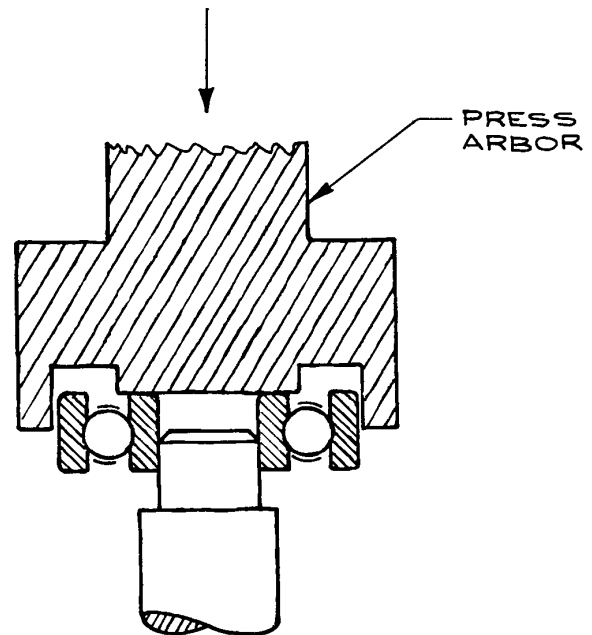
Ball bearings are made of steel or stainless steel. They are graded 1, 3, 5, 7, or 9 depending on manufacturing tolerances. Grades 7 and 9 have ground races and are made to the closest tolerances. They cost little more than lower-grade ones and should be specified for instrument applications.

Proper installation is required to obtain good performance from a ball bearing. The rotating race should be given a firm interference fit, and the stationary race given a light “push fit” to permit some rotational creep. This slight movement of the stationary race helps prevent the maximum load from always bearing on the same spot.

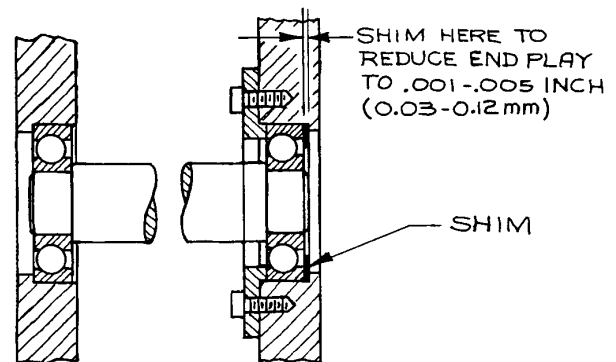
Press fitting changes the internal clearances in a bearing. Bearing manufacturers specify the amount of interference that should be used. As shown in Figure 1.58, the press arbor used to drive a bearing onto a shaft or into a housing should be designed so that the thrust is not transmitted through the balls. Never hammer a bearing into place.

The surface quality and diameter tolerance of the shaft that is to be fitted into a bearing, or the hole that is to house a bearing, must be of the same quality as the bearing. For high-quality bearings the mounting surfaces should be ground. Centerless-ground precision shafting is available to fit all standard bearings. Bearing mounts and pillow blocks with premounted bearings are similarly available to eliminate the need for precision machine work when installing ball bearings.

Ball bearings are designed with both radial and axial clearances. This play is intended to allow for axial misalignment and for dimensional changes that occur upon installation or because of thermal stresses. For the best locational precision and to obtain smooth, vibration-free operation, a ball bearing should be *preloaded* to remove most of this play. A preloading force that displaces one race axially with respect to the other will remove both radial and axial play by causing the balls to roll up the sides of their grooves. The use of a shim to take up the play in a bearing is illustrated in Figure 1.59. For light-duty



**Figure 1.58** Installation of a bearing. The press arbor should bear on the race that is being fitted.



**Figure 1.59** Installation of a shim to remove end play in a shaft mounted on ball bearings.

applications, the endplay can also be taken up by installing a spring washer (Section 1.7.5) instead of a shim. With proper installation, a ball bearing will locate a revolving axis to within a few thousandths of a millimeter (ten-thousandths of an inch).

Bearings must be protected from effects that damage the race surface on which the balls roll. Ball bearings are most likely to fail because of occasional large static loads that produce an indentation in the race. Such a dent is called a *brinell*. Dynamic loads are distributed around the race and are less likely to cause damage; however, a hard vibration can cause brinelling in the form of a series of dents or waves on the surface of a race. Of course, a bearing is also damaged by the introduction of foreign matter that abrades or corrodes the bearing surfaces.

Bearings should be lubricated with petroleum oils or greases. For high speeds and light loads the lightest, finest grades of machine oil can be used. Ball bearings require very little oil. Lubrication is sufficient if there is enough oil to produce an observable meniscus at the point where each ball contacts a race.

Cleanliness is important. In a dirty environment, bearings with built-in side shields should be used. It is probably wise to use enclosed bearings in all instrument applications to keep the bearings clean and to prevent oil from contaminating the environment.

The chief advantages of ball bearings are their low starting friction and very low running friction. They are well suited to high-speed, low-load operation. Relative to plain bearings, ball bearings are noisy and occupy a large volume. The cost of quality ball bearings is so small that economic considerations are usually not important in choosing between rolling bearings and plain bearings for instrument use.

### 1.7.4 Linear-Motion Bearings

A linear-motion bearing may be of either the sliding or rolling type. A plain journal bearing can be used to locate a shaft that is to move axially. A V-shaped or dovetail groove sliding over a mating rail can serve as a bearing between a heavily loaded, slowly moving carriage and a stationary platform. This is the type of bearing used between the carriage and the bed of a lathe.

Linear ball bearings are commercially available. In a bearing for use with an axially moving shaft, the balls that carry the load between the outer race and the shaft move in grooves that run parallel to the axis of the shaft. The balls are recirculated through a return track when they roll to the end of a groove. Linear-motion ball bearings will locate a

shaft to within  $\pm 0.005$  mm ( $\pm .0002$  in.) of a reference axis. Their cost is comparable to conventional rotating ball bearings. Complete roller-slide assemblies are also available. These employ balls rolling in V-grooves. Roller slides will maintain straight-line motion to within 0.005 mm (.0002 in.) per inch of travel.

### 1.7.5 Springs

In many instances it is desirable for a motion to be constrained by a flexible element such as a spring. Springs are used to hold two parts in contact when zero clearance is required, to absorb shock loads, to damp vibrations, and to measure forces.

A spring is characterized by the ratio of the magnitude of applied force to the resulting deflection,  $d$ . This is the *spring rate*:

$$k = \frac{F}{d} \quad (1.34)$$

As is the case for any flexible system, an assembly consisting of a spring and attached load has a natural frequency of vibration:

$$f_n = \frac{1}{2\pi} \left( \frac{k}{m} \right)^{1/2} \text{ Hz} \quad (1.35)$$

where  $m$  includes the mass of both the spring and any load that is affixed to it. Since  $m = W/g = F/a$ , we have, upon substituting for the spring rate in this equation

$$f_n = \frac{1}{2\pi} \left( \frac{g}{d_{st}} \right)^{1/2} \text{ Hz}, \quad (1.36)$$

where  $d_{st}$  is the static deflection produced by the weight of the spring plus the attached load.

In most applications it is desirable to choose a spring that will not resonate with any other part of the apparatus into which it is installed. If the spring is expected to damp a vibratory motion, its natural frequency should differ from that of the disturbance by more than an order of magnitude.

A spring will exert an uneven force when it is subjected to a periodically varying load whose frequency is close to the natural frequency of the spring. When the vibration of

the spring is in phase with the load, the reactive force of the spring will be less than its static force for any given deflection. When the spring is out of phase with the load, it will exert a greater force than expected. This phenomenon is called *surge*. Surge can be reduced or eliminated by using two springs with different natural frequencies. In the case of helical springs, they can be placed one inside the other.

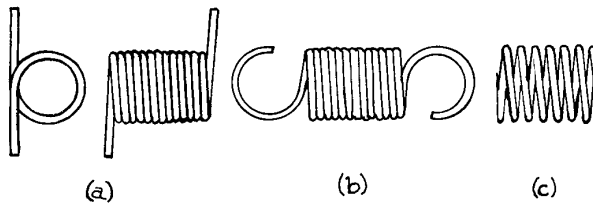
In instrument work, helical springs are most often used. *Helical torsion*, *extension*, and *compression* springs are illustrated in Figure 1.60. The number of coils in a helical spring must be sufficient to ensure that the spring wire remains within its elastic limit when the spring is at maximum deflection. The number of coils in a compression spring also determines the minimum length that is realized when successive coils come into contact. A long compression spring may buckle under stress. This tendency is discouraged if the ends of the spring are square. It can be prevented by placing a rod through the center or by installing the spring in a hole. In general, a compression spring must be supported by some means if its length exceeds its diameter by more than a factor of five.

The spring rate of a helical spring made of round wire is:

$$k = \frac{Gt^4}{8D^3N} \quad (1.37)$$

where  $G$  is the shear modulus of the spring material (about 1/3 of the elastic modulus),  $t$  the diameter of the wire,  $N$  the number of coils, and  $D$  the mean diameter (the average of the inner and outer diameters) of the spring. The natural frequencies for free vibrations are:

$$f_n = \frac{nt}{4\pi ND^2} \left( \frac{gG}{\rho} \right)^{1/2} \text{ Hz} \quad (1.38)$$



**Figure 1.60** Springs: (a) torsion spring; (b) helical extension spring; (c) helical compression spring.

where  $\rho$  is the weight density of the spring wire and  $n$  is an integer. For steel wire:

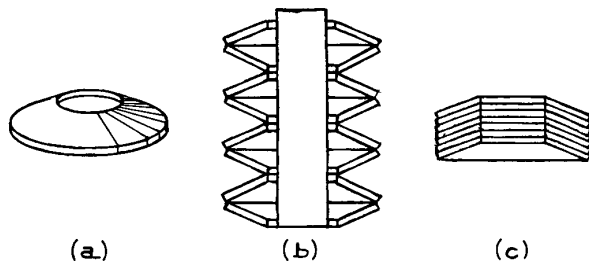
$$\begin{aligned} f_n &= \frac{14000nt}{ND^2} \text{ Hz} \quad (D \text{ and } t \text{ in inches}) \\ &= \frac{550nt}{ND^2} \text{ Hz} \quad (D \text{ and } t \text{ in millimeters}) \end{aligned} \quad (1.39)$$

A variety of steels and bronzes are used for spring manufacture. High-carbon steel wire works well. If necessary the spring can be wound in the annealed state and hardened after forming. Music wire, or piano wire, is one of the best materials for one-off construction of small springs. It is available in diameters of 0.1 to 3 mm (.004 to .103 in.). This wire is very strong and hard because of the drawing process used in its production, and does not need to be hardened after forming. Type-302 stainless-steel wire is useful for springs that are subject to a corrosive environment. Springs of beryllium-copper wire are especially useful in applications where spring deflection is used to gauge a force, since this material maintains a linear stress-strain relation almost to the point of permanent deformation. As mentioned earlier, quartz fiber torsion springs can also be used in these applications.

Winding wire on a mandrel as it is rotated in a lathe conveniently forms helical coil springs. The wire must be kept under tension as it is pulled onto the mandrel. When the tension is released, the formed spring will expand so the mandrel must be somewhat smaller than the desired inner diameter of the finished spring. The production of a small number of springs of a given size and spring rate is probably best carried out by cut-and-try.

Commercially manufactured springs are readily available. They are convenient to use because the supplier specifies such properties as the spring rate and free length.

There are hundreds of possible spring configurations; however, *disc* springs are the only form that we shall mention other than coil springs. A disc spring, also known as a Belleville spring washer, is a cone-shaped disc with a hole in the center [Figure 1.61(a)]. When loaded, the cone flattens. This is a very stiff spring that can absorb a large amount of energy per unit length. Stacking disc springs as in Figure 1.61(b) can create a spring of any desired travel. They must be aligned by a rod passing through their centers or else by stacking them in a hole slightly larger than the



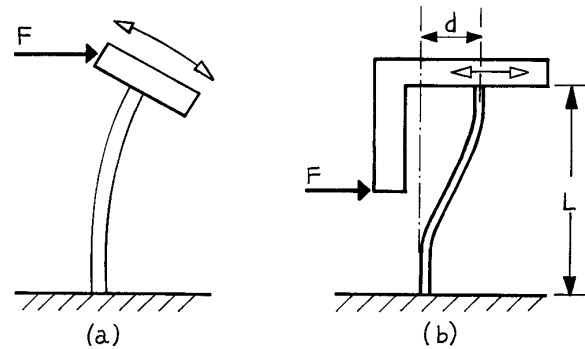
**Figure 1.61** Bellville spring washers; (a) a disc spring; (b) disc springs stacked on a shaft; (c) disc springs stacked in parallel.

outer diameter of the discs. If the discs are stacked in parallel as shown in Figure 1.61(c), they will provide a great deal of damping owing to friction between the faces of the discs.

### 1.7.6 Flexures

Flexures are beams made of a flexible spring material, such as beryllium–copper or spring steel, that are intended to provide a small, controlled, displacement along or about a known axis. They are basically flexible hinges. A clock pendulum is often suspended from a flexure. Owing to their uncomplicated nature and ease of manufacture they are particularly attractive for the provision of controlled motion in small devices. They have the advantages of requiring no lubrication and having no hysteresis since there is no friction and there are no clearances. It is often possible to create a flexure by machining a thin section in a single monolithic mechanism rather than forming and mounting a separate flexible element to the machine. This is not only economical, but avoids problems with alignment between the mechanism and the flexible hinge, as well as movement about the fastener that joins the hinge to the machine. A flexure is typically a simple homogeneous shape so the force–displacement relation can be accurately calculated from elementary physical principles (described in section 1.6.1). With careful design, a linear force–displacement relation can be achieved for small displacements.

Figure 1.62 illustrates, at least conceptually, two basic motions obtainable with a flexible beam. In both cases the flexible element is rigidly mounted to a stationary base at one end and to a moveable platform at the free end. In Figure 1.62(a) a force applied to the platform produces a rotation of the platform. For small displacements the angle



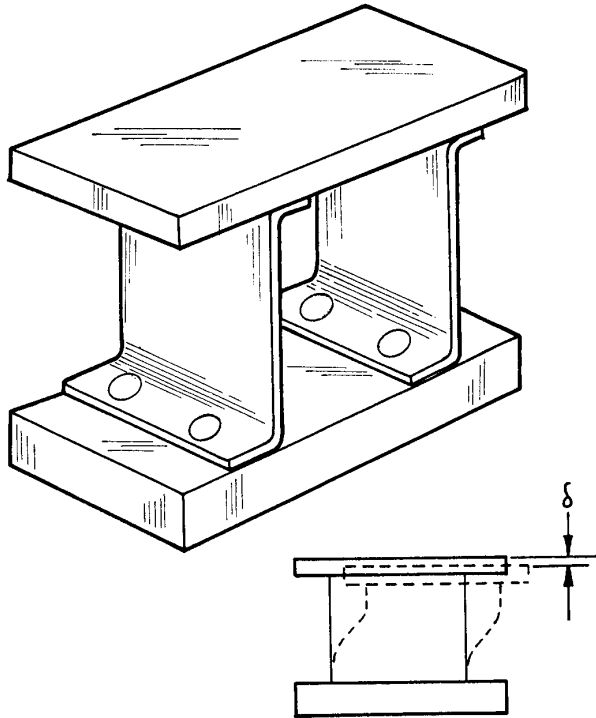
**Figure 1.62** A platform mounted to a flexure attached to a solid base: (a) a transverse force applied at the end of the flexure produces a rotation of the platform; (b) a force applied normal to the center of the flexure produces a linear translation of the platform.

of rotation is proportional to the applied force. In Figure 1.62(b) a bracket is attached to the platform so that the force can be applied along a line normal to the center of the flexure. The flexed spring bends near its ends and assumes an “S” shape with a straight section in the middle. The bending moment, being a function of the curvature of the beam, is zero at the midpoint in line with the applied force and the platform undergoes a linear displacement ( $d$ ) that, for small displacements, is proportional to the force ( $F$ ):

$$d = \frac{FL^3}{12EI} \quad (1.40)$$

where  $I$  is the centroidal moment of inertia of the spring section and  $E$  the modulus of elasticity of the spring material. The spring constant, or *stiffness*, is  $F/d$ .

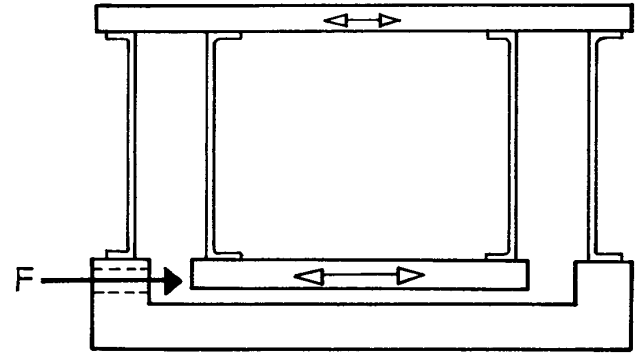
In practice neither of the arrangements in Figure 1.62 is satisfactory since any misalignment of the applied force leads to pitching or twisting of the platform. A practical design of a linear translation platform is illustrated in Figure 1.63. The wide springs illustrated offer resistance to twisting and provide a wider platform for the attachment of other components of the apparatus. Using two springs in a parallelogram arrangement offers resistance to pitching. In many applications of this design, the force is applied directly to the platform; however, maximum resistance to pitching is gained if a bracket is attached, as in



**Figure 1.63** A practical design for a linear translation platform mounted on leaf springs. The inset illustrates the vertical error in the linear motion.

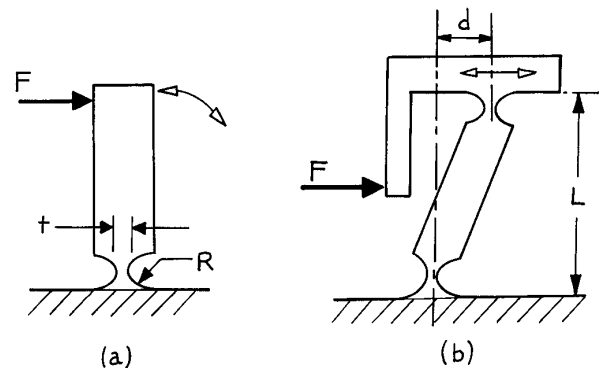
Figure 1.62(b) so that the force acts through the midpoint of the springs. With two springs, the stiffness of the mechanism is doubled and the displacement for a given force is halved. The parallelogram arrangement results in a slightly curved path of the platform; the platform moves downward with increasing displacement. This error in the linear motion has been analyzed in detail by Jones<sup>12</sup> and must be accounted for in precision applications.

The vertical displacement error in the design in Figure 1.63 is corrected in the compound flexure design shown in Figure 1.64. Here the platform in the simple flexure becomes the base for a second flexure. The downward displacement of the intermediate platform is exactly compensated by an upward displacement of the lower platform. The driving force is applied to the lower platform. The overall stiffness is half that of the simple flexure in Figure 1.63. It is worth noting that the compound design also compensates for temperature effects on the lengths of the springs.



**Figure 1.64** A compound flexure design to correct for the error shown in Figure 1.63.

A leaf spring with clamped ends moving in antiparallel directions bends into an “S” shape as noted above. Nearly all the flexion occurs near the clamped ends with the center section remaining straight as if it were rigid. This flexion near the clamped end of a spring is quite similar to what would occur when a bending moment is applied to a cantilever that is notched near its fixed end, as shown in Figure 1.65(a). This is a *notch hinge*. A beam with a notch hinge at each end can simulate a leaf spring fixed at one end to a base and at the other end to a transversely moving platform. The obvious extension of this idea is a flexure in which the base, platform, and spring are all machined from a single piece of material. Figure 1.65(b) shows the monolithic notch-hinge analogy to the basic linear-motion flexure



**Figure 1.65** Monolithic flexures analogous to those shown in Figure 1.62.

illustrated in Figure 1.62(b). The force–displacement relation for the notch hinge has been derived analytically by Paros and Weisbord<sup>13</sup> and by finite element methods by Smith, Chetwynd, and Bowen.<sup>14</sup> The latter analysis gives:

$$d = \frac{3FKRL^2}{EBt^3} \quad (1.41)$$

for the basic linear-motion flexure illustrated in Figure 1.65(b), where  $R$  is the radius of each of the notches,  $t$  is the width of the web at each notch,  $B$  is the width of the beam (as in Figure 1.48), and the factor  $K$  accounts for the effective length of the flexible part of the web.  $K$  varies from about 0.2 for a relatively thin, flexible web, where  $t$  is much less than  $R$ , to about 0.7 for a relatively stiff spring, with  $t$  comparable to  $R$ .

Flexures provide for relatively small displacements, especially flexures employing notch hinges. Linear force–displacement response, free of hysteresis, requires that the elastic limit of the spring material not be exceeded. For the linear-motion notch-hinge flexure in Figure 1.65(b), maximum displacement, in the model of Smith, *et al.*,<sup>14</sup> is approximated by:

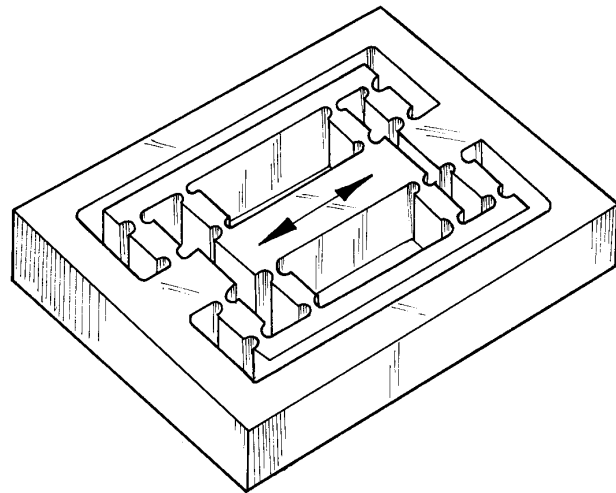
$$d_{\max} = 4K \left( \frac{R}{t} \right) \left( \frac{s_{\max}}{E} \right) L \quad (1.42)$$

where  $s_{\max}$  is the maximum tolerable stress in the material of the spring. For design purposes,  $s_{\max}$  can be taken to be the yield strength of the material, or, conservatively, some fraction thereof. The ratio of the yield strength to the modulus of elasticity ( $E$ ) for mild steel, stainless steel, or tempered aluminum alloy is about 0.003 and for hard plastics such as polyamide (Nylon) or polyimide (Vespel) the ratio is about 0.03. In a simple notch-hinge linear-motion flexure, the ratio ( $R/t$ ) is typically about 5; this gives a maximum displacement of no more than about  $0.01L$  for a simple metal flexure and  $0.1L$  for a plastic flexure.

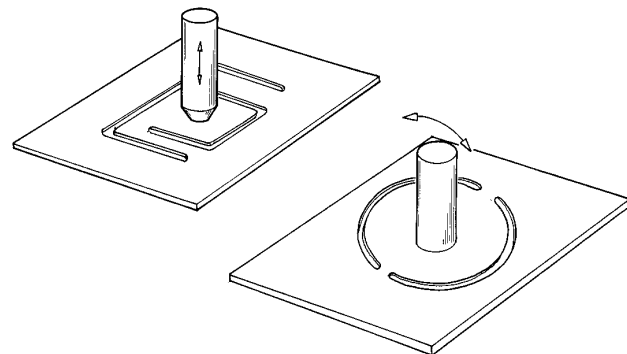
As suggested by the contour of the notch hinges shown in Figure 1.65, fabrication of a monolithic flexure requires only milling and boring operations. The precision of the motion, as well as the stiffness, of a monolithic flexure is determined primarily by the precision in the location of the centers of the holes that create a notch. The relative ease of manufacture of precise monolithic flexures allows for quite complex designs. An elegant,

stable, and precise, double-compound, rectilinear spring designed by Smith, *et al.*,<sup>14</sup> is shown in Figure 1.66.

There are many variations on the basic notch spring. On a beam, a closely spaced pair of notches at right angles provides for two degrees of freedom. Notch springs machined into thin metal stock can serve the function of a Belleville spring to absorb shock or take up end-play in a mechanism or to limit motion to a single plane. Examples are shown in Figure 1.67. In these flexures the material in the notch is stressed in torsion.



**Figure 1.66** Monolithic, double-compound, rectilinear spring (reference 14).



**Figure 1.67** Monolithic flexures in a thin material.



## Cited References

1. Richard P. Pohandish (Ed.), *Machinery's Handbook Pocket Companion*, Industrial Press Inc., New York, 2000.
2. *ASM Handbook Volume 04: Heat Treating*, ASM International, Ohio, 1991.
3. J. W. Dally and W. F. Riley, *Experimental Stress Analysis*, McGraw-Hill, New York, 1965, pp. 165–221.
4. F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, Addison-Wesley, Reading, MA, 1965.
5. A comprehensive list of brazing alloys has been compiled by W. H. Kohl, in *Handbook of Vacuum Physics*, Vol. 3, *Technology*, A. H. Beck (Ed.), Pergamon Press, Elmsford, NY, 1964; *Materials and Techniques for Electron Tubes*, Reinhold, N, 1959. See also reference 6.
6. *The Brazing Book*, Lucas-Milhaupt/Handy & Harman of Canada, Cudahy, WI, USA.
7. R. J. Roark, *Formulas for Stress and Strain*, 4th edn., McGraw-Hill, New York, 1965.
8. Clifford Mathews, A. S. M. E. *Engineer's Data Book*, 2nd edn., ASME Press, NY, 2001, pp. 75, 188–189.
9. D. E. Newland, *Mechanical Vibration Analysis and Computation*, Longman Scientific & Technical/John Wiley & Sons, Inc., New York, 1989; S. S. Rao, *Mechanical Vibrations*, 2nd edn., Addison-Wesley, Reading, MA, 1990; W. Weaver, Jr., S. P. Timoshenko, and D. H. Young, *Vibration Problems in Engineering*, 5th edn., John Wiley & Sons, Inc., New York, 1990.
10. A useful review of kinematic design is given by J. E. Furse, *J. Phys. E.*, **14**, 264, 1981.
11. S. T. Smith and D. G. Chetwynd, *Foundations of Ultraprecision Mechanism Design*, Gordon and Breach Science Publishers, Philadelphia, 1992.
12. R. V. Jones, *J. Sci. Instr.*, **28**, 38, 1951; *J. Sci. Instr.*, **33**, 11, 1956.
13. J. M. Paros and L. Weisbord, *Machine Design*, **151**, 25 Nov. 1965.
14. S. T. Smith, D. G. Chetwynd, and D. K. Bowen, *J. Phys. E: Sci. Instrum.*, **20**, 977, 1987.

## General References

### Design of Moving and Rotating Machinery

- R. M. Phelan, *Dynamics of Machinery*, McGraw-Hill, New York, 1967.

### Vibration Analysis

- D. E. Newland, *Mechanical Vibration Analysis and Computation*, 5th edn., Longman Scientific & Technical, Essex, England and John Wiley & Sons, New York, 1989.
- S. S. Rao, *Mechanical Vibrations*, 2nd edn., Addison-Wesley, Reading, MA, 1990.
- W. Weaver, Jr., S. P. Timoshenko, and D. H. Young, *Vibration Problems in Engineering*, John Wiley & Sons, New York, 1990.

### Mechanical Design Texts

- A. D. Deutschman, W. J. Michels, and C. E. Wilson, *Machine Design*, Macmillan, New York, 1975.
- V. M. Faires, *Design of Machine Elements*, 4th edn., Macmillan, New York, 1965.
- R. E. Parr, *Principles of Mechanical Design*, McGraw-Hill, New York, 1970.
- R. M. Phelan, *Fundamentals of Mechanical Design*, 3rd edn., McGraw-Hill, New York, 1970.
- S. T. Smith and D. G. Chetwynd, *Foundations of Ultraprecision Mechanism Design*, Gordon and Breach Science Publishers, Philadelphia, 1992.

### Mechanical-Drawing Texts

- T. E. French and C. J. Vierck, *Engineering Drawing and Graphic Technology*, 11th edn., McGraw-Hill, New York, 1972; (entitled *A Manual of Engineering Drawing* in its first 10 editions).
- F. E. Ciesecke, A. Mitchell, H. C. Spencer, and I. L. Hill, *Technical Drawing*, 5th edn., Macmillan, New York, 1967.

### Mechanical Engineering Handbooks

- R. Timings, *Mechanical Engineer's Pocket Book*, 3rd edn., Newnes/Elsevier, Oxford, 2006.
- Richard P. Pohandish (ed.), *Machinery's Handbook Pocket Companion*, Industrial Press Inc., New York, 2000.
- Clifford Mathews, A. S. M. E. *Engineer's Data Book*, 2nd edn., ASME Press, NY, 2001.

### Properties of Materials

- A. J. Moses, *The Practicing Scientist's Handbook*, Van Nostrand Reinhold, New York, 1978; F. Rosebury, *Handbook of Electron*

*Tube and Vacuum Techniques*, Addison-Wesley, Reading, MA, 1965.

*Goodfellow Catalog*, Berwyn, PA and, Goodfellow Cambridge Limited, Cambridge, web catalog <http://www.goodfellow.com>.

## **Brazing**

*The Brazing Book*, Lucas-Milhaupt/Handy & Harman of Canada, Cudahy, WI and online at <http://www.handyharmancanada.com>

**Appendix 1.1 Number drills with inch and metric equivalents and tap sizes for approximately 75% thread depth.**

SIZE No or Letter	inches	mm	Tap Size American National	Tap Size ISO Metric Coarse Thread Series
80	.0135	0.34		
79	.0145	0.37		
78	.0160	0.41		
77	.0180	0.46		
76	.0200	0.51		
75	.0210	0.53		
74	.0225	0.57		
73	.0240	0.61		
72	.0250	0.64		
71	.0260	0.66		
70	.0280	0.71		
69	.0292	0.74		
		0.75		M1
68	.0310	0.79		
67	.0320	0.81		
66	.0330	0.84		
65	.0350	0.89		
64	.0360	0.91		
63	.0370	0.94		
		0.95		M1.2
62	.0380	0.97		
61	.0390	0.99		
60	.0400	1.02		
59	.0410	1.04		
58	.0420	1.07		
57	.0430	1.09		
		1.15		M1.4
56	.0465	1.18		
	3/64		0-80	
		1.30		M1.6
55	.0520	1.32		
54	.0550	1.40		
		1.50		M1.8
53	.0595	1.51	1-64	
52	.0635	1.61		
		1.65		M2
51	.0670	1.70		
50	.0700	1.78	2-56	
		1.8		M2.2
49	.0730	1.85		
48	.0760	1.93		
47	.0785	1.99	3-48	
46	.0810	2.06		

---

**Appendix 1.1. (contd.)**


---

SIZE No or Letter	inches	mm	Tap Size American National	Tap Size ISO Metric Coarse Thread Series
45	.0820	2.08		
		2.10		M2.5
44	.0860	2.18		
43	.0890	2.26	4-40	
42	.0935	2.37		
41	.0960	2.44		
40	.0980	2.49		
39	.0995	2.53		
		2.55		M3
38	.1015	2.58	5-40	
37	.1040	2.64		
36	.1065	2.71	6-32	
35	.1100	2.79		
34	.1110	2.82		
33	.1130	2.87		
32	.1160	2.95		M3.5
31	.1200	3.05		
30	.1285	3.26		
		3.40		M4
29	.1360	3.45	8-32	
28	.1405	3.57		
27	.1440	3.66		
26	.1470	3.73		
25	.1495	3.80	10-24	M4.5
24	.1520	3.86		
23	.1540	3.91		
22	.1570	3.99		
21	.1590	4.04	10-32	
20	.1610	4.09		
19	.1660	4.22		
		4.30		M5
18	.1695	4.31		
17	.1730	4.39		
16	.1770	4.50	12-24	
15	.1800	4.57		
14	.1820	4.62		
13	.1850	4.70		
12	.1890	4.80		
11	.1910	4.85		
10	.1935	4.91		
9	.1960	4.98		
8	.1990	5.05		
		5.10		M6

---

---

**Appendix 1.1. (contd.)**


---

SIZE No or Letter	inches	mm	Tap Size American National	Tap Size ISO Metric Coarse Thread Series
7	.2010	5.11	1/4-20	
6	.2040	5.18		
5	.2055	5.22		
4	.2090	5.31		
3	.2130	5.41	1/4-28	
2	.2210	5.61		
1	.2280	5.79		
A	.234	5.94		
B	.238	6.05		
		6.10		M7
C	.242	6.15		
D	.246	6.25		
E	.250	6.35		
F	.257	6.53	5/16-18	
G	.261	6.63		
H	.266	6.76		
		6.90		M8
I	.272	6.91		
J	.277	7.04		
K	.281	7.14		
L	.290	7.37		
M	.295	7.49		
N	.302	7.67		
		7.90		M9
	5/16		3/8-16	
O	.316	8.03		
P	.323	8.20		
Q	.332	8.43		
		8.60		M10
R	.339	8.61		
S	.348	8.84		
T	.358	9.09		
U	.368	9.35	7/16-14	
V	.377	9.58		
		9.60		M11
W	.386	9.80		
X	.397	10.08		
Y	.404	10.26		
		10.40		M12
Z	.413	10.49		
	27/64		1/2-13	
	29/64		1/2-20	
	.480	12.20		M14

---

**Appendix 1.2 American Standard Threads. Unified National Coarse (UNC) and Unified National Fine (UNF) threads with tap drill and clearance drill sizes. The tap drill size is for approximately 75% thread depth. A range of clearance hole drills is given depending upon whether a close fit or a free fit is desired.**  
(Number and letter drill diameters are given in Appendix 1.1.)

<i>Size</i> (Nominal Diameter, Inches)	<i>UNC</i>		<i>UNF</i>		<i>Clearance Hole</i> <i>Drill Close-Free</i>
	<i>Threads per Inch</i>	<i>Tap Drill Size</i>	<i>Threads per Inch</i>	<i>Tap Drill Size</i>	
0 (0.060)			80	3/64	52–50
1 (0.073)	64	53	72	53	48–46
2 (0.086)	56	50	64	50	43–41
3 (0.099)	48	47	56	45	37–33
4 (0.112)	40	43	48	42	32–30
5 (0.125)	40	38	44	37	30–29
6 (0.138)	32	36	40	33	27–25
8 (0.164)	32	29	36	29	18–16
10 (0.190)	24	25	32	21	9–7
12 (0.216)	24	16	28	14	2–1
1/4	20	7	28	3	F–H
5/16	18	F	24	I	P–Q
3/8	16	5/16	24	Q	W–X
7/16	14	U	20	25/64	29/64–15–32
1/2	13	27/64	20	29/64	33/64–17/32

Note: ASA B1.1-1989.

### Appendix 1.3 ISO (Metric) Coarse Threads.

<i>Size M(mm)</i>	<i>Pitch (mm)</i>	<i>Tap Drill (mm)</i>	<i>Size M(mm)</i>	<i>Pitch (mm)</i>	<i>Tap Drill (mm)</i>
M1	0.25	0.75	M3.5	0.6	2.90
M1.2	0.25	0.95	4	0.7	3.30
M1.4	0.30	1.10	5	0.8	4.20
M1.6	0.35	1.25	6	1	5.00
M1.8	0.35	1.45	8	1.25	6.80
M2	0.4	1.60	10	1.5	8.50
M2.2	0.45	1.75	12	1.75	10.50
M2.5	0.45	2.05	14	2	12.00
M3	0.5	2.50	16	2	14.00
M3.5	0.6	2.90	18	2.5	15.50

---

**Appendix 1.4 American Standard Taper Pipe Threads.**


---

<i>Nominal Bore Size of Pipe (in.)</i>	<i>Actual O.D. of Pipe (in.)</i>	<i>Threads per Inch</i>	<i>Length of Engagement by Hand (in.)</i>	<i>Length of Effective Thread (in.)</i>
1/8	0.405	27	0.180	0.260
1/4	0.540	18	0.200	0.401
3/8	0.675	18	0.240	0.408
1/2	0.840	14	0.320	0.534
3/4	1.050	14	0.340	0.546
1	1.315	11 1/2	0.400	0.682
1 1/4	1.660	11 1/2	0.420	0.707
1 1/2	1.900	11 1/2	0.420	0.724
2	2.375	11 1/2	0.436	0.756
2 1/2	2.875	8	0.682	1.136
3	3.500	8	0.766	1.200

---

Note: ASA B2.1-1989.

---

**Appendix 1.5 British Standard Pipe Thread Tapered.**  
 (BSPT, BS EN 10226 equivalent to ISO 7)
 

---

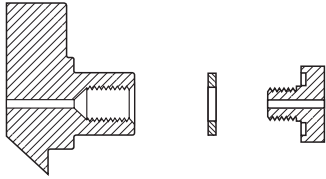
<i>Nominal Bore Size of Pipe<sup>a</sup> (in.)</i>	<i>Nominal Bore Size of Pipe<sup>a</sup> (mm)</i>	<i>Pipe OD (mm)</i>	<i>Threads per Inch</i>	<i>Pitch (mm)</i>	<i>Length of Engagement by Hand (mm)</i>	<i>Effective Thread Length (mm)</i>
1/8	6	10.2	28	0.907	4.0	6.5
1/4	8	13.5	19	1.337	5.0	8.7
3/8	10	17.5	19	1.337	6.4	10.1
1/2	15	21.3	14	1.814	8.2	13.2
3/4	20	26.9	14	1.814	9.5	14.5
1	25	33.7	11	2.309	10.4	16.8
1 1/4	32	42.4	11	2.309	12.7	19.1
1 1/2	40	48.3	11	2.309	12.7	19.1

---

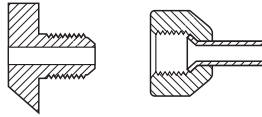
<sup>a</sup> Nominal bore sizes in inches and mm are equivalents not conversions

Appendix 1.6 CGA connections for high pressure gas cylinders.

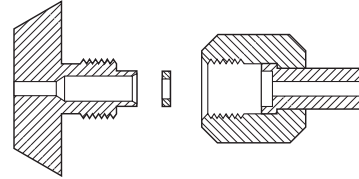
**CGA 110**  
.3125-32UNEF-2B-RH-INT



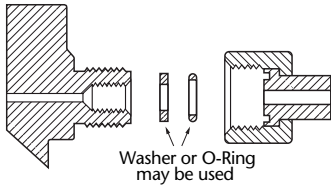
**CGA 165**  
.4375-20UNF-2A-RH-EXT (1/4" SAE Flare)



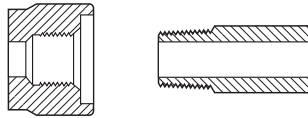
**CGA 170**  
.5625-18UNF-2A-RH-EXT



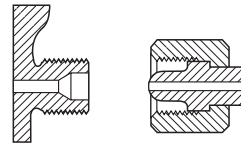
**CGA 180**  
.625-18UNF-2A-RH-EXT



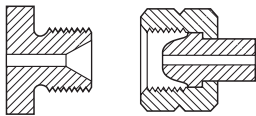
**CGA 240**  
.375-18NGT-RH-INT



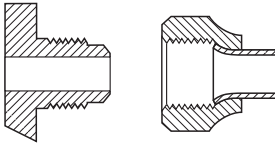
**CGA 280**  
.745-14NGO-RH-EXT



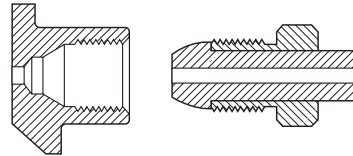
**CGA 290**  
.745-14NGO-LH-EXT



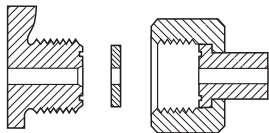
**CGA 295**  
.750-16UNF-2A-RH-EXT (1/2" SAE Flare)



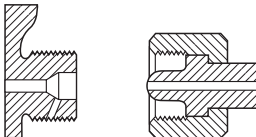
**CGA 296**  
.803-14UNS-2B-RH-INT



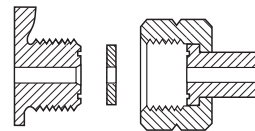
**CGA 320**  
.825-14NGO-RH-EXT (Flat Nipple)



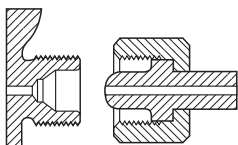
**CGA 326**  
.825-14NGO-RH-EXT (Small Round Nipple)



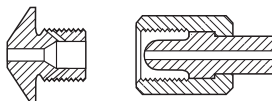
**CGA 330**  
.825-14NGO-LH-EXT (Flat Nipple)



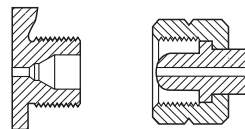
**CGA 346**  
.825-14NGO-RH-EXT (Large Round Nipple)



**CGA 347**  
.825-14NGO-RH-EXT (Long Round Nipple)



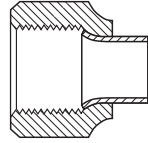
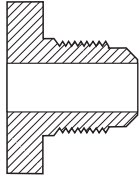
**CGA 350**  
.825-14NGO-LH-EXT (Round Nipple)



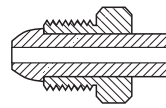
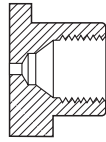


Appendix 1.6 (contd.)

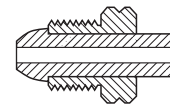
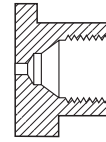
**CGA 440**  
.875-14UNF-2A-RH-EXT (5/8" SAE Flare)



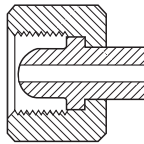
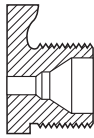
**CGA 500**  
.885-14NGO-RH-INT (Bullet Nipple)



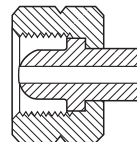
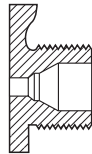
**CGA 510**  
.885-14NGO-LH-INT (Bullet Nipple)



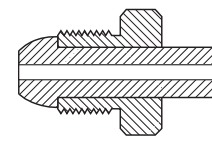
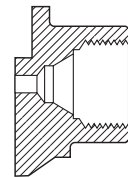
**CGA 540**  
.903-14NGO-RH-EXT



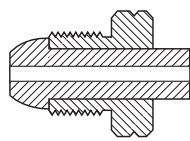
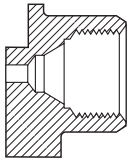
**CGA 555**  
.903-14NGO-LH-EXT



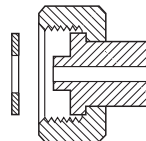
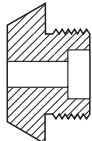
**CGA 580**  
.965-14NGO-RH-INT



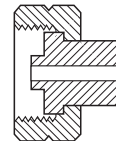
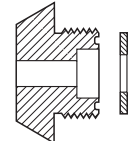
**CGA 590**  
.965-14NGO-LH-INT



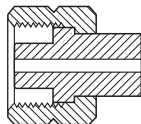
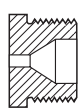
**CGA 660**  
1.030-14NGO-RH-EXT (Face Washer)



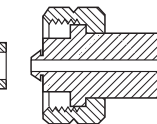
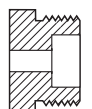
**CGA 670**  
1.030-14NGO-LH-EXT (Face Washer)



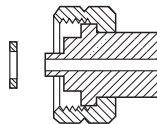
**CGA 677**  
1.030-14NGO-LH-EXT (Round Nipple)



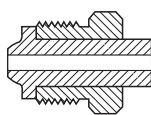
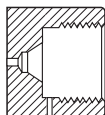
**CGA 678**  
1.030-14NGO-LH-EXT (Recessed Washer)



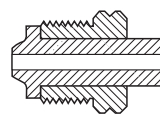
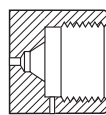
**CGA 679**  
1.030-14NGO-LH-EXT (Tipped Nipple)



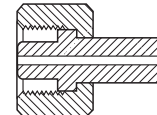
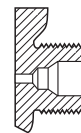
**CGA 680**  
1.045-14NGO-RH-INT



**CGA 695**  
1.045-14NGO-LH-INT



**CGA 701**  
1.103-14NGO-RH-EXT

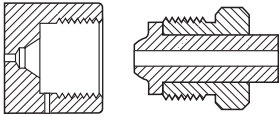


---

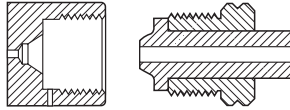
**Appendix 1.6 (contd.)**

---

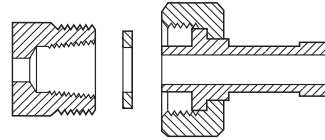
**CGA 702**  
1.125-14NGO-RH-INT



**CGA 703**  
1.125-14NGO-LH-INT



**CGA 705**  
1.125-14UNS-2A-RH-EXT



## WORKING WITH GLASS

Glass has been called the miraculous material. The ubiquity of glass in the modern laboratory certainly confirms this. Because glass is chemically inert, most containers are made of it. Glass is transparent to many forms of radiation, and its transmission properties can be varied by controlling its composition; all sorts of windows and lenses are made of glass. Because glass can be polished to a high degree and is dimensionally stable, most mirrors are supported on glass surfaces. Glass is strong and stiff and is often used as a structural material. Considering its mechanical rigidity and density, it is a reasonably good thermal insulator. It is an excellent electrical insulator. Perhaps the greatest virtue of this material is that many glasses are inexpensive and can be cut and shaped in the laboratory with inexpensive tools.

Thirty-five years ago, most glass laboratory apparatus was produced by the scientist or technician *in situ* by blowing molten glass or by grinding, cutting, and polishing hard glass. Today the glass industry has grown to such an extent that nearly all components of a glass apparatus are available from commercial sources at low cost. These include all sorts of containers, chemical labware, vacuum-system components, mirrors, windows, and lenses. It is often only necessary for laboratory scientists to acquaint themselves with the range of components available and to acquire the skills needed to assemble an apparatus from these components.

### 2.1 PROPERTIES OF GLASSES

The chemical composition of glass is infinitely variable, and so therefore are the thermal, electrical, mechanical,

and chemical properties of glass. Furthermore, glass is a fluid that retains a memory of its past history. It is possible, however, to review the general properties of glass and to specify the properties of glasses of a particular composition and method of manufacture.

#### 2.1.1 Chemical Composition and Chemical Properties of Some Laboratory Glasses

The chief constituent of any commercial glass is silica ( $\text{SiO}_2$ ). All laboratory ware is at least 3/4 silica, with other oxides added to obtain certain thermal properties or chemical resistance.

The least expensive and, until the last quarter century, the most common glass used for laboratory ware is known as *soda-lime glass* or *soft glass*. This glass contains 70–80% silica, 5–10% soda ( $\text{Na}_2\text{O}$ ), 5–10% potash ( $\text{K}_2\text{O}$ ), and 10% lime ( $\text{CaO}$ ). A particular advantage of this glass is that it can be softened in a natural-gas–air flame.

*Borosilicate glass* has supplanted soda-lime glass for the manufacture of labware. In this glass the alkali found in soft glass is replaced by  $\text{B}_2\text{O}_3$  and alumina ( $\text{Al}_2\text{O}_3$ ). Borosilicate glass is superior to soda-lime glass in its resistance to chemical attack and thermal or mechanical shock. It softens at a higher temperature than soft glass, however, and is more difficult to work. Borosilicate glasses of many different compositions are manufactured for various laboratory applications, but far and away the most common laboratory glass is the borosilicate glass designated by Corning as *Pyrex 7740* and by Kimble as *Kimax KG-33*. The composition is:  $\text{SiO}_2$ , 80.5%;  $\text{B}_2\text{O}_3$ , 12.9%;  $\text{Na}_2\text{O}$ ,

3.8%;  $\text{Al}_2\text{O}_3$ , 2.2%;  $\text{K}_2\text{O}$ , 0.4%. Borosilicate glass tubing and rod, as well as a wide range of components, are available from many suppliers, including Corning Glass Works, Kimble-Kontes, Wilmad-Labglass, ChemGlass and Ace Glass.

Glassware of extremely good chemical resistance is produced from borosilicate glass by thermal and chemical processing: a heat treatment causes the glass to separate into two phases – one high in silica and the other rich in alkali and boric oxides. This second phase is leached out with acid, and the remaining phase is heated to give a clear consolidated glass that is nearly pure silica. This glass is known as *96% silica glass* and is designated by Corning as Vycor No. 7900.

Glass composed only of silica is known as *vitreous* or *fused silica* or simply as “quartz.” Because of its refractory properties and chemical durability, this material would be the most desirable glass were it not for the high cost of making it and the extremely high temperatures required to work it. On the other hand, the relative cost of quartz is decreasing to the extent that the market for 96% silica glass is rapidly disappearing.

Most laboratory glasses are transparent to visible light, making visual distinction between the various glass compositions impossible. For this reason it is important to label glass materials before storing and to avoid mixing different kinds of glass. When necessary it is possible to distinguish different glasses from one another by differences in thermal or optical properties. The gas–air flame of a Bunsen burner will soften soda-lime glass, but not borosilicate glass. A natural-gas–oxygen flame is required to soften borosilicate glass, but this flame will not affect fused silica. Fused silica must be raised to a white heat in an oxyhydrogen flame before it will soften. Glasses of different composition cannot generally be successfully fused together because of differences in the amount of expansion and contraction on heating and cooling. An unknown piece of glass may be compared with a known piece by placing the two side by side with their ends coincident. The two ends are softened together in a flame and pressed together with tweezers. Then the fused ends are reheated and drawn out into a long fiber about 0.5 mm in diameter and permitted to cool. If the fiber remains straight, the two pieces have the same coefficient of expansion. If the fiber curves, the two pieces are of different composition.

A measurement of the refractive index is usually a sensitive and reliable test of glass composition. A piece of glass placed in a liquid of exactly the same refractive index will become invisible. For example, a test solution for Pyrex 7740 can be made of 16 parts by volume of methanol in 84 parts benzene. This test solution should be kept in a tightly covered container so that its composition does not change as a result of evaporation.

As can be judged from the extreme chemical environments to which glasses are routinely subjected, glass is indeed resistant to chemical action. However, one need only observe windows clouded by the action of rainfall in a polluted atmosphere or glassware permanently stained by laboratory chemicals to confirm that glass is not entirely impervious to chemical attack. Glass is attacked most readily by alkaline solutions, and all types are affected about equally. Water has an effect. The soft glasses are most susceptible, borosilicate glass is only slightly affected, and 96% silica hardly at all, since most of its soluble components were leached out in manufacture. Acids attack glass more readily than water, although, again, borosilicate glass is more resistant to acids than soda glass, and 96% silica is more resistant yet.

### 2.1.2 Thermal Properties of Laboratory Glasses

One of the most important properties of glass is its very low coefficient of thermal expansion. It is this property that permits glass to be formed at high temperatures in a molten state and then cooled without changing shape or breaking. The coefficients of linear expansion of several types of glass are given in Table 2.1. As can be seen from these data, borosilicate glass is much more resistant to thermal shock than soft glass. Fused silica is so stable that a white-hot piece can be immersed in liquid air without fracturing.

Glass does not have a melting point. Instead, the working properties of a glass are specified by particular points on its viscosity–temperature curve. The *softening point* is approximately the temperature at which a glass can be observed to flow under its own weight. The *annealing point* is the temperature at which internal stresses can be relieved (annealed) in a few minutes. The *strain point* is the temperature below which glass can be quickly cooled

**Table 2.1 Thermal properties of glass**

<i>Glass</i>	<i>Linear Expansion Coefficient (<math>\text{cm}^{-1} \text{K}^{-1}</math>)</i>	<i>Strain Point (<math>^{\circ}\text{C}</math>)</i>	<i>Annealing Point (<math>^{\circ}\text{C}</math>)</i>	<i>Softening Point (<math>^{\circ}\text{C}</math>)</i>	<i>Working Point (<math>^{\circ}\text{C}</math>)</i>
Soda-lime (typical)	$8\text{--}10 \times 10^{-6}$	500	550	700	1000
Pyrex 7740 (borosilicate)	$3.3 \times 10^{-6}$	510	555	820	1250
Vycor 7900 (96% silica)	$0.75 \times 10^{-6}$	890	1020	1500	—
Fused Silica	$0.55 \times 10^{-6}$	950	1100	1600	—

without introducing additional stress. The *working point* is the temperature at which the glass is formed by a skilled glassblower.

The specific heat of glass is about  $0.8 \text{ J g}^{-1} \text{ K}^{-1}$ , and the thermal conductivity of glass is about  $0.008 \text{ J cm sec}^{-1} \text{ cm}^{-2} \text{ K}^{-1}$ . For reference, this thermal conductivity is about an order of magnitude less than that of graphite, two orders less than that of metal, and about an order of magnitude greater than that of wood.

### 2.1.3 Optical Properties of Laboratory Glassware

In many experiments light must be transmitted through the wall of a glass container. The transmission as a function of wavelength depends upon the composition of the glass. The composition of the soda-lime glasses varies considerably, and the optical properties of each piece should be determined by spectroscopic analysis before it is used in an experiment requiring light transmission. Spectrophotometric curves for some lab glasses of well-defined composition are given in Figure 2.1. As can be seen, fused silica is transparent over the widest wavelength range.

### 2.1.4 Mechanical Properties of Glass

The tensile, compressive, and shear strength of a piece of glass depends upon its shape and history and upon the time over which it is loaded. Glass appears to be much stronger in compression than tension. This is in part due to the fact that a glass surface can be made very smooth in order to uniformly distribute a compressive load. Values of  $400\text{--}1200 \text{ MN/m}^2$  ( $60\ 000\text{--}180\ 000 \text{ psi}$ ) are quoted for the compression strength. ( $1 \text{ psi}$  (pound per square inch) =  $6895 \text{ N/m}^2$ .)

Glass is nearly perfectly elastic. When it breaks it does so without plastic deformation. Springs made of glass or quartz fibers behave almost ideally. The modulus of elasticity of glass is high and, of course, depends upon composition. Young's modulus (the modulus of elasticity) for Pyrex 7740 is  $63 \text{ GN/m}^2$  ( $9.1 \times 10^6 \text{ psi}$ ) at  $0 \text{ }^{\circ}\text{C}$  and  $65 \text{ GN/m}^2$  ( $9.4 \times 10^6 \text{ psi}$ ) at  $100 \text{ }^{\circ}\text{C}$ . For fused silica the corresponding values are  $72 \text{ GN/m}^2$  ( $10.5 \times 10^6 \text{ psi}$ ) and  $74 \text{ GN/m}^2$  ( $10.7 \times 10^6 \text{ psi}$ ).

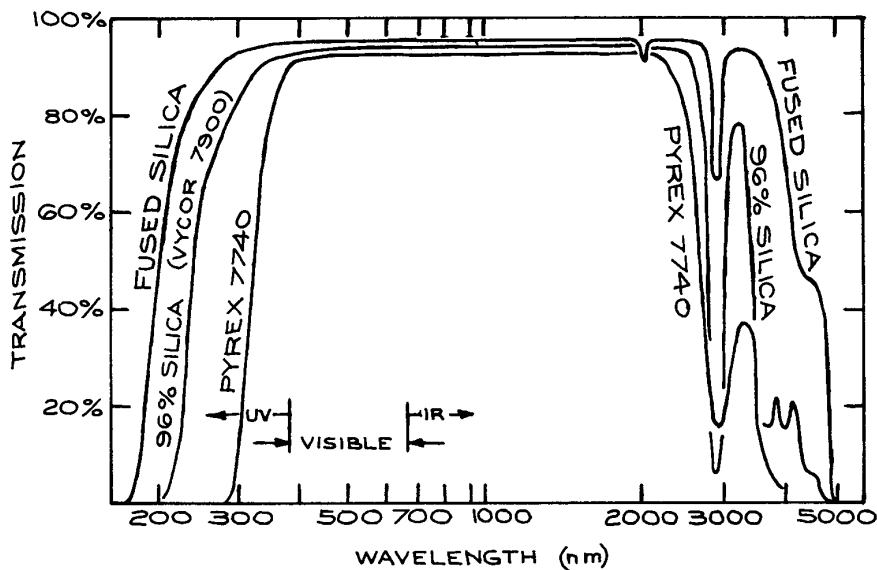
Glass fibers are very strong. A quartz fiber  $0.3 \text{ mm}$  ( $.01 \text{ in.}$ ) in diameter has a tensile strength of  $0.3 \text{ GN/m}^2$  ( $50\ 000 \text{ psi}$ ); a fiber  $0.03 \text{ mm}$  ( $.001 \text{ in.}$ ) in diameter a strength of at least  $1.3 \text{ GN/m}^2$  ( $200\ 000 \text{ psi}$ ); and a fiber  $0.003 \text{ mm}$  ( $.0001 \text{ in.}$ ) in diameter a strength in excess of  $7 \text{ GN/m}^2$  ( $1\ 000\ 000 \text{ psi}$ ).

## 2.2 LABORATORY COMPONENTS AVAILABLE IN GLASS

The list of laboratory apparatus components produced commercially in glass is nearly endless. The laboratory scientist should make a careful survey of lab suppliers' literature before embarking on the design and construction of a glass apparatus. Very often a complex device can be assembled in the lab entirely from inexpensive components without requiring the services of a skilled glassblower. Some of the most frequently used glass components are described below.

### 2.2.1 Tubing and Rod

Most laboratory supply houses carry a wide range of tubing and rod of Pyrex 7740 or Kimax KG-33 in  $1 \text{ m}$  lengths at a cost of only a few euros per kilogram. Outer diameters



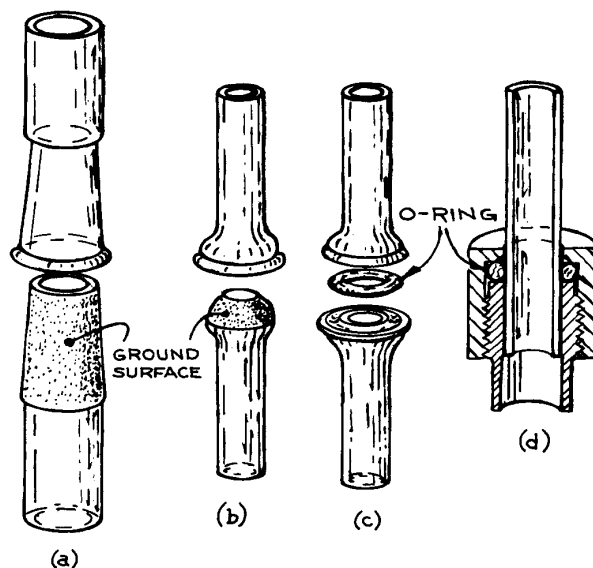
**Figure 2.1** Optical-transmission curves for Pyrex 7740, Vycor 7900 (96% silica), and fused silica.

between 3 and 178 mm are readily available. The standard wall thickness ranges from 0.5 mm for the smallest tubing to 3.5 mm for the largest. Heavy-wall tubing with wall thickness ranging from 2 to 10 mm is also available. In addition, heavy-wall tubing of very small bore diameter (0.5–4 mm), known as *capillary tubing*, is available. Finally, solid rod of 3 to 30 mm diameter is a standard item.

## 2.2.2 Demountable Joints

Laboratory apparatus can be quickly assembled if components are connected with joints of the type illustrated in Figure 2.2.

A gas- or liquid-tight seal between two close-fitting pieces of glass can be achieved if the mating surfaces are lightly coated with a viscous lubricant before assembly. A number of low-vapor-pressure lubricants such as Apiezon vacuum grease or Dow-Corning silicone vacuum grease are especially formulated for this purpose. The taper joint and the ball-and-socket joint in Figure 2.2(a) and (b) are assembled in this manner. The mating surfaces must fit together very well. In general this requires that they be



**Figure 2.2** Demountable joints: (a) standard-taper joint; (b) ball and socket; (c) O-ring joint; (d) a quick-connect. (a), (b), and (c) are used for attaching glass components to one another; (d) is used for joining a glass tube to a metal component.

lapped together. This is accomplished by coating the surfaces with a fine abrasive such as Carborundum in water, fitting the pieces together, and rotating one with respect to the other. The rotation should not be continuous, but rather after half a turn or so the pieces should be pulled apart and then assembled again so as to redistribute the abrasive. Fortunately, taper and ball-and-socket joints with ground mating surfaces are now commercially produced in standard sizes with sufficient precision that lapping and grinding in the lab is seldom necessary.

The standard taper (indicated  $\nabla$ ) for ground joints is 1:10. Standard-taper joints on tubing with outer diameters between 8 and 50 mm are available. These joints are identified by figures that indicate the diameter of the large end of the taper and the length of the ground zone in millimeters. For example,  $\nabla$  10/30 indicates a ground zone 10 mm in diameter at the large end and 30 mm in length. Standard-taper ground joints are available in borosilicate glass, quartz, and type-303 stainless steel. These components are interchangeable, and thus it is often convenient to make a transition from glass to quartz or from glass to metal with a taper joint.

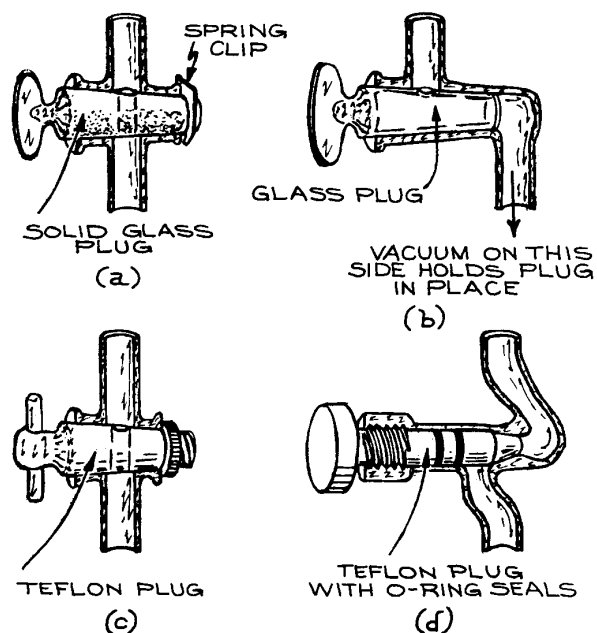
Ball-and-socket ground joints do not seal as reliably as taper joints; however, their design permits misalignment and even some slight motion between joined parts. This type of joint must be secured with a suitable clamp. Ball-and-socket joints are designated by a two-number code (i.e., 28/15). The first number gives the diameter of the ball in millimeters, the second number the inner diameter of the tubing to which the ball and socket are attached. Ball-and-socket joints of standard sizes are commercially available in borosilicate glass, quartz, and stainless steel.

The O-ring joint illustrated in Figure 2.2(c) is rapidly replacing the ground joint as a means of making demountable joints in glass vacuum apparatus. These joints require no grease. Furthermore, the mating pieces are sexless, since each member of the joint is grooved to a depth of less than half the thickness of the O-ring. To date, these joints are only available in borosilicate glass.

*Quick-connects* of the type shown in Figure 2.2(d) are also sealed with an O-ring. They are suited to joining either glass or metal tubing to a metal container. Quick-connects are often used to join glass ion-gauge tubes to metal vacuum apparatus.

### 2.2.3 Valves and Stopcocks

A glass stopcock of one of the types illustrated in Figures 2.3(a)–(c) has traditionally been used for controlling fluid flow in glass apparatus. These consist of a hollow tapered body and a plug with a hole bored through that can be aligned by rotation with inlet and outlet ports in the body. The simplest and least expensive version uses a glass plug with a ground surface that mates with a ground surface on the interior of the stopcock body. The surface of the plug must be lightly lubricated with stopcock grease to ensure a good seal and freedom of rotation. A stopcock of this type may be used to control liquid flow or gas flow at pressures down to a few millitorr. For use at pressures above 1 atm the end of the plug may be fitted with a spring-loaded clip or collar to prevent its being blown out. Stopcocks of this type are now available with a Teflon plug. These are well suited for use with liquids.



**Figure 2.3** Stopcocks and valves for controlling fluid flow: (a) glass stopcock with solid glass plug; (b) high-vacuum stopcock with vacuum cup and hollow plug; (c) glass stopcock with Teflon plug; (d) threaded glass vacuum valve with O-ring-sealed Teflon stem.

No lubrication is required, and the plug seldom freezes in the body of the stopcock.

In glass vacuum systems, the stopcock illustrated in Figure 2.3(b) is far more reliable than the simple solid plug design. The plug is hollow, and the small end of the body is closed off by a vacuum cup so that the interior of the stopcock can be evacuated. The plug is held securely in place by atmospheric pressure.

Valves with threaded borosilicate glass bodies and threaded Teflon plugs [Figure 2.3(d)] are rapidly replacing conventional stopcocks for most applications. These valves are no more expensive than good vacuum stopcocks. They require no lubrication; only glass and Teflon are exposed to the interior of the system. They are suitable for use with most liquids and gases and may be used at pressures from  $10^{-6}$  torr to 15 atm.

### 2.2.4 Graded Glass Seals and Glass-to-Metal Seals

In general, glasses of different composition have different coefficients of thermal expansion. As a consequence, two glasses of different composition usually cannot be joined, since differing rates of expansion and contraction produce destructive stresses as the fused joint cools. In practice two glasses can be successfully joined if their coefficients of thermal expansion differ by no more than about  $1 \times 10^{-6}$  K. The problem of joining two glasses with significantly different coefficients of thermal expansion is solved by interposing several layers of glass, each with a coefficient only slightly different from its neighbors. This stack of glass when fused together is called a *graded seal*. Tubing with graded seals joining different glasses is manufactured commercially.

Borosilicate-to-soft-glass seals are available in tube sizes from 7 to 20 mm. Seals graded from borosilicate glass (for example Pyrex 7740) to quartz (Vycor 7913) are available in tube sizes from 7 to 51 mm.

Glass tubing can, with care, be joined to metal tubing. Ideally the glass and the metal should have the same coefficient of expansion. It has been demonstrated, however, that glass tubing can be joined to metal tubing that has a very different coefficient provided that the end of the metal tubing has been machined to a thin, sharp, feathered edge.<sup>1</sup> The stresses of thermal expansion and contraction can then

be taken up by stretching of the metal. The stresses created by joining glass to metal can also be reduced by interposing glasses whose coefficient of thermal expansion is intermediate between that of the metal and the final glass.

Direct tube seals between borosilicate glass and copper or stainless steel are available in diameters from 6 mm to 50 mm. Graded glass seals between borosilicate glass and Kovar, an iron–nickel–cobalt alloy, are available in sizes from 3 mm to 75 mm. These graded seals are robust, and far and away the most common glass-to-metal seal. Kovar is easily silver-soldered or welded to most steels and brasses.

## 2.3 LABORATORY GLASSBLOWING SKILLS

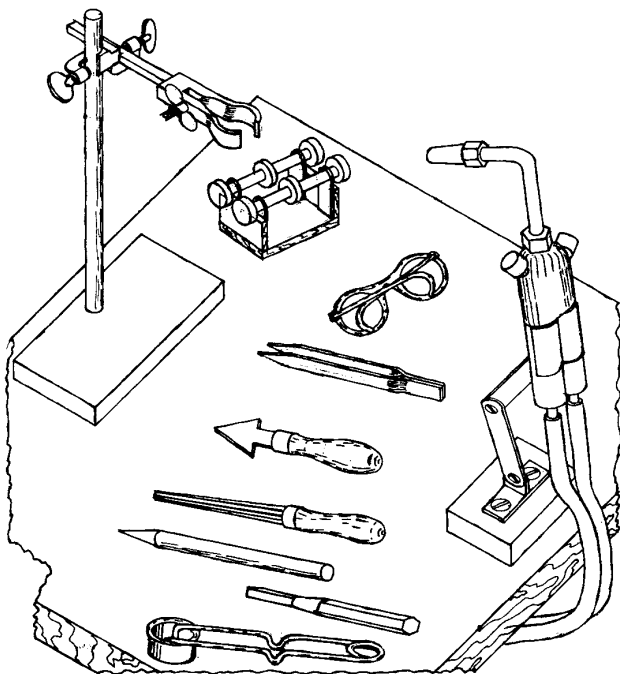
Considering the range of components now available, it is usually only necessary for the laboratory worker to be able to join commercially produced components to assemble a complete apparatus. The required skills consist primarily of joining tubes in series or at right angles, cutting and sealing-off tubing, and making simple bends. It is easiest and often necessary to make a joint between a piece of tubing held in the hand and a piece attached to an apparatus or otherwise rigidly clamped. In the following sections only the simplest glassblowing operations with borosilicate glass are described. There are certainly many more elegant and complex manipulations that can be carried out by a skilled glassblower using only hand tools. Workers who find a need for these operations or discover in themselves a flair for glassblowing may consult a more complete text on the subject.

NOTE: Many of the simplified methods described in Section 2.3 were developed by John Trembly of the University of Maryland for use in his remarkably successful introductory glassblowing classes.

### 2.3.1 The Glassblower's Tools

The necessary equipment for laboratory glasswork is illustrated in Figure 2.4. The basic tool is, of course, a torch. The National type-3A blowpipe with #2, #3, or #4 tip is the standard in the trade. A torch stand is essential to hold the lighted torch when both hands are required to manipulate a piece of glass. A source of natural gas and low-pressure





**Figure 2.4** Glassblower's equipment.

oxygen is required, along with two lengths of flexible rubber or Tygon tubing for connecting the gas supplies to the torch. A rigid ring stand and appropriate clamps with fiberglass-covered jaws are needed to support glass being worked upon. The torch and stand should be set up on a fireproof bench top. Also required are a number of small hand tools, inexpensively obtained from a scientific supply house. These include a torch lighter, a wax pencil, a glass-cutting knife, a swivel, a mouthpiece (an old tobacco-pipe bit will do), glassblowing tweezers, a collection of corks and one-hole stoppers, and a variety of reaming tools. These reaming tools consist of flat triangular pieces of brass fitted with wooden file handles, and tapered round or octagonal carbon rods in various sizes. Didymium eyeglasses are necessary to filter out the intense sodium D-line emission produced when glass is exposed to the flame. Without these it is impossible to see hot glass as it is worked. Clip-on didymium lenses are available for these who wear prescription eyeglasses.

A little preliminary practice with the hand torch is advisable. The gas inlet may be connected directly to a

low-pressure natural-gas outlet in the lab. The outlet valve is turned completely open and gas flow is controlled by the valve in the body of the torch. Oxygen is usually obtained from a high-pressure cylinder. A pressure regulator for the oxygen tank is essential. The outlet pressure should be between 40 and 70 kN/m<sup>2</sup> (6 and 10 psi). Once again, the flow of oxygen gas is controlled by the valve in the body of the hand torch. The proper procedure is to open the gas control and ignite the gas first. The gas should be adjusted to give a flame 5 to 8 cm long before the oxygen is introduced. If the flame blows out, the oxygen valve should be closed and the gas permitted to flow for a few moments to purge the torch of the explosive gas-oxygen mixture before reignition is attempted.

With a #3 torch tip it is possible to produce a range of flames suitable for work on tubing from a few millimeters up to 50 mm in diameter. Opening the gas valve wide and adding only a little oxygen gives a large, bushy yellow flame. This is a cool reducing flame suitable for preheating large pieces of glass and for annealing finished work. At the opposite extreme, a sharp blue oxidizing flame can be obtained by restricting the gas flow. For work on fine tubing and for cutting tubing it is possible to make a flame no more than an 3 cm long and 3 mm in diameter. The hottest flame is made with an excess of oxygen. The hottest part of the flame is the tip of the blue inner cone.

Some practice will be required to master even the simple operations described below. The neophyte should obtain the necessary tools and a supply of clean dry borosilicate glass tubing of 8–12 mm diameter. About ten hours of practice is required before one can profitably embark on any serious apparatus construction.

### 2.3.2 Cutting Glass Tubing

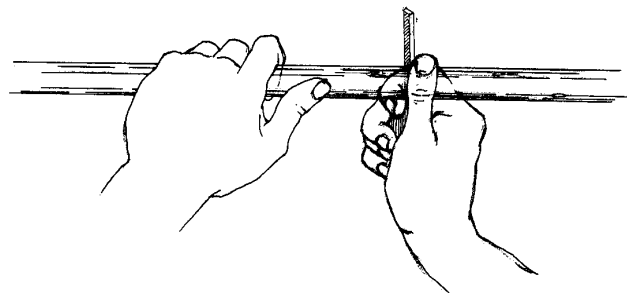
Tubing up to about 12 mm diameter can be broken cleanly by hand. The location of the desired break may be indicated by marking the tubing with a wax pencil. The glass is then scored with a glass knife or the edge of a triangular file, as shown in Figure 2.5. The scratch should be perpendicular to the axis of the tube and need only be a few millimeters in length. Use considerable pressure and make only one stroke. Do not saw at the glass. Wet the scratch and then, holding the tubing as shown in Figure 2.6 with the scratch toward you, push the ends apart. Applying a

force that tends to bend the ends of the tube away from you, while simultaneously pulling, will facilitate a clean break.

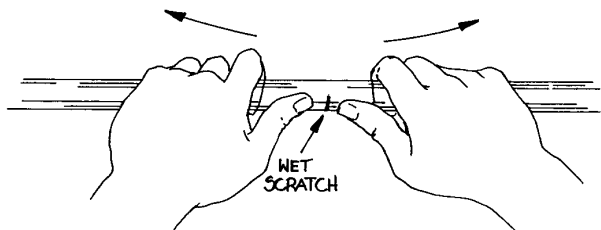
Tubing of 12 to 25 mm diameter can be cut by cracking it with a flame. First score the glass around approximately one third of its circumference with the glass knife. Then wet the scratch and touch the end of the scratch with a very fine sharp flame that is oriented tangentially to the circumference, as shown in Figure 2.7.

Large-diameter tubing can be parted by cracking it using a resistively heated wire held in a yoke, as illustrated in Figure 2.8. The tubing is placed on the resistance wire with one end butted against a stop. The wire is heated red-hot by passing current through it, and the tube is rotated to heat a narrow zone around its circumference. After a moment the tube is quickly removed and the hot zone is brushed with a wet pipe cleaner. The thermal shock thus produced should result in a clean break.

The edge of the glass at a fresh break is sharp and fragile. After cutting, the end of a tube can be *fire-polished*



**Figure 2.5** Scoring glass tubing in preparation for breaking. Use only one stroke with considerable force.

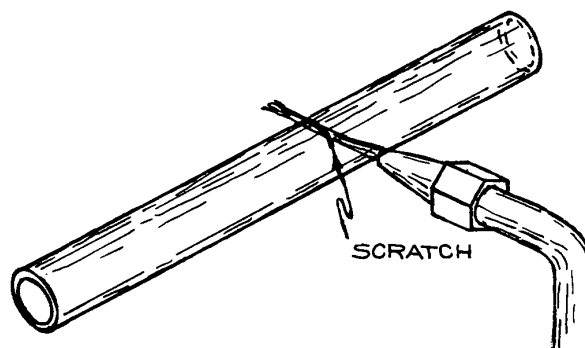


**Figure 2.6** Breaking glass tubing. The tubing is bent and simultaneously pulled apart.

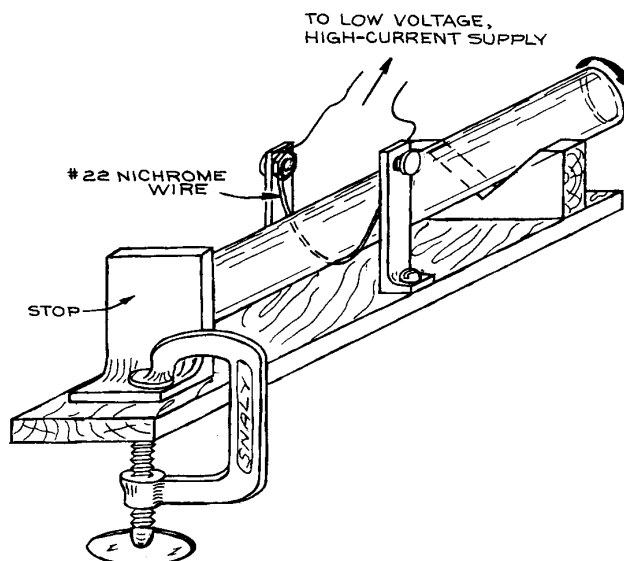
to relieve this hazardous condition. This is accomplished simply by heating the end of the tube in the flame until the glass is soft enough that surface tension rounds and thickens the sharp edge.

### 2.3.3 Pulling Points

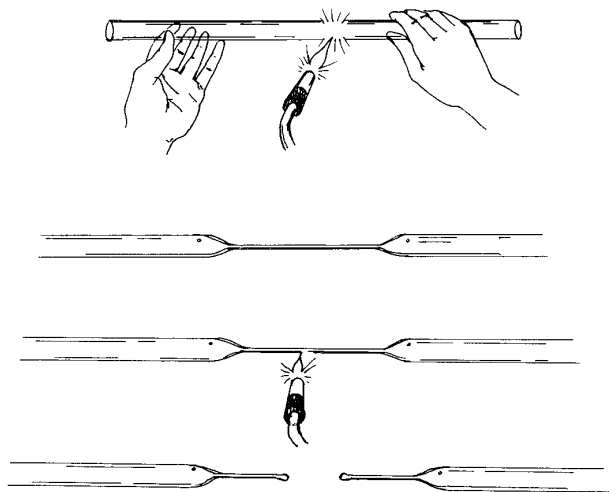
*Points* are elongations on the end of a tube produced by heating the glass and stretching it. These points serve as



**Figure 2.7** Cracking tubing with a flame.



**Figure 2.8** Device for cutting large-diameter glass tubing.



**Figure 2.9** Pulling points.

handles for manipulating short pieces of tubing. Pulling a point is the first step in closing the end of a tube.

The procedure for pulling points depends upon whether the tubing can be rotated or not. If the tubing is free, then place the torch in its holder with the flame pointed away from you and hold the glass in both hands as illustrated in Figure 2.9. Adjust the torch to give a fairly full, neutral flame. Then, rotating the glass at a uniform rate, pass it through the flame to heat a zone about as wide as the tube diameter. When the glass softens, remove it from the flame and pull the heated section to a length of about 20 cm. Then quickly heat the center of the stretched section and pull the glass in two, leaving a small bead at the end of each point. It is important that the points be on the axis of the original tube. This requires that the circumference of the tube be heated uniformly and that the ends of the tube be pulled straight away from one another. Some practice is required to synchronize the motion of the hands so that both ends of the tubing are rotated in such a way that the tube does not twist or bend when it becomes pliable in the flame. Beginners tend to overheat glass, thus exacerbating this problem of misalignment. When glass is sufficiently plastic for proper manipulation, it is still rigid enough to help support the end sections.

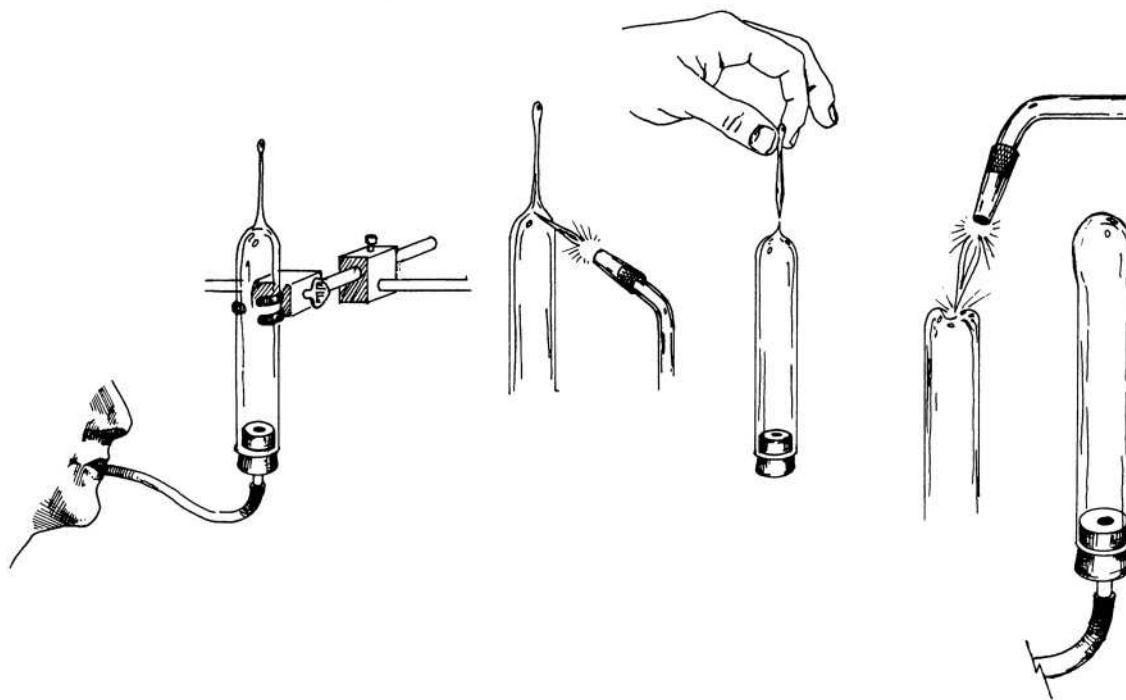
If one end of a tube is attached, so that the tube cannot be rotated, then it becomes necessary to swing the torch around the tubing rather than rotate the glass before the

stationary torch. Some practice is required to achieve a smooth motion with the torch so that the tube is heated uniformly around its circumference. The unattached end of the tubing is supported with the free hand until the glass is sufficiently hot to be pulled.

### 2.3.4 Sealing Off a Tube: The Test-Tube End

Tubing is closed off by forming a hemispherical bubble similar to that found at the end of a test tube. In order for the *test-tube end* to form a strong closure, it must be smooth and of a uniform thickness. The hemispherical shape is formed by closing off the tubing and then softening the glass in the flame and blowing into the tube.

Begin by pulling a point on the tubing that is to be closed off. If the tubing is not rigidly mounted, place it in a clamp secured to a ring stand on the workbench. If possible the work should be at chest height with the point upward. Press a one-hole stopper, with a piece of soft rubber tubing attached, into the open end of the tube as shown in Figure 2.10(a). Warm the glass at the base of the point with a soft bushy flame. Adjust the gas–oxygen mixture to the torch to give a fairly small flame with a well-defined, blue inner cone. Heat the point close to the shoulder where it joins the tube. Swing the torch around the work so that the glass is heated uniformly. When the glass melts, pull the point off the tube. Only a small bead of glass should be left behind. A large blob of glass will result in a closure that is too thick in relation to the wall of the tubing. A test-tube end that is thick in the middle and thin at the sides will develop destructive strains, since it does not cool at a uniform rate. The key is to heat only a narrow zone of the point near the shoulder and to perform the whole operation rather quickly. If a large globule of glass is left after pulling off the point, the excess must be removed. With a sharp flame, heat the globule until it melts. Remove the flame, touch the molten bead with the end of a piece of glass rod (“cane”), and quickly pull the rod away. Excess glass will be pulled away from the soft bead into a fiber which can be broken or melted off. The test-tube end is completed by heating the closed-off end as far as the shoulder with a somewhat larger flame until it is soft and the glass has flowed into a fairly uniform thickness. The flame should be directed downward at the end of the tube with a circular motion. Remove the flame and gently blow the end into a hemispherical shape.



**Figure 2.10** Blowing a test-tube end.

If a flat end is desired, reheat the round end and gently blow while pressing a flat carbon block against the end.

### 2.3.5 Making a T-Seal

A tube is joined at a right angle to the side of another tube with a *T-seal*. To make a T-seal, close one end of a tube with a cork and attach the blow tube to the other. If possible, mount the tube horizontally at chest height. With a sharp flame heat a spot on the top side of the tube. Using a circular motion of the torch, soften an area about the size of the cross-section of the tube that is to be joined. Then move the flame away and gently blow out a hemispherical bulge, as shown in Figure 2.11(b). Reheat the bulge and blow out a thin-walled bubble. Finally, while blowing to pressurize the inside of the tube, touch the bubble with the tip of the flame so that a hole is blown through it. Now heat the edge of this hole so that surface tension pulls the glass back to the edges of the original bubble. The result should

be a round opening slightly smaller than the diameter of the tube that is to be joined, as shown in Figure 2.11(f). If the opening is too small or if the edge is irregular, the hole can be reamed and shaped by softening the glass and shaping the hole with a tapered carbon. A tool for this job can be made by sharpening a 1/4 in. diameter carbon rod in a pencil sharpener.

With a cork or a point, close off the end of the tube that is to be sealed to the horizontal tube. Then, holding this tube above the hole and tipped slightly back, simultaneously heat the edges of the tube and the hole until they become soft and tacky. Bring the tube into contact with the back edge of the hole, remove the flame, and tip the tube forward until it rests squarely on the hole. The tacky edges should join together to produce an airtight seal. Blow to check the seal. Small holes can be closed by heating the seal all around until the glass is soft, and tipping the vertical tube in the direction of the leak. It also is possible to “stitch” a gap closed with cane. This is done by pulling a

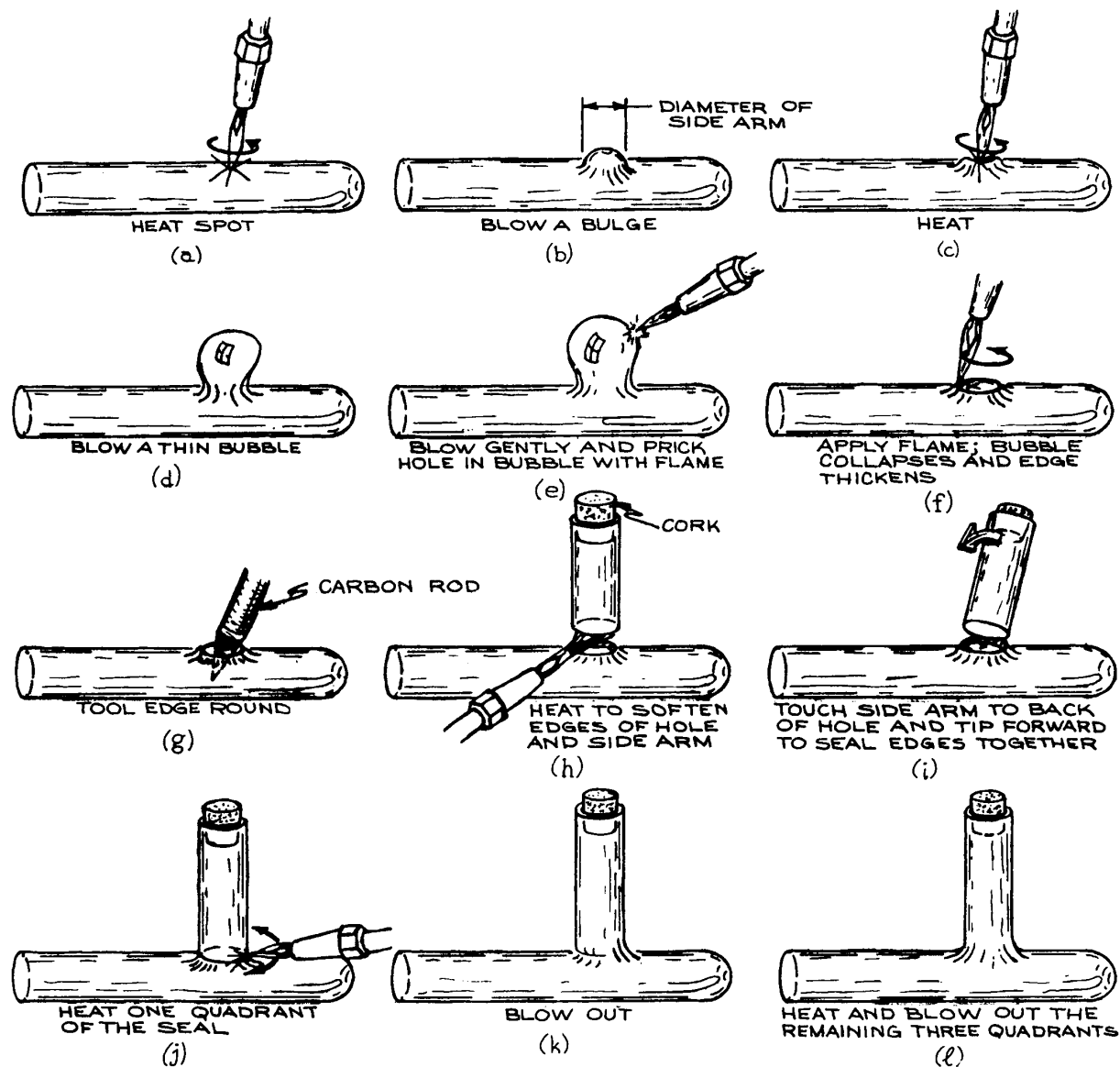
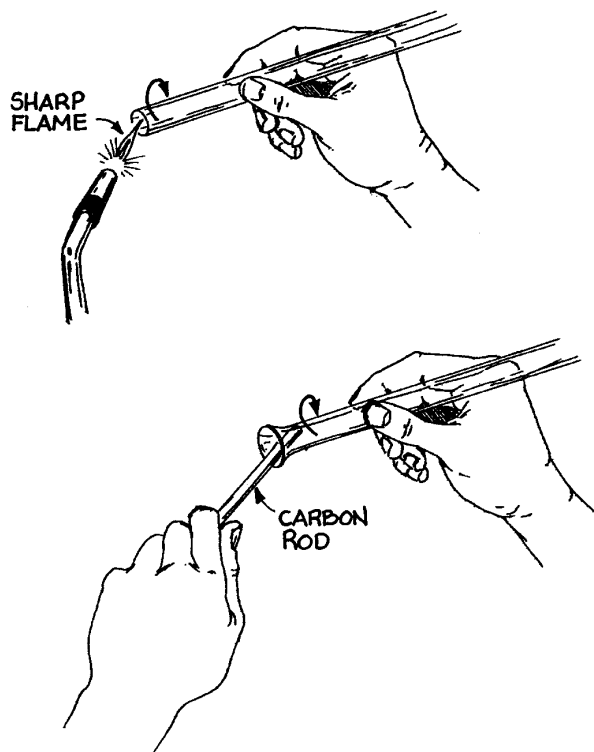


Figure 2.11 Steps in making a T-seal.

point on a piece of glass rod. Then warm the gap which is to be repaired, heat the point of the cane until it is molten and deposit the tiny blob of molten glass thus produced into the gap. The seal at this point is quite fragile because the glass is too thick around the seal. Cracks will develop if

the glass is permitted to cool before the seal is blown out in order to reduce the wall thickness.

Glassblowing around the seal proceeds in four steps. Begin with a sharp flame parallel to the axis of the horizontal tube. Swing the torch back and forth to heat one quadrant



**Figure 2.12** Flaring the end of a tube.

of the seal. When the glass softens, remove the flame and blow gently. Timing is important here. Blowing while the glass is in the flame will cause the thinnest parts to bulge. If, however, one waits a moment after removing the flame, the thinnest parts will cool and blowing will preferentially thin out the thick portions of the seal, yielding a more uniform wall thickness. The joint between the vertical and horizontal tubes should then be a smooth curve of glass. During the blowing operation the vertical tube must be held with the free hand to prevent its tipping or sagging into the horizontal tube. Next repeat the heating and blowing operation on the side diametrically opposite. Finally, in a third and fourth step, blow out the remaining two quadrants.

### 2.3.6 Making a Straight Seal

A *straight seal* joins two tubes coaxially. There are two simple procedures for making such a seal.

If the tubes are quite different in diameter, close off the larger one with a test-tube end and mount it vertically with the closed end up. Then blow a bubble of the appropriate size in the test-tube end and proceed as in making a T-seal.

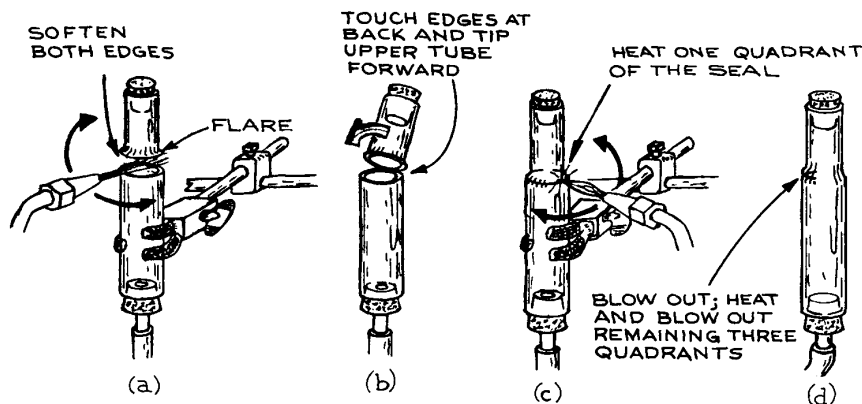
Alternatively, the smaller tube can be flared out to the diameter of the larger one. The flaring operation is illustrated in Figure 2.12. With the torch in its stand, heat the edge of the tube by rotating it in a sharp flame. Remove it from the flame, and while continuing to rotate the tube, insert a tapered carbon and gently bring the soft edge of the glass up against the taper. Do not attempt to produce a flare of the desired diameter in one rotation of the tube. It will not come out round. The flaring operation is much easier if the tube is placed on a set of glassblower's rollers.

As shown in Figure 2.13(a), the larger tube should be mounted vertically if possible. Suspend the smaller tube with the flare just above the larger one. With the upper tube tipped back slightly, heat the edges of the larger tube and the rim of the flare until the glass is tacky. Bring them into contact at the back of the seal, remove the flame, and tip the upper piece forward to join the two pieces. Proceed to blow out the seal in four steps as for the T-seal. Minimize the distance above and below the seal over which the glass is heated.

### 2.3.7 Making a Ring Seal

A *ring seal* is used for sealing a tube inside a larger tube, as in Figure 2.14(f), or for passing a tube through the wall of a container. There are many procedures for making such seals. The simplest involves bulging the smaller-diameter tube sufficiently for it to close off the hole through which the tube must pass. A bulge in a tube is known as a *maria*. Making a maria is a freehand operation and requires some practice.

Begin with the torch in its holder and the gas supply adjusted to give a long, narrow flame. Then rotating the tubing in both hands, heat a very narrow zone at the desired location. As soon as the glass softens, drop the tube out of the flame, and, with your elbows braced on the bench top or against your sides, push the ends of the tube toward one another. If the tube has been heated uniformly around and the ends are held coaxial, a bulge should develop around the tube. If the heating is not uniform, the tube will bend. If the glass is too soft, the ends



**Figure 2.13** Steps in making a straight seal.

of the tube are liable to move out of alignment and an eccentric bulge will result. When a uniform bulge is attained, reheat it, remove it from the flame, and push again to increase the diameter of the maria. The bulge should be solid glass. If too wide a zone is heated, the maria will be hollow and dangerously fragile.

To make a coaxial ring seal between two tubes, form on the smaller tube a maria of the same diameter as the larger. Mount the larger tube vertically, fire-polish the top edge, and rest the smaller tube on top, as in Figure 2.14(e). With a small flame, heat the junction uniformly between the maria and the edge of the larger tube. The weight of the smaller tube should cause it to settle and become fused to the larger one. Finally blow out the seal in four steps, as in making a conventional seal. Exercise care to prevent the wall of the outer tube from sagging into the inner tube. If the inner tube is not centered, it can be realigned by softening the seal with a cool, bushy flame and manipulating its protruding end.

### 2.3.8 Bending Glass Tubing

Bends are best made beginning with the tubing vertical [Figure 2.15(a)]. Close off the upper end with a cork or a point and attach the blowtube to the other end. While supporting the upper end of the tube with the free hand, apply a large bushy flame at the location of the desired bend. Heat the glass uniformly over a length equal to about

four times the diameter of the tube. When the glass becomes pliable, remove the flame and quickly bend the upper end over to the desired angle. As soon as the bend is completed, blow to remove any wrinkles which may have developed and to restore the tubing to its original diameter. Some reheating may be necessary.

A sharp right-angle bend can be produced by first making a T and then pulling off one running end of the T close to the seal [Figure 2.15(b)]. Blow out a test-tube end where the tubing is pulled off.

### 2.3.9 Annealing

Even when a seal is carefully blown out to give a uniform wall thickness, harmful stresses are inevitably frozen into the glass. Worked glass must be annealed by heating to a temperature above the annealing point (see Table 2.1) and then slowly cooling to the strain point. In a large professional glass shop, special ovens are employed for this operation. In the lab, however, glass must be annealed in the flame of the torch. After a piece of glass has been blown, it is heated uniformly over a wide area surrounding the work with a large, slightly yellow, bushy flame. For Pyrex 7740 the glass should be brought to an incipient red heat. Care must be taken to prevent the glass from becoming so hot that it begins to sag under its own weight. The heated work is then cooled by slowly reducing the oxygen to the torch until the flame becomes sooty.

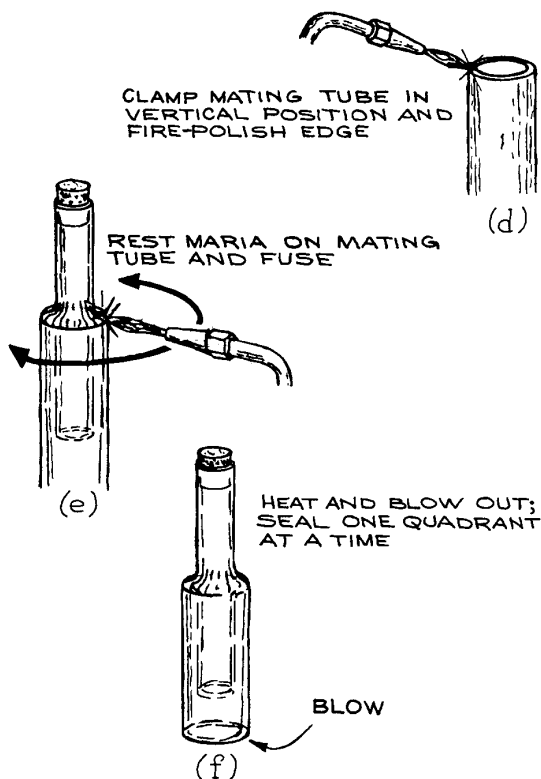
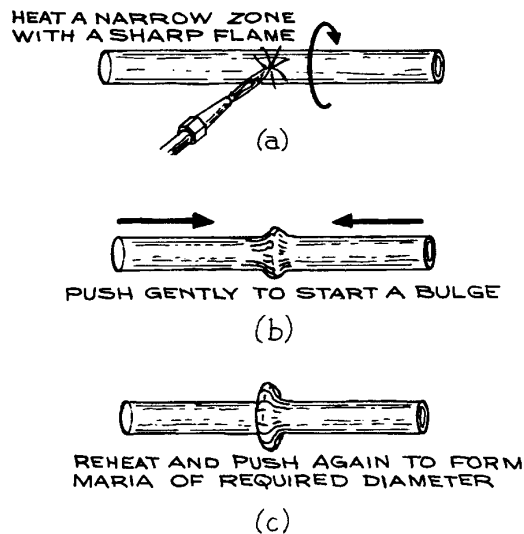


Figure 2.14 Steps in making a ring seal.

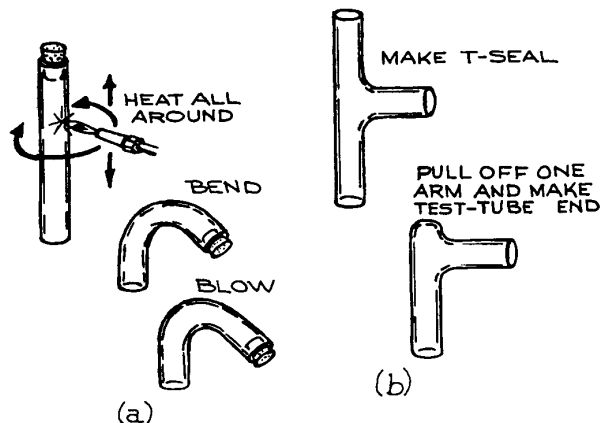


Figure 2.15 Bending tubing.

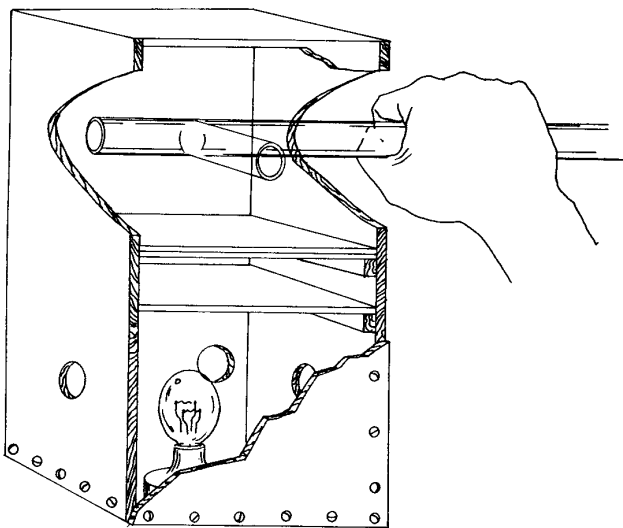
When soot begins to adhere to the glass, the flame can be removed.

Areas in glass which are under stress will rotate the plane of polarization of transmitted light. A *polariscope* can be used to observe strain in glass in order to test the effectiveness of the annealing operation. A polariscope can be purchased from a laboratory supplier or can be constructed with high-extinction sheet polarizers, as shown in Figure 2.16. The polarizers are crossed so that unstrained glass appears dark when placed between them. Areas of strain rotate the plane of polarization and appear bright, as in Figure 2.17. A large stationary piece of glasswork, such as a vacuum manifold, can be tested by illuminating it from behind with vertically polarized light and then viewing the work through Polaroid sunglasses (which transmit horizontally polarized light). A suitable source can be constructed by placing a 100 watt lamp behind a diffuser in a ventilated, open-top box, and covering the box with a sheet of polarizer sandwiched between glass plates.

### 2.3.10 Sealing Glass to Metal

Glass-to-metal seals are fragile and difficult to produce in the lab. The simplest means of attaching metal tubing to a glass system or of passing a metallic electrode through a glass wall is to make use of a commercially produced glass-to-metal seal [Figure 3.28(b)]. If necessary a direct

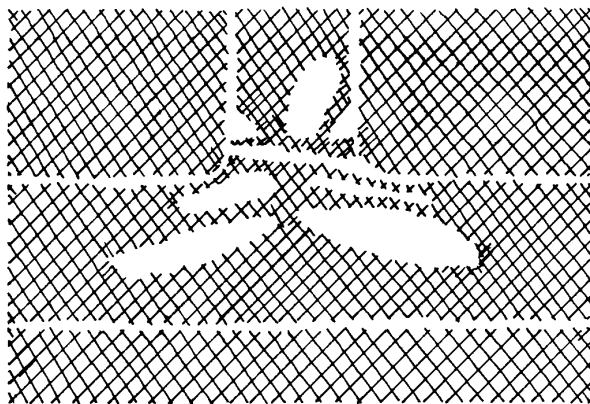




**Figure 2.16** A polariscope.

seal between a variety of metals and glasses can be produced with careful attention to design and fabrication.<sup>2</sup> Of these, the tungsten-to-Pyrex seal is most easily and reliably fabricated in the lab.

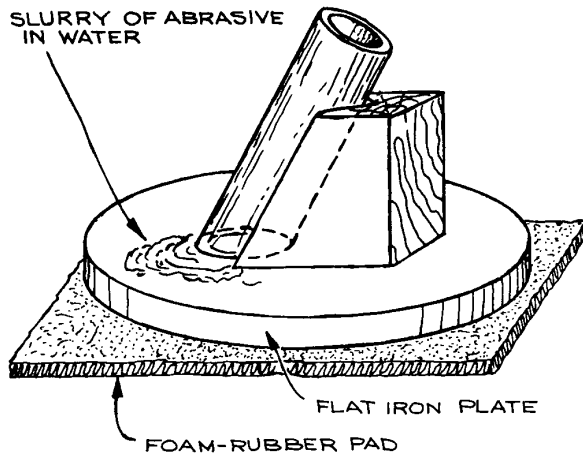
Although the linear coefficient of expansion of tungsten is about 50% greater than that of 7740 borosilicate glass, it is possible to make a strong, vacuum-tight seal



**Figure 2.17** The appearance of strain in unannealed glass as seen in the polariscope.

between Pyrex and tungsten wire or rod of up to a millimeter diameter.<sup>3</sup> Since the glass will wet tungsten oxide, but not tungsten, it is necessary that the metal surface be properly prepared. Tungsten wire or rod usually has a longitudinal grain by virtue of the method of manufacture. The metal stock should be cut to length by grinding to prevent splintering. To prevent gas from leaking through the seal along the length of the wire it is good practice to fuse the ends of the wire in an electric arc. Alternatively, a copper or nickel wire can be welded to each end of the tungsten wire. Tungsten electrodes with copper or nickel wires attached are also available from commercial sources. Holding the wire in a pin vise, heat the area to be sealed to a dull red and immediately rub the hot metal with a stick of sodium nitrite. The salt combines with the tungsten oxide to leave a clean metal surface. Rinse the metal with tap water to remove the salt and then rinse with distilled water. The metal must then be reoxidized in an oxygen-rich flame to produce an oxide layer of the proper thickness. If the layer is too thick it will pull away from the metal; if it is too thin it may all dissolve into the glass. The clean tungsten should be gently heated only until the metal appears blue-green when cool. A piece of 7740 glass, 2 cm long with a wall thickness of about 1 mm and an inner diameter slightly larger than the tungsten wire, is slipped over the wire and slid along to the sealing position. Be careful not to scratch the oxide layer. The glass tube is fused to the metal using a small sharp flame. Melting should proceed from one end to the other with continuous rotation to prevent trapping air. If the metal is properly wetted by the glass, the wire will appear oversize where it passes through the seal. A copper or amber color after the seal is cool indicates a good seal. The beaded tungsten rod is then sealed into the apparatus, as in making a ring seal (Section 2.3.7). If necessary, a shoulder can be built up around the middle of the bead by fusing on a winding of 1 mm glass rod.

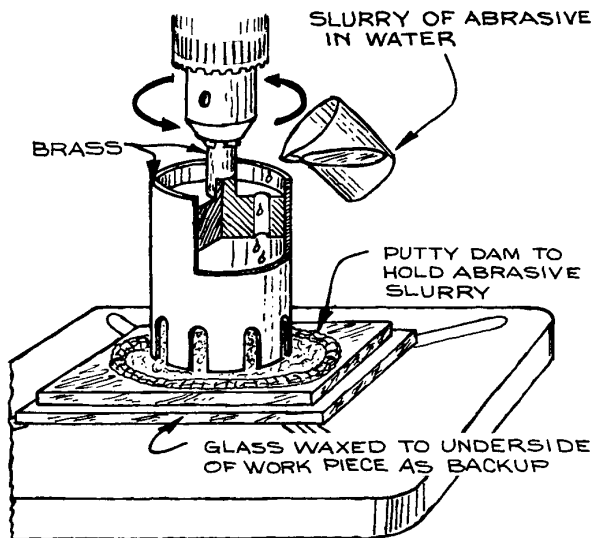
A tungsten-to-Pyrex seal makes an excellent electrical feedthrough. Wires can be brazed or spot-welded to either end of the tungsten rod. Since tungsten is not a good electrical conductor, care must be taken to ensure that electrical currents do not overheat the feedthrough. The safe current-carrying capacity of the tungsten in this application is about 8 A/mm<sup>2</sup>.



**Figure 2.18** A wooden jig used in grinding the face of a piece of tubing.

### 2.3.11 Grinding and Drilling Glass

In many instances the ends of a tube must be ground square or at an angle, or a tube must be ground to a specific length. The ends of a laser tube, for example, must be cut and ground to Brewster's angle before optical windows are



**Figure 2.19** Drilling glass with a tube drill.

glued in place. Grinding a plane surface is performed in a water slurry of abrasive on a smooth iron plate. It is done with a circular motion while applying moderate pressure. The work should be lifted from the surface frequently in order to redistribute the abrasive. If the angle of the ground surface relative to the workpiece is not critical, the work can be grasped in the hand; otherwise, a wooden jig, such as is shown in Figure 2.18, may be fabricated to orient the work with respect to the grinding surface.

Carborundum or silicon carbide grit of #30, 60, 120, and 220 mesh is used for coarse grinding, and emery of #220, 400, and 600 mesh is used for fine grinding. In any grinding operation, begin with the grit required to remove the largest irregularities and proceed in sequence through the finer grits. Before changing from one grit to the next finer, the work and the grinding surface must be scrupulously washed so that coarse abrasive does not become mixed with the next finer grade. Before changing from one grit to the next, it is useful to inspect the ground surface under a magnifying glass. If the abrasive in use has completed its job, the scratch marks left by the abrasive should be of a fairly uniform width.

Holes can be cut in glass, and round pieces of glass can be cut from larger pieces, with a slotted brass tube drill – a “cookie cutter” – in a drill press (Figure 2.19). These drills are easily fabricated in sizes from several millimeters to many centimeters in diameter. Either flat or curved glass can be drilled. Cutting is accomplished by continuously feeding the drill with a slurry of #60 or #120 mesh abrasive. It is usually most convenient to form a dam of soft wax or putty around the hole location to hold the slurry of abrasive around the drill. Work slowly with a light touch at a drill speed of 50–150 rpm. The tool should be backed out frequently to permit the abrasive to flow under. Because grinding action occurs along both the leading edge and the side of the drill tube, the finished hole will be somewhat larger than the drill body. Frequently the edges of the hole will chip where the drill breaks through on the under side of the work. This can be prevented when drilling flat glass by backing up the workpiece with a second piece of glass waxed onto the underside.

Glass can be sawn as well as drilled. Many glass shops use a powered, diamond-grit cutting wheel for precision cutting. In the absence of such a machine, it is possible to cut glass in much the same manner as it is drilled. A hacksaw with a strip of 20-gauge copper as a blade is used. The

work is clamped in a miter box and sawn with a light touch while feeding an abrasive slurry to the blade.

## Cited References

1. W. G. Housekeeper, *Trans, A.I.E.E.*, **42**, 870, 1923.
2. F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, Addison-Wesley, Reading, MA, 1965, pp. 54–66.
3. J. Strong, *Procedures in Experimental Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1938, p. 24; G. W. Green, *The Design and Construction of Small Vacuum Systems*, Chapman and Hall, London, 1968, pp. 100–102.

## General References

- W. E. Barr and J. V. Anhorn, *Scientific and Industrial Glass Blowing and Laboratory Techniques*, Instruments Publishing, Pittsburgh, 1959.
- Corning Glass Works, *Laboratory Glass Blowing with Corning's Glasses*, Publ. No. B-72, Corning Glass Works, Corning, NY.
- J. E. Hammesfahr and C. L. Strong, *Creative Glass Blowing*, Freeman, San Francisco, 1968.
- W. Morey, *The Properties of Glass*, Reinhold, New York, 1954.
- F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, Addison-Wesley, Reading, MA, 1965.

## VACUUM TECHNOLOGY

In the modern laboratory, there are many occasions when a gas-filled container must be emptied. Evacuation may simply be the first step in creating a new gaseous environment. In a distillation process, there may be a continuing requirement to remove gas as it evolves. Often it is necessary to evacuate a container to prevent air from contaminating a clean surface or interfering with a chemical reaction. Beams of atomic particles must be handled *in vacuo* to prevent loss of momentum through collisions with air molecules. Many forms of radiation are absorbed by air and thus can propagate over large distances only in a vacuum. A vacuum system is an essential part of laboratory instruments such as the mass spectrometer and the electron microscope. Far infrared and far ultraviolet spectrometers are operated within vacuum containers. Simple vacuum systems are used for vacuum dehydration and freeze-drying. Nuclear particle accelerators and thermonuclear devices require huge, sophisticated vacuum systems. Many modern industrial processes, most notably semiconductor device fabrication, require carefully controlled vacuum environments.

### 3.1 GASES

The pressure and composition of residual gases in a vacuum system vary considerably with its design and history. For some applications a residual gas density of tens of billions of molecules per cubic centimeter is tolerable. In other cases no more than a few hundred thousand molecules per cubic centimeter constitutes an acceptable vacuum: “One man’s vacuum is another man’s sewer”.<sup>1</sup> It is necessary to understand the nature of a vacuum and of

vacuum apparatus to know what can and cannot be done, to understand what is possible within economic constraints, and to choose components that are compatible with one another as well as with one’s needs.

#### 3.1.1 The Nature of the Residual Gases in a Vacuum System

The pressure below one atmosphere is loosely divided into vacuum categories. The pressure ranges and number densities corresponding to these categories are listed in Table 3.1. Vacuum science has been slow in adopting the SI unit of pressure – the Pascal or Pa – preferring instead torr or bar. We will follow this preference. For the purposes of the discussion here, a pressure of 1 torr, 1 millibar, and 100 Pa can be taken as equivalent. As a point of reference for vacuum work, it is useful to remember that the number density at 1 mtorr is about  $3.5 \times 10^{13}/\text{cm}^3$  and that the number density is proportional to the pressure.

The composition of gas in a vacuum system is modified as the system is evacuated because the efficiency of a vacuum pump is different for different gases. At low pressures molecules desorbed from the walls make up the residual gas. Initially, the bulk of the gas leaving the walls is water vapor and carbon dioxide; at very low pressures, in a container that has been baked, it is hydrogen.

#### 3.1.2 Gas Kinetic Theory

In order to understand mass flow and heat flow in a vacuum system it is necessary to appreciate the immense change in

Table 3.1 Air at 20 °C

	Pressure (torr) <sup>a</sup>	Number Density (/cm <sup>3</sup> )	Mean Free Path (cm)	Surface Collision Frequency (/cm <sup>2</sup> /s <sup>1</sup> )	Time for Monolayer Formation <sup>b</sup> (s)
One atmosphere	760	$2.7 \times 10^{19}$	$7 \times 10^{-6}$	$3 \times 10^{23}$	$3.3 \times 10^{-9}$
Lower limit of:					
Rough Vacuum	$10^{-3}$	$3.5 \times 10^{13}$	5	$4 \times 10^{17}$	$2.5 \times 10^{-3}$
High Vacuum	$10^{-6}$	$3.5 \times 10^{10}$	$5 \times 10^3$	$4 \times 10^{14}$	2.5
Very High Vacuum	$10^{-9}$	$3.5 \times 10^7$	$5 \times 10^6$	$4 \times 10^{11}$	$2.5 \times 10^3$
Ultrahigh Vacuum	$10^{-12}$	$3.5 \times 10^4$	$5 \times 10^9$	$4 \times 10^8$	$2.5 \times 10^6$

<sup>a</sup> 1 torr = 1.33 mbarr = 133 Pa

<sup>b</sup> assumes unit sticking coefficient and a molecular diameter of  $3 \times 10^{-8}$  cm

freedom of movement experienced by a gas molecule as the pressure decreases.

The average velocity of a molecule can be deduced from the Maxwell – Boltzmann velocity distribution law:

$$\bar{v} = \left( \frac{8kT}{\pi m} \right)^{1/2} \quad (3.1)$$

For an air molecule (molecular weight of about 30 amu) at 20 °C:

$$\bar{v} \approx 5 \times 10^4 \text{ cm/s} = \frac{1}{2} \text{ km/s} \quad (3.2)$$

Each second a molecule sweeps out a volume with a diameter twice that of the molecule and a length equal to the distance traveled by the molecule in a second. As shown in Figure 3.1, this molecule collides with any of its neighbors whose center lies within the swept volume. The number of collisions per second is equal to the number of neighbors within the swept volume. On average, the number of collisions per second ( $Z$ ) is the number density of molecules ( $n$ ) times the volume swept by a molecule of velocity  $\bar{v}$  and diameter  $\xi$ . More precisely:

$$Z = \sqrt{2}n\pi\xi^2\bar{v} \quad (3.3)$$

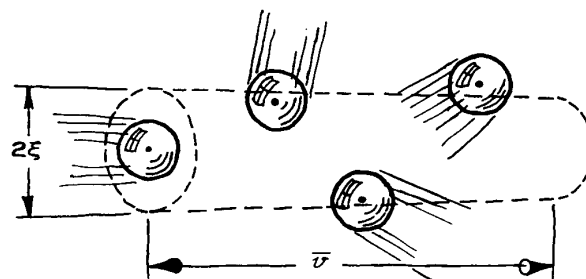
where the  $\sqrt{2}$  accounts for the relative motion of the molecules. The time between collisions is the reciprocal of this

collision frequency, and the average distance between collisions, or *mean free path*, is:

$$\begin{aligned} \lambda &= \bar{v}Z^{-1} \\ &= \frac{1}{\sqrt{2}n\pi\xi^2} \end{aligned} \quad (3.4)$$

Note that the mean free path is independent of the temperature and the molecular weight.

The mean free path is inversely proportional to the pressure. For an N<sub>2</sub> or O<sub>2</sub> molecule,  $\xi \approx 3 \times 10^{-8}$  cm. The number density at 1 mtorr is about  $3.5 \times 10^{13}$ /cm<sup>3</sup>. Thus the mean free path in air at 1 mtorr is about 5 cm. Recalling



**Figure 3.1** The volume swept in one second by a molecule of diameter  $\xi$  and velocity (*"vee bar ital"*) (cm/s). The molecule will collide with any of its neighbors whose center lies within the swept volume.

the relationship between  $\lambda$  and  $P$ , a valuable rule of thumb is that for air at 20 °C:

$$\lambda \approx \frac{5}{P(\text{mtorr})} \text{ cm} \quad (3.5)$$

### 3.1.3 Surface Collisions

The frequency of collisions of gas molecules with a surface is:

$$Z_{\text{surface}} = \frac{n\bar{v}}{4} \text{ (/s/cm}^2\text{)} \quad (3.6)$$

The sticking probability for most air molecules on a clean surface at room temperature is between 0.1 and 1.0. For water the sticking probability is about unity for most surfaces. Assuming unit sticking probability and a molecular diameter  $\xi \approx 3 \times 10^{-8}$  cm, the time required to form a monolayer of adsorbed air molecules at 20 °C is:

$$t = \frac{2.5 \times 10^{-6}}{P(\text{torr})} \text{ s} \quad (3.6)$$

Thus to maintain a clean surface for a useful period of time may require a gas pressure over the surface less than  $10^{-9}$  torr.

### 3.1.4 Bulk Behavior versus Molecular Behavior

The pressure of gas within a vacuum system may vary over 10 or more orders of magnitude as the system is evacuated from atmospheric pressure to the lowest attainable pressure. At high pressures, when the mean free path is much smaller than the dimensions of the vacuum container, gas behavior is dominated by intermolecular interactions. These interactions, resulting in viscous forces, ensure good communication among all regions of the gas. At high pressures, a gas behaves as a homogeneous fluid. When the gas density is decreased to the extent that the mean free path is much larger than the container, molecules rattle around like the balls on a billiard table (not a pool table, where the ball density can be high), and gas behavior is determined by the random motion of the molecules as they bounce from wall to wall. Gas flow in the high-pressure

regime is characterized as *viscous flow* and in the low-pressure regime as *molecular flow*. In the typical laboratory vacuum system the transition between the two flow regimes occurs at pressures in the vicinity of 100 mtorr.

In the viscous-flow region where the mean free path is relatively small, gas flow improves with increasing pressure because gas molecules tend to queue up and push on their neighbors in front of them. Gas flow is impeded by turbulence and by viscous drag at the walls of the pipe that is conducting the gas. In the viscous region the coefficients of viscosity and thermal conductivity are independent of pressure.

In the molecular flow region, momentum transfer occurs between molecules and the wall of a container, but molecules seldom encounter one another. In the molecular flow regime, a gas is not characterized by a viscosity. Gas flows from a region of high pressure to one of low pressure simply because the number of molecules leaving a unit of volume is proportional to the number of molecules within that volume. Gas flow is a statistical process. At very low pressure a molecule does not linger at a surface for a sufficient time to reach thermal equilibrium. Thus thermal conductivity at low pressures is a function of gas density, and the coefficient of thermal conductivity depends upon the pressure and upon the condition of the surface.

## 3.2 GAS FLOW

### 3.2.1 Parameters for Specifying Gas Flow

Before discussing vacuum apparatus it is necessary to define the parameters used by vacuum engineers to characterize gas flow.

The volume rate of flow through an aperture or across the cross section of a tube is defined as the *pumping speed* at that point:

$$S \equiv \frac{dV}{dt} \quad [\text{L/s; m}^3\text{/s; ft}^3\text{/min(cfm)}] \quad (3.7)$$

The capacity of a vacuum pump is specified by the speed measured at its inlet:

$$S_p \equiv \frac{dV}{dt} \quad (\text{at pump inlet}) \quad (3.8)$$

The mass rate of flow through a vacuum system is proportional to the *throughput*:

$$Q \equiv PS \quad [(\text{torr L/s; Pa m}^3/\text{s; std cm}^3/\text{s} \\ \text{(a standard cubic centimeter} \\ \text{or std cm}^3 \text{ or scc is 1 cm}^3 \text{ at} \\ \text{273 K and 1 atm or 1.01 bar)}] \quad (3.9)$$

To determine the throughput it is necessary that the pressure and speed be measured at the same place, since these quantities vary throughout the system.

The ability of a tube to transmit gas is characterized by its *conductance*  $C$ . The definition of conductance is analogous to Ohm's law for electrical circuitry. The throughput of a tube depends upon the conductance of the tube and the driving force, which in this case is the pressure drop across the tube:

$$Q = (P_1 - P_2)C \quad (P_1 > P_2) \quad (3.10)$$

Notice that conductance has the same units as pumping speed.

### 3.2.2 Network Equations

A complicated network of tubes can be reduced to a single equivalent conductance for the purpose of analysis. In analogy to electrical theory once again, a number of tubes in series can be replaced by a single equivalent tube with conductance  $C_{\text{series}}$  given by:

$$C_{\text{series}} = \left[ \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \dots \right]^{-1} \quad (3.11)$$

A parallel network can be replaced by an equivalent conductor with conductance:

$$C_{\text{parallel}} = C_1 + C_2 + C_3 + \dots \quad (3.12)$$

### 3.2.3 The Master Equation

By use of the network equations given above, an entire vacuum system can be reduced to a single equivalent conductance leading from a gas source to a pump, as shown in Figure 3.2. The net speed of the system is:

$$S = \frac{Q}{P_1} \quad (3.13)$$

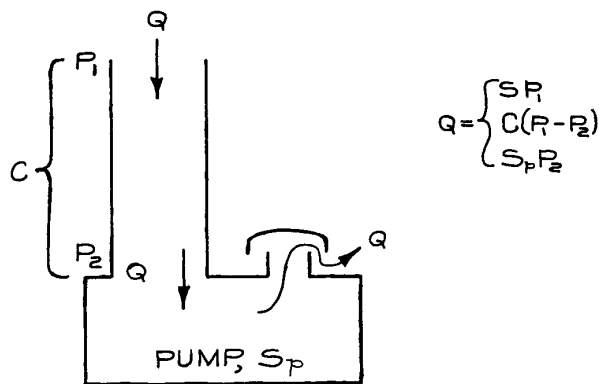


Figure 3.2 Analytical representation of a vacuum system.

and the speed at the pump inlet is:

$$S_p = \frac{Q}{P_2} \quad (3.14)$$

Notice that the throughput is the same at every point in this system, since there is only one gas source. The throughput of the conductance is:

$$Q = (P_1 - P_2)C \quad (3.15)$$

Upon substitution for  $P_1$  and  $P_2$  from the previous two equations, this equation, after some rearrangement, becomes:

$$\frac{1}{S} = \frac{1}{S_p} + \frac{1}{C} \quad (3.16)$$

This is the master equation that relates the net speed of a system to the capacity of the pump and the conductors leading to the pump. From this equation we see that both the conductance of a system and the speed of the pump exceed the net speed of the system. In vacuum-system design, economics should be considered. The cost of a vacuum pump is usually greater than the cost of constructing a tube leading from a vacuum container to the pump. In order to achieve a given net speed for a system, it is most economical to design a system whose equivalent conductance exceeds the required net speed by a factor of three or four so that the speed of the pump need only exceed the net speed by a small amount.

### 3.2.4 Conductance Formulae

In the viscous flow region the conductance of a tube depends upon the gas pressure and viscosity.<sup>2</sup> For a tube of circular cross-section and in the absence of turbulence, the conductance from the Hagen – Poiseuille derivation is:

$$C = 32\,600 \frac{D^4}{\eta L} P_{av} \text{ L/s} \quad (3.17)$$

when the diameter  $D$  and length  $L$  of the tube are in centimeters, the average pressure  $P_{av}$  is in torr, and the viscosity of the gas,  $\eta$ , is in micropoise ( $\mu\text{P}$ ). For air at 20 °C the conductance is:

$$C = 180 \frac{D^4}{L} P_{av} \text{ L/s} \quad (3.18)$$

These equations are used for determining the dimensions required for the lines connecting to a pump that operates in the rough vacuum region. A complex system is treated as a series of simple cylindrical tubes; the conductance of each tube is calculated from the equation above and the conductance of the whole system is determined using the network equations. Joints and elbows between simple tubes will reduce the conductance of the system so it is wise to specify tube diameters somewhat oversize to offset conductance losses in the transitions between simple cylindrical shapes.

In the molecular-flow region, conductance is independent of pressure. The conductance of a tube of circular cross-section is:

$$C = 2.6 \times 10^{-4} \bar{v} \frac{D^3}{L} \text{ L/s} \quad (3.19)$$

when the diameter and length are measured in centimeters and the average molecular velocity  $\bar{v}$  is in cm/s. For air at 20 °C the conductance is:

$$C = 12 \frac{D^3}{L} \text{ L/s} \quad (3.20)$$

These equations are useful in determining the dimensions of a high-vacuum chamber and the size of the tubes leading to and from the chamber. As an example of poor design consider the case of a vacuum chamber connected to a 100 L/s diffusion pump by a tube 2.5 cm in diameter

and 10 cm long. The net speed of the pump and connecting tube is:

$$S = \left( \frac{1}{100} + \frac{1}{19} \right)^{-1} = 16 \text{ L/s} \quad (3.21)$$

The connecting tube is strangling the pump. Consider also the pressure drop across the tube. Since the throughput is a constant, one can write:

$$SP_1 = S_p P_2 = Q \quad (3.22)$$

thus:

$$\frac{P_1}{P_2} = \frac{S_p}{S} = 6 \quad (3.23)$$

This result indicates the importance of proper location of a pressure gauge. In this case a gauge at the mouth of the pump would indicate a pressure six times lower than the pressure in the vacuum container.

There are frequently occasions when a vacuum system is divided into compartments at very different pressures separated by a wall with an aperture whose diameter exceeds its length (a so-called “thin aperture”). In the viscous flow regime, the conductance of an aperture at 20 °C is approximately:

$$C \approx 20A \text{ L/s} \quad (3.24)$$

where  $A$  is the area in  $\text{cm}^2$ . In the molecular-flow region, the conductance of an aperture for a gas of molecular weight  $M$  is:

$$C = 3.7 \left( \frac{T}{M} \right)^{1/2} A \text{ L/s} \quad (3.25)$$

where  $A$  is the area in  $\text{cm}^2$ . These equations are useful for determining the rate of gas flow into a vacuum chamber through an aperture in a gas-filled collision chamber or ion source.

There is a *transition region* between the viscous and molecular flow regions where the mean free path approximately matches the dimensions of the container. For laboratory-scale apparatus, the transition region falls in the range of 500 mtorr down to 5 mtorr. A mathematical description of this region is difficult. Fortunately one finds in most vacuum systems that the transition region is encountered only briefly during pumpdown from



atmosphere to high vacuum. Conductance formulae, such as given above, for flow in the viscous (high-pressure) region compared to conductance formulae in the molecular (low-pressure) region show two important differences: gas flow in the viscous region depends upon the gas pressure and the square of the cross-section of the conductor (i.e.,  $D^4$ ), whereas gas flow in the molecular flow region is independent of pressure and goes as the  $3/2$  power of the cross-section. It follows that the formulae for conductance in the molecular flow region will give a conservative estimate of the conductance in the transition region.

### 3.2.5 Pumpdown Time

The time required to pump a chamber of volume  $V$  from pressure  $P_0$  to  $P$  is:

$$t = \frac{V}{S} \ln\left(\frac{P_0}{P}\right) \quad (3.26)$$

assuming a constant net pumping speed and considering only the gas that resided in the container at the outset. A typical high-vacuum system is rough-pumped to about 50 mtorr with a mechanical pump and then pumped to very low pressures with a diffusion pump or some other pump that works effectively in the molecular-flow region. The pumpdown calculation is performed in two steps corresponding to these two operations. In the low-pressure regime, gas desorbing from the container walls may contribute to the gas load and the rate of desorption will need to be considered in estimating the pumpdown time.

### 3.2.6 Outgassing

The rate of at which the pressure falls in a chamber at pressures below about  $10^{-5}$  torr is largely determined by the rate of evolution of gas from the walls of the chamber rather than by the speed of the pump. The rate of gas evolution – *outgassing* – from solid materials depends both on the nature of the solid and the adsorbate. Water and carbon dioxide from the air are adsorbed most tenaciously. A rough, porous, or polar surface adsorbs more gas than a clean polished surface. The outgassing rate from a surface decreases slowly, roughly as  $e^{-at}$  ( $a$  a property of the surface and the adsorbate, as well as a strong function of temperature). Baking a vacuum chamber to 150 to

200 °C under vacuum will encourage outgassing and result in a lower ultimate pressure after the chamber cools.

Following exposure to air, stainless steel outgases at roughly  $10^{-8}$  torr L/s/cm<sup>2</sup> of surface area after an hour of pumping at room temperature. A day of pumping may be required to reduce this rate below  $10^{-9}$  torr L/s/cm<sup>2</sup>. The outgassing rate from aluminum is more than 10 times greater than for stainless steel. Outgassing from plastics is roughly 10 times worse than from aluminum, and outgassing from elastomers (rubber) is 10 times worse yet. Plastics and rubbers have no use in a vacuum system that is to achieve pressures below about  $10^{-7}$  torr.

Consider, for example, the problem of maintaining a base pressure of  $10^{-7}$  torr in a stainless-steel chamber. If the outgassing rate is  $10^{-9}$  torr L/s/cm<sup>2</sup>, a pumping speed of  $10^{-2}$  L/s is required for every cm<sup>2</sup> of surface. For a container 1 m in diameter and 1 m high, a pumping speed of 400 L/s is needed. This speed is typical of a high vacuum pump (a diffusion pump or turbomolecular pump) with a nominal 4 in. (10 cm) diameter inlet port. Without baking to further reduce the outgassing rate, a pump 10 times larger (and nearly 10 times as expensive) would be required to reduce the pressure to  $10^{-8}$  torr; ultrahigh vacuum is out of the question.

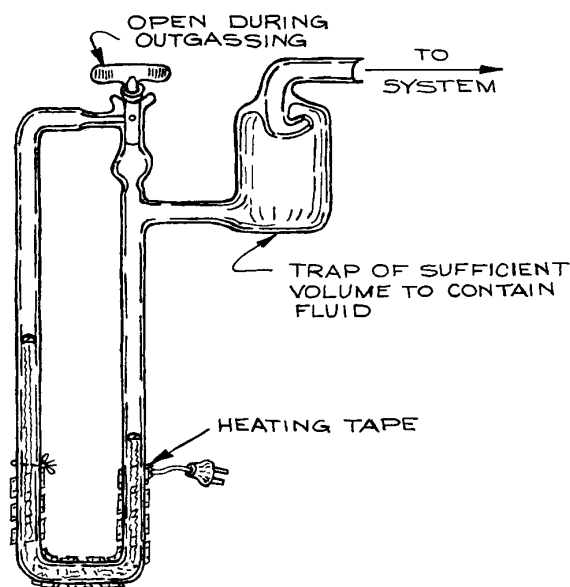
## 3.3 PRESSURE AND FLOW MEASUREMENT

The pressure within a vacuum system may vary over more than 13 orders of magnitude. No one gauge will operate over this range, so most systems are equipped with several different gauges.

The pressure ranges in which the various gauges are useful are shown in Figure 3.7.

### 3.3.1 Mechanical Gauges

**Hydrostatic Gauges.** The simplest pressure gauges are hydrostatic gauges such as the oil or mercury manometer. A closed-end U-tube manometer filled with mercury is useful down to 1 torr. Diffusion-pump oils, such as Octoil, di-*n*-butyl phthalate, or Dow Corning DC704 or DC705 silicone oils, may also be used as

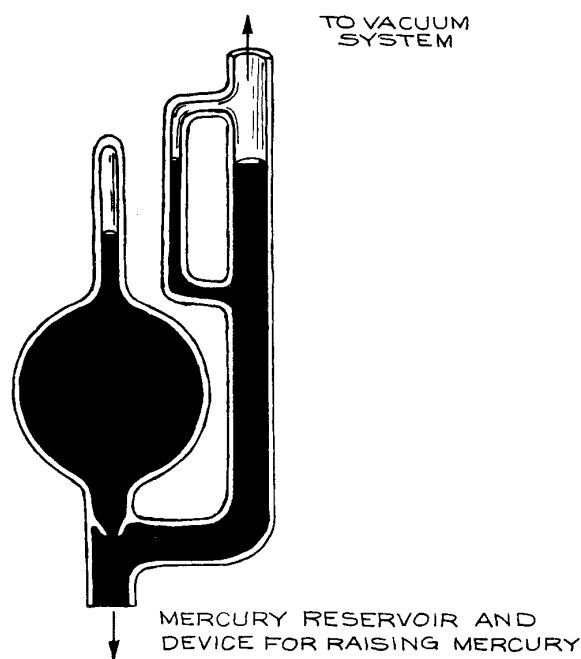


OIL DENSITIES (20°C)

di- <i>n</i> -butyl phthalate	1.044 g cm <sup>3</sup>
Octoil	0.983
Octoil-S	0.912
DC-704	1.07
DC-705	1.09

**Figure 3.3** Oil manometer. The trap is intended to contain the oil charge in the event that the low-pressure arm of the manometer is accidentally open to the atmosphere.

manometer fluids. Because the density of these oils is much less than that of mercury, oil-filled manometers are useful down to about 0.1 torr. As shown in Figure 3.3, an oil manometer should be made of glass tubing of at least 1 cm diameter in order to reduce the effects of capillary depression. To prevent oil adhesion to the walls, the manometer should be cleaned with dilute hydrofluoric acid before filling. A heater is usually provided to drive dissolved gas from the oil prior to use.<sup>3</sup> A number of manometers have been designed to increase the surface area of oil exposed to the vacuum in order to hasten outgassing.<sup>4</sup> Both oil and mercury manometers will contaminate a vacuum system with vapor of the working fluid. If this is a problem, a cold trap is placed between the gauge and the system to condense the offending vapor. Mercury in a manometer can be isolated and protected



**Figure 3.4** McLeod gauge.

from chemical attack by floating a few drops of silicone oil on top of the mercury column.

The McLeod gauge shown in Figure 3.4 is a complex hydrostatic gauge and is sensitive to much lower pressures than a simple U-tube manometer. In the McLeod gauge, a sample of gas is trapped and compressed by a known amount (typically 1000:1) by displacing the gas with mercury from the original volume into a much smaller volume. The pressure of the compressed gas is measured with a mercury manometer and the original pressure determined from the general gas law. Of course, it is not possible to use a McLeod gauge to measure the pressure of a gas that condenses upon compression. With care, a high-quality McLeod gauge will provide accuracy of a few % down to about  $10^{-4}$  torr. Some gauges can be used at even lower pressures, but it is necessary to consider the error caused by the pumping action of mercury vapor streaming out of the gauge.<sup>5</sup>

The McLeod gauge is constructed of glass and is easily broken. If the mercury is raised into the gas bulb too quickly, it can acquire sufficient momentum to shatter

the glass. When a gauge is broken, the external air pressure frequently drives a quantity of mercury into the vacuum system. For this reason it is wise to place a ballast bulb of sufficient volume to contain the mercury charge of the gauge between the gauge and the system.

**Mechanical Gauges.** A number of different gauges depend upon the flexure of a metal tube or diaphragm as a measure of pressure. In the Bourdon gauge, a thin-wall, curved tube, closed at one end, is attached to the vacuum system. Pressure changes cause a change in curvature of the tube. A mechanical linkage to the tube drives a needle to give a pressure reading on a curved scale. Another type of mechanical gauge contains a chamber divided in two by a thin metal diaphragm. The volume on one side of the diaphragm is sealed, while the volume on the other side is attached to the system. A variation in pressure on one side relative to the other causes the diaphragm to flex, and this movement is sensed by a system of gears and levers that drives a needle on the face of the gauge. The precision of these gauges is limited by hysteresis caused by friction in the linkage. To overcome this friction it is helpful to gently tap the gauge before making a reading. Unlike the liquid manometer, mechanical gauges are not absolute gauges. The pressure scale and zero location on these gauges must be calibrated against a McLeod gauge or U-tube manometer. Mechanical gauges are useful down to 1 torr with an accuracy of about 0.5 torr. They offer the advantage of being insensitive to the chemical or physical nature of the gas.

**Piezoelectric Pressure Transducers.** A piezo transducer is an absolute, direct-reading device that senses the gas pressure on a piezoelectric crystal. They function in the  $10^{-1}$  to 1000 torr range. Typical gauges include the transducer and the sensor electronics in a single compact unit (for example, MKS Instruments). They are useful for monitoring roughing and foreline pressures, and for inclusion in an interlock system to control vacuum system operation (see Sections 3.6.1, 6.6.9, 6.6.10).

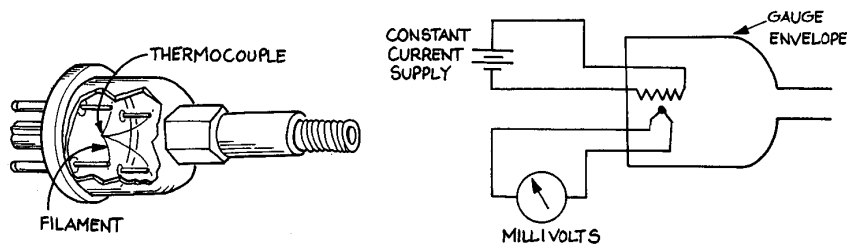
**Capacitance Manometers.** A capacitance manometer is a diaphragm manometer wherein the position of the diaphragm is determined from a measurement of

electrical capacitance. The gauge head is divided into two chambers by the thin metal diaphragm that serves as one plate of a capacitor; the second plate of the capacitor is fixed in one of the chambers. A change in the pressure difference between the two chambers results in a change in capacitance as the flexible diaphragm moves relative to the fixed plate. A sensitive capacitance bridge measures the capacitance between the two plates and the capacitance is converted to a reading of the pressure difference. Capacitance manometers are inherently differential pressure gauges; most, however, are manufactured to read absolute pressure by sealing and evacuating one of the chambers. This arrangement does not provide a truly absolute gauge since the readout must be set to zero by connecting the open chamber to a true zero pressure (i.e., several orders of magnitude less than the full-scale reading). A capacitance manometer is designed so that the flexible diaphragm touches the opposing capacitor plate at a small overpressure. This protects the fragile diaphragm from a sudden inrush of gas. As might be expected, the capacitance measurement is very sensitive to the temperature of the flexible diaphragm. High precision gauge heads incorporate a heater and temperature control that maintains the gauge head at a constant temperature.

Capacitance manometers are available to measure pressures from 1000 torr down to less than  $10^{-4}$  torr. A single gauge head is useful over a range of three to four orders of magnitude with an accuracy of 0.01% full scale at the low end of the pressure range and 0.5% or better at the high end (e.g., a 1 torr gauge head reads  $1 \times 10^{-4} \pm 1 \times 10^{-4}$  torr at the low end and  $1.00 \pm 0.005$  torr at the high end). Gauges are calibrated by the manufacturer and have a reputation for holding their calibration for years given careful use. The electronics have a response time of about 10 ms. This makes the capacitance manometer especially useful for pressure control in a feedback loop.

### 3.3.2 Thermal-Conductivity Gauges

The thermal conductivity of a gas decreases from some constant value above about 10 torr to essentially zero at about  $10^{-3}$  torr. This change in thermal conductivity is used as an indication of pressure in the Pirani gauge and the thermocouple gauge. In both gauges a wire filament is



**Figure 3.5** Thermocouple gauge.

heated by the passage of an electrical current. The temperature of the filament depends on the rate of heat loss to the surrounding gas. In the thermocouple gauge, the pressure is determined from the e.m.f. produced by a thermocouple in contact with the filament heated by the passage of a constant current (Figure 3.5). In the Pirani gauge, the heated filament is one arm of a Wheatstone bridge. A change in temperature of the filament produces a change of resistivity and hence a change in the voltage across the filament. The resulting imbalance of the bridge gives an indication of pressure. Alternatively, a voltage can be applied to keep the bridge in balance and this voltage becomes the indication of pressure.

The pressure indicated by a thermocouple gauge or a Pirani gauge depends upon the thermal conductivity of the gas. They are usually calibrated by the manufacturer for use with air. For other gases these gauges must be recalibrated point-by-point over their entire range, since thermal conductivity is a nonlinear function of pressure. In routine use, a thermal-conductivity gauge can only be expected to be accurate to within a factor of two. This accuracy is adequate when the gauge is used to sense the foreline pressure of a diffusion pump or to determine whether the pressure in a system is sufficiently low to begin diffusion-pumping. The principal advantages of these gauges are their ease of use, ruggedness, and low cost.

### 3.3.3 Viscous-Drag Gauges

Above about 1 torr, the viscosity of a gas is constant, but, in the molecular flow regime, the viscous drag exerted by a gas on a moving surface depends upon the gas density and the mean velocity of the gas. It follows that below about 1 torr, a first-principles calculation yields gas pressure from

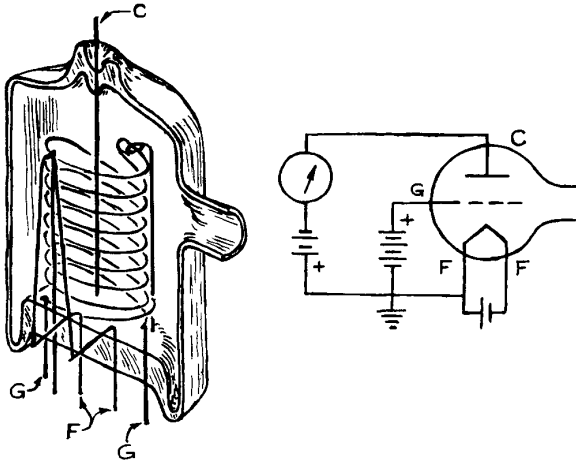
a measurement of viscous drag for a gas of known temperature and composition. This approach to absolute pressure measurement is applied in the *spinning-rotor gauges* manufactured by Leybold AG and MKS Instruments.

In the gauge head of a spinning-rotor gauge, a magnetically suspended steel ball is spun up to a rotational speed of about 415 Hz by a rotating magnetic field. The drive is interrupted and the ball is permitted to decelerate to a preset lower speed, typically about 405 Hz. The pressure is determined from a measurement of the time required for a complete revolution of the coasting ball. The rate of deceleration is very small – at  $10^{-4}$  torr the deceleration from 415 to 405 Hz takes many minutes. The measurement of the ball rotation rate is repeated many times to obtain satisfactory precision. A microprocessor in the controller computes the statistical average of the deceleration rate after many cycles.

Spinning-rotor gauges provide absolute pressure measurement between  $10^{-1}$  and  $10^{-6}$  torr with an accuracy of a few %. At the lower end of this range, they are the only available means for making highly accurate, repeatable, absolute-pressure measurements. Spinning-rotor gauges are used as a transfer standard in calibrating other types of gauges. These instruments must be maintained with great care. They are expensive.

### 3.3.4 Ionization Gauges

In the region of molecular flow, pressure is usually measured with an ionization gauge or “ion gauge.” In this type of gauge, gas molecules are ionized by electron impact and the resulting positive ions are collected at a negatively biased electrode. The current to this electrode is a function



**Figure 3.6** Bayard – Alpert ionization gauge.

of pressure. There are two basic types of ion gauges, differing primarily in the mechanism of electron production. In the hot cathode ionization gauge, electrons are produced by thermionic emission from an electrically heated wire filament. In cold cathode gauges, electrons and ions are produced in a discharge initiated by a high-voltage electrical discharge.

The most familiar configuration of the hot cathode ionization gauge is that devised by Bayard and Alpert and illustrated in Figure 3.6. The gauge electrodes are mounted in a glass envelope with a side arm that is attached to the vacuum system. Electrons from an electrically heated filament (F) are accelerated through the gas toward a positively biased, helical wire grid (G). Collisions between the accelerated electrons and gas molecules create ions that are collected at a central wire electrode (C), and the positive ion current is measured by a sensitive electrometer. These gauges are useful in the  $10^{-3}$  to  $10^{-10}$  torr range. The ion current measured with an ionization gauge is a linear function of pressure; the proportionality between ion current depends upon the electron emission current from the filament and upon the ionization efficiency of the gas in the gauge. Commercial gauge controllers are calibrated for use with air at a specified emission current, typically about 10 mA. An adjustment on the controller permits the user to set the emission current. For gases other

**Table 3.2 Sensitivity Factors For Ionization Gauges**

Gas	Formula	Multiplicative Factor <sup>a</sup>
Air		1.0
Nitrogen	N <sub>2</sub>	1.0
Oxygen	O <sub>2</sub>	0.9
Hydrogen	H <sub>2</sub>	2.4
Carbondioxide	CO <sub>2</sub>	0.7
Methane	CH <sub>4</sub>	0.7
Hydrocarbons	C <sub>4</sub> and up	0.2
Argon	Ar	0.9
Neon	Ne	3.5
Helium	He	6.8

<sup>a</sup> Calibrated for air. Recommended values based primarily upon the survey published by R. L. Summers, NASA Tech.

Note: TN D-5285, and an assumed nominal acceleration potential of 150V

than air, the indicated pressure is multiplied by a sensitivity factor that is inversely proportional to the electron-impact ionization cross-section for the gas in question. Sensitivity factors for some common gases are given in Table 3.2.

Ion gauges have the sometimes useful and sometimes detrimental property of functioning as pumps. There are two mechanisms for pumping. Ionized gas molecules accelerated to and embedded in the collector are effectively removed from the system. In addition, metal evaporated from the filament deposits on the glass envelope of the gauge to produce a clean, chemically active surface that adsorbs nitrogen, oxygen, and water. At pressures below  $10^{-7}$  torr this pumping action may create a significant pressure gradient between the gauge tube and the system to which it is attached, resulting in a pressure reading that is lower than that in the system. For ultrahigh-vacuum systems the pumping problem is circumvented by the use of “nude” gauges that consist of the gauge electrodes mounted to a flange that is installed so that the electrodes protrude into the system proper; there is no envelope or tubulation joining the gauge to the system.

The pressure indicated by an ionization gauge is also affected by the emission of gas adsorbed on the gauge electrodes and envelope. To overcome this problem the gauge is periodically degassed by heating the electrodes

and the envelope. This is accomplished by the gauge controller in the “degas” mode in which an AC current is passed through the grid to heat it to incandescence. Radiative heating of the surrounding components encourages outgassing. A hot cathode ionization gauge is activated when the pressure falls below  $10^{-3}$  torr – turn-on at higher temperatures will cause the filament to burn out. The gauge should be degassed for a few minutes when the pressure falls to about  $10^{-4}$  torr and the degassing should be repeated with each decrease of two orders of magnitude in pressure.

The electron-emissive filament in the ion gauge is made of tungsten or thoria-coated iridium. Tungsten filaments operate at  $1800^{\circ}\text{C}$ . They provide the most stable operation, however, their lifetime is significantly shortened by exposure to hydrocarbons at relatively high pressures. Thoria-coated filaments operate at about  $1200^{\circ}\text{C}$  and can be expected to be longer-lived than tungsten filaments. Many gauge tubes provide two filaments with separate external connections. Only one filament is activated at a time. When the first filament burns out the second can be used so that it is not necessary to let the system up to air pressure following the first failure. When the filaments have failed in a glass-enclosed gauge, the entire gauge is replaced. Individual filaments are replaceable in nude gauges.

Cold cathode ionization gauges in various configurations are identified as Penning gauges, magnetrons, inverted magnetrons, and Philips gauges. In these gauges an electrical discharge is struck in a low-pressure gas between two electrodes maintained at a potential difference of several kV. The range of stability of the discharge is extended by a magnetic field provided by a permanent magnet that is an integral part of the gauge head. The magnetic field confines the electrons in the discharge so as to increase their pathlength through the gas. The discharge current measured at one of the electrodes provides a measure of the gas pressure. The measured current is relatively much larger than that obtained with a hot-cathode gauge – of the order of 1 A torr – thus greatly simplifying the controller electronics. Cold cathode gauges operate in the  $10^{-2}$  to  $10^{-7}$  torr range with an accuracy of a factor of two at best. The pressure range is extended down to less than  $10^{-9}$  torr in some more complex designs. Without the fragile filament of the hot cathode gauge, cold cathode gauges are robust and economical, suffering primarily only from their limited range of applicability.

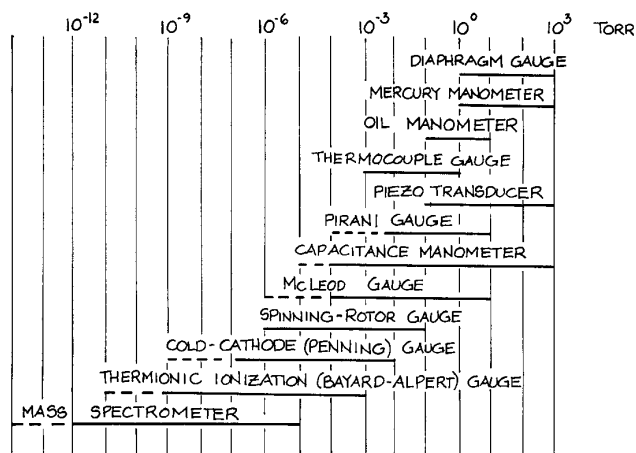


Figure 3.7 Operating ranges of pressure gauges.

The operating pressure ranges of the vacuum gauges is shown in Figure 3.7.

### 3.3.5 Mass Spectrometers

A mass spectrometer can be used to determine the partial pressures of residual gases in a system and to detect leaks. Small, relatively inexpensive mass spectrometers designed specifically for these purposes are called residual-gas analyzers (RGAs). These instruments employ quadrupole or magnetic-deflection type analyzers and typically cover the mass range from 2 to 100 amu. They are sensitive to partial pressures as low as  $5 \times 10^{-14}$  torr. Fixed-focus mass spectrometers set to detect helium are widely used as sensitive leak detectors. Leaks are located by monitoring the helium concentration within a vacuum system while the exterior of the system is probed with a small jet of helium. The minimum detectable leak rate of commercial helium leak detectors is less than  $10^{-11}$  torr L/s.

### 3.3.6 Flowmeters

Gas flow can be measured as a throughput or as a mass flow. Flow rate, as a throughput, is determined from a measurement of the pressure drop across a tube or system of tubes of known conductance. The mass flow rate can be determined from a measurement of the power required to

maintain a given temperature gradient across a tube through which a gas flows.

**Differential Pressure Flowmeters.** For gas flowing through a tube, the throughput depends upon the pressure difference between the entrance and exit of the tube and the conductance of the tube [from Section 3.2.1,  $Q = (P_1 - P_2)C$ ]. Obviously two pressure gauges are required to determine the throughput. In addition the gas temperature must be measured and the dimensions of the tube must be chosen so that the gas flow remains in a definable flow regime (either viscous flow or molecular flow) so that the conductance of the tube can be calculated from the tube dimensions. The length of the tube should exceed the diameter by at least a factor of 10 so that end effects on the conductance can be ignored. A capillary of 1 mm inner diameter and several centimeters in length would assure viscous flow conditions at downstream pressures above 1 torr and the conductance could be calculated from the Hagen – Poiseuille relation (Section 3.2.4). To operate in the molecular-flow regime, a tube of macroscopic dimensions would be only useful at very low flows and low downstream pressures. In practice, the flowrate in the molecular-flow regime is measured using a parallel array of capillary tubes, the dimension of each tube being sufficiently small to assure molecular flow at pressures up to 1 torr. Capillary arrays with thousands of parallel channels each 1 mm long with diameters as small as 25 micrometers ( $\mu\text{m}$ ) are available for this purpose.

The differential pressure flowmeter is appealing in that flow rate is accurately determinable from first-principles calculation; in practice, however, these gauges are used primarily for calibration purposes.

**Thermal Mass Flowmeters.** Commercial flowmeters for routine use are of the thermal mass flow type. The mass flow of gas through a tube can be calculated from the temperature rise in the flowing gas as heat is added or from the power required to maintain a given temperature rise; the latter scheme is the basis of most commercially available gauges. Thermal mass flow gauges operate in the viscous flow regime where thermal conductivity of a gas is essentially constant. The flowmeter consists of a tube that is surrounded by a series of heaters that establish a temperature profile

along the tube. Gas flowing through the tube modifies the temperature profile and this change of temperature can be used as a measure of the gas flow rate. Alternatively, the power to the heaters is adjusted by a sensitive bridge to keep the temperature profile fixed. The power to the heaters then becomes a measure of the mass flow rate. These gauges are usually calibrated for air. For other gases, a correction factor, inversely proportional to the heat capacity and density of the gas, is applied.

Thermal mass flowmeters determine mass flow from a measurement of heat flow; the instruments are, however, calibrated in throughput units of standard cubic centimeters per minute (sccm) of air or  $\text{N}_2$ . One standard cubic centimeter is the amount of gas contained in  $1 \text{ cm}^3$  at 1 atm pressure and  $0^\circ\text{C}$ . Flowmeters are available with full-scale ranges of 10 to 50 000 sccm and an accuracy of about 1% of the full-scale reading. Because thermal mass flowmeters operate in the laminar viscous-flow regime, they are always placed on the delivery side of a flow system. In choosing a gauge one must consider the delivery pressure and the pressure drop across the gauge element. For condensable gases it may be necessary to choose a gauge with a range larger than required, in order to keep the delivery pressure below the vapor pressure of the sample. Thermal mass flowmeters have a response time of the order of 1 s and can be used in a feedback loop to control gas flow rate. Controllers incorporating a flowmeter and an electrically driven metering valve are commercially available.

### 3.4 VACUUM PUMPS

Vacuum pumps are classified according to the chemical or physical phenomenon responsible for moving gas molecules out of the vacuum vessel. Like pressure gauges, every vacuum pump operates in a limited pressure range. The useful range of a pump is also limited by the vapor pressure of the materials of construction and the working fluids within the pump. In general, a pump that operates in the viscous-flow region will not operate in the molecular-flow region and vice versa. To achieve very low pressures, two or more pumps are used in series. In addition, some pumps are more efficient for high-molecular-weight gases than others, and some pump low boiling gases (e.g.,  $\text{H}_2$  and He) more efficiently than others. To achieve ultrahigh

vacuum it is common practice to employ pumps with complimentary characteristics in parallel.

Mechanical pumps operating in the rough vacuum region with ultimate low-pressure limits of 5 torr to 5 mtorr serve two important functions in attaining high and ultra-high vacuum. A pump that will achieve a high vacuum at its inlet will invariably have a maximum tolerable outlet pressure; this is the *critical backing pressure*. A mechanical pump is employed to maintain the high vacuum pump exhaust pressure below the critical pressure and in this function the mechanical pump is referred to as a *backing pump or forepump*. In addition, it is necessary for the pressure in a vacuum chamber to be below some pressure before the high-vacuum pump can be put into operation; this pressure is sometimes referred to as the *stall pressure*. A mechanical pump is used to attain this maximum starting pressure in the chamber and in this function the pump is referred to as a *roughing pump*.

### 3.4.1 Mechanical Pumps

**Oil-Sealed Rotary Vane Pumps.** The pump most commonly used for attaining pressures down to a few millitorr is the oil-sealed rotary vane pump shown schematically in Figure 3.8. In this pump a rotor turns off-center within a cylindrical stator. The interior of the pump is divided into two volumes by spring-loaded vanes attached to the rotor. Gas from the pump inlet enters one of these volumes and is compressed and forced through a one-way valve to the exhaust. A thin film of oil maintains the seal between the vanes and the stator. The oil used in these pumps is good-quality hydrocarbon oil from which the high-vapor-pressure fraction has been removed. These pumps are also made in a two-stage version in which two pumps with rotors on a common shaft operate in series. Rotary pumps to be used for pumping condensable vapors, water vapor in particular, are provided with a gas ballast. This is a valve that admits air to the compressed gas just prior to the exhaust cycle. This additional air causes the exhaust valve to open before the pressures of condensable vapors exceed their vapor pressure and thus prevents these vapors from condensing inside the pump.

Oil-sealed rotary vane pumps will operate for years without attention if the inner surfaces do not rust and the

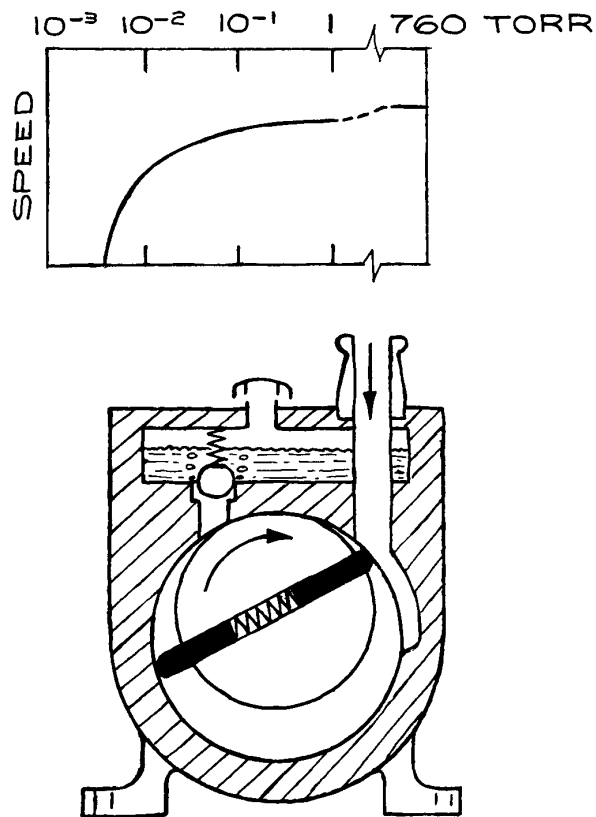


Figure 3.8 Oil-sealed rotary vane pump.

oil maintains its lubricating properties. It is wise to leave these pumps operating continuously so that the oil stays warm and dry. In use, an increase in the lowest attainable pressure (the base pressure) indicates that the oil has been contaminated with volatile materials. The dirty oil should be drained while the pump is warm. The pump should be filled with new oil, run for several minutes, drained, and refilled with a second charge of new oil. For storage, a pump should be filled with new oil and the ports sealed.

Rotary pumps are available with capacities of 1 to 500 L/s. A single-stage pump is useful down to 50 mtorr, and a two-stage pump to 5 mtorr. Typical performance of a single-stage pump is indicated by the pumping-speed curve in Figure 3.8. A two-stage pump is to be preferred as a backing pump for a diffusion pump. With a two-stage pump, a base pressure of  $10^{-4}$  torr can be achieved after a long



pumping time, if the back diffusion of oil vapor from the pump is suppressed by use of a sorption or liquid-air trap on the pump inlet. This is a simple scheme for evacuating small spectrometers and Dewar flasks or other vacuum-type thermal insulators.

The exhaust gases from an oil-sealed mechanical vacuum pump contain a mist of fine droplets of oil. This oil smoke is especially dense when the inlet pressure exceeds 100 torr. The oil droplets are extremely small, usually less than 5 microns. Over the course of time, this oil settles on the pump and its surroundings and collects dirt and grime. Furthermore, breathing the finely dispersed oil may injure the operator's lungs. In addition, over the course of weeks or months, this oil represents a significant loss from the oil charge in the pump, so a pump failure may result if the oil level in the pump is not faithfully monitored. Most pump manufacturers market filters that cause the exhaust oil mist to coalesce and run back into the pump; their use is recommended. In addition, modern standards of laboratory hygiene require venting a mechanical pump into the laboratory fume hood. If toxic gases are being pumped it is of course essential to pipe the exhaust out of the lab. Sometimes, a pretreatment involving passage of the exhaust through a neutralizing chemical bath is required. In most cases exhaust lines can be made of PVC drainpipe available from plumbing suppliers. The exhaust line should rise vertically from the pump so that oil in the exhaust will collect on the walls of the line and run downward back to the pump. An exception to this scheme arises in the case of a wet vacuum process. If considerable water is exhausted from the pump, there is the danger that water will condense in the exhaust line and run back into the pump.

**Roots Blowers.** A Roots blower is a displacement pump. As illustrated in Figure 3.9, these pumps consist of a pair of counter-rotating two-lobed rotors on parallel shafts. Rotational speeds are about 3000 rpm. There is a clearance of a few thousandths of an inch between the rotors themselves and between the rotors and the housing. A volume of gas is trapped at the inlet and compressed as it is moved to the exhaust. There is no oil in the body of the pump to maintain a high-pressure seal, but, owing to the high speed of the rotors, a compression ratio as great as 40:1 can be achieved. The virtue of the Roots blower is its relatively high pumping speed in the 1 to  $10^{-3}$  torr region. To achieve

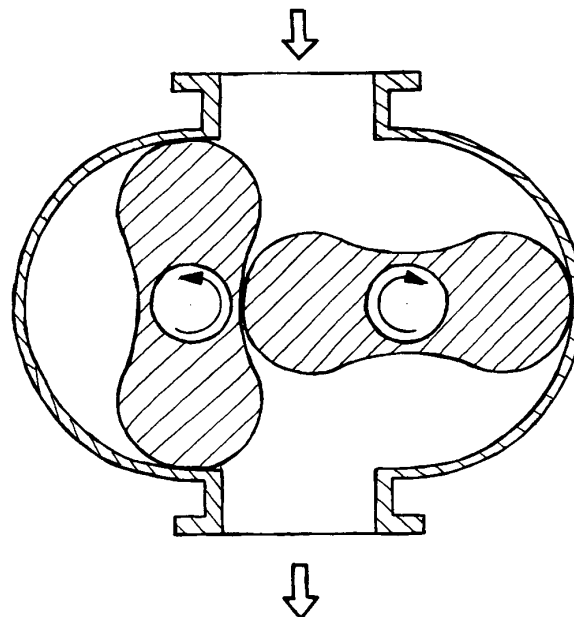


Figure 3.9 Roots blower.

an ultimate vacuum of  $10^{-3}$  torr, a Roots blower is used in series with a rotary pump at its exhaust. The required speed of the rotary pump is smaller than that of the Roots pump by a factor of the inverse of the compression ratio of the Roots pump. Roots blowers are available with speeds of a few hundred to many thousand L/s.

**Piston Pumps.** There are now available piston pumps with Teflon-sealed pistons that do not contaminate the vacuum environment with oil, as do oil-sealed rotary pumps. The Leybold EcoDry pump employs a novel diaphragm exhaust valve that permits a higher compression ratio than pumps with poppet valves. The low-pressure limit of these pumps is about 50 mtorr. Piston pumps are robust and reliable, but they are also relatively expensive in comparison to rotary pumps. They find use as backing pumps for pumps such as turbomolecular pumps, when contamination with hydrocarbons is not permissible.

**Diaphragm Pumps.** Diaphragm pumps are displacement pumps employing the flexure of a thin

metal or Teflon diaphragm to compress gas and drive it out through an exhaust valve. These are small pumps of low pumping speed and relatively modest ultimate vacuum. The low pressure limit is typically about 0.3 torr. Diaphragm pumps can be made completely free of hydrocarbons so they are used as roughing pumps and, in some applications, as backing pumps in ultraclean systems.

**Molecular Drag Pumps.** A molecular drag pump incorporates a rotating cylinder within a closely fitting, stationary cylinder. The clearance between rotor and stator is of the order of 0.01 in. Pumping occurs when the surface velocity of the rotor approaches the velocity of the molecules being pumped, so that molecules striking the rotor acquire a significant velocity component tangential to the rotor – they are dragged along with the rotor. Tangentially accelerated gas molecules rebounding to the stator are decelerated: they “pile up,” thus yielding compression. The process is repeated continuously around the gap between rotor and stator. Helical grooves on the surface of the rotor or stator direct the flow of continually compressing gas toward the pump outlet. Drag pumps provide efficient pumping at inlet pressures from nearly 1 torr down to about  $10^{-6}$  torr. The *compression ratio* for air (that is, the ratio of outlet to inlet pressure) is typically  $10^7$  at the low-pressure end of this range. As a consequence, the exhaust pressure, that is, the critical backing pressure, falls in the 1 to 10 torr range. The practical result is that a very small pump is required to transfer the drag pump throughput to the atmosphere. Small diaphragm pumps serve well as backing pumps for molecular drag pumps.

The chamber pressure at the inlet must be below about 1 torr for a drag pump to begin to operate efficiently; a roughing pump is necessary. The backing pump, pumping through the drag pump, can serve to rough down the chamber, however, the gap between rotor and stator is small and restrictive in the viscous-flow regime. A separate roughing pump or a bypass (with a valve) from the chamber to the backing pump should be provided.

Molecular drag pumps provide modest pumping speeds at high vacuum – the inlet aperture is effectively the gap between rotor and stator. Drag pumps are currently available with pumping speeds of up to about 30 L/s. The great virtue of the drag pump is the relatively high throughput in

the 1 to  $10^{-3}$  torr pressure range. Displacement pumps such as rotary-vane pumps, piston pumps, and Roots blowers, lose efficiency below 0.1 torr. High vacuum pumps such as diffusion pumps, turbomolecular pumps and ion/getter pumps, “stall” at pressures above 0.01 torr. If pump-down time is an important issue, a molecular drag pump may be a good choice for a small vacuum system with an ultimate pressure of  $10^{-6}$  torr. With a 1 L/s roughing/backing pump and a drag pump of 30 L/s, a volume of 100 L can be evacuated to less than  $10^{-4}$  torr in a matter of minutes.

**Turbomolecular Pumps.** Turbomolecular pumps operate in the molecular-flow regime. The construction is similar to that of an aircraft-type jet turbine engine. A series of bladed turbine rotors on a common shaft turn at 20 000 to 90 000 rpm. The edge speed of a rotor approaches molecular velocities. The rotor blades are canted so that a molecule striking a blade receives a significant component of velocity in the direction of the pump exhaust (Figure 3.10). Bladed stators are interleaved between the rotors. The stator blades are canted in the opposite direction from that of the rotors in order to decelerate the molecules and compress the flowing gas before it is delivered downward to the next rotor – stator pair. These pumps provide roughly the same pumping speed for all gases; however, the compression ratio

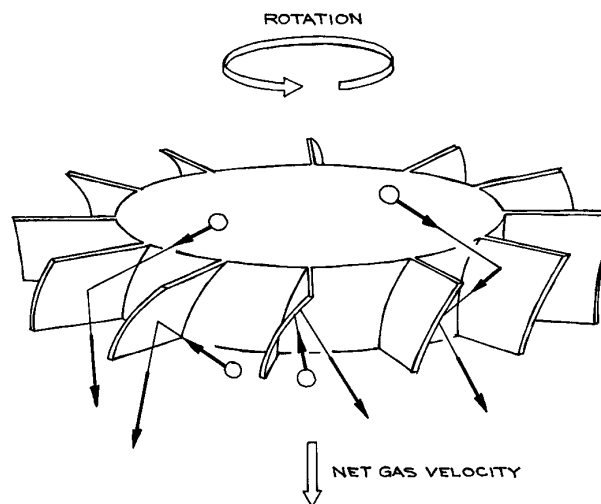


Figure 3.10 Rotor of turbomolecular pump.

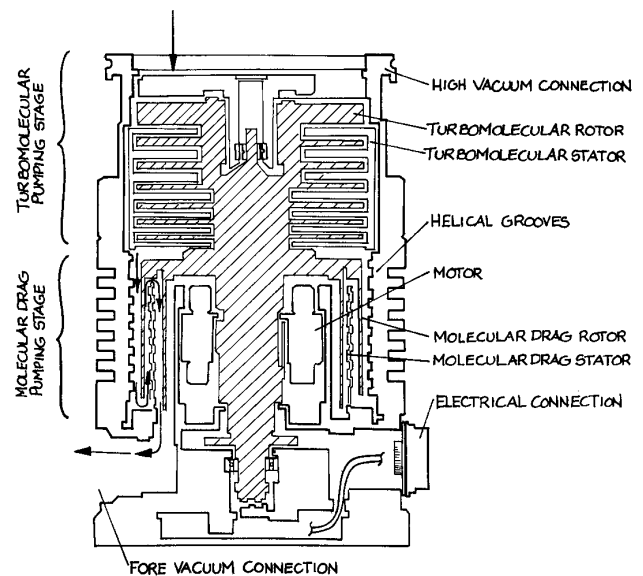
depends upon the nature of the gas being pumped. A single rotor – stator pair typically provides a compression ratio of about 10 for  $N_2$ . To a first approximation, the logarithm of the compression ratio is proportional to the square root of the molecular weight of the gas; thus, for example, a multistage pump with eight rotors may have a compression ratio of  $10^8$  for  $N_2$ , but only  $10^2$  to  $10^3$  for  $H_2$ . A desirable consequence of the strong dependence upon molecular weight is that the compression ratio is very high for oil vapor backstreaming from the pump bearings or from the pump exhaust; thus a turbopump provides an essentially oil-free vacuum. The compression ratio is extremely sensitive to the pressure at the outlet of the pump. A typical turbopump may have a compression ratio of  $10^8$  for air if the pressure at the exhaust is maintained below 0.1 torr, but if the exhaust pressure rises to 1.0 torr, the compression ratio falls to 10.

A turbomolecular pump is usually run in series with a conventional oil-sealed rotary pump as a backing pump. The backing pump also serves as a forepump. With the vacuum chamber at atmospheric pressure, the rotary pump is activated to evacuate the chamber, drawing gas directly through the body of the turbopump. Depending on the manufacturer's specification, the turbo is activated at the same time or somewhat later as the chamber pressure falls so that the turbo is coming up to its operating speed as the chamber pressure reaches a pressure of a few millitorr.

The critical elements of a turbomolecular pump include the motor and the bearings. Typically, 400 Hz synchronous motors are used. Active electronic control is employed to maintain rotor speed over a wide range of loads and to protect the pump in the event of an overload. The motor driver and control represent a significant fraction of the total cost of a turbomolecular pumping system. Many turbopumps use water-cooled, grease-packed bearings. Some use oil as a bearing lubricant with the oil being circulated to a water-cooled heat exchanger. Magnetic suspension systems are the most recent development (Balzers, Osaka, and Varian). There is no mechanical contact in a magnetic bearing; there is essentially no friction and lubrication is unnecessary. In some pumps the rotor is completely magnetically suspended, while in others a magnetic bearing is used at the inlet end of the rotor with a conventional grease-lubricated bearing at the outlet, relying upon the high pumping speeds for hydrocarbons to suppress oil

backstreaming. Axial and a radial magnetic fields to support the rotor are provided by electromagnets in a feedback system that senses the position of the rotor and makes corrections to offset disturbing forces. A mechanical bearing is provided to catch the rotor in the event of catastrophic imbalance and to support the rotor when the pump is not in operation. A backup electrical system is required to protect against a sudden loss of power to the magnet system when the pump is in operation. This may be simply a battery, however in sophisticated systems, the pump motor is operated as a generator when external power is lost, in order to provide electricity for the magnetic bearings.

The most recent innovation is a compound pump design incorporating a multistage turbomolecular pump mounted on a common axis above a molecular drag pump, as shown in Figure 3.11. These are sometimes called hybrid pumps or compound turbomolecular drag pumps; a complete system effectively has three pumps operating in series: a turbomolecular pump, a molecular drag pump, and a displacement-type backing pump (Balzers). The hybrid arrangement increases the compression ratio of the high vacuum pumping stage by several orders of magnitude, with a concomitant increase in the outlet pressure.



**Figure 3.11** Compound turbomolecular/molecular drag pump.

The tolerable backing pump pressure of a hybrid pump is in the 1 to 10 torr range. With the increase in outlet pressure comes a corresponding decrease in the volume flow rate at the exhaust. The important consequence is that these pumps require only a small dry-diaphragm backing pump that, with the inclusion of a magnetic rotor suspension, creates an ultrahigh vacuum pumping system that is completely free of hydrocarbons. Complete turnkey systems are available that include a compound pump with a motor controller, a backing pump, and all necessary interlocks to control start up and shutdown. Pressures below  $10^{-10}$  torr are attainable.

The high rotational speeds and close tolerances in a turbomolecular pump place a premium upon careful mounting, cleanliness, and proper maintenance. Owing to the very small clearances between moving and stationary components, twisting and bending forces that would cause distortion cannot be applied to the pump body. The pump should not be used as the mounting platform for the vacuum system. The pump should be suspended by its inlet flange. Small particles can have disastrous effects upon bearings and rotors at high speeds. The vacuum container pumped by a turbopump must be kept scrupulously clean. A protective screen, known as a splinter shield, over the inlet is essential. When the vacuum container is to be let up to atmospheric pressure, gas should be admitted slowly so that turbulence does not carry debris into the turbopump.

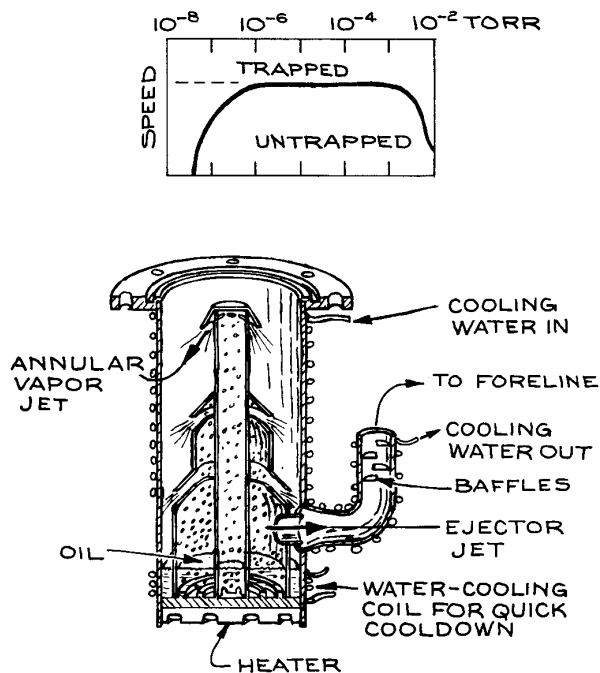
Careful and systematic startup and shutdown procedures are essential to protect the turbopump and to exclude oil from the vacuum system. As mentioned above, turbos are usually operated without a valve at the inlet, however, a valve is required in the *foreline* between the turbo outlet and its backing pump. At startup with the turbo and the vacuum system at atmospheric pressure, the backing pump is activated and the foreline valve is opened. The rush of dense gas into the backing pump prevents oil from backstreaming. As the pressure falls, the turbo activation is timed so that the turbo is up to at least 50% of its maximum speed before the pressure at the pump inlet falls below a few hundred millitorr. At this point the turbo compression should be sufficient to prevent oil vapor from backstreaming to the vacuum system. At shutdown, the foreline valve is closed immediately after power is cut from the turbo. The vacuum system should be vented before the rotor falls

below 50% of its maximum speed – again so that the compression ratio is sufficient to prevent backstreaming of oil from the bearings. When the pressure has risen above several hundred torr, backstreaming is essentially impossible. It follows that a pump that is not operating should not be stored under vacuum. It also should be emphasized that the pump must be vented above the outlet. Many manufacturers sell turbopumps and controllers as a complete system that incorporates the necessary valves and timing circuits to automatically perform the startup and shutdown procedures.

A turbomolecular pump offers many advantages for high- or ultrahigh-vacuum systems. Of course these must be weighed against the disadvantages, the main one being the relatively high initial cost and the cost of major servicing; a complete turbomolecular-pump system may cost two or more times as much as a comparable diffusion-pump system. Turbomolecular pumps with pumping speeds from 10 to 1000 L/s are available. A turbopumped system can achieve pressures below  $10^{-10}$  torr. Most gases can be pumped, although, as noted, turbos do not efficiently pump hydrogen and helium. Most corrosive gases are acceptable, providing the bearing lubricant does not come under attack. Turbopumps are compact and relatively light in weight; they need not be mounted on the underside of a vacuum system. For many types of pumps, the mounting orientation is not critical – these can be mounted with the axis horizontal. Turbos can be baked to reduce outgassing; many are equipped with heating jackets for this purpose. A turbo can be used on a system that is to be baked, provided the inlet temperature does not exceed 100 – 120 °C. Magnetic fields can be a problem, as eddy currents induced in the aluminum rotor assembly cause heating. The specification for most pumps calls for magnetic fields not to exceed 50 to 100 gauss at the inlet. A turbomolecular pump can run for one to three years without attention. The major maintenance procedure usually involves replacing the bearings and rebalancing the rotor. These procedures often require the pump to be returned to the manufacturer.

### 3.4.2 Vapor Diffusion Pumps

In a diffusion pump, gas molecules are moved from inlet to outlet by momentum transfer from a directed stream of oil



**Figure 3.12** Diffusion pump.

or mercury vapor. As shown in Figure 3.12, the working fluid is evaporated in an electrically heated boiler at the bottom of the pump. Vapor is conducted upward through a tower above the boiler to an array of nozzles from which the vapor is emitted in a jet directed downward and outward toward the pump walls. The walls of the pump are cooled, so that molecules of the working-fluid vapor condense before their motion is randomized by repeated collisions.

The diffusion pump walls are usually water-cooled, although in some small pumps they are air-cooled. The condensate runs down the pump wall to return to the boiler. Efficient operation of a diffusion pump requires a temperature gradient from top to bottom. The pump wall should be lowest at the top of the pump – less than 30 °C – to effectively condense oil vapor, and highest at the bottom to drive absorbed gas out of the oil condensate. For this reason the cooling water flows from top to bottom and the flow rate is regulated; too high a flow rate results in too low a temperature at the bottom of the pump.

Because the pumping action of a diffusion pump relies on momentum transferred in collisions between the mole-

cules of the working fluid vapor and the molecules being pumped, the speed of the pump is greater for light gases than for heavy gases. The inlet area essentially determines the net speed of the pump. For air the pumping speed is about 4 L/s/cm<sup>2</sup> of inlet area.

As indicated by the pumping-speed curve in Figure 3.12, the pumping action of a diffusion pump begins to fail when the inlet pressure increases to the point where the mean free path of the molecules being pumped is less than the distance from the vapor-jet nozzle to the wall. When this occurs the net downward momentum of vapor molecules is lost and the vapor begins to diffuse upward into the vacuum system. Diffusion pumping of a vacuum chamber may be initiated at 50 to 100 mtorr, but the system pressure should quickly fall below 1 mtorr or the chamber will become significantly contaminated with the vapor of the working fluid. Oil diffusion pumps can run against an outlet pressure of 300 to 500 mtorr, and mercury pumps can tolerate an outlet pressure of a few torr. A backing pump, usually an oil-sealed rotary pump, is required. The speed of the diffusion pump is insensitive to foreline pressure up to some critical pressure. If this *critical backing pressure* is exceeded, the pump is said to stall. This is a disaster, because hot pump-fluid vapor is flushed backwards through the pump into the chamber. The usual practice is to maintain the oil-diffusion-pump foreline pressure below 50 mtorr.

The primary difficulty with diffusion pumps is the backstreaming of oil vapor from the pump inlet. Without appropriate precautions this can amount to as much as a microgram per minute for each cm<sup>2</sup> of pump inlet area. Backstreaming is much worse if the foreline pressure of a hot pump accidentally rises above the critical backing pressure of a few hundred millitorr. It is essential to maintain the pump cooling system. Most diffusion pumps are water-cooled. A filter in the cooling-water line is necessary to reduce the possibility of clogging. Cooling water should enter at the top of the pump so that the uppermost part of the pump wall is at the lowest temperature. The next line of defense is a “cold cap” typically consisting of a conical shield over the top of the jet tower. The cold cap is cooled by water flow or by solid thermal contact with the pump body. Diffusion pumps usually are supplied with a cold cap installed. In most installations a water- or refrigerant-cooled baffle (Figure 3.26) is mounted on top of the pump. The cold cap and baffle together reduce backstreaming by

**Table 3.3 Properties of Diffusion-Pump Fluids**

<i>Trade Names (chemical composition)</i>	<i>Boiling Point at 1 torr (°C)</i>	<i>Approximate Vapor Pressure @20°C (torr)</i>
(Mercury)	120	$2 \times 10^{-3}$
(Di- <i>n</i> -butyl phthalate)	140	$2 \times 10^{-5}$
Octoil, Difoil (di-2-ethylhexylphthalate)	200	$3 \times 10^{-7}$
Octoil-S, Difoil-S (di-2-ethylhexylsebacate)	210	$3 \times 10^{-8}$
Convoil-20 (saturated hydrocarbon)	190	$5 \times 10^{-7}$
Neovac Sy (alkyldiphenylether)	240	$1 \times 10^{-8}$
Convalex-10 (polyphenyl ether)	280	$7 \times 10^{-10}$
Santovac-5 (pentaphenyl ether)	280	$5 \times 10^{-10}$
D.C.704 (tetraphenyl tetramethyl trisiloxane)	240	$3 \times 10^{-8}$
D.C.705 (pentaphenyl trimethyl trisiloxane)	250	$5 \times 10^{-10}$
FomblinY-VAC 25/9 (perfluoropolyether)	230	$2 \times 10^{-9}$
Krytox1625 (perfluoropolyether)	250	$2 \times 10^{-9}$

as much as three orders of magnitude with a loss of pumping speed of perhaps 50%. For many purposes, the vacuum above the baffle can be considered oil-free. Oil can reach the vacuum chamber as a liquid film that creeps up the walls above the pump. To eliminate this problem, an anti-creep barrier consisting of an annular ring made of Teflon, a material not wetted by oil, can be installed on top of the baffle. The ultimate performance and the cleanest vacuum are obtained with the installation of a liquid-nitrogen-cooled trap (Figure 3.25) above the baffle.

The properties of diffusion pump fluids are given in Table 3.3; included are hydrocarbon, perfluorocarbon and silicone oils, and mercury. The various diffusion pump oils can be used interchangeably in an oil diffusion pump. The choice depends upon the desired ultimate pressure, the

demands of the particular application, and the cost. The matter of economics is significant; the cost of the oils listed in Table 3.3 ranges over more than a factor of 20. The perfluorocarbon and polyether oils are sufficiently costly that it is economical to reclaim them; many vacuum hardware companies offer this service. The absolute lowest pressure attainable with an untrapped diffusion pump is the room-temperature vapor pressure of the working fluid.

The silicone oils are the least expensive of the high-performance (i.e., low ultimate pressure) pump fluids. They are exceptionally resistant to oxidation and chemical attack, except in the presence of  $\text{BCl}_3$  and, to a lesser extent,  $\text{CF}_4$  and  $\text{CCl}_4$ . The main drawback to silicone oils is that the residue of backstreaming oil found on surfaces in the vacuum system will decompose and polymerize to produce insulating films upon exposure to heat and charged-particle bombardment; silicone oils are unsuitable for diffusion pumps on instruments such as mass spectrometers or electron microscopes. The phthalate and sebacate oils are inexpensive fluids suitable for many applications. Decomposition of these hydrocarbon oils yields carbonaceous deposits that are conductive. Polyphenylether oils offer exceptional chemical and thermal stability. These are well suited for use with mass spectrometers and other charged-particle-beam instruments. The perfluorocarbon oils are suitable for pumping corrosive gases, with the exception of Lewis acids such as  $\text{BCl}_3$  and  $\text{AlF}_3$ . Their performance under charged-particle bombardment is probably the best of all pump fluids.

Mercury diffusion pumps find only a few specialized uses that rely upon the chemical inertness of mercury vapor. They are occasionally used on mass spectrometers because of the simple, easily identified spectrum of mercury vapor. Mercury diffusion pumps are most often made of Pyrex glass and are intended for use with glass vacuum systems used to handle and distill small quantities of volatile chemicals. Mercury diffusion pumps tolerate inlet pressures 10 times greater than the maximum inlet pressure tolerated by oil pumps. In addition the critical foreline pressure for a mercury pump is much higher. Because the room-temperature vapor pressure of mercury is about  $10^{-3}$  torr, an inlet cold trap is required to prevent bulk migration of mercury into the vacuum system. Mercury vapor is toxic, and care must be taken to properly trap and vent the backing-pump exhaust.

Diffusion pumps with speeds of 50 to 50 000 L/s and with nominal inlet port diameters of 2 to 100 cm are available. Diffusion pumps are constructed of Pyrex glass, mild steel, or stainless steel. The jets and towers of pumps with steel barrels are aluminum. The choice of glass or steel usually depends upon the material that is used to construct the vacuum container, although it is also relatively simple to mate glass to steel with an elastomer gasket or through a metal-to-glass seal. For those systems that are intended to handle reactive gases, glass is the preferred material; however, glass pumps are available only in small sizes.

Diffusion pump maintenance involves assuring the quality and quantity of the pump fluid and checking that the heaters are operating. Some reaction of the pumped gas with the pump fluid is inevitable, so periodic inspection of the pump innards is essential. Moderate discoloration of the oil is not significant, however, if the oil is opaque it should be replaced. The interior of the pump must be thoroughly cleaned before adding new oil, especially if a different type of oil is being installed. Scrubbing with acetone, followed by an acetone and an ethanol rinse will remove hydrocarbon oils and, with persistence, silicone oils. Deposits of badly decomposed oil can be removed with naphthalene. Polyphenylether oils are soluble in trichloroethylene and 1,1,1-trichloroethane. Both these solvents should be used with care in a well-ventilated environment. Fluorocarbon oils are soluble in fluorinated alkanes such as trichlorotrifluoroethane and perfluorooctane. Severe decomposition of the oil results in black, carbonaceous deposits on the jet tower. These must be removed mechanically by scrubbing with fine abrasive or by glass-bead blasting. The tower must be carefully cleaned before reinstallation in the pump. There is a close fit between the tower and the pump body. A residue of abrasive cleaning materials will result in the tower becoming jammed in the pump body.

Diffusion pumps use flat “pancake” heating elements bolted to the bottom of the pump or cylindrical “cartridge” heaters inserted in close-fitting holes in the bottom of the pump. As many as six cartridge heaters wired in parallel may be used. The advantage to multiple heaters is that the pump continues to operate if one or more heaters burn out, albeit with some loss of efficiency. On the other hand, the loss of one or more heaters may go unnoticed until significant contamination of the vacuum has occurred. The individual cartridge heaters may be disconnected from one

another and individually checked for electrical continuity. It is essential that diffusion pump heaters be in good thermal contact with the pump over their entire mating surfaces or the heaters will burn out prematurely. A coating of milk of magnesia (a slurry of MgO in water) will assure good contact and facilitate removal at a later time by preventing the heater from sticking to the pump because of corrosion or cold welding.

A diffusion-pumped system is the most economical route to high and even ultrahigh vacuum in chambers of moderate size. With no moving parts and no critical dimensions, diffusion pumps are robust and long-lived. Decades-old pumps can be perfectly serviceable. This is a boon for researchers on a limited budget, since used pumps are available at little or no cost.

### 3.4.3 Entrainment Pumps

A variety of vacuum pumps remove gas from a system by chemically or physically tying up molecules on a surface or by trapping them in the interior of a solid. Two of the principal advantages of these pumps are that they require no backing pump and they contain no fluids to contaminate the vacuum.

**Sorption Pumps.** The simplest of the entrainment pumps is the sorption pump illustrated in Figure 3.13. The sorbent material is activated charcoal or one of the synthetic zeolite materials known as molecular sieves. These materials are effective sorbents partly because of their huge surface area – of the order of thousands of square meters per gram. The most common molecular sieves are zeolite 5A or 13X. The number in the sieve code specifies the pore size: 5A is preferred for air pumping, while 13X is used for trapping hydrocarbons. All of these materials will pump water and hydrocarbon vapors at room temperature, but they must be cooled to liquid-nitrogen temperature to absorb air. Sorbent materials do not trap hydrogen or helium at liquid-nitrogen temperature. Without some auxiliary means of removing hydrogen and helium, the partial pressures of these gases in a system will establish the lowest attainable pressure. Sorption pumps must be provided with a poppet valve because the sorbent material releases all of its absorbed air as it warms to room temperature.

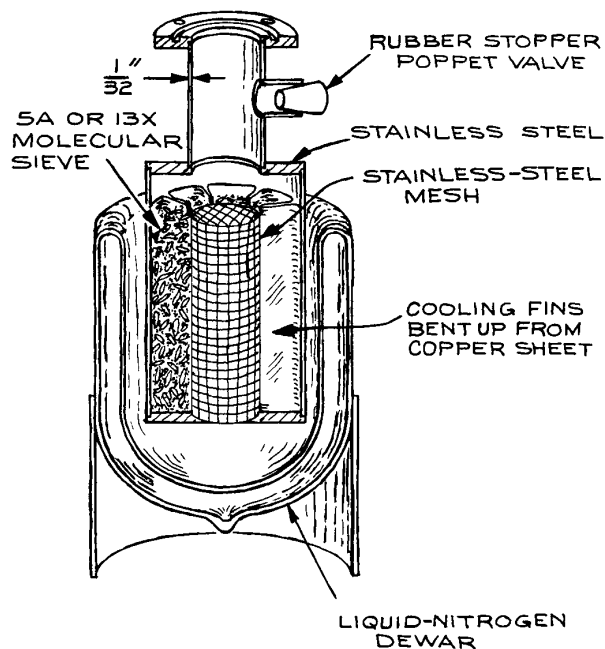


Figure 3.13 Sorption pump.

The sorbent is initially activated by baking to 250 °C. After several pumping cycles the pores of the sorbent material will become clogged with water and the efficiency of the pump will deteriorate. Water is removed by baking the pump to 250 °C with the poppet valve open and then cooling to room temperature with the valve closed so that moisture from the room is not reabsorbed. Baking can be accomplished by wrapping the pump with heating tape. Custom-made heating mantles can be obtained inexpensively from the Glas-Col Company.

In a well-designed pump, 1 kg of dry Linde 5A molecular sieve will pump a 100 L volume from atmosphere to less than  $10^{-2}$  torr in about 20 minutes. The pump must be designed to provide good thermal contact between the sieve and the coolant to obtain the maximum pumping speed. A thin-walled inlet tube will minimize heat conduction into the pump. Sorption pumps similar to that in the figure are available from suppliers of vacuum hardware. Pressures below  $10^{-6}$  may be attained in a system that has been roughed down. A simple way to achieve low pressures with sorption pumping is to use two or three pumps on a manifold. Each pump is equipped with a shutoff valve.

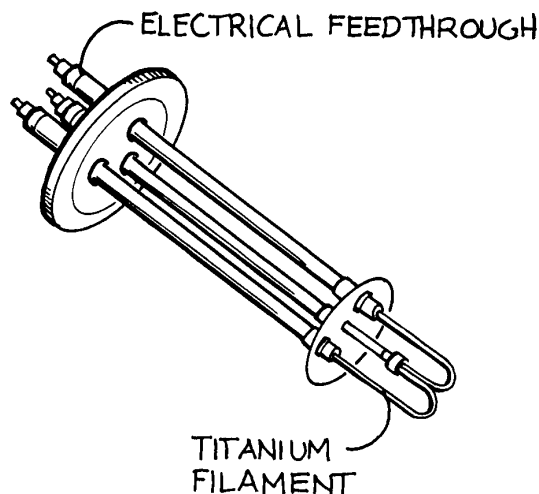
Starting at atmospheric pressure, one pump valve is opened to rough down the system. This pump should be of a size such that the pressure falls to less than 1 torr. This pump is valved off and the next pump is brought on line and so on. To attain the lowest possible pressure with this arrangement, the vacuum chamber can be purged of hydrogen and helium before evacuation proceeds. This can be accomplished by flushing the vacuum chamber with dry air or dry nitrogen before initiating pumping with the sorption pumps.

**Getter Pumps.** Clean surfaces of refractory metals such as titanium, molybdenum, tantalum, or zirconium will pump most gases by chemisorption. This process is called *gettering*. In a getter pump, the active metal surface is produced *in vacuo*. Titanium or a titanium/molybdenum alloy is the preferred metal. A getter pump employing a freshly evaporated titanium surface is called a *titanium sublimation pump*. At room temperature a titanium surface pumps  $H_2$ ,  $N_2$ ,  $O_2$ ,  $CO_2$ , and  $H_2O$  at rates of several L/s per  $cm^2$  of exposed surface. Cooling the surface to liquid nitrogen temperatures may double or triple the pumping speed for  $N_2$  and  $H_2$ . Methane and the noble gases, such as Ar and He, are not pumped at all.

A titanium sublimation pump consists of a titanium sublimator and the surface surrounding the sublimator. The surface may or may not be cooled. These pumps operate effectively between  $10^{-3}$  and  $10^{-11}$  torr, however, the inert gases must be removed before ultrahigh vacuum can be achieved. This can be accomplished with a small ion pump in parallel with the getter pump. An alternative method to (at least initially) remove rare gases is to purge the system with pure dry nitrogen prior to evacuation. The titanium surface must be renewed at a rate roughly equal to the rate at which a monolayer of gas is adsorbed. Down to about  $10^{-7}$  torr the titanium surface is continuously deposited. Below  $10^{-7}$  torr the titanium need only be deposited periodically. The capacity of a titanium sublimation pump is about 30 torr L per gram Ti. A forepump is not required, but a mechanical or sorption roughing pump is needed to reach a starting pressure of about  $10^{-3}$  torr.

Titanium sublimation pumps are quite simple in design and correspondingly inexpensive. Commercially available sublimators consist of two to four electrical feedthroughs in a flange with titanium filaments suspended between the feedthroughs, as illustrated in Figure 3.14. A controller

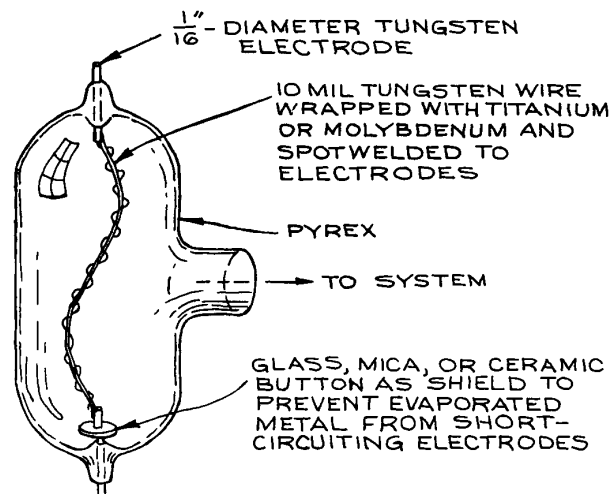




**Figure 3.14** Titanium sublimator on a UHV flange.

supplies electrical current through the filaments at a rate appropriate for the vacuum conditions. The sublimator can be inserted in a side arm on the vacuum system with baffles or an elbow arranged to prevent titanium vapor from entering the main part of the vacuum chamber. The walls of the sidearm then serve as the pump. The pumping speed with this arrangement will usually correspond to the conductance of the sidearm with its baffles. Vacuum housings with water- or liquid nitrogen-cooled baffles are available to contain the sublimator. These are designed to be mounted directly atop a turbomolecular pump in an ultra-high vacuum system. For a glass vacuum system, a simple pump can be made by wrapping titanium wire around a tungsten wire sealed in a glass bulb, as shown in Figure 3.15. The tungsten filament is electrically heated to evaporate the titanium that in turn condenses on the walls of the bulb.

A new class of getter pumps employing so-called “non-evaporable” getters has recently come into use (see, for example, SAES Getters, S.p.A.). These employ a porous, sintered, zirconium alloy that strongly absorbs and reacts with active gases. The getter is activated by heating to 700 to 900 °C *in vacuo* to remove any surface oxide. In operation the getter is maintained at a temperature of about 400 °C. CO, CO<sub>2</sub>, N<sub>2</sub>, and O<sub>2</sub> are irreversibly adsorbed. Water vapor and hydrocarbons are dissociatively adsorbed.



**Figure 3.15** Simple titanium sublimation pump as an appendage on a glass vacuum system.

H<sub>2</sub> is reversibly adsorbed; there is no sorption of the rare gases. Nonevaporable getter pumps operate down to 10<sup>-12</sup> torr. The pumping speed is highest for H<sub>2</sub> and H<sub>2</sub>O. These pumps are an excellent adjunct to a turbomolecular pump in a UHV system.

**Cryopumps.** Any surface will act as a pump for a gas that condenses at the temperature of the surface. A pump that relies primarily upon condensation on a cold surface is called a cryopump. Commercial versions of these pumps employ a closed-circuit helium refrigerator to cool the active surfaces. Working temperatures are typically below 20 K. These pumps usually employ two stages. In the first stage a metal surface is maintained at 30 to 50 K to trap water vapor, carbon dioxide, and the major components of air. In some designs a liquid-nitrogen-cooled baffle to reduce the load of water vapor on the subsequent stages precedes this first stage. The second stage, maintained at 10 to 20 K, is coated with a cryosorbent material such as activated charcoal to provide pumping of neon, hydrogen, and helium. Cryopumps provide a high pumping speed for the easily condensed gases. Their performance with helium depends critically upon the quality and recent history of the cryosorbent surface. The gas captured by a cryopump is not permanently bound. The pump must periodically be

warmed to regenerate the pumping surfaces. Because of the limited capacity of the pump, a roughing pump is necessary, as is a valve between the cryopump and the vacuum chamber so that the vacuum system can be rough pumped to a pressure less than about  $10^{-3}$  torr before cryopumping is initiated.

**Ion Pumps.** An ion pump combines getter pumping with the pumping action exhibited by an ionization gauge. Within these pumps a magnetically confined discharge is maintained between a stainless-steel anode and a titanium cathode. The discharge is initiated by field emission when a potential of about 7 kV is placed across the electrodes. After the discharge is struck, a current-limited power supply maintains the discharge rate at 0.2 to 1.5 A, depending on the size of the pump. Inert-gas molecules, and other molecules as well, are ionized in the discharge and accelerated into the cathode with sufficient kinetic energy that they are permanently buried. Active gases are chemisorbed by titanium that has been sputtered off the cathode by ion bombardment and deposited on the anode.

Ion pumps function between  $10^{-2}$  and  $10^{-11}$  torr. A forepump is needed to reach the starting pressure. Since ion pumps are completely free of hydrocarbons, an oil-free rough pump is called for. A well-trapped oil-sealed rotary pump can be used, but roughing with sorption pumps will guarantee that the vacuum chamber remains clean. An ion pump has a limited lifetime of operation due largely to the rate at which the cathode erodes during operation. The operational lifetime is an inverse function of pressure. Under continuous operation at  $10^{-3}$  torr, the lifetime of a typical pump is only 20 to 50 hours, whereas operation at  $10^{-6}$  torr can continue for 20 000 to 50 000 hours (6 yrs). At ultrahigh vacuum the lifetime is essentially infinite. Ideally, an ion pump is only operated below  $10^{-5}$  torr, but because this pressure is difficult to attain with most roughing systems, pumping can be initiated at  $10^{-3}$  torr, provided that the gas load is such that the pressure quickly falls into the desirable range. Ion pumps are frequently used in conjunction with titanium sublimation pumps. In fact, some ion pump designs include an integral sublimation pump. The sublimation pump can serve to lower the pressure from  $10^{-3}$  torr to  $10^{-5}$  torr before power is applied to the ion pump. At lower pressures the titanium

sublimation pump significantly increases the pumping speed for getterable gases. To attain the lowest possible pressures, an ion pump can be baked to at least 300 °C. Most pumps are equipped with integral heaters for this purpose. Baking should be initiated only while the pump is in operation.

The positive-ion current to the cathode of an ion pump is a function of pressure; thus the pump serves as its own pressure gauge. The power supply and control unit for most ion pumps includes a pressure readout. This is a significant consideration in the design of a very small ultra-high vacuum system; the cost of a 1 to 5 L/s ion pump with its control unit is little more than the cost of an ion gauge and controller.

Ion pumps require no cooling water or backing pump. They continue pumping by getter action even if the power fails. Ion pumps do not introduce hydrocarbon or mercury vapors into the vacuum. Pumps with speeds from one to many thousands of liters per second are available. The limited capacity in the  $10^{-3}$  torr to  $10^{-5}$  torr range can be a disadvantage. Stray magnetic and electric fields may also present a problem; ion pumps may produce a magnetic field of a few gauss at a distance of one inlet diameter. There may also be a problem with ions escaping the pump; ion emission can be suppressed by placing a suitably biased screen above the pump inlet. An ion pump with controller can cost two to three times as much as a comparable diffusion pump with an appropriately sized liquid-nitrogen-cooled baffle.

## 3.5 VACUUM HARDWARE

### 3.5.1 Materials

**Glass.** Two glasses are used in vacuum work: borosilicate glass or "hard glass" such as Pyrex or Kimax that is about 70%  $\text{SiO}_2$ , and quartz glass that is pure silica. Both materials have a low coefficient of thermal expansion and are remarkably resistant to fracture under large temperature gradients (thermal shock). Borosilicate glass has a broad softening temperature range allowing glass vacuum apparatus to be constructed and modified *in situ* by a moderately competent glassblower, as described in Chapter 2. Quartz has a very high softening temperature

and a narrow softening temperature range and requires a skillful glassblower. Quartz can be used in continuous operation at temperatures up to 1000°C. Quartz is used primarily when its high temperature strength or broad range of optical transmission is to be exploited. The permeability of helium through quartz is surprisingly large – about  $10^{-11}$  torr L/s mm/cm<sup>2</sup>/torr [throughput (torr L/s) through unit thickness (mm) per unit area (cm<sup>2</sup>) per unit helium pressure differential (torr)] at room temperature and increasing rapidly with increasing temperature. This can be a problem with quartz windows in UHV systems. Even when the permeation of atmospheric helium is not significant, the permeation of helium through a quartz window can give rise to false signals when helium leak-testing.

Hard glass tubing and glass vacuum accessories are inexpensive. Stopcocks and Teflon-sealed vacuum valves; ball joints, taper joints, and O-ring-sealed joints; traps; and diffusion pumps are all easily obtained at low cost. In most cases only the glassblowing ability to make straight butt joints and T-seals is required to make a complete system from standard glass accessories.

Glass pipe and pipe fittings are manufactured for the chemical industry. A variety of standard shapes, such as elbows, tees, and crosses, are available in pipe diameters of 1/2 to 6 in. Glass process pipe with Teflon seals can be used to make inexpensive vacuum equipment for use down to  $10^{-8}$  torr. Couplings are also available for joining glass pipe to metal pipe and to standard metal flanges and pipe fittings so it is easy to make a system that combines glass and metal vacuum components.

Glass and metal parts can be mated through a graded glass seal. Typically a graded seal appears to be simply a section of glass tubing butt-sealed to a section of Kovar metal tube. In fact, the glass tubing consists of a series of short pieces of glass whose coefficients of thermal expansion vary in small increments from that of hard glass to that of the metal. Graded seals are useful for joining glass accessories such as ion-gauge tubes to metal systems. Alternatively, a graded seal can be employed to insert metal hardware, such as a valve or an electrical feedthrough, into a glass vacuum line. Inexpensive graded seals in sizes up to 5 cm diameter are available from commercial sources.

Windows in vacuum systems are often made of quartz. For high-vacuum work the window can be sealed to the

vacuum wall with an elastomer gasket or O-ring. For ultrahigh vacuum, the edge of the quartz window is metalized so that the window can be brazed into a metal flange for installation in a vacuum wall. Window/flange assemblies are available from most vacuum equipment suppliers.

**Ceramics.** Ceramics are used as electrical insulators and thermal isolators. There are literally hundreds of different kinds of ceramic materials; two find frequent use in vacuum apparatus: steatite, a magnesium silicate, and alumina ceramics containing at least 96% Al<sub>2</sub>O<sub>3</sub>. Fabrication of parts from these materials requires specialized grinding equipment not ordinarily immediately available to a laboratory scientist, however, there are many specialized shops that provide grinding services under contract. Standard shapes, such as rod, tubing, plate, and spheres, are available in alumina (Coors Ceramics Co., McDanel Refractory Porcelain Co., Industrial Tectonics, Inc.). It is often possible to design an insulator that incorporates these simple shapes into a more complex assembly. Threaded steatite standoff insulators inexpensively available from electronics suppliers can be used as structural elements in vacuum apparatus. Corning produces MACOR, a machinable ceramic that can be fashioned into complex shapes with ordinary shop machine tools.

**Brass and Copper.** Brass has the advantage of being easily machined, and brass parts can be joined by either soft solder or silver solder. Unfortunately, brass contains a large percentage of zinc whose volatility limits the use of brass to pressures above about  $10^{-6}$  torr. Heating brass causes it to lose zinc quite rapidly. The vapor pressure of zinc is about  $10^{-6}$  torr at 170°C and  $10^{-3}$  torr at 300°C. Forelines for diffusion pumps are conveniently constructed of brass or copper tubing and standard plumbing elbows and tees. Ordinary copper water pipe is acceptable for forelines and other rough vacuum applications; however, oxygen-free high-conductivity (OFHC) copper should be used for high-temperature and high-vacuum work. A skilled technician can weld copper.

**Stainless Steel.** The most desirable alloys for the construction of high-vacuum and ultrahigh-vacuum apparatus are the (American AISI) 300 series austenitic

stainless steels, especially type 304 and type 316 stainless steels; type 303 is not ordinarily used as it contains volatile elements added to improve machinability. Stainless steel is strong, reasonably easy to machine, bakeable, and easy to clean after fabrication. 304 and 316 stainless steels are essentially nonmagnetic with magnetic permeabilities less than about 1.01. Stainless-steel parts may be brazed or silver-soldered. Low-melting (230 °C) silver – tin solders are useful for joining stainless-steel parts in the laboratory; however, it is best if stainless parts are fused together by arc welding using a nonconsumable tungsten electrode in an argon atmosphere. This process, known commonly as heliarc fusion welding or tungsten-inert-gas (TIG) welding, produces a very strong joint, and, because no flux or welding rod is used, such a joint is easily cleaned after welding. Most machine shops are prepared to do TIG welding on a routine basis.

The bulk of commercial vacuum fittings (flanges and shapes such as tees, crosses, and nipples) are fabricated of type 304 stainless steel. In addition, type 304 stainless is widely used in the milk- and food-processing industry. As a result, many stock shapes are commercially available at low cost. Elbows, tees, crosses, Ys, and many other fittings in sizes up to at least 6 in. diameter may be purchased from the Ladish Company or Alloy Products Company.

**Aluminum.** Aluminum alloys, particularly the 6000 series alloys, are used for vacuum apparatus. Aluminum has the advantage over stainless steel of being lighter and stiffer per unit weight, and much easier to machine. It is nonmagnetic. The chief disadvantages of aluminum stem from its porosity and the oxide layer that covers the surface. The rate of outgassing from aluminum is five to ten times greater than for stainless steel. Aluminum can be anodized black (black chrome) to reduce reflectivity in optical devices, however, anodizing may significantly increase outgassing rates. Welds in aluminum are not as reliable as welds in stainless and tend to outgas volatile materials that have been occluded in the weld; nevertheless, welded aluminum vacuum containers are used down to at least  $10^{-7}$  torr. The hard oxide layer that forms instantly on a clean aluminum surface is an electrical insulator. This insulating surface tends to collect electrons or ions to produce an electric field that

is undesirable in some applications. Copper or gold plating on aluminum or a coating of colloidal graphite (see Section 5.6.1) should eliminate the accumulation of surface charge.

**Plastics.** Plastics are used as fixtures, electrical insulators, bearings, and windows in vacuum systems at pressures down to  $10^{-7}$  torr. Their use is limited to varying degrees because they outgas air, water, and plasticizers, and because they cannot be heated to high temperatures. Polyamide (e.g., Nylon) and acetal-resin-based plastics (e.g., Delrin) are used for mechanical elements in vacuum systems at temperatures below 100 °C. These plastics are hard and machinable. Nylons are self-lubricating and make excellent bearings. Nylon is hygroscopic and outgasses water vapor after exposure to air. Outgassing is less of a problem with Delrin. Windows in vacuum systems may be made of polymethylmethacrylate (e.g., Plexiglas or Lucite) or polycarbonate (e.g., Lexan) plastics. Lexan is especially strong, tough, and machinable, well suited for use as a structural material or as an electrical insulator. Fluorocarbon polymers have superior high-temperature characteristics. Polytetrafluoroethylene (PTFE) (e.g., Teflon) can withstand temperatures up to 250 °C. The outgassing rate falls below  $10^{-8}$  torr L/s/cm after an initial pumpdown of a day at 100 °C. Unfortunately, Teflon is relatively soft, and cold-flows under mechanical pressure, thus limiting its usefulness as a structural element. Various polyimides, although expensive, are finding use in vacuum systems that are to be baked. Polyetherimide (Ultem) can be used at temperatures to 200 °C. Polyamide/imide (Torlon) is the highest strength thermoplastic. Polyimide (Vespel, Kapton) is the ultimate material for both mechanical elements and electrical insulation at high temperatures. Vespel has been used continuously at temperatures approaching 300 °C in high vacuum. Polyimide is somewhat hygroscopic, but can be baked to reduce outgassing. Kapton is polyimide sheet used for electrical insulation. Kapton-insulated wire is supplied for high-temperature use in high vacuum and even ultrahigh vacuum.

A variety of low-vapor-pressure sealers and adhesives have vacuum applications. Apiezon M grease, Dow-Corning silicone high-vacuum grease, and FOMBLIN perfluorinated grease are used to seal stopcocks and ground-glass taper

joints and, in some instances, as low-speed lubricants. Apiezon compounds are used to seal joints and windows as well as to fill small gaps in vacuum systems. These wax-like materials soften at temperatures of 45 to 100 °C and have vapor pressures between  $10^{-4}$  and  $10^{-8}$  torr at room temperature. A variety of resin-based sealers in spray cans are marketed for sealing very small leaks in high-vacuum and even ultrahigh-vacuum systems while they are under vacuum.

Epoxy resins are particularly useful. Epoxy cements consist of a resin and a catalytic hardener that are combined immediately before use. The proportions of resin and hardener and subsequent curing must be carefully controlled to prevent excessive outgassing from the hardened material. Small epoxy kits with resin and catalyst prepackaged in the correct proportions are available, as are silver-filled epoxy formulations that conduct both electricity and heat.

### 3.5.2 Demountable Vacuum Connections

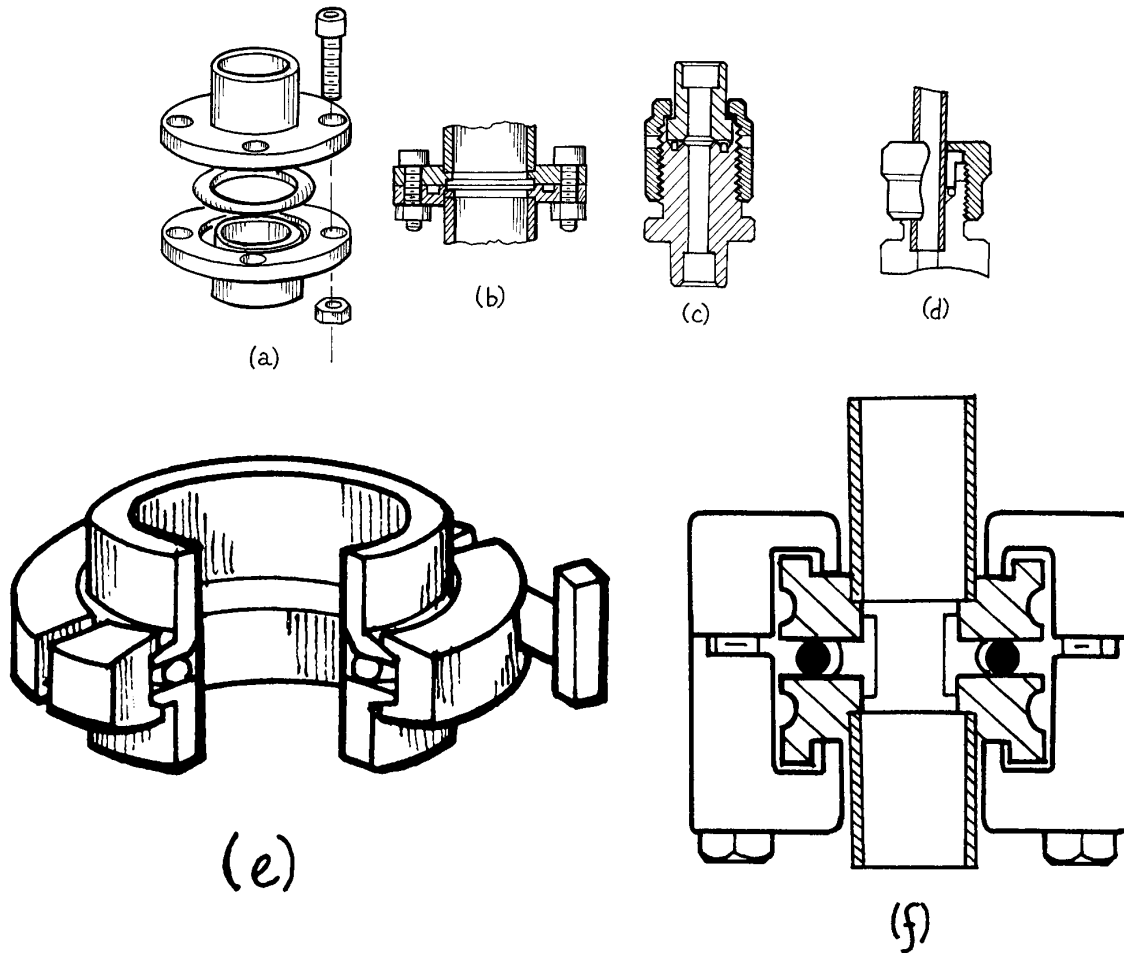
Vacuum systems require detachable joints for convenience in assembling and servicing. There are basically three means of creating a vacuum-tight seal between mating connectors: pipe threads with a sealant in the joint, elastomer gaskets or “O”-rings, and deformable metal gaskets.

**Pipe Threads.** The threaded parts of a pipe joint are tapered so that the joint becomes tighter and tighter as the two are screwed together. The two parts are not made with sufficient precision to assure a vacuum tight seal even though significant distortion of the metal threads can occur as the parts are joined. To assure a seal, the male part is covered with a thin layer of Teflon tape prior to assembly. Teflon thread tape, or thread “dope,” is sold in hardware stores for use with plumbing fixtures. The tape is wrapped around the threaded end. No more than two layers are required and the tape should be stretched as it is wound on to make a very thin layer. Looking at the end of the male part, the tape should be wound in a counterclockwise direction so that the exposed tag end does not get caught up in the female threads as the joint is assembled. The tape must be replaced upon reassembly. A significant volume of gas is trapped in an assembled pipe thread joint. Pipe thread joints are not appropriate for ultrahigh-vacuum

systems, and their use is to be avoided as much as possible in high-vacuum systems.

**“O”-Ring Joints.** For pressures down to about  $10^{-7}$  torr, vacuum connections are most often sealed with rubber O-rings. Several O-ring-sealed joints are illustrated in Figure 3.16. O-rings are circular gaskets with a round cross-section. They are available in hundreds of sizes from 3 mm (.125 in.) I.D. (inner diameter) with a cord diameter of 2 mm [.070 in. (nominally 1/16 in.)] to 60 cm (2 ft.) I.D. with a cord diameter of 7 mm [.275 in. (nominally 1/4 in.)]. Very large rings may be made from lengths of cord stock with the ends butted together and glued with cyanoacrylate adhesive (Eastman 910). O-rings are made of a variety of elastomers. The most common are Buna-N, a synthetic rubber, and Viton-A, a fluorocarbon polymer. Buna-N may be heated to 80 °C and will not take a set after long periods of compression. Unfortunately this material outgasses badly, particularly after exposure to cleaning solvents. Viton-A has a low outgassing rate and will withstand temperatures up to 250 °C, but will take a set after baking. Generally, the advantages of Viton-A O-rings offset their higher cost.

O-ring-sealed flanges and “quick connects” are easily fabricated and they are also available ready-made. Mating flanges have a groove that contains the ring after the flanges have been pulled into contact with one another. Usually, one flange is flat and a rectangular-cross-section groove is cut into the mating flange; flanges can be made sexless by cutting a groove of half the required depth in both flanges. Groove design is critical to obtaining a reliable seal, as is the surface finish of the flange and inside the groove. Surface roughness should not exceed  $1 \mu\text{m}$  (32 microinches). The flange surface and groove should be cut on a lathe so that tool marks run circumferentially. Radial scratches are one of the most common causes of seal failure. To obtain the correct compression of the O-ring material, the groove depth should be about 80% of the actual cord diameter for static seals and 85% of the diameter for dynamic seals. Rubber is deformable, but incompressible; the width of the groove should be such that the cross-sectional area of the groove exceeds that of the O-ring by 30 to 40%. The ID of the groove should match the O-ring ID; the O-ring is not to be stretched into



**Figure 3.16** O-ring-sealed vacuum connections: (a) an exploded view of an O-ring-sealed flange joint; (b) assembled ASA-style flange joint; (c) O-ring tube coupling (Cajon VCO); (d) a quick connect; (e) QF(or KF) joint; (f) ISO joint with K-style flanges.

its groove. It is sometimes convenient to undercut the sides of the groove slightly to give a closed dovetail cross-section rather than a rectangular cross-section, so that the ring is retained in the groove during assembly. For rings up to 30 cm (12 in.) in diameter, the nominal cord diameter should be 3 mm (1/8 in.) or less. Choose a ring to fit in a groove that is as close to the flange ID as possible, in order to minimize the amount of gas trapped in the narrow space between the mated flanges. The bolts or clamps that pull the flanges together should be as close to the groove as

possible to prevent flange distortion. For static seals, O-rings should be used dry. Before assembly, the groove should be cleaned with solvent and the ring wiped free of mold powder with a dry lint-free cloth. O-rings should not be cleaned with solvents. For rotating seals a very light film of vacuum grease on the ring will prevent abrasion. Occasionally a film of grease may be required on a static O-ring to help make a seal on an irregular surface. Grease is a liability, however, as it tends to pick up debris that may disrupt the seal.

For very high temperatures, elastomer O-rings can be replaced by metal O-rings composed of a thin metal sleeve with an internal spring to provide resiliency. These are manufactured by Helicoflex in aluminum, silver, copper, nickel, and other materials suitable for temperatures from 300 to 600 °C and in sizes comparable to those of standard rubber O-rings. The groove design is essentially the same as for rubber O-rings; detailed design instructions are available from the manufacturer.

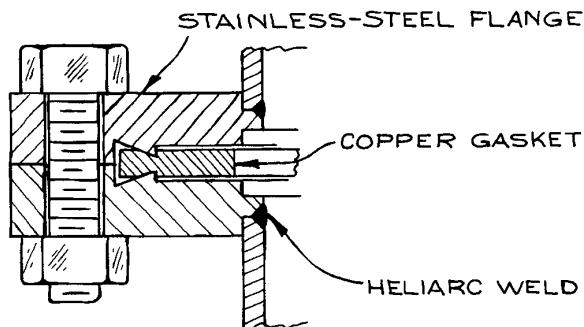
There are three standard types of O-ring sealed joints: ASA, QF (or KF), and ISO. Examples are shown in Figure 3.16. The ASA design dates back to an early steamfitters standard flange. ASA design consists of two flat flanges with an O-ring groove in one flange and a circle of bolt holes well outside of the groove. These flanges are bulky and require fasteners much larger than needed to resist atmospheric pressure. The sizing of ASA flanges refers to the nominal ID of the flange or the nominal OD of the tube to which it is attached: the bore of a 4 in ASA flange is about 4 in; the OD of the flange is 9 in. Some commercial vacuum hardware, including large diffusion pumps, large mechanical pumps and Roots blowers, still use ASA flanges.

The QF (quick flange) or KF (Klein flange) design employs two, flat, and therefore sexless, flanges with an outer diameter only slightly larger than that of the O-ring. The O-ring is mounted around a separate, aluminum “centering ring” that prevents the O-ring from being forced inward by external pressure and whose thickness determines the extent of compression of the O-ring. The centering ring has a lip on both sides that engages the inner circumference of each flange so that the O-ring is correctly located on the flanges. A jointed, circular clamp is installed around the assembled pair of flanges and a thumbscrew is tightened to draw the two flanges together, compressing the O-ring. They are compact and, with only a single screw to tighten, quite convenient to use; a QF fitting can be assembled with one hand. QF flanges are available in brass, aluminum, and type 304 and 316 stainless steel in configurations suitable for a brazed or welded butt joint or socket joint to tube sizes from 10 mm (1/2 in) to 50 mm (2 in). A remarkable range of components are available with QF fittings, including nipples, elbows, tees, 4-, 5-, and 6-way crosses, as well as adapters to mate QF with other styles of connectors. In the spirit of “Tinker Toys,” it

is often possible to assemble a large part of a vacuum system using only these standard components.

ISO flanges are similar to QF flanges but are sized for larger tubing: nominal bore sizes range from 63 to 630 mm. The ISO fitting is comprised of two flat flanges with a combination centering ring and O-ring mounted between them. The centering ring has a lip on each side that engages a counterbore near the ID of the flange. ISO flanges are designed for one of two methods of joining a pair of flanges. K-style ISO flanges employ clamps that hook over the outer edges of the mating flanges. F-style flanges have holes for bolts to draw the two flanges together. K-style flanges are more compact than F-style flanges. F-style flanges make a stronger structure appropriate for joining tubing that must support large transverse forces. An all-metal C-Flex assembly manufactured by EG&G Engineered Products can replace the centering ring/O-ring assembly in a QF or ISO joint. These employ a springy, tin-coated, Inconel element to replace the elastomer O-ring. They can be baked at modest temperatures and are said to be usable at UHV pressures. The A&N Corporation ULRIC™ flange system is similar in configuration to the ISO flange system and can incorporate a copper gasket to make a fitting that is bakeable and useable for ultrahigh vacuum.

**Metal Gaskets.** Metal sealing materials are required for ultrahigh-vacuum (UHV). Elastomers are unacceptable because they outgas and because they cannot be baked to high temperatures. The ConFlat design developed by Varian has become the industry standard. Flanges of this design, now generally known as CF (for ConFlat-compatible) flanges, are now manufactured by many vacuum equipment suppliers. The CF sealing system is composed of two identical flanges and a flat OFHC copper gasket, as shown in Figure 3.17. Annular knife-edge ridges in each flange cut into the gasket to make a seal as the flanges are drawn together by bolts that pass through holes in the flanges outboard of the copper gasket. The entire assembly can be baked to 450 °C. A great deal of force is required to make the seal, so many high-tensile-strength bolts are used, closely spaced around the flange. The knife-edge is designed so that the correct deformation of the gasket is obtained when the flanges are drawn into contact. Careful assembly is important to avoid warping

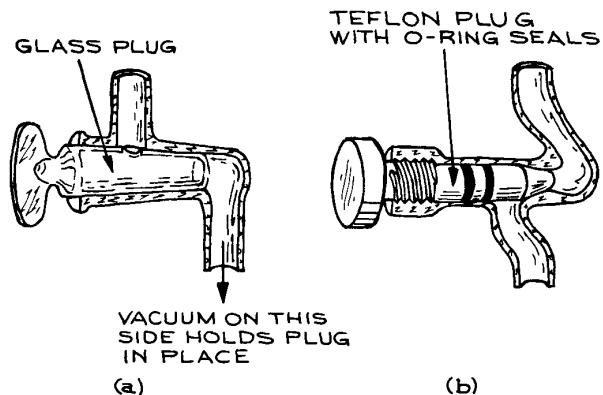


**Figure 3.17** Detail of a bakeable, ultrahigh-vacuum CF flange.

the flanges. The joint is initially assembled with all the bolts just finger tight. Then, using good-fitting wrenches and working in sequence around the flange, each bolt is tightened no more than one half turn. Going round and round the flange, the bolts are drawn up just until the flange surfaces touch. The copper gaskets are permanently deformed during installation and hence are used only once. Bolts wear and stretch and must be replaced after perhaps a dozen assembly cycles. Baking a CF flange encourages the bolts to seize in their nuts. A very thin coating of molybdenum disulfide grease helps avoid this problem. Silver-plated bolts are available to help reduce galling and seizing in systems that are baked frequently to high temperatures. CF flanges are specified by the flange OD in sizes that run from 1.33 in (or 34 mm) to 12 in (305 mm) to 16 1/2 in. These are used with tube sizes from .75 in (or 16 mm in metric sizes) to 10 in (or 250 mm) to 14 in. In UHV systems, circular cross-section tubes larger than 14 in diameter are joined with Wheeler flanges that employ a 2 mm diameter OFHC copper wire as a deformable gasket. Wire-sealed flanges are heavy and very large forces are required to achieve a seal. They are available from many manufacturers in sizes up to about 30 in.

### 3.5.3 Valves

A wide variety of valves are available commercially for use in glass systems and in high-vacuum and ultrahigh-vacuum metal systems. Characteristics to be considered in choosing a valve (roughly in order of importance) include:



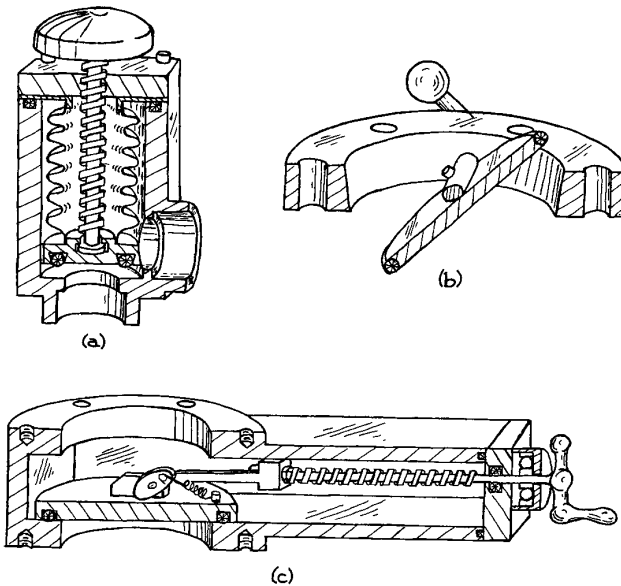
**Figure 3.18** Glass vacuum valves: (a) a stopcock; (b) a glass valve with Teflon plug.

(i) conductance; (ii) operating temperature; (iii) leak rate around the valve-operating mechanism and the valve seat; (iv) material of construction [related to (ii)]; (v) configuration; (vi) cost; (vii) reliability; (viii) ease of maintenance.

Metal valves are always used in metal vacuum systems. The construction of glass systems is easiest if glass valves are used; however, glass valves (stopcocks) have relatively low conductance. The clear passage through an open stopcock is typically only a few millimeters in diameter. When a large-bore, high-conductance, valve is required in a glass system, a metal valve of suitable size can be inserted using a glass-to-metal transition on each side of valve. The two most common glass valves are illustrated in Figure 3.18. One (a) is a vacuum stopcock consisting of a tapered glass plug that is fitted into a glass body by lapping. The mating parts are sealed with a film of high-vacuum grease. As shown, these stopcocks are designed so that atmospheric pressure forces the plug into the body of the valve. The more modern type of glass valve (b) has a Teflon plug threaded into a glass body. The plug is sealed to the body with an O-ring. It is generally advisable to apply a light film of vacuum grease to the O-ring.

There are three basic designs for metal vacuum valves – an example of each is shown in Figure 3.19. The differences between the three designs have to do with the motion of the sealing plate as it approaches the valve seat. The valve in Figure 3.19(a) is a piston-type, or poppet, valve in which an actuator moves the sealing plate axially as the





**Figure 3.19** Metal high-vacuum valves: (a) bellows-sealed valve; (b) swing or “butterfly” valve; (c) sliding-gate valve.

valve is opened and closed. The piston valve is inherently a “right-angle” valve with the inlet and outlet ports at  $90^\circ$  to one another; the design may, however, be modified to a “quasi-straight-through,” or “in-line,” valve by introducing a bend in the downward-facing inlet port, so that it enters on the side opposite the outlet port with its centerline parallel to that of the outlet port. The conductance of inline valves is not as high as might be expected, since the flow through the valve encounters two right-angle turns. The valve shown in Figure 3.19(b) is a swing valve or “butterfly” valve. The sealing plate rotates about an axis along its diameter as the valve is open and closed. The actuator mechanism need only rotate one quarter of a turn. The obvious advantage of the butterfly valve is the low profile that provides the shortest possible path for gas and hence the highest possible conductance. The butterfly valve must be installed so that there is clearance for the valve plate in the open position. The valve shown in Figure 3.19(c) is a sliding-gate valve. The sealing plate moves radially with respect to the valve seat and the direction of gas flow; a cam drives the sealing plate against the valve seat once the plate is positioned above the seat. Vacuum

valves of the three basic designs are available with bores ranging from less than a centimeter to more than 20 cm.

Vacuum valves are fabricated of brass, cast aluminum, machined aluminum, or stainless steel. The cost increases, and the low-pressure limit of usability decreases, in the same order. Brass contains the volatile metal zinc and hence should not be used above room temperature. Small piston-type valves are often manufactured of brass. These are suitable for rough vacuum applications, such as diffusion pump forelines. Cast aluminum is porous and out-gases in vacuum. Inexpensive cast aluminum valves can be used in high vacuum. Valves manufactured of high-quality, vacuum-cast aluminum can be used down to at least  $10^{-9}$  torr. Valves made of aluminum machined from solid stock, with the appropriate seals, find some use at ultrahigh vacuum. Stainless steel is the standard for valves used in high vacuum and UHV. In addition to a much lower outgassing rate (compared to aluminum), stainless steel withstands higher bakeout temperatures and is much more resistant to corrosive gases such as are used in semiconductor processing. The vacuum seal between inlet and outlet of most vacuum valves is made by an O-ring installed in the valve plate coming into contact with a smooth seating surface machined into the valve body. Viton is the standard O-ring material. A valve with Viton seals can be baked to  $200^\circ\text{C}$ . Valves with a polyimide seal are available for use up to  $300^\circ\text{C}$ . The sealing materials in a true UHV valve must be metal. Most designs employ a copper pad on the actuator plate that is driven onto a stainless-steel knife-edge seat in the valve body. An all-metal UHV valve can be baked to  $450^\circ\text{C}$ . The soft copper sealing pad is effective for a limited number of cycles and must be replaced frequently. This problem, along with the relatively high cost of UHV valves, encourages UHV system designs with as few valves as possible.

All valves require an actuator mechanism that passes from the atmosphere to the evacuated innards of the valve. In butterfly valves this is a rotating shaft; in some piston-type valves and gate valves, a rotating shaft passes through the vacuum wall to drive a screw mechanism that provides the linear motion required to move the valve plate; in other designs the linear motion is transmitted through the vacuum wall by a linearly translating shaft. In butterfly valves and inexpensive piston and gate valves, the actuator shaft is sealed with an O-ring. Leakage around this seal is

inevitably the source of vacuum failure in these valves. In high-quality piston and gate valves, linear motion is transmitted through the vacuum wall and the drive mechanism is located outside the vacuum. A bellows that flexes to follow the driver in and out as the valve is closed and opened seals the actuator shaft. The life of a bellows shaft seal ranges from 10 000 to more than 100 000 cycles.

The drive mechanism in a vacuum valve may be mechanical, electromagnetic, or pneumatic. The most robust actuators are driven by a rotating handwheel or crank. These designs make for slow opening and closing, since the valve plate in a vacuum valve must be moved over a relatively large distance to provide good conductance when the valve is open. A large gate valve may require as many as 30 turns to open or close. Some piston and gate valves employ a toggle mechanism. The toggle handle swings through 90° or 180°. Considerable force is required at the end of the swing to affect a seal so the valve body must be mounted very securely. Valves with toggle actuators are rarely used in the laboratory. Valves with electromagnetic actuators are available for use with tubing of up to 40 mm or 1.5 in. OD. These valves are actuated by an electrical current to a solenoid that drives a plunger that opens the valve. The plunger is spring-loaded so that the valve automatically closes if power is lost. Piston-type valves and gate valves of all sizes are available with pneumatic actuators. Compressed air is admitted to a cylinder to drive a piston that is connected to the valve actuator. The motion is reversed by venting air on one side of the piston and admitting compressed air to the opposite side. An electromagnetic valve in the compressed air circuit controls the airflow. The electromagnetic valve is designed so that air is directed to the side that closes the valve when electrical power is lost. Some small valves use a simpler pneumatic circuit employing a spring to close the valve. Pneumatic valves typically require compressed air at 0.5 to 0.7 MPa (80 to 100 psi).

It is often necessary to admit gas at a low rate into a vacuum system. The gas source may be at a high pressure. For example, the pressure in a commercial gas cylinder may exceed 10 MPa (1500 psi). Metering valves or “leak” valves are available for this purpose. Valves providing flow rates spanning several decades and with minimum controllable flow rates as low as  $10^{-10}$  torr L/s are available. There are a variety of designs. Some are simply needle

valves employing a long, gently tapering needle moving in a conical seat. Another design incorporates a metal or sapphire plate moving against a metal knife-edge. All provide a more or less precise micrometer drive to control the moving element. In general, a leak valve should never be used to shut off the flow of gas, as deformation of one of the two elements will result in a loss of low flow control. A shutoff valve, if required, should go on the upstream side of a leak valve and as close to the inlet as possible to minimize the “dead” volume of gas that must be pumped through the leak valve. A sintered metal filter should always be used at the inlet of a leak valve. The smallest piece of solid material compressed against the seat of a leak valve will certainly cause irreversible damage.

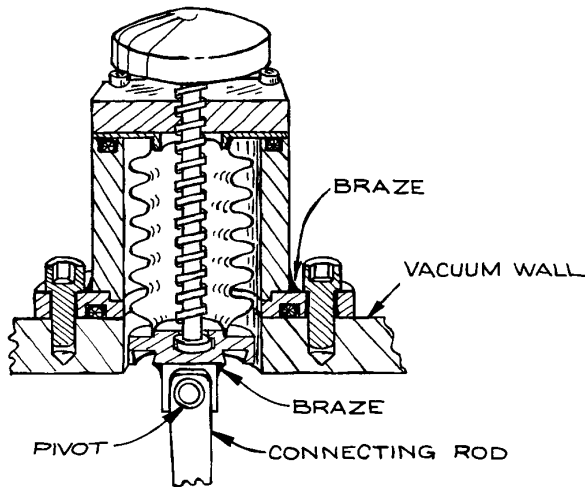
### 3.5.4 Mechanical Motion in the Vacuum System

Mechanical motion can be transmitted through the vacuum wall using a sliding seal, a flexible wall, or a magnetic coupling.

Rotary motion at speeds up to 100 rpm may be transmitted through an O-ring-sealed shaft. These seals are inexpensive, but may fail if the O-ring becomes abraded. In addition, lubricants and the elastomer O-ring material are exposed to the vacuum.

A bellows seal to a shaft is the most common means to permit motion to pass through the vacuum wall. A simple and inexpensive linear-motion feedthrough can be made from the actuator mechanism of a small bellows-sealed vacuum valve. As shown in Figure 3.20, the valve body is truncated above the seat, a mounting flange is brazed to the body, and a fixture is brazed or screwed to the valve plate for attaching the mechanism to be driven inside the vacuum. Rotary motion may be transmitted through a bellows-sealed “wobble drive,” illustrated in Figure 3.21. Relatively inexpensive rotary drives of this type are available from most vacuum-hardware suppliers. Much more expensive and more complex drives are also available, including drives that provide three-axis rotation and three-axis translation.

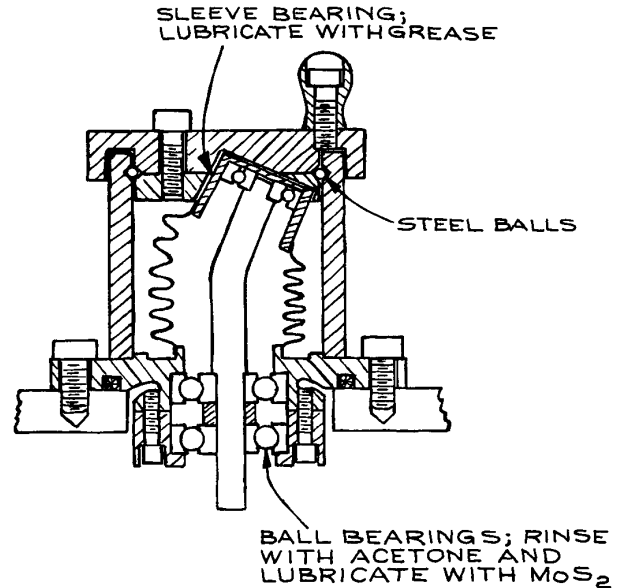
Moving parts can be magnetically coupled through a vacuum wall. This scheme is particularly useful for the transmission of linear motion through the wall of a glass vacuum system. The driven element is attached to a block



**Figure 3.20** A bellows-sealed valve [Figure 3.19(a)] converted to a linear-motion feedthrough.

of magnetic material such as iron or nickel that is placed against the inside of the vacuum wall. A magnet can then drag the driven element along the wall. A more exotic application of magnetic forces is found in the use of ferro-magnetic fluids as shaft seals held in place by magnetic fields. Ferrofluidic shaft seals rated for 5000 rpm at  $10^{-9}$  torr are commercially available (A&N Super-Seal™).

Metal surfaces become very clean *in vacuo*, particularly after baking, and metals in close contact tend to cold-weld. Because of this, unlubricated bearing surfaces within a vacuum system often become very rough after only a little use. Liquid lubricants can be used if they are thoroughly outgassed before installation. The vapor pressure of the lubricant will be the limiting factor. Diffusion pump oils can be used in diffusion-pumped vacuum systems, since the ultimate pressure is determined by the room-temperature vapor pressure of the oil in the pump. Polyphenylether oils lubricate well and are among the lowest-vapor-pressure oils. Perfluorinated oils, such as DuPont Krytox 143AZ, are an excellent choice for liquid lubrication in a vacuum. A difficulty with liquid lubricants is their tendency to creep out of a bearing and along the surface of a shaft. A creep barrier can be placed between a lubricated bearing surface and the surroundings. This may consist of a coating or a plastic lip of a material such as Teflon that is not wetted by the lubricant.



**Figure 3.21** Bellows-sealed, wobble-drive, rotary-motion feedthrough.

There are a number of methods of improving bearing performance in a vacuum system without introducing high-vapor-pressure oils into the vacuum. The tendency for a bearing to gall is reduced if the two mating bearing surfaces are made of different metals. For example, a steel shaft rotating without lubrication in a brass or bronze journal will hold up better than in a steel bushing. A solid lubricant may be applied to one of the bearing surfaces. Silver, lead – indium, and molybdenum disulfide have been used for this purpose. Graphite does not lubricate in a vacuum;  $\text{MoS}_2$  is probably best. The lubricant should be burnished into the bearing surface. The part to be lubricated is placed in a lathe. As the part turns, the lubricant is applied and rubbed into the surface with the rounded end of a hardwood stick. By this means, the lubricant is forced into the pores. After burnishing, the surface should be wiped free of loose lubricant.

One component of a bearing may be fabricated of a self-lubricating material such as nylon, Delrin, Teflon, or polyimide. Teflon is good for this purpose, but its propensity to cold-flow will cause the bearing to become sloppy with time. Polyimide can be used in ultrahigh vacuum after baking to 250–300 °C. Brown, Sowinski, and Pertel<sup>6</sup> have

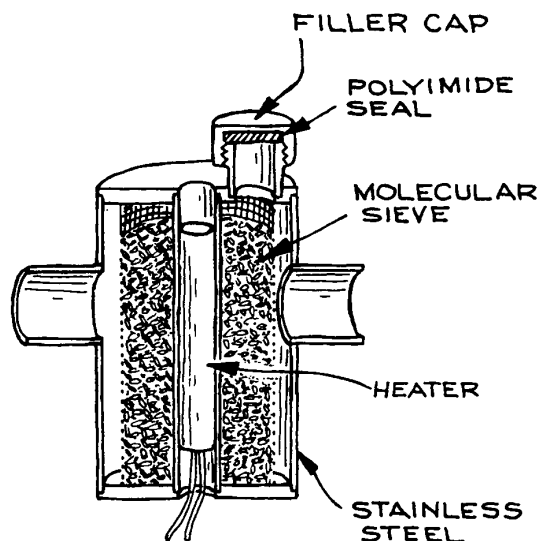
overcome the cold-flow problem in the design of a drive screw for use in a vacuum. Both the screw and its nut are steel, and lubrication is accomplished by placing a Teflon key in a slot cut in the side of the screw. Teflon is then continuously wiped onto the screw threads as the screw turns.

For very precise location of rotating parts and for high rotational speeds, ball bearings are required. Precision, bakeable ball bearings can be fabricated using sapphire balls running in a stainless-steel race. Inexpensive precision sapphire balls are available from Industrial Tectonics. A ball retainer is required to prevent the balls from rubbing against one another (see Section 1.7.3). Without grease or oil lubrication, it is helpful to burnish the race with  $\text{MoS}_2$ .

For high-speed applications a fluid lubricant is necessary. The following process has been developed at NASA-Goddard Space Flight Center.<sup>7</sup> Purchase high quality stainless-steel ball bearings with side shields and phenolic ball retainers, such as the New Hampshire PPT series or Barden SST3 series. Remove the shields and leach clean the phenolic by boiling the bearings in chloroform – acetone, and then vacuum-dry at 100 °C. The bearings are then lubricated by impregnating the phenolic with DuPont Krytox 143AZ, a fluorinated hydrocarbon. Impregnation is accomplished by immersing the bearing in the lubricant and heating to 100 °C at a pressure of 1 torr or less until air stops bubbling out of the phenolic. After this process the bearing must be wiped almost dry of lubricant. The Tex-wipe Company makes ultraclean foam cubes for this process. The amount of lubricant in the bearing is determined by weighing before and after impregnation. About 25 mg of Krytox is required for an R4 (1/4 in.) bearing, and about 50 mg is needed for an R8 (1/2 in.) bearing. This lubrication process should be carried out in a very clean environment, and the bearing should be inspected under a microscope for cleanliness before the side shields are replaced. Bearings treated in this manner have been run at speeds up to 60 000 rpm *in vacuo*.

### 3.5.5 Traps and Baffles

Traps are used in vacuum systems to intercept condensable vapors by means of chemisorption or physical condensation. When an oil-sealed mechanical pump is used as the backing pump for a diffusion pump or a turbomolecular



**Figure 3.22** A foreline trap filled with molecular sieve.

pump, a trap is placed in the foreline between the backing pump and the high vacuum pump to prevent backstreaming of mechanical pump oil vapor. This is absolutely essential for the creation of a hydrocarbon-free vacuum. In some applications, such as vacuum processes employed in the semiconductor industry, the foreline trap is intended to condense corrosive gases from the process chamber in order to prevent harm to the forepump. In addition, a trap is placed between a diffusion pump and a vacuum chamber to pump water vapor and to remove diffusion-pump fluid vapors that migrate backwards from the pump toward the chamber.

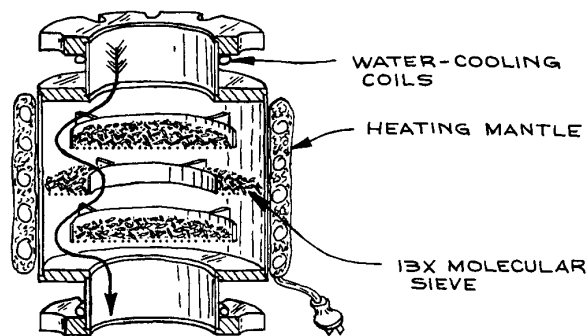
In many foreline traps, an ultraporous artificial zeolite known as molecular sieve is employed to adsorb oil vapor. These traps are similar in design to the molecular-sieve sorption pumps previously described, except, of course, that a trap must have both an inlet and an outlet. A typical foreline trap filled with zeolite 13X molecular sieve is shown in Figure 3.22. The molecular sieve is initially activated by baking to 300 °C for several hours. In use the sieve material is regenerated at intervals of about two weeks by baking under vacuum to drive out absorbed oil and water. Care must be taken to prevent the oil driven out of the trap from condensing upstream of the trap, thus

doing more harm than good. A valve on the upstream side of the trap can be closed to prevent oil vapor from back-streaming, provided that the valve and the connection to the trap are themselves heated to prevent condensation while the trap is being baked. It is also possible that vapor driven from the trap will contaminate the forepump. Some trap designs provide a third connection for a separate pump to be used during the bakeout. Another trap design provides a removable basket for the sieve charge that can be baked in a separate oven. For a system that must be kept in continuous operation, two traps in parallel are used. Valves on each side of both traps allow one trap to be isolated for regeneration. An auxiliary pump (with suitable valving) is used to evacuate the trap being regenerated. Some traps are provided with a third port for connection to this auxiliary pump.

The lowest pressure attainable with a two-stage oil-sealed mechanical pump is initially determined by the vapor pressure of the oil in the pump. By using a molecular-sieve trap in series with a mechanical pump to remove oil vapor, it is possible to achieve pressures as low as  $10^{-4}$  torr in a small system with no gas load. This is an effective scheme for maintaining a vacuum for thermal insulation.

Molecular sieve may also be used in a high-vacuum trap over the inlet of an oil diffusion pump. As shown in Figure 3.23, these traps are designed to be optically opaque, so that a molecule cannot pass through the trap in a straight line. This precaution is necessary because this type of trap is intended for use at pressures where the mean free path of a molecule is very long. To ensure high conductance, the inlet and outlet ports should have the same cross-sectional area as the inlet of the attached pump. Also, the cross-section perpendicular to the flow path through the trap (indicated by an arrow in Figure 3.23) should be at least as large as that of the inlet and outlet ports. The molecular sieve is activated by baking under vacuum to  $300^{\circ}\text{C}$  for at least six hours. The trap may be heated with heavily insulated heating tape or a custom-made heating mantle such as those obtained from the Glas-Col Apparatus Company. The flanges of the trap are water-cooled to prevent overheating the seals.

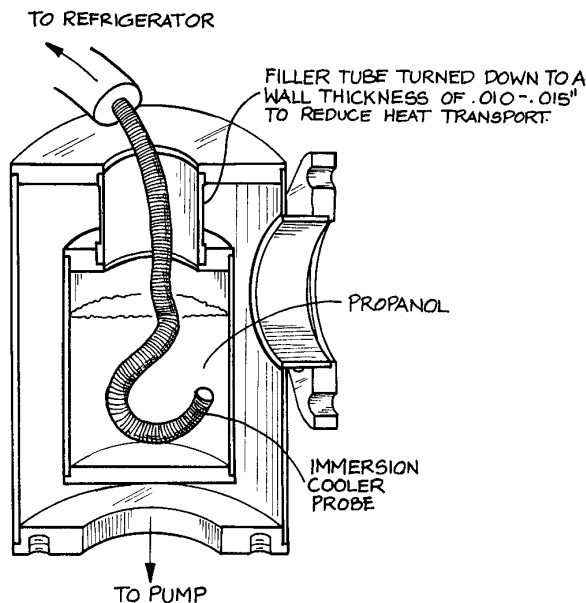
The maintenance of a high-vacuum trap is different from that of a foreline trap. A high-vacuum molecular-sieve trap is placed above an oil diffusion pump, and a gate valve is located above the trap to permit isolation of



**Figure 3.23** High vacuum molecular-sieve trap for use over the inlet of a diffusion pump.

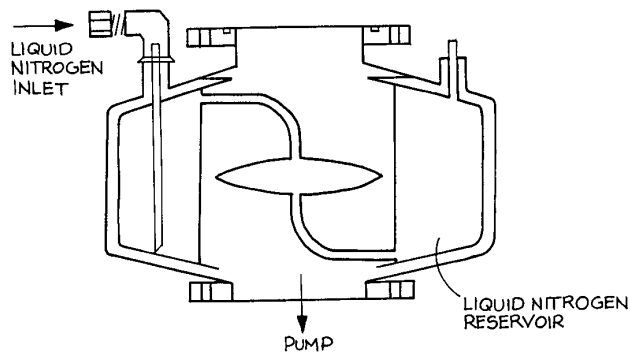
the trap and pump stack from the vacuum chamber. After the initial bakeout the pump is run continuously, so that the trap is always under vacuum. The isolation valve is closed whenever it is necessary to open the chamber to the atmosphere, and the chamber is rough-pumped before reopening the isolation valve. The molecular sieve cannot be regenerated after the first bakeout. The initial bake of the sieve results in the evolution of water vapor, but subsequent baking would drive out absorbed pump-fluid vapor. This oil vapor would condense on the bottom of the isolation-valve gate and be exposed to the vacuum when the valve is open. After the initial bakeout, the molecular sieve will trap oil vapor effectively for a period of several months if it is not exposed to moist air during that period. When the sieve becomes clogged with absorbed vapor, the base pressure of the pump-and-trap combination will begin to rise. At this time the molecular-sieve charge should be replaced. If it is absolutely necessary to stop the diffusion pump, the pump and trap should be filled with argon or dry nitrogen and isolated from the atmosphere in order to preserve the molecular sieve.

Condensation traps employing cold surfaces – cold traps – are used as foreline traps as well as inlet traps over diffusion pumps. The coolant may be simply a flow of cold water, an actively cooled refrigerant (Freon), or liquid nitrogen. In principal, a trap cooled to liquid nitrogen temperature would seem most effective, however, the vapor pressure of most pump fluids is below  $10^{-12}$  torr at  $-40^{\circ}\text{C}$ , a temperature easily achieved with a Freon refrigerator.

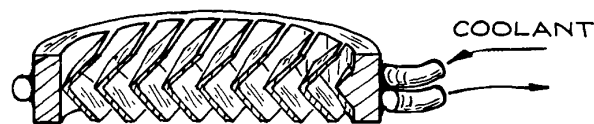


**Figure 3.24** A thimble trap cooled by the flexible, cooled probe of an immersion cooler.

A simple “thimble” trap is shown in Figure 3.24. These traps are used as foreline traps or over a small diffusion pump. For periodic operation, the thimble can be filled with liquid nitrogen. The coolant must be replenished every few hours. For continuous, long-term operation, an alternative, shown in the figure, is to fill the trap with a low-melting liquid, such as isopropyl alcohol, and refrigerate the liquid with an immersion cooler (Neslab Instruments; FTS Systems). These coolers consist of a refrigerator housed in a small console, with a hose to carry refrigerant to a flexible probe that is inserted in the bath in the trap. A temperature of  $-40^{\circ}\text{C}$  is easily attained with the smallest unit since the only heat load is thermal leakage from the surroundings into the trap. A liquid-nitrogen-cooled cryotrap for use over a large diffusion pump is illustrated in Figure 3.25. These traps are connected to a large liquid-nitrogen Dewar with a valve operated by a sensor that detects the coolant level in the trap. When a cold trap is permitted to warm up, it must be isolated from the vacuum chamber so that condensed material does not migrate into the chamber. Also the trap should be vented so that the evaporating condensate does not build up a dangerously high pressure.



**Figure 3.25** Liquid-nitrogen-cooled cryotrap for a diffusion pump.



**Figure 3.26** Cooled baffle.

An optically dense “chevron” baffle of the type shown in Figure 3.26 is usually placed over the inlet of a diffusion pump. These baffles may be air-cooled, water-cooled, or cooled by refrigerant from a small refrigerator. Recently, baffles with thermoelectric coolers operating at  $-35^{\circ}\text{C}$  have become available. In addition to reducing backstreaming of pump-fluid vapor, a baffle serves as a radiation barrier between the hot pump and the cold trap or vacuum system above the baffle.

### 3.5.6 Molecular Beams and Gas Jets

Often it is necessary to introduce a gaseous sample into a vacuum system. The sample may be contained in a small chamber with holes suitably located to admit probes such as light beams or electron beams; however, the walls of such a chamber may prove a hindrance to the proposed experiment. Alternatively, a gas can be introduced as an uncontained, but directed, beam of atoms or molecules. The absence of walls is only one of the advantages to using a gas beam. Because of the directed velocities of the particles in the beam, it is possible to maintain the sample in a

collisionless environment while still obtaining useful densities. The beam can be crossed with another beam or directed at a surface to obtain collisions of a specified orientation. Under appropriate conditions, the expansion of a jet of gas results in extreme cooling to give a very narrow population distribution among quantum states of molecules in the gas. Lucas succinctly stated the case for atomic or molecular beams when he observed that "beams are employed in experiments where collisions . . . are either to be studied, or to be avoided."<sup>8</sup>

Permitting gas to flow into a vacuum through a tube creates a gas beam. The nature of the beam depends upon whether the flow exiting the tube is in the molecular-or the viscous-flow regime. Apertures placed downstream of the channel through which the gas flows into the vacuum can establish the shape of the beam.

We begin with the molecular-flow case, where the mean free path,  $\lambda$ , at the outlet of the gas channel is greater than the diameter  $d$  of the channel. For the case of flow from a region of relatively high pressure into a vacuum through a round aperture (that is, a tube of length  $l = 0$ ), the flux distribution in the direction specified by the angle  $\theta$  relative to the normal to the aperture surface is:

$$I(\theta) = I(\theta = 0)\cos \theta \quad (\text{atoms/s}^1/\text{sr}^1) \quad (3.27)$$

The beam width  $H$  specified as the full angular width measured at the angle where the flux has fallen by half is  $120^\circ$ . Increasing the length  $l$  of the tube or channel through which the gas flows into the vacuum can narrow the beam. Any description of the gas beam issuing from such a channel must consider the molecular diameter and the nature of the scattering of molecules from the walls of the tube. Both theoretical and experimental investigations have been carried out in an effort to describe a gas beam in terms of easily measured parameters. In comparing experiment and theory it has turned out that the observed flux and the sharpness of the beam fall short of theoretical expectations by a factor of two to five.

In general, both experimental and theoretical results imply that to obtain useful fluxes and reasonable collimation, a beam source consisting of a tube or channel must be at least 10 times greater in length than its diameter, and the gas pressure behind this channel should be such that  $d < \lambda < l$ .

Lucas has devised a succinct description of gas beams in the molecular-flow regime in terms of a set of reduced parameters.<sup>8</sup> The gas pressure  $P$  behind the channel (in torr), the beam width  $H$  (degrees), and the gas particle flux  $I$  [atoms/s/sr (on axis)] and throughput  $q$  (atoms/s) are related to the corresponding reduced parameters by:

$$\begin{aligned} P &= \frac{P_R}{l\sigma^2} \\ H &= \frac{H_R d}{l} \\ I &= \left(\frac{T}{295M}\right)^{1/2} \frac{d^2 I_R}{l\sigma^2} \\ q &= \left(\frac{T}{295M}\right)^{1/2} \frac{d^3 q_R}{l^2 \sigma^2} \end{aligned} \quad (3.28)$$

where  $T$  is the absolute temperature,  $M$  is the molecular weight,  $d$  and  $l$  are in cm, and  $\sigma$  is the molecular diameter in  $\text{\AA}$  ( $10^{-8}$  cm). For pressures roughly in the range where  $d < \lambda < l$ , the reduced half angle, flux, and throughput are related to the reduced pressure by:

$$\begin{aligned} H_R &= 2.48 \times 10^2 \sqrt{P_R} \\ I_R &= 1.69 \times 10^{20} \sqrt{P_R} \\ q_R &= 2.16 \times 10^{21} P_R \end{aligned} \quad (3.29)$$

From model calculations, Lucas found that to obtain maximum intensity for a given angular width  $H$ , the reduced pressure  $P_R$  should not be less than unity.

For design work, a useful "optimum" equation can be derived by combining the above and setting  $P_R = 1$  to give:

$$I = 6.7 \times 10^{17} \left(\frac{T}{295M}\right)^{1/2} \frac{dH}{\sigma^2} \quad (3.30)$$

the optimum axial intensity for a given channel diameter and desired angular width. The tube length is fixed by the equations for  $H$  and  $H_R$  ( $P_R = 1$ ) above, and the input pressure by the equation for  $P$ . Operating at somewhat higher pressures gives some control over the flux and throughput without significantly departing from the optimum condition.

The construction of low-pressure, single-channel gas-beam sources is fairly straightforward. An excellent source can be made of a hypodermic needle cut to the appropriate

length. These needles are made of stainless steel, they are available in a wide range of lengths and diameters, and they come with a mounting fixture that is reasonably gas tight.

The goal of high intensity in a sharp beam is incompatible with a single-channel source because, for a given beam width, the gas load (throughput) increases more rapidly than the flux as the channel diameter is increased. The solution to this problem is to use an array of many tubes, each having a small aspect ratio (i.e.,  $d/l \ll 1$ ).<sup>9</sup> Tubes with diameters as small as  $2 \times 10^{-4}$  cm and lengths of  $1 \times 10^{-1}$  cm arranged in an array a centimeter or more across are commercially available in glass (Burle Electro-Optics, Pegasus Glassworks, Hamamatsu).

When the gas pressure behind an aperture or nozzle leading to a vacuum is increased to the extent that the mean free path is much smaller than the dimensions of the aperture, a whole new situation arises. Not only is the density of the resultant gas jet much greater than in the molecular-flow case, but also the shape of the jet changes and the gas becomes remarkably cold. This is a direct result of collisions between molecules in the gas as it expands from the orifice into the vacuum. Because of collisions, the velocities of individual molecules tend toward that of the bulk gas flow, just as an individual in a crowd tends to be dragged along with the crowd. The translational temperature of the gas, defined by the width of its velocity distribution, decreases, while the bulk-flow velocity increases. The conversion of random molecular motion into directed motion continues until the gas becomes too greatly rarefied by expansion, at which point the final temperature is frozen in. Because the mass-flow velocity increases while the local speed of sound, proportional to the square root of the translational temperature, decreases, the Mach number rises and the flow becomes supersonic. Inelastic molecular collisions in the expanding gas also cause internal molecular energy to flow into the kinetic energy of bulk flow, with the result that there may be substantial rotational and vibrational relaxation. Rotational temperatures less than 1 K and vibrational temperatures less than 50 K have been obtained. At these low temperatures only a small number of quantum states are occupied, an ideal condition for a variety of spectroscopic studies.

The low temperatures obtained in a supersonic jet can present some problems. Chief among these is that of condensation. Collisions in the high-density region of the jet

cause dimer or even polymer formation. With high-boiling samples, bulk condensation may occur. Of course, if one wishes to study dimers or clusters, this condensation is desirable. To avoid dimer formation, the sample gas is mixed at low concentration with helium. This *seeded gas* sample is then expanded in a jet. Adjusting its concentration in the helium carrier controls the degree of clustering of the seed-gas molecules.

The cooling effect, as well as the directionality of the jet, is lost if the expanding gas encounters a significant pressure of background gas in the vacuum chamber. Unfortunately, optimum operating conditions require a large throughput of gas into the chamber. The extent of cooling depends upon the probability of binary collisions, which is proportional to the product  $P_0 d$  of the pressure behind the nozzle and the diameter of the nozzle. Furthermore, to minimize condensation, the ratio  $d/P_0$  should be as large as possible. The result, in practice, is that a large-capacity vacuum pumping system is needed. In typical continuously operating supersonic-jet apparatus, source pressures of 10 to 100 atm have been used with nozzle diameters of 0.01 cm down to 0.0025 cm. The conductance of an aperture under these conditions is roughly:

$$C = 15d^2 \text{ L/s} \quad (3.31)$$

when the diameter is in centimeters. This implies a throughput on the order of 10 torr L/s. To obtain a mean free path of several tens of centimeters, a base pressure of about  $10^{-4}$  torr is required. A pump speed approaching 10 000 L/s may be necessary. Typically, several large diffusion pumps are used. When possible the jet is aimed straight down the throat of a pump.

In many experiments the cold molecules in a supersonic jet are only probed periodically – as, for example, when doing laser spectroscopy with a pulsed laser. Operation of the jet in a synchronously pulsed mode can reduce the gas load on the pump by several orders of magnitude compared to that in the continuous mode. A number of fast valves for this purpose have been devised.<sup>10</sup> The simplest are based upon inexpensive automobile fuel injector valves.<sup>11</sup> General Valve produces a system consisting of a pulsed valve and driver that operates at pulse rates as high as 250 Hz with a pulse width as small as 160  $\mu$ s.

Another efficient means of dealing with the background-gas problem has been demonstrated.<sup>12</sup> This relies upon the



fact that interaction of the expanding gas with the background gas gives rise to a shock wave surrounding the gas jet. If the pressure  $P_0$  behind the nozzle is increased, it is possible to achieve a mode of operation where a rarefied region is created behind the shock wave. In the region upstream of the shock front the gas behaves like a free jet expanding into a perfect vacuum. The distance from the nozzle to the shock front is:

$$\ell = 0.67d \left( \frac{P_0}{P} \right)^{1/2} \quad (3.32)$$

where  $P_0$  is the pressure in the nozzle,  $P$  the background pressure, and  $d$  the nozzle diameter. When the nozzle pressure is sufficiently high to achieve a free jet length of usable dimensions, the background pressure will increase greatly. This state of affairs has a distinct advantage, since at a high background pressure, a large throughput can be achieved with a pump of moderate speed. For example, to obtain a free length  $\ell = 1$  cm with a 0.01 cm nozzle and a source pressure of 10 atm, the background pressure should be about 400 mtorr. The throughput in this case is about 10 torr L/s and the required pumping speed, 25 L/s could be achieved by a large rotary mechanical pump or a Roots pump of modest capacity.

The fabrication of a nozzle is straightforward, although some difficulty may be encountered in making the necessary small hole. A skillful mechanical technician can drill a hole as small as 0.01 cm diameter. Smaller holes can be made by spark or electrolytic erosion, or by swaging a hole closed on a piece of hard wire and then withdrawing the wire. One of the first small, high-pressure nozzles<sup>12</sup> was made by drilling down the axis of a stainless-steel rod to within 0.1 mm of the end with a drill bit having a sharp conical point. A 0.0025 cm diameter hole was then made through the remaining metal by spark erosion.

For many experiments, the gas flowing into a vacuum system through an aperture or a channel produces a beam that is too broad or too divergent for the intended application. In this case one or more apertures placed downstream from the source can define the beam shape. Often it is advantageous to build these apertures into partitions that separate the vacuum housing into a succession of chambers. Each chamber can be evacuated with a separate pump. The pressure in the first will be highest, so that

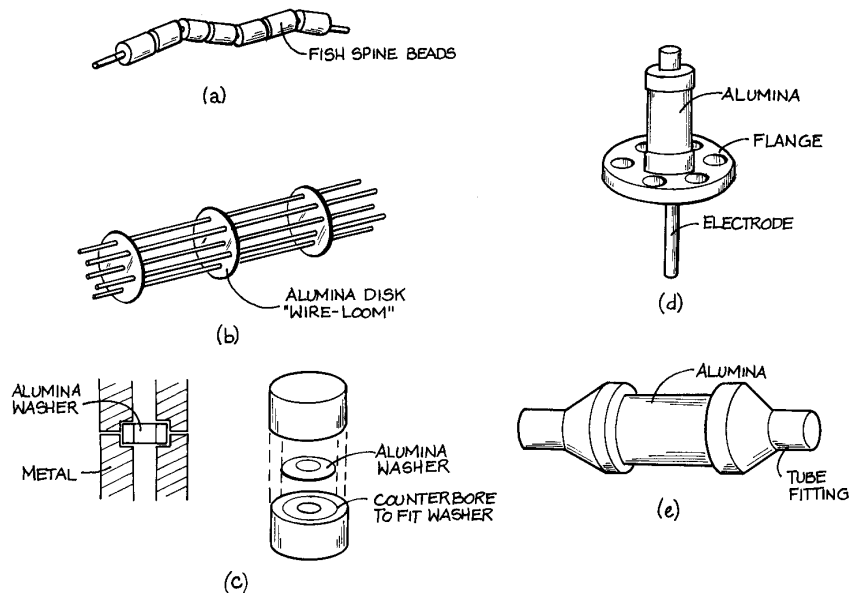
the large throughput obtained here will significantly reduce the gas load on the next pump. This scheme is called *differential pumping* (see Section 3.6.2).

An aperture downstream of a supersonic jet, but close enough to be in the viscous-flow region, is called a *skimmer*. The design of these skimmers is critical, since they tend to produce turbulence that will destroy the directionality of the flowing gas. A skimmer is usually a cone with its tip cut off and the truncated edge ground to knife sharpness. The details of the design, location, and fabrication of a skimmer are beyond the scope of this book, and the interested reader should refer to a specialized text on this subject.<sup>13</sup>

### 3.5.7 Electronics and Electricity in *Vacuo*

At low frequencies the vacuum between electrodes or wires provides excellent insulation. At pressures below  $10^{-4}$  torr, the breakdown voltage between smooth, gently rounded surfaces is in excess of 1000 V for each millimeter of separation. Occasionally sparking will occur between newly made parts because of high field gradients around whiskers of metal. These whiskers quickly evaporate and sparks do not recur. The challenge in making electrical devices function in vacuum is in the design of spacers and other elements that locate electrodes and wires that are at different electrical potentials. Insulating materials must possess adequate mechanical strength and be compatible with the vacuum environment and available fabrication processes. In addition, insulators must be very clean. Electrical breakdown in the vacuum is usually initiated by a discharge traveling along the surface of an insulator.

Ceramic materials are the best insulators and, among these, alumina (at least 96%  $\text{Al}_2\text{O}_3$ ) is probably the best. Simple shapes such as rods, tubes, and balls are readily available to be incorporated in insulator assemblies (see Figures 5.40 and 5.41). More complex shapes require specialized grinding techniques. Alumina circuit board substrate is available as plate .010 in. to .060 in. thick and 6 in. square. Owing to the scale of the circuit board industry, the facilities for custom fabrication by laser cutting of alumina plate are widely available. The designer of vacuum apparatus can now specify quite complex shapes cut from alumina plate at a cost of just a few dollars a piece. A simple scheme employing alumina washers to locate and insulate electrically isolated parts is illustrated in Figure 3.27(c).



**Figure 3.27** Applications of ceramics for electrical insulation in vacuum systems: (a) ceramic “fish spine” beads threaded on a wire as insulation; (b) ceramic disk with holes used as a wire-loom to keep wires in a bundle from touching one another; (c) design for electrodes located and electrically isolated with spacers laser-cut from alumina circuit-board substrate; (d) a ceramic-insulated electrical feedthrough; (e) a ceramic break in a vacuum line mounted between two standard vacuum flanges.

Alumina can be metalized and subsequently joined by brazing to stainless steel or Kovar structures. It is impractical to attempt this process in the lab; however, this technology is employed in many standard vacuum components, such as electrical feedthroughs (Cerameal, Insulator Seal) to be inserted in the vacuum wall [Figure 3.27(d)] and ceramic “breaks” built into the middle of a length of metal vacuum tubing [Figure 3.27(e)].

A few precautions should be observed in designing and mounting ceramic parts. Ceramic materials are strongest in compression and cannot be plastically deformed. The best design of an assembly with ceramic components captures each ceramic part between smooth surfaces that are aligned with the face of the part so that shear forces are not applied to the ceramic material.

Ceramics are often used simultaneously for electrical insulation and thermal isolation. In this case it is important to be aware of the strong dependence of electrical resistivity on temperature. The resistivity of a ceramic material

falls by roughly a factor of 10 for every 100 °C increase in temperature.<sup>14</sup> For example, the resistivity of alumina is greater than  $10^{17}$  ohm cm at 100 °C and falls below  $10^7$  ohm cm at 1000 °C.

The surface of a ceramic part is more or less porous and susceptible to trapping contaminants. In handling ceramic parts, inorganic salts from skin oils may be transferred from the fingers to the surface of the ceramic. Salts absorb water and encourage electrical conduction across the surface. A part should be cleaned with organic solvents to remove oils followed by a final rinsing in hot distilled water to remove these salts. Rinsing with pure methanol and heating can dry the part. Surgical gloves should be worn to handle a part after it has been cleaned.

If the spacing between wires in a vacuum system cannot be reliably maintained, then insulation is required. Ordinary Teflon-insulated, solid, hookup wire may be used down to  $10^{-7}$  torr, although air bleeding out from under the insulation will slow pumpdown. Stranded wire should

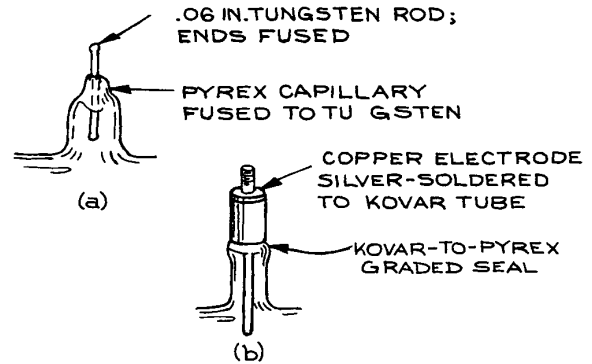
never be used, as gas is trapped between the strands. At very low pressures and high temperatures, wires can be insulated by stringing ceramic “fish spine” beads (from Coors) [Figure 3.27(a)] or pieces of Pyrex tubing over them. The spacing between wires can be maintained with ceramic wire looms [Figure 3.27(b)] available from many suppliers of vacuum apparatus. Wire insulated with a Kapton (polyimide) film or a ceramic coating is available from vacuum-equipment suppliers for use in UHV. The latter is better for use at elevated temperatures, but has the disadvantage of shedding bits of ceramic material when flexed. Electrical solder should not be used for electrical connections in a vacuum system because the lead in solder and soldering flux contaminate the vacuum. Mechanical connections are preferable. Wrapping a wire under the head of a screw and tightening the screw can make connections to electrodes. Wires may be jointed by slipping the ends into a close-fitting piece of tubing and crimping the tubing onto the wire. A variety of wire connectors are available from vacuum equipment suppliers. Tungsten, molybdenum, nichrome, or stainless-steel wires may be spot-welded together. In welding refractory metals, a more secure weld is obtained if a piece of nickel foil is interposed between the wires.

Electronic devices tend to overheat in a vacuum, since the only cooling is by radiation. Electronic components such as power transistors and integrated circuits that must dissipate more than about 2 watt are particularly unreliable. Such devices can be placed in vacuum-tight, gas-filled boxes that are thermally connected to the vacuum wall.

Electrical feedthroughs for wires passing through the wall of a glass vacuum system are illustrated in Figure 3.28. As described in Section 2.3.10, an electrical feedthrough can be made by sealing a tungsten rod directly into a hard glass wall. An easier solution is to braze an electrode into the metallic end of a standard Kovar-to-glass graded seal and join the glass end of the seal to the glass vacuum wall.

### 3.6 VACUUM-SYSTEM DESIGN AND CONSTRUCTION

Before beginning the design of a vacuum system, a number of parameters must be specified in at least a semiquantitative manner. The size and shape of the vacuum chamber must be determined. The desired ultimate pressure and the



**Figure 3.28** Electrical feedthroughs in a glass vacuum wall employing (a) a tungsten-to-metal seal; (b) a standard Kovar-to-glass seal.

composition of the residual gas in the chamber must be specified. One must estimate the gas load on the pumps. The amount of money and time available are also important considerations. It is instructive to spend a few evenings leafing through vacuum-equipment manufacturers' catalogs to become familiar with the specifications and cost of commercial vacuum components. The author has found the Kurt J. Lesker, Co., Leybold AG, and Varian Vacuum catalogs to be especially informative.

The ultimate pressure requirement dictates the choice of the primary pump. Cost and limitations on hydrocarbon contamination will determine the choice of roughing and backing pumps. An ultimate pressure down to about 1 mtorr can be achieved with a mechanical pump with an appropriate trap. Sorption pumps can be used on a system requiring an ultimate pressure of 1 mtorr if there is not a large continuous gas load. Ultimate pressures in the high-vacuum range down to about  $10^{-9}$  torr can be achieved with a diffusion pump, a turbomolecular pump, or a cryopump. True ultrahigh-vacuum systems ordinarily use ion pumps and titanium sublimation pumps. Ultrahigh vacuum can also be achieved with turbomolecular pumps. To reach UHV starting from a backing pressure near 1 torr requires a compression ratio approaching  $10^{12}$ . This can be accomplished with two turbomolecular pumps operating in series, a turbo backed with a drag pump, or with a compound turbomolecular/drag pump.

The cost of a vacuum system is about equally divided between the cost of the pumps and the cost of the vacuum

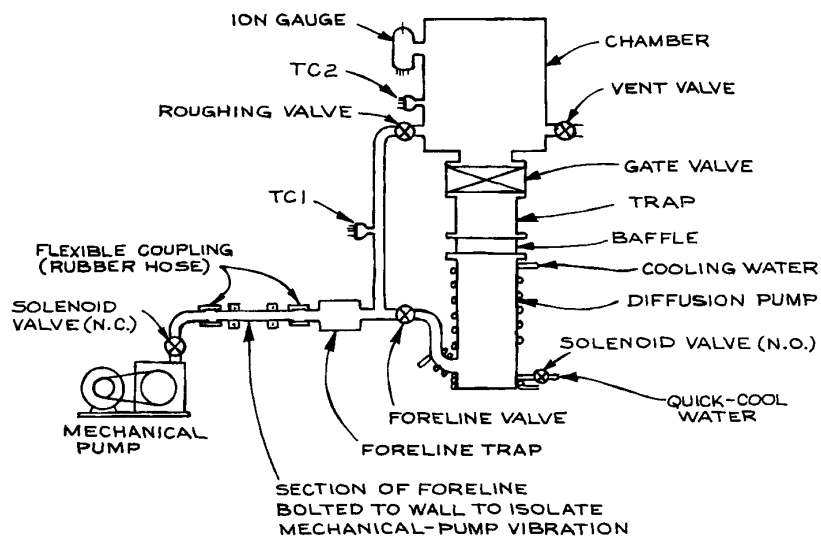
vessel and ancillary hardware, such as gauges and valves. In terms of cost per unit pumping speed, mechanical pumps are most expensive – on the order of \$100 per L/s. Most systems require a mechanical pump as either the primary pump or a backing pump; however, since they operate at relatively high pressures and correspondingly high throughputs, one ordinarily finds that for laboratory apparatus only modest pumping speed is required of a mechanical pump – a few L/s to a few tens of L/s. For high-vacuum systems, diffusion pumps are most cost effective although sorption or cold traps needed to prevent hydrocarbon contamination will add to the total cost. The cost of a diffusion pump amounts to \$2 to \$5 per L/s of pumping speed. A turbomolecular pump with its controller, a cryopump with a refrigerator, or an ion pump and controller will cost five to ten times as much as a diffusion pump of comparable pumping speed.

There are three sources of gas that must be pumped out of a vacuum vessel: molecules adsorbed on the walls of the container, as well as the components therein, gas intentionally admitted as part of the process under study, and leaks. The rate of outgassing of adsorbed water, carbon dioxide, and hydrocarbons, and, ultimately, hydrogen, from the surface of materials is difficult to measure. A summary of

published data is given in O'Hanlon's book.<sup>15</sup> With care in the choice and processing of the materials of construction of vacuum apparatus, it is generally possible to reduce outgassing rates below  $10^{-9}$  torr L/s cm; lower yet with baking. Similarly, actual leaks can be eliminated by careful design and construction; so-called *virtual leaks* are, however, more or less of a problem in all systems. Virtual leaks are the result of gas being trapped in small occlusions within the vacuum vessel. These typically occur where parts are fitted together. Gas will be trapped between mated surfaces and the conductance through the space between such surfaces will be very small. The trapped gas may take weeks to be pumped away and thus acts as a continual gas load on the pump. Every effort should be made to vent these spaces with pumpout holes (Figure 3.40).

### 3.6.1 Some Typical Vacuum Systems

**Diffusion Pumped System.** A diffusion-pumped vacuum system is shown schematically in Figure 3.29. For illustration, suppose the vacuum chamber is constructed of stainless steel and is 30 cm in diameter and 30 cm high with a surface area of 5000 cm<sup>2</sup>. If



**Figure 3.29** Schematic diagram of diffusion-pumped high-vacuum system.

the pump stack consists of a 2 in. oil diffusion pump with a speed of 150 L/s, and a trap and baffle with a conductance of 200 L/s, the net speed of the pump stack will be:

$$\begin{aligned} S &= \left( \frac{1}{S_p} + \frac{1}{C} \right)^{-1} \\ &= \left( \frac{1}{150 \text{ L/s}} + \frac{1}{200 \text{ L/s}} \right)^{-1} \\ &= 85 \text{ L/s} \end{aligned} \quad (3.33)$$

Assuming an initial outgassing rate of  $10^{-9}$  torr L/s/cm<sup>2</sup>, the vacuum vessel will pump down to:

$$\begin{aligned} P &= \frac{Q}{S} = \frac{(5000 \text{ cm}^2) \times (10^{-8} \text{ torr L/s/cm}^2)}{85 \text{ L/s}} \\ &= 6 \times 10^{-7} \text{ torr} \end{aligned} \quad (3.34)$$

within a few hours. After a day of pumping and perhaps a light bake to 100 °C, the outgassing rate should fall below  $10^{-9}$  torr L/s/cm<sup>2</sup> and an ultimate pressure in the  $10^{-8}$  torr range should be achievable.

The size of the backing pump is determined from the maximum throughput of the diffusion pump. This occurs after the vacuum vessel is rough pumped and the gate valve over the diffusion pump is first opened. The pressure in the vessel at this time should not be greater than about 50 mtorr. The maximum throughput will then be:

$$\begin{aligned} Q_{\max} &= P_{\max} S = (5 \times 10^{-2} \text{ torr}) \times (85 \text{ L/s}) \\ &= 4 \text{ torr L/s} \end{aligned} \quad (3.35)$$

The foreline pressure must always remain below about 200 mtorr so the speed of the backing pump should be:

$$\begin{aligned} S_{p,\text{backing}} &= \frac{Q_{\max}}{P_{\text{backing}}} = \frac{4 \text{ torr L/s}}{2 \times 10^{-1} \text{ torr}} \\ &= 20 \text{ L/s.} \end{aligned} \quad (3.36)$$

A small two-stage mechanical pump can meet this requirement.

The conductance of the foreline should be at least twice the speed of the forepump so that the forepump is not strangled by the foreline. To achieve a conductance of 40 L/s at a pressure of 200 mtorr in a foreline a

half meter long, the diameter (from Section 3.2.4) must be:

$$\begin{aligned} D &= \left( \frac{CL}{180P} \right)^{1/4} \text{ cm } (L \text{ in cm}) \\ &= \left( \frac{40 \times 50}{180 \times 2 \times 10^{-1}} \right)^{1/4} = 2.7 \text{ cm} \end{aligned} \quad (3.37)$$

In this case 1 in. copper water pipe would make an excellent foreline.

To activate a system of the type shown in Figure 3.29, starting with all valves closed and the pumps off, proceed as follows:

- (1) Turn on the mechanical pump.
- (2) When the pressure indicated by thermocouple gauge 1 (TC1) is below 200 mtorr, open the foreline valve.
- (3) When TC1 again indicates a pressure below 200 mtorr, turn on the cooling water and activate the diffusion pump. The pump will require about 20 minutes to reach operating temperature. When operating it makes a crackling sound.
- (4) Before pumping on the chamber with the diffusion pump, the pressure in the chamber must be reduced to a rough vacuum. Close the foreline valve and open the roughing valve. When the pressure indicated by TC2 is below 50 mtorr, close the roughing valve and open the foreline valve. The diffusion pump should not be operated for more than two minutes with the foreline valve closed. If necessary, interrupt the roughing procedure, close the roughing valve, wait for TC1 to indicate less than 200 mtorr, and reopen the foreline valve for a moment to ensure that the diffusion-pump outlet pressure does not rise above about 200 mtorr.
- (5) Slowly open the gate valve that isolates the diffusion pump from the chamber. In less than 10 s the pressure indicated by TC2 should fall below 1 millitorr and the ionization gauge can be turned on.

A diffusion-pumped system of the type shown in Figure 3.29 is one of the most convenient and least expensive systems for obtaining high or even ultrahigh vacuum. The system can be fabricated of metal or glass. A baffled, but untrapped, oil diffusion pump on this system should result in an ultimate pressure slightly below  $10^{-6}$  torr.

With a cold trap an ultimate pressure of  $10^{-9}$  torr is possible if the vacuum chamber is baked and only low-vapor-pressure materials are exposed to the vacuum.

A number of inexpensive safeguards are incorporated in the design in Figure 3.29 to avoid contamination or damage in the event of a loss of electrical power. A normally open (N.O.) solenoid-operated water valve in parallel with the pump heater admits water to the diffusion pump, quick-cooling coils if the power fails. The diffusion pump is isolated from the mechanical pump by a normally closed (N.C.) solenoid valve to prevent air or oil from being sucked through the mechanical pump into the foreline. Special mechanical pump inlet valves are available for this purpose (HPS Vacuum Sentry® and Leybold SECUVAC®). These valves close automatically when power is lost and vent the pump to atmosphere; when power is regained, the valve does not reopen until the pressure on the pump side falls to a few hundred millitorr. The SECUVAC valve is also designed to close in the event of a sudden rise in inlet pressure. It is helpful to place a ballast volume in the foreline to maintain a low foreline pressure while the diffusion pump cools after a power outage. A relatively large foreline trap will also serve as a ballast. A ballast tank of 5 to 10 L in the foreline will permit the continued operation of the diffusion pump for a period of perhaps 10 to 20 minutes if the backing pump is turned off and the valve at the inlet to the backing pump is closed. This intermittent mode of operation can be useful in the event that vibration of the mechanical pump has an adverse effect on an instrument housed in the high vacuum chamber. For example, this mode of operation is used on some electron microscopes.

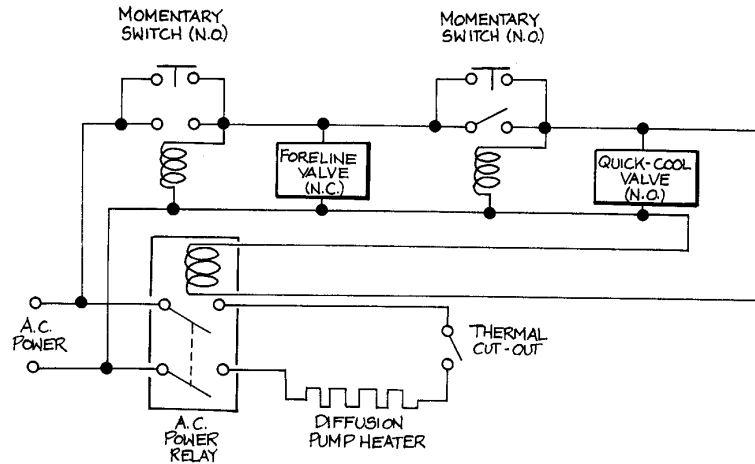
A useful fail-safe mechanism, not shown in Figure 3.29, would be a water-pressure sensor in the water line to interrupt the electrical power if the water fails. Alternatively, a heavy-duty, normally closed (N.C.), bimetallic temperature switch wired in series with the diffusion pump heater may be mounted to the bottom turn of the water-cooling coil. Then if the cooling-water flow is inadequate, the rising temperature of the pump barrel will interrupt the electricity to the pump heater. Many diffusion pumps are equipped with a mounting plate for a temperature switch that can be obtained from the pump manufacturer.

A power interruption in the operator's absence threatens serious contamination of the vacuum vessel, both at the

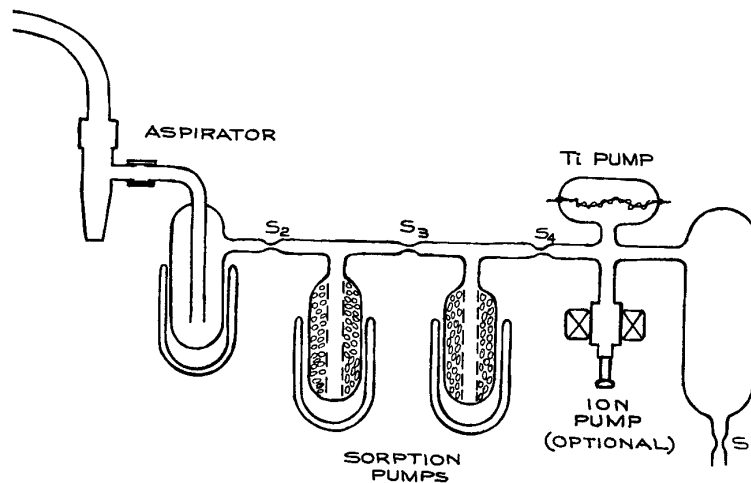
time the power goes off and when it comes back on again. When the power goes off it is essential that the diffusion pump stops operating before the foreline pressure rises above the critical backing pressure. This is the purpose of the quick-cool water valve described above. When power is restored in the operator's absence there is no provision to rough out the vacuum vessel or otherwise follow the orderly initiation of pumping described above. As a minimum safeguard against a power failure, the power to the system's components should be supplied through power relays that are wired to latch "off" until reactivated by the operator. A latching relay circuit is shown in Figure 3.30.

For a vacuum system that operates for long periods of time without close supervision, it is worthwhile incorporating a control system that monitors the pressure in the chamber and the foreline. Many high-quality commercial thermocouple-gauge and ionization-gauge controllers include relays that can be set to trip at a predetermined pressure. The relay controlled by the foreline thermocouple gauge (TC1) can be set to trip at 200 – 300 mtorr. This relay can be wired to operate a second, heavy-duty relay that removes power from the diffusion pump and simultaneously opens the quick-cool water valve and closes the foreline solenoid valve at the mechanical pump. The overpressure relay controlled by the ion gauge can be employed to isolate the high-vacuum chamber by means of an electrically controlled, pneumatically activated gate valve when the pressure in the chamber rises above a few millitorr. In lieu of a gauge controller with control relays, there are now available inexpensive, modular, Pirani vacuum sensors with built-in pressure switches (for example, the HPS Moducell). These can be set to trip in the range  $5 \times 10^{-3}$  to 30 torr, making them suitable for monitoring either the high-vacuum chamber or the foreline. The design of a logical interlock system employing temperature and pressure sensors to protect a vacuum system is discussed in Sections 6.6.9 and 6.6.10.

**A Small, Inexpensive Ultrahigh-Vacuum System.** A small, inexpensive system, capable of producing a vacuum of about  $10^{-10}$  torr in a 1 L volume, is illustrated in Figure 3.31. The entire system can be made of Pyrex. Constrictions in the glass tubing joining



**Figure 3.30** Relay circuit to latch "off" a diffusion-pumped vacuum system in the event of a power failure.



**Figure 3.31** A small inexpensive ultrahigh-vacuum system made of Pyrex glass. Constrictions (labeled "S") between components are heated with a glassblower's torch until they collapse and fuse under the external atmospheric pressure to seal-off and isolate components as evacuation proceeds.

components of the system are fused closed with a glassblower's torch to seal off each part of the system as evacuation proceeds. Rough pumping is initiated with a water aspirator or alternatively a compressed-air Venturi pump. High vacuum is achieved with sorption pumps operated in sequence and ultrahigh vacuum is reached

with a titanium sublimation pump, possibly with the aid of a small ion pump.<sup>16</sup> To achieve UHV, special precautions are taken to displace rare gases from the system and to remove water and carbon dioxide adsorbed on the walls of the system. Neither sorption pumping nor gettering efficiently pumps rare gases,

helium in particular. The rare gases are swept from the system by a flow of dry nitrogen before evacuation is begun. The entire system is maintained at a moderately high temperature throughout the entire pumpdown procedure in order to drive adsorbed gases off the walls and into the pumps.

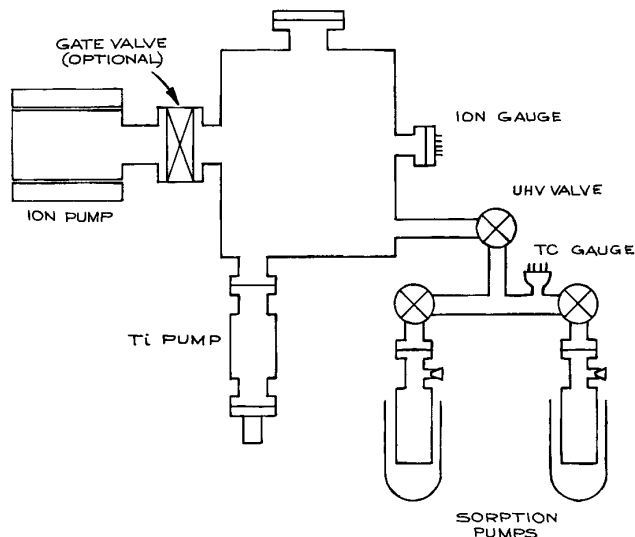
The pumpdown procedure for this vacuum system is as follows:

- (1) Flow pure dry nitrogen through the system from  $S_1$  to the aspirator. Do not activate the aspirator. Pure dry nitrogen gas can be obtained from the headspace over liquid nitrogen in a Dewar.
- (2) As nitrogen flows through the system, the molecular-sieve sorption pumps are heated to  $300^\circ\text{C}$  and the trap is cooled with liquid nitrogen. The remainder of the system should be heated to about  $100^\circ\text{C}$  with heating tape or by brushing the glass with a cold flame or with hot air from a heat gun.
- (3) After about an hour, close pinchoff  $S_1$  by heating the glass with a torch and turn on the aspirator. The pressure will quickly fall to about 20 torr.
- (4) Seal off constriction  $S_2$  and refrigerate the first molecular-sieve pump. The pressure will fall to about  $10^{-4}$  torr. Continue to heat the system.
- (5) Seal off  $S_3$  and cool the second sorption pump. The pressure will fall to about  $10^{-7}$  torr. Continue heating the system for a short time to drive off the remaining adsorbed gases.
- (6) Pass a current through the filament of the titanium pump to evaporate a layer of titanium onto the glass envelope, and activate the ion pump.
- (7) Seal off  $S_4$  and deposit a fresh layer of titanium. The pressure should fall well below  $10^{-9}$  torr.

When sealing off a constriction, it is wise to proceed slowly so that the pumps have time to take up any gas that is liberated.

A similar system can be made with the pinchoffs replaced by all-metal high-vacuum valves. In this case it would also be possible to construct the entire system of stainless steel.

**An Oil-Free Ultrahigh-Vacuum System.** A large, oil-free ultrahigh-vacuum system is illustrated in Figure 3.32. The system is roughed down with sorption pumps,



**Figure 3.32** An oil-free ultrahigh-vacuum system.

and the ion pump is depended upon for the removal of rare gases. The system must be baked. A number of variations on this system are possible. The rough-pumping operation can be accelerated and the gas load on the sorption pumps reduced if the system is first roughed down with one of the compressed-air aspirators (Venturi pumps) sold for this purpose. Alternatively, one can employ one of the oil-free mechanical pumps that have recently appeared on the market. Varian, for example, manufactures a reciprocating piston pump with Teflon seals that would be suitable as a roughing pump. The pumping at high vacuum can be carried out with a cryopump rather than a getter pump. The choice of the high-vacuum pump or combination of pumps depends upon the composition of the residual gas, as well as the composition of gases that may be admitted to the system at high vacuum.

**Small, Oil-Free, Turbomolecular Pumped UHV System.** With appropriate precautions, hydrocarbon-free vacuum below  $10^{-11}$  torr can be attained with a turbomolecular pump. The compression ratio of a turbomolecular pump for air is typically  $10^8$  to  $10^9$ , so a foreline pressure well below  $10^{-2}$  torr is required to achieve UHV. This is accomplished with a small high-vacuum pump



between the turbomolecular pump and the mechanical foreline pump. One possibility is to back the turbo with a diffusion pump backed by a mechanical pump. The intermediate pump might just as well be another turbo or a molecular-drag pump. These intermediate pumps can be very small: in principle, the speed of these pumps need only be that of the primary turbopump divided by its compression ratio in order to deal with the throughput of the primary turbo!

The compression ratio of a turbomolecular pump exceeds  $10^{16}$  for gases with molecular weights greater than 100, so when the pump is running there is essentially no backstreaming of hydrocarbons from oil-lubricated bearings or from an oil-sealed mechanical forepump; however, hydrocarbon contamination of the UHV chamber is possible when the turbo is not operating. There are two essential precautions. When the turbo is turned off it should be vented to atmospheric pressure. Dense gas in the pump will prevent backstreaming from bearings at the bottom of the pump into the upper stages. There is usually no need for a valve between a turbo pump and the vacuum chamber so the chamber is also vented to atmospheric pressure when the pump is vented. On the other hand, there must be a valve at the turbopump exhaust. This valve must be kept closed except when the pump is activated. In a turbomolecular-pumped system, the chamber is rough-pumped directly through the turbopump. To initiate pumping, the mechanical backing pump is turned on, the valve at the turbo exhaust is opened, and the turbo is then immediately activated so that roughing and turbo spin-up occur simultaneously. The intermediate pump can also be activated at this time. By following this procedure, the initial flow of dense gas from the chamber will prevent backstreaming while the turbo is coming up to speed. Of course, the best, albeit most expensive, means for eliminating hydrocarbons is to employ pumps that contain no hydrocarbons: turbomolecular pumps and molecular-drag pumps with magnetically suspended rotors and a diaphragm pump as a forepump (see Section 3.4.1).

The ultimate pressure in a UHV chamber with a turbomolecular pump is determined by the compression ratio for  $H_2$  and the partial pressure of  $H_2$  in the foreline. The compression ratio for  $H_2$  in modern pumps is in the range from  $10^3$  to  $10^4$ . The exclusion of hydrocarbon oils and elastomer seals in the foreline helps reduce the hydrogen partial

pressure in the foreline. In any event it has been found that hydrogen represents 80 to 90% of the residual gas in a turbopumped chamber at pressures below about  $10^{-10}$  torr. When this is a problem for a particular experiment, a titanium sublimation pump can be installed above the turbo. It would not be necessary for titanium to be continuously evaporated once UHV conditions are achieved.

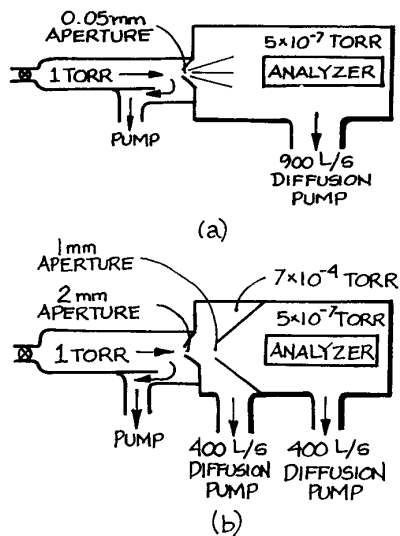
Custom-made UHV chambers can be quite expensive especially when many access ports are required for components, such as mechanical and electrical feedthroughs, and windows. When a relatively small volume is required (a few hundred cubic inches or a few thousand cubic centimeters), it is often possible to make up the chamber of standard UHV components, such as tees and crosses. In the author's laboratory a 6 in. cube cross (Kurt J. Lesker Company) is used as a small UHV chamber. The pump is mounted on one face and the remaining five faces provide ports for mounting components. The cube was ordered 0.5 in. oversize (i.e., 6.5 in.  $\times$  6.5 in.  $\times$  6.5 in.). With the additional size it is possible to bore .375 in. diameter holes through the cube close to each vertical edge. Cartridge heaters inserted in the holes are used for bakeout.

### 3.6.2 Differential Pumping

One is occasionally required to design a system wherein gas is maintained at two very different pressures. Often two or more pumps with interconnecting tubing must be balanced against one another to obtain the desired pressure differential.

As an example we will consider the problem of making a mass-spectrometric analysis of trace constituents in argon flowing through a tube at a pressure  $P_0$  of 1.0 torr. As illustrated schematically in Figure 3.33(a), the sample is to be withdrawn through a hole in the flow tube. The mass spectrometer must be operated at a pressure  $P_2$  of  $5 \times 10^{-7}$  torr. To begin, let us suppose that a 6 in. diffusion pump is to be used to maintain the pressure  $P_2$ . A properly trapped 6 in. diffusion pump has a speed of about 900/L over the range  $10^{-3}$ – $10^{-9}$  torr (see Figure 3.12). The throughput is then:

$$\begin{aligned} Q &= P_2 S = (5 \times 10^{-7} \text{ torr})(900/\text{L}) \\ &= 4.5 \times 10^{-4} \text{ torr/L} \end{aligned} \quad (3.38)$$



**Figure 3.33** (a) Schematic representation of an analyzer in high vacuum ( $5 \times 10^{-7}$  torr) used to sample a gas at high pressure (1 torr). (b) Differential pumping arrangement that permits a much larger sampling aperture with the same total pumping speed for the analyzer.

The conductance of the sampling aperture that provides this throughput is:

$$C = \frac{Q}{P_0 - P_2} = \frac{4.5 \times 10^{-4} \text{ torr L/s}}{1.0 \text{ torr}} = 4.5 \times 10^{-4} \text{ L/s} \quad (3.39)$$

The flow through this aperture is in the viscous-flow regime. The area of the aperture that provides this conductance (from Section 3.2.4) is approximately:

$$A = \frac{C(\text{L/s})}{20} \text{ cm}^2 = \frac{4.5 \times 10^{-4}}{20} = 2.3 \times 10^{-5} \text{ cm}^2 \quad (3.40)$$

corresponding to a diameter of 0.05 mm for a circular aperture. There are two problems associated with this small aperture. Firstly, there is the mechanical difficulty of creating such a small hole. Secondly, this tiny hole may impose a severe limitation upon the sensitivity of the meas-

urement, particularly if the analyzer samples only atoms moving along the line of sight from aperture to analyzer. A solution to these problems is to incorporate several stages of pumping – a scheme known as *differential pumping*.

A two-stage differentially pumped system is illustrated in Figure 3.33(b). An intermediate-pressure chamber with its own pump is interposed between the sample and analyzer volumes. There are two apertures now and the pressure drop occurs in two steps. For comparability with the previous example, suppose the intermediate chamber and the analyzer chamber are each evacuated with 4 in. diffusion pumps, each with a speed of 400 L/s. Take the ratio of pressures across each aperture to be the same, so that  $P_0/P_1 = P_1/P_2$  and the pressure  $P_1$  in the intermediate chamber is then  $7 \times 10^{-4}$  torr. The throughput of the pump on the lowest pressure chamber is:

$$\begin{aligned} Q_2 &= P_2 S \\ &= (5 \times 10^{-7} \text{ torr})(400 \text{ L/s}) \\ &= 2 \times 10^{-4} \text{ torr L/s} \end{aligned} \quad (3.41)$$

The conductance  $C_2$  of the aperture separating the region of pressure  $P_1$  from the region of pressure  $P_2$  must maintain this throughput:

$$\begin{aligned} C_2 &= \frac{Q_2}{P_1 - P_2} = \frac{2 \times 10^{-4} \text{ torr L/s}}{7 \times 10^{-4} \text{ torr}} \\ &= 0.3 \text{ L/s} \end{aligned} \quad (3.42)$$

The flow through this aperture is in the molecular-flow regime. The area of the aperture that provides this conductance (from Section 3.2.4) for argon at 300 K is:

$$\begin{aligned} A_2 &= C \left[ 3.7 \left( \frac{T}{M} \right)^{1/2} \right]^{-1} \text{ cm}^2 \\ &= 0.03 \text{ cm}^2 \end{aligned} \quad (3.43)$$

corresponding to a diameter of about 2 mm for a circular aperture. Similarly, the conductance of the aperture separating the high- and intermediate-pressure regions must match the throughput of the intermediate pump:

$$\begin{aligned} Q_1 &= P_1 S \\ &= (7 \times 10^{-4} \text{ torr})(400 \text{ L/s}) \\ &= 0.3 \text{ torr L/s} \end{aligned} \quad (3.44)$$

so that:

$$\begin{aligned} C_1 &= \frac{Q_2}{P_0 - P_1} = \frac{0.3 \text{ torr L/s}}{1 \text{ torr}} \\ &= 0.3 \text{ L/s} \end{aligned} \quad (3.45)$$

The flow through this aperture is in the viscous-flow regime. The area of the aperture that provides this conductance (from Section 3.2.4) is approximately:

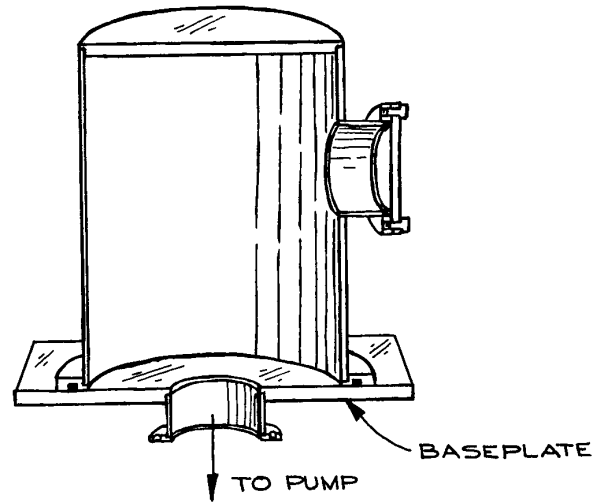
$$\begin{aligned} A_1 &= \frac{C(\text{L/s})}{20} \text{ cm}^2 \\ &= \frac{0.3}{20} = 0.015 \text{ cm}^2 \end{aligned} \quad (3.46)$$

corresponding to a diameter of about 1 mm for a circular aperture. In this example, we find that, by pumping in stages, and without an increase in total pumping speed, the required pressure differential can be maintained across apertures that are hundreds of times larger than if all the available pumping speed had been applied in a single stage. Furthermore, assuming that the gas flow through first aperture is directional, the sensitivity of the measurement is greatly increased by aligning the second aperture and analyzer with the flow through the first aperture, as suggested in the figure.

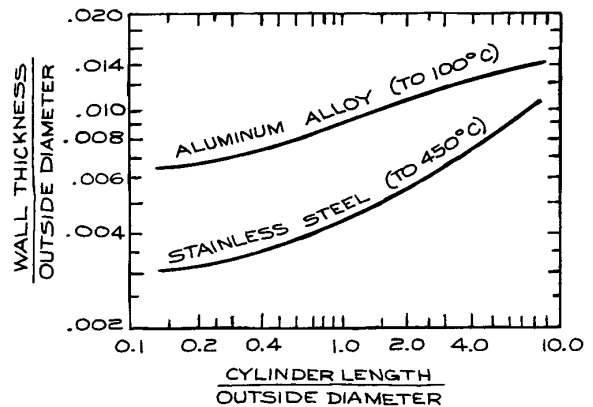
### 3.6.3 The Construction of Metal Vacuum Apparatus

Metal vacuum chambers are often built up of one or more cylindrical sections. Tube stock is readily available and a cylindrical shape is second only to a spherical shape in strength to withstand external pressure. Even when tube stock is unavailable, a cylindrical section is easily fabricated by rolling up sheet stock. The ends of a cylindrical chamber are most conveniently closed with flat plates, as shown in Figure 3.34. A flat end plate must be quite thick to withstand atmospheric pressure, as discussed below. Domed or spherical end caps are much stronger, but more expensive to fabricate.

The ability of a cylinder to resist collapsing under external pressure depends upon the length of the cylinder between supporting flanges, the diameter, the wall thickness, and the strength of the material. The ASME has



**Figure 3.34** Typical aluminum or steel vacuum chamber resting on a baseplate. The pump, gauges, and electrical and mechanical feedthroughs are most conveniently mounted to the baseplate.



**Figure 3.35** Recommended minimum wall thickness (scaled to diameter) for evacuated aluminum-alloy and stainless-steel tubes under external atmospheric pressure, as a function of tube length (scaled to diameter) between supporting flanges.

published standards for the wall thickness of cylindrical vessels under external pressure.<sup>17</sup> The data in Figure 3.35 have been abstracted from the ASME specifications for type 3003-T0 aluminum alloy and type 304 stainless steel

respectively. These recommendations should be more than adequate for any of the 4000, 6000, or 7000 series tempered aluminum alloys (such as 2024-T4, 6061-T6 or 7075-T6) and adequate for most stainless steels. Soft aluminum, such as the 1000-series alloys, should not be used for vacuum chambers. Holes cut for ports in the side of a cylinder reduce the strength of the cylinder and should be avoided. If side ports are necessary, the wall thickness should be increased over the recommendations of the code. Dents and out-of-roundness in a cylinder weaken it and should be avoided.

A flat end plate will bend a surprising amount under atmospheric pressure. Assuming the plate caps a cylindrical shape, the nature of the stress in the plate and the maximum deflection depend upon whether or not the edges of the plate are clamped to the cylinder. In this case “clamped” implies that the plate is welded to the cylinder. A removable end plate bolted in place should be considered “unclamped” when designing a vacuum chamber. The deflection at the center of a flat circular end plate clamped at its edges is:

$$\delta = \frac{3PR^4(1 - \mu^2)}{16Ed^3} \quad (\text{edge clamped}) \quad (3.47)$$

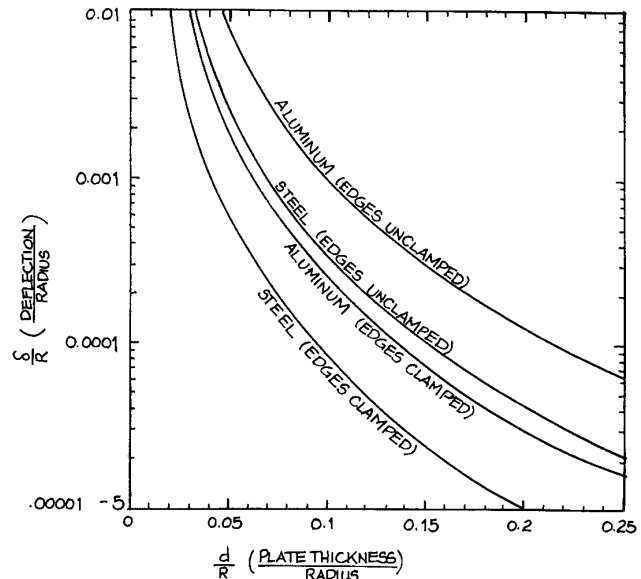
where  $P$  the external pressure,  $R$  is the radius of the plate,  $d$  the thickness,  $\mu$  Poisson’s ratio, and  $E$  the modulus of elasticity or Young’s modulus of the material.<sup>18</sup> In this case, the maximum tensile stress is at the edge of the plate (see Section 1.6.3).

The deflection of a demountable end plate is greater than that of a permanent end cap. For a circular plate with unclamped edges:

$$\delta = \frac{3PR^4(5 + \mu)(1 - \mu)}{16Ed^3} \quad (\text{unclamped}) \quad (3.48)$$

The maximum stress is at the center of the plate (see Section 1.6.3).

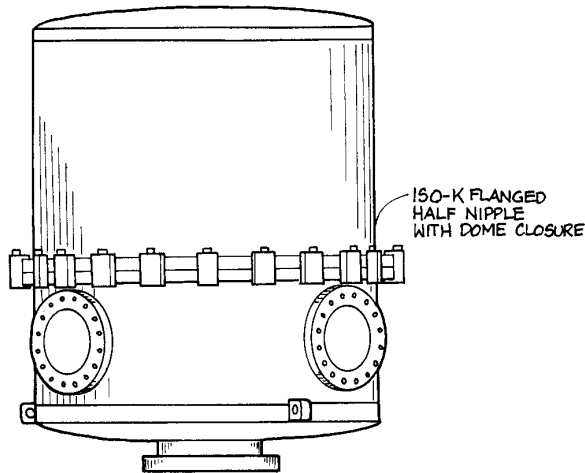
The deflection of flat end plates of steel or aluminum under atmospheric pressure is plotted in Figure 3.36. In most cases, relative deflection,  $(\delta/R)$ , in the range 0.001 to 0.005 should be acceptable, otherwise the plate may be too heavy. Steel is about three times more dense than aluminum. For a given maximum deflection, an aluminum plate is less than half as heavy as a steel plate. A vacuum-chamber design that incorporates a stainless-steel cylinder with a demountable



**Figure 3.36** Deflection of flat circular plates under atmospheric pressure. A clamped end plate caps a cylindrical tube and is welded in place, otherwise the plate should be considered as unclamped. These curves assume a Poisson’s ratio of 0.3 and a Young’s modulus of  $210 \text{ GN m}^{-2}$  ( $3 \times 10^7$  psi) for steel and  $70 \text{ GN m}^{-2}$  ( $1 \times 10^7$  psi) for aluminium.

aluminum end plate has much to say in its favor. Welds in the stainless cylinder are much stronger and more reliable than would be the case if the cylinder were made of aluminum. A removable aluminum end plate is lighter and much less expensive to machine than one of steel.

The deflection of the walls of a vacuum chamber must be accounted for in designing the mounting for a vacuum system. If both the chamber and the pump are securely mounted, severe stresses will develop as the chamber is evacuated. The correct practice is to securely mount the vacuum chamber and leave the pump suspended from a flange on the chamber, or, alternatively, to mount the pump and permit the chamber to rest on top of the pump with no further constraint. In the latter case, it is important that forces generated by the apparatus resting on the top of the pump do not distort the pump. This is particularly important for pumps requiring a high degree of mechanical precision, such as a turbomolecular pump or a molecular drag pump. It is generally recommended that turbopumps

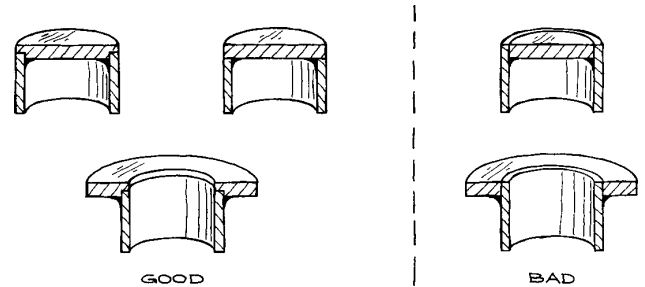


**Figure 3.37** A vacuum chamber composed of a pair of ISO-500 (500 mm diameter) half nipples with domed end caps.

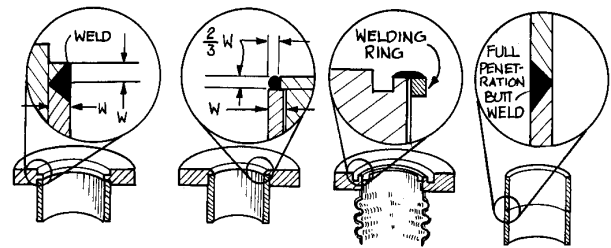
and drag pumps be suspended by their inlet flange with no other attachment.

Using standard, commercially available components saves both time and money. In fact, even quite large systems can be built entirely of standard parts. In the author's laboratory a chamber 50 cm (20 in.) in diameter and 100 cm (40 in.) high was created from a pair of ISO 500 half nipples that were ordered with domed end caps. Ports were added to the lower half nipple for feedthroughs and for a pump, as shown in Figure 3.37. The total cost was about \$5000 in 1994.

Metal parts may be joined by soldering, brazing, or welding. Tungsten – inert-gas welds (so-called “TIG” or “heliarc”) welds produce the cleanest, strongest joints in stainless steel. Soldering and brazing result in less distortion of the joined parts and are more convenient operations in the laboratory. As shown in Figure 3.38, a joint that is to be soldered or brazed should be designed so that the metal parts take the thrust of the atmosphere and the solder or brazing alloy serves only as a sealant. If possible, the joint should be designed to provide positive location of mating parts. The surfaces to be joined should be cleaned before assembly by sandblasting or by polishing with sandpaper so that the solder or brazing metal will flow into and completely fill the joint. This is necessary to achieve maximum strength and to eliminate a narrow gap that is difficult to



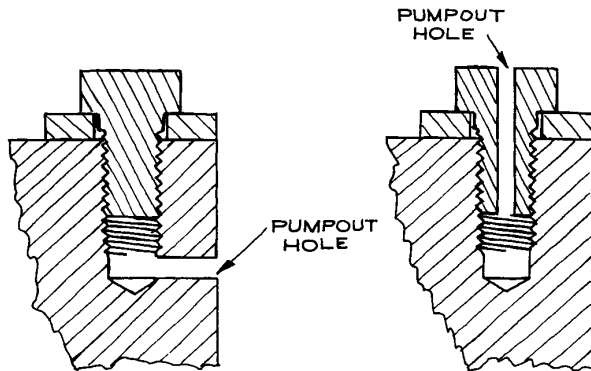
**Figure 3.38** The design of joints to be soldered or brazed.



**Figure 3.39** Welded joints in a vacuum system.

pump out and is liable to collect flux and other material that will contaminate the vacuum system. Tin-alloy solders (plumbing solder) should only be used for rough vacuum applications such as a diffusion-pump foreline. Brazed joints are acceptable in high vacuum. The copper–silver eutectic alloy is the preferred brazing material. Brazing is an excellent way of attaching thin metal parts, such as bellows, that can easily be burned through in a welding operation.

Welded joints are designed to minimized distortion in the welding process and to avoid occlusions on the vacuum side of the joint that may trap contaminants or act as virtual leaks. Parts to be joined by welding are designed so that the rate of heat dissipation while welding is the same to each part. This is necessary to prevent destructive thermal stresses from being frozen into the finished joint. When joining a heavy flange to a thin tube, a notch or a groove is cut in the thick part near the joint to control the rate at which heat is dissipated, as shown in Figure 3.39. To join a very thin part to a flange, extra material, such as the welding ring shown in the figure, is added to back up the weld. Whenever possible, welds should be on the vacuum side,



**Figure 3.40** Pumpout holes for blind screw holes.

that is, on the inside of the vacuum wall. Locational welds or “tack welds” should be on the outside and should be intermittent. For material up to about 3 mm (1/8 in.) thick, full-penetration welds should be specified. A weld from the outside, often necessary when joining a small diameter tube to a flange, should be a full penetration weld if possible. When an outside weld is unavoidable, the joint should be designed to minimize the trapped volume on the inside, as shown in the figure.

Metal parts inside a vacuum system may be joined by welding or brazing, but in most cases nut-and-bolt assembly is more convenient. Special care must be taken to prevent air from being trapped under a screw in a blind hole. The placement of pumpout holes for blind screw holes is illustrated in Figure 3.40.

### 3.6.4 Surface Preparation

Surfaces that are to be exposed to a vacuum should be free of substances that have a significant vapor pressure, and they should be as smooth as possible to minimize the microscopic surface area and thus minimize the amount of adsorbed gas. The substances that must be removed from the surfaces of vacuum apparatus are mainly hydrocarbon oils and greases, and inorganic salts that are hygroscopic and outgas water vapor.

The two preferred treatments for metal vacuum-system surfaces are electropolishing and bead blasting. Conventional wheel polishing and buffing is unsatisfactory, since these processes tend to flatten surface burrs and trap gas

underneath. Electropolishing is conveniently carried out in the laboratory, although most machine shops are prepared to do it routinely. An electropolishing solution for stainless steel recommended by Armco Steel consists of 50 parts by volume of citric acid and 15 parts by volume of sulfuric acid, plus enough water to make 100 parts of solution. The solution should be used at a temperature of 90 °C. A current density of about 0.1 A/cm<sup>2</sup> at 6 to 12 V is required. A copper cathode is used with the piece to be polished serving as the anode.

Blasting with 20 – 30 micron glass beads effectively reduces the adsorbing area of a metal surface and thus reduces the rate of outgassing from the surface. Glass beading can be carried out in a conventional sandblasting apparatus. The machine should be carefully cleaned of coarse grits before use.

Shop processes such as welding, plasma cutting, and high-speed turning can leave surfaces with a heavy layer of oxides and other contaminants. Bead blasting or, alternatively, the chemical process known as pickling may remove this layer. A pickling solution for stainless steel consists of 2% (by volume) concentrated nitric acid (HNO<sub>3</sub>) and 25% hydrochloric acid (MCI) in distilled water at 60 °C. This solution is known as “bright dip” in the shop. After dipping in this solution, the part should be thoroughly rinsed and its surface neutralized by immersion in an alkaline solution. Aluminum is pickled in a 10% solution of sodium hydroxide (NaOH) saturated with common salt (NaCl) at 80 °C. After about half a minute in this solution the part is transferred to a 10% solution of hydrochloric acid to obtain a bright finish.

An ultrasonic cleaner is particularly effective for preparing vacuum parts. The cleaner should be fired with a detergent solution. If two cleaners are available, fill one with detergent solution and the other with distilled water, and use them in sequence with a hot-water rinse between.

Diffusion-pump oil is especially difficult to remove. Nevertheless, a diffusion pump must be thoroughly cleaned when replacing old oil with new or when changing from one type of oil to another. Hydrocarbon oils, polyphenylethers, and silicone fluids can be dissolved in trichloroethylene followed by acetone and then ethyl alcohol. Trichloroethylene is somewhat toxic, and acetone and ethanol are flammable, so this cleaning should be carried out carefully in a fume hood. Avoid skin contact with these

solvents. Fluorinated diffusion-pump fluids, such as the perfluoropolyethers (Fomblin), are soluble in fluorinated solvents such as chlorotrifluoroethane.

Glass can usually be effectively cleaned by washing in a detergent solution followed by a rinse in hot water and immersion in a 10–20% solution of hydrofluoric acid. The glass is then rinsed with water, neutralized with an alkaline solution, and rinsed with distilled water. Protect eyes and skin during these operations. Hydrofluoric acid burns are especially painful and persistent.

After surface treatment as required for particular components, the final cleaning procedure for all components of a vacuum system should proceed as follows:

- (1) Scrub with a strong solution of detergent (Alconox or liquid dishwashing detergent is fine).
- (2) Rinse with very hot water.
- (3) Rinse with distilled water.
- (4) Rinse with pure methanol.

Do not touch clean parts with bare hands. Disposable plastic gloves (free of talc) are convenient for handling vacuum apparatus after cleaning.

### 3.6.5 Leak Detection

A leak that raises the base pressure of a system above about  $10^{-6}$  torr can usually be found by probing suspected locations on the outside of the vacuum chamber with a liquid or vapor for which the gauge sensitivity or pump speed is very different from that for air. A squeeze bottle of acetone or a spray can of a Freon cleaner is a useful tool. These liquids will usually cause a very abrupt increase in indicated pressure as they flow through a leak, but sometimes rapid evaporation of a liquid through a leak will cause the liquid to freeze and temporarily plug the leak, causing the pressure to fall. A disadvantage of this method is that the solvent may contaminate O-rings. A small jet of helium is also a useful leak probe because an ionization gauge is quite insensitive to helium. The indicated pressure will fall when helium is introduced into a leak.

In a glass system, a leak that raises the pressure into the 10 mtorr to several torr range can be located with a Tesla coil. The surface of the glass is brushed with the discharge from the Tesla coil. The discharge is preferentially directed

toward the leak, and a bright white spot will reveal the location as the discharge passes through. Avoid very thin glass walls and glass-to-metal seals, as the Tesla discharge can hole the glass in these fragile areas.

For very small leaks in high-vacuum systems, a mass-spectrometer leak detector is needed (see Section 3.3.5).

### 3.6.6 Ultrahigh Vacuum

To achieve pressures much below  $10^{-7}$  torr, baking is required, to remove water and hydrocarbons from vacuum-system walls. Heating to 50–100 °C for several hours will improve the ultimate pressure of most systems by an order of magnitude. A true ultrahigh-vacuum system must be baked to at least 250 °C for several hours during the initial pump out. A system contaminated with hydrocarbons may require a bakeout at temperatures approaching 400 °C. After a hard preliminary bake *in vacuo* to remove deeply adsorbed material, the system can be baked more gently after subsequent exposures to air.

A small system can be wrapped with heating tape and then covered with fiberglass insulation. For larger systems, an oven constructed of sheet metal insulated with fiberglass can be erected around the vacuum chamber and heated with bar heaters. Batts of fiberglass insulation intended for use in automobile engine compartments are readily available. Do not use home-insulating fiberglass materials. Metal seals must be used in a metal vacuum system that is to be baked. It is, however, often necessary to join the vacuum chamber to its pumping station with an O-ring-sealed flange. In this case, a cooling coil should be installed around the flange (as in Figure 3.23) to prevent damage to the O-ring while the chamber is being baked. Alternatively a metal O-ring can be used, as described in Section 3.5.2.

Quartz lamp heaters can be installed inside a vacuum system to directly heat the surfaces that need to be out-gassed. This method is frequently faster and more economical than heating from the outside. A recent innovation for removing water absorbed on the walls of vacuum apparatus employs ultraviolet radiation rather than heat. A lamp known as a Phototron, manufactured by Danielson, is inserted in the vacuum system. This lamp radiates at a wavelength absorbed by water molecules and imparts sufficient

energy to break the bond holding the water to the walls. This method is not so effective as a high-temperature bakeout, since the radiation is fairly specific for water and, in addition, the radiation cannot be expected to penetrate into small involutions where water may be trapped. The method is faster than a thermal bakeout and can be used in a system that contains materials that are degraded by heat. It is claimed to be 80 to 90% as effective as a bakeout.

Hydrocarbons from pump oils, cleaning fluids and shop processes are ubiquitous contaminants in vacuum systems. Hydrocarbons create insulating layers on electrodes in electron microscopes and mass spectrometers. Their presence is a particular problem in mass spectrometer experiments since the hydrocarbon spectrum is very rich. XEI Scientific has developed a device (the Evactron) that produces a low-pressure plasma in oxygen to make radicals that react with hydrocarbons in the vacuum system leaving only low-vapor-pressure residues ("ash").

To reduce outgassing to a minimum in an ultrahigh-vacuum system it is common practice to prefire ("stove") components before they are assembled, in order to clean the surfaces and to drive out dissolved gases. The usual shop practice is to fire components in a hydrogen atmosphere such as is provided by a hydrogen furnace or a dissociated-ammonia furnace. This has the advantage of effectively removing oxides from surfaces. The disadvantage to this procedure is that hydrogen is left to outgas at lower temperatures. Copper parts containing trapped or dissolved oxygen are also subject to embrittlement as a result of hydrogen firing. It is for this reason that oxygen-free high-conductivity (OFHC) copper should be used in constructing vacuum apparatus. High-temperature baking in vacuum or under an inert atmosphere is the preferred treatment. Iron, steel, stainless-steel, and nickel-alloy parts are fired at a temperature of about 900 °C for up to eight hours. Copper is fired at 500 °C.

## Cited References

1. N. Milleron, *Res. Dev.*, **21**(9), 40, 1970.
2. Derivations of the conductance formulae are given by C. M. VanAtta, *Vacuum Science and Engineering*, McGraw-Hill, New York, 1965, pp. 44–62 and J. F. O'Hanlon, *A User's Guide to Vacuum Technology*, 2nd edn., John Wiley & Sons, Inc., New York, 1989, Chap. 3.
3. M. A. Biondi, *Rev. Sci. Instr.*, **24**, 989, 1953.
4. A. T. J. Hayward, *J. Sci. Instr.*, **40**, 173, 1963; C. Veillon, *Rev. Sci. Instr.*, **41**, 489, 1970.
5. H. Ishii and K. Nakayama, *Transactions of the 2nd International Congress on Vacuum Science and Technology*, Vol. 1, Pergamon Press, Elmsford, NY, 1961, p. 519; R. J. Tunnicliffe and J. A. Rees, *Vacuum*, **17**, 457, 1967; T. Edmonds and J. P. Hobson, *J. Vac. Sci. Tech.*, **2**, 182, 1965.
6. G. R. Brown, P. J. Sowinski, and R. Pertel, *Rev. Sci. Instr.*, **43**, 334, 1972.
7. An application of this process is described by J. H. Moore and C. B. Opal, *Space Sci. Instr.*, **1**, 377, 1975.
8. C. B. Lucas, *Vacuum*, **23**, 395, 1973.
9. R. H. Jones, D. R. Olander, and V. R. Kruger, *J. Appl. Phys.*, **40**, 4641, 1969; D. R. Olander, *J. Appl. Phys.*, **40**, 4650, 1969; D. R. Olander, *J. Appl. Phys.*, **41**, 2769, 1970; D. R. Olander, R. H. Jones, and W. J. Siekhaus, *J. Appl. Phys.*, **41**, 4388, 1970; W. J. Siekhaus, R. H. Jones, *J. Appl. Phys.*, **41**, 4392, 1970.
10. M. G. Liverman, S. M. Beck, D. L. Monts, and R. E. Smalley, *J. Chem. Phys.*, **70**, 192, 1979; W. R. Gentry and C. F. Giese, *Rev. Sci. Instr.*, **49**, 595, 1978; C. M. Lovejoy and D. J. Nesbitt, *J. Chem. Phys.*, **86**, 3151, 1987.
11. D. Bassi, S. Iannotta, and S. Niccolini, *Rev. Sci. Instr.*, **52**, 8, 1981; F. M. Behlen, *Chem. Phys. Lett.*, **60**, 364, 1979.
12. R. E. Smalley, D. H. Levy, and L. Wharton, *J. Chem. Phys.*, **64**, 3266, 1976.
13. J. B. Anderson, R. P. Andres, and J. B. Fenn, in *Advances in Chemical Physics*, Vol. 10, *Molecular Beams*, J. Ross (ed.), John Wiley & Sons, Inc., New York, 1966, Chapter 8; H. Pauly and J. P. Toennies, in *Methods of Experimental Physics*, Vol. 7, Part A, B. Bederson and W. L. Fite, (eds.), Academic Press, New York, 1968, Chapter 3.1; D. R. Miller, in *Atomic and Molecular Beam Methods*, Vol. I, G. Scoles, (Ed.), Oxford University Press, Oxford, 1988, pp. 14–53.
14. W. D. Kingery, H. K. Bowen, and D. R. Uhlmann, *Introduction to Ceramics*, 2nd edn., John Wiley & Sons, Inc., New York, 1976, chapter 17.
15. J. F. O'Hanlon, *A User's Guide to Vacuum Technology*, 3rd edn., Wiley, Hoboken, NJ, 2003, pp. 366–369.
16. This system is an adaptation of a design described by N. W. Robinson, *Ultra-high Vacuum*, Chapman and Hall, London, 1968, pp. 72–73.
17. *A.S.M.E. Boiler and Pressure Vessel Code*, Section VIII, Division 1, Appendix V, 1974.
18. Clifford Mathews, *A.S.M.E. Engineer's Data Book*, 2nd edn., ASME Press, NY, 2001, p. 75; J. F. Harvey, *Pressure Vessel Design*, Van Nostrand, Princeton, NJ, 1963, pp. 89–91.



## General References

### Comprehensive Texts on Vacuum Technology

- S. Dushman, *Scientific Foundations of Vacuum Technique*, 2nd edn., J. M. Lafferty (Ed.), John Wiley & Sons Inc., New York, 1961.
- M. Hablanian, *High-Vacuum Technology*, Dekker, New York, 1990.
- G. Lewin, *Fundamentals of Vacuum Science and Technology*, McGraw-Hill, New York, 1965.
- J. F. O'Hanlon, *A User's Guide to Vacuum Technology*, 3rd edn., Wiley, Hoboken, NJ, 2003.
- Foundations of Vacuum Science and Technology*, J. M. Lafferty (Ed.), John Wiley & Sons, Inc., New York, 1998.
- C. M. Van Atta, *Vacuum Science and Engineering*, McGraw-Hill, New York, 1965.
- "Vacuum Technology: Its Foundations, Formulae and Tables", until 1996 published as an appendix of the Leybold AG catalog, *Product and Vacuum Technology Reference Book*, Leybold, Inc., San Jose, CA.

### Detailed Calculation of Gas Flow

- R. A. Roth, *Vacuum Technology*, 2nd edn., North-Holland, Amsterdam, 1982.
- R. G. Livesey, "Flow of Gases through Tubes and Orifices" in *Foundations of Vacuum Science and Technology*, J. M. Lafferty (Ed.), John Wiley & Sons, Inc., New York, 1998.

### Design of Vacuum Systems

- N. T. M. Dennis and T. A. Heppell, *Vacuum System Design*, Chapman and Hall, London, 1968.
- G. W. Green, *The Design and Construction of Small Vacuum Systems*, Chapman and Hall, London, 1968.
- R. P. LaPelle, *Practical Vacuum Systems*, McGraw-Hill, New York, 1972.

### Outgassing Data

- W. A. Campbell, Jr., R. S. Marriott, and J. J. Park, *A Compilation of Outgassing Data for Spacecraft Materials*, NASA Technical Note TND-7362, NASA, Washington, DC, 1973.

### Properties of Materials Used in Vacuum Systems

- W. Espe, *Materials of High Vacuum Technology*: Vol. 1, *Metals and Metalloids*; Vol. 2, *Silicates*; Vol. 3, *Auxiliary Materials*, Pergamon Press, Oxford, 1968 (a translation of the original German published in 1960).

### Sealing Ceramics and Glass to Metal, Heat-Treating, Cleaning, Building Joints, and Feedthroughs

- F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, Addison-Wesley, Reading, Mass., 1969.
- A. Roth, *Vacuum Sealing Techniques*, Pergamon Press, New York, 1966 and American Vacuum Society Classics, American Institute of Physics, New York, 1994.

### Ultrahigh Vacuum

- P. A. Redhead, J. P. Hobson, and E. V. Kornelsen, *The Physical Basis of Ultrahigh Vacuum*, Chapman and Hall, London, 1968.
- R. W. Roberts and T. A. Vanderslice, *Ultrahigh Vacuum and Its Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- W. Robinson, *The Physical Principles of Ultrahigh Vacuum Systems and Equipment*, Chapman and Hall, London, 1968.
- G. F. Weston, *Ultrahigh Vacuum Practice*, Butterworths, London, 1985.
- J. T. Yates, Jr., *Experimental Innovations in Surface Science*, Springer-Verlag, New York, 1998.

## Manufacturers and Suppliers

### General Vacuum Equipment and Instruments (Pumps, Valves, Gauges, Fittings)

- Alcatel Vacuum Products, 7 Pond St., Hanover, MA 02339
- Balzars, 8 Sagamore Park Rd., Hudson, NH 03051.
- CVC Products, Inc., 525 Lee Rd. P.O. Box 1886, Rochester, NY 14603.
- Edwards High Vacuum, Inc., 3279 Grand Island Blvd., Grand Island, NY 14072.
- Kurt J. Lesker Co., 1515 Northington Ave., Clairton, PA 15025.
- Leybold Vacuum Products, 5700 Mellon Rd., Export, PA 15632.

MDC Vacuum Products, 23842 Cabot Blvd., Hayward, CA 94545-1651.  
Nor-Cal Products, Inc., 1967 South Oregon St., P.O. Box 518, Yreka, CA 96097.  
Osaka Vacuum, Ltd., Keihan-Yodoyabashi Bldg., 2-25, Kitahama 3-chome, Chuo-ku, Osaka 541, Japan, and, 911 Bern Court, San Jose, CA 95112.  
Varian Vacuum Division, 611 Hansen Way, Palo Alto, CA 94303

## Ceramics and Ceramic Fittings

Arenco Products, P.O. Box 429, Ossining, NY 10562, (Machinable ceramics)  
Ceramaseal, P.O. Box 260, New Lebanon, NY 12125, (terminal end bushings, electrical feedthroughs, vacuum breaks)  
Coors Ceramics Co., 600 Ninth Street, P.O. Box 4025, Golden, CO 80401, (alumina shapes, fishspine)  
Coors Ceramics Co., 2449 River Rd, Grand Junction, CO 81505, (Alumina thick film substrate)  
The Materials Business, Corning Glass Works, Corning, NY 14830, (MACOR machinable ceramic)  
Industrial Tectonics, Inc., P.O. Box 1128, Ann Arbor, MI 48106, (Ceramic balls)  
Insulator Seal Incorporated, 23874-B Cabot Blvd., Hayward, CA 94545, (electrical feedthroughs, vacuum breaks, viewports)  
Latronics Corporation, 1001 Lloyd Ave., Latrobe, PA 15650, (Ceramic feedthroughs)  
McDanel Refractory Porcelain Co., 510 Ninth Ave., Beaver Falls, PA 15010, (Ceramic rod and tube)  
Specialty Ball Co., 951 West St., Rocky Hill, CT 06067, (Ceramic balls)

## Cleaning Materials

Miller-Stephenson Chemical Co., P.O. Box 628, Danbury, CT 06810, (Freon degreasers)  
Texwipe Company, Hillsdale, NJ 07642

## Components

Alloy Products Company, 1045 Perkins Ave., Waukesha, WI 53186, (Stainless-steel vacuum fittings)  
Cajon Company, 9760 Shepard Rd., Macedonia, OH 44056, (Small O-ring- and metal-sealed tube fittings)  
Combination Pump Valve Company, 851 Preston St., Philadelphia, PA 19104, (Small O-ring-sealed tube fittings)

Duniway Stockroom Corporation, 1600 Stierlin Rd., Mountain View, CA 94043, (replacements for OEM parts)  
HPS Division of MKS, 5330 Sterling Dr., Boulder, CO 80301  
Kimball Physics, Inc., 311 Kimball Hill Road, Wilton, NH 03086-9742, (small, monolithic UHV chambers and fittings)  
Ladish Company, Cudaby, WI 53110, (Stainless-steel vacuum fittings)  
Tyler Griffin, 46 Darby Rd., Paoli, PA 19301  
U-C Components, Inc., 18700 Adams Court, Morgan Hill, CA 95037, (vented screws, vacuum-baked O-rings)  
VACOA, 390 Central Ave., Bohemia, NY 11716

## Glass and Glass Components

Ace Glass Inc., P.O. Box 688, 1 430 Northwest Blvd., Vineland, NJ 08360  
Corning Glass Works, Corning, NY 14830  
Fischer & Porter Company, Warminster, PA  
Kimble Glass, P.O. Box 10350, Toledo, OH 43666  
Labglass, Northwest Blvd., Vineland, NJ 08360  
Wilmad Glass Company, Route 40 & Oak Rd., Buena, NJ 08310, (Precision glass tubing)

## Heating Mantles and Bakeout Lamps

Briskheat, P.O. Box 628, Columbus, OH 43216, (Heating mantles)  
Danielson Associates, Inc., 1989A University Lane, Lisle, IL 60532, (Bakeout lamps)  
Glas-Col Apparatus Co., 709 Hulman St., Terre Haute, IN 47802, (Heating mantles)

## Glass Microchannel Arrays

Burle Electro-Optics, Inc. (formerly Galileo), Galileo Park, Sturbridge, MA 01518  
Minitubes-Grenoble, 7 Ave de Grand Chatelet, 38100 Grenoble, France

## Nonevaporable Getter Pumps

SAES Getters, S.p.A., Group Headquarters, Viale Italia, 77, 20020 Lainate (Milano Italy)  
SAES Getters USA Inc., 1122 East Cheyenne Mountain Blvd., Colorado Springs, CO 80906

## O-Rings

EG&G Engineered Products, 11642 Old Baltimore Pike,  
Beltsville, MD 20705, (metal seals to replace O-rings)

Helicoflex Company, Components Division, 2770 The  
Boulevard, P. O. Box 9889, Columbia, SC 29290, (metal  
O-rings)

Parker Hannifin Corporation, 2360 Palumbo Dr., Lexington,  
KY 40509, (Rubber O-rings)

## Pressure Measuring Instruments

Granville-Phillips, 5675 E. Arapahoe Ave., Boulder, CO 80303,  
(Ion and thermocouple gauges and controllers)

Hastings-Raydist, P.O. Box 1275, Hampton, VA 23661,  
(Thermocouple gauge and controllers)

HPS Division of MKS, 5330 Sterling Dr., Boulder, CO 80301,  
(Pirani gauges)

MKS Instruments, 6 Shattuck Rd., Andover, MA 01810,  
tel: 978-975-2350, (Capacitance manometers)

Wallace & Tiernan, 25 Main Street, Belleville, NJ 07109,  
(Mechanical gauges)

## Pulsed Gas Sources

General Valve Corporation, 19 Gloria Lane, P. O. Box 1333,  
Fairfield, NJ 07004

## Refrigerators and Refrigerated Baffles

CVC, (Thermoelectric baffles)

FTS Systems, P.O. Box 158, Stone Ridge, NY 12484-0158,  
(Immersion coolers)

Naslab Instruments Inc., 871 Islington St., P. O. Box 1178,  
Portsmouth, NH 03801, (Immersion coolers)

## OPTICAL SYSTEMS

Physical, chemical, and biological phenomena are regularly studied or induced optically. Such experiments can involve light absorption, light emission, or light scattering. We can characterize any experimental arrangement where light is used, produced, measured, modified, or detected as an *overall optical system*. Any such optical system will always be reducible to three parts: a source of light, a detector of that light, and everything in between. We will frequently refer to this important and varied intermediary arrangement as the *optical system*. Consequently, our discussion of overall optical system design and construction will involve three key topics: sources, optical systems, and detectors (to be discussed in detail in Chapter 7). The light source may be a laser, lamp, light-emitting diode, or the sun. The detector may be a vacuum tube, solid-state device, or even the eye. Light intensity may vary from continuous wave (CW) to pulsed, and these pulses may have durations as short as a few femtoseconds. Passive elements in the system may transmit, reflect, combine, polarize, or separate light according to its spectral content. Nonlinear optical elements change the spectral content of light.

It is our aim in this chapter to explain the basic concepts that need to be understood by the experimentalist who uses optical techniques. In addition, we will provide examples of useful techniques for producing, controlling, analyzing, and modulating light.

### 4.1 OPTICAL TERMINOLOGY

Light is one form of electromagnetic radiation, the many categories of which make up the electromagnetic spec-

trum. Electromagnetic radiation, which transports energy from point to point at the velocity of light, can be described in terms of both *wave* and *particle* “pictures” or “models”. This is the famous “wave-particle” duality of all fields or particles in our model of the universe. In the electromagnetic-wave picture, waves are characterized by their frequency  $\nu$ , wavelength  $\lambda$ , and the velocity of light  $c$ , which are related by  $c = \nu\lambda$ . A propagating electromagnetic wave is characterized by a number of field vectors, which vary in time and space. These include the electric field  $\mathbf{E}$  (measured in Volt/m), the magnetic field  $\mathbf{H}$  (measured in Amp/m), the displacement vector  $\mathbf{D}$  (measured in Coulomb/m<sup>2</sup>), and the magnetic flux density  $\mathbf{B}$  (measured in Webers/m<sup>2</sup> or Tesla). For a complete description the *polarization* state of the wave must also be specified. *Linearly polarized* waves have fixed directions for their field vectors, which do not re-orient themselves as the wave propagates. *Circular* or *elliptically* polarized waves have field vectors that trace out circular or elliptic helical paths as the wave travels along. In the particle picture, electromagnetic energy is carried from point to point as quantized packets of energy called *photons*. The energy of a photon of frequency  $\nu$  is  $h\nu$ , where  $h$  is Planck’s constant –  $6.626 \times 10^{-34}$  J s. Photons have zero mass, travel at the velocity of light and also carry both linear and angular momentum. The linear momentum of a photon of wavelength is  $p = h/\lambda$ , the angular momentum depends on the equivalent polarization state of the corresponding wave. Circularly polarized photons have angular momentum  $h/2\pi = \hbar$ .

Our everyday experience of “light” generally only encompasses the small part of the electromagnetic spectrum to which the human eye is sensitive, a wavelength range

running roughly from 400–700 nm. The full electromagnetic spectrum, going from low to high frequencies, is divided into radiowaves (0–1 GHz), microwaves (1–300 GHz), infrared waves ( $\lambda$  0.7–1000  $\mu\text{m}$ )(300 GHz–430 THz)<sup>a</sup> visible light ( $\lambda$  400–700 nm), ultraviolet light ( $\lambda$  10–400 nm), X-rays ( $\lambda$  0.1–10 nm), and  $\gamma$  waves ( $\lambda < 0.1$  nm).

The wavelength region between 10 nm and 200 nm is often called the vacuum-ultraviolet region because these wavelengths are absorbed by air and most gases. Optical systems can somewhat arbitrarily be classified as systems that handle light in the spectral region between 10 nm and 100  $\mu\text{m}$ . Beyond the far-infrared, between 100 and 1000  $\mu\text{m}$  (the submillimeter wave region of the spectrum), lies a spectral region where conventional optical methods can still be used, but become difficult, as does the extension of microwave techniques for the centimeter- and millimeter-wave regions. The use of optical techniques in this region, and even in the millimeter region, is often called *quasi-optics*.<sup>1–4</sup> Only a few experimental techniques and devices that are noteworthy in this region will be mentioned.

Table 4.1 summarizes the important parameters that are used to characterize light and the media through which it passes. A few comments on the table are appropriate. Although the velocity of light in a medium depends both on the relative magnetic permeability  $\mu_r$  and dielectric constant  $\epsilon_r$  of the medium, for all practical optical materials  $\mu_r = 1$ , so the refractive index and dielectric constant are related by:

$$n = \sqrt{\epsilon_r} \quad (4.1)$$

When light propagates in an anisotropic medium, such as a crystal of lower than cubic symmetry,  $n$  and  $\epsilon_r$  will, in general, depend on the direction of propagation of the wave and its polarization state. The velocity of light *in vacuo* is currently the most precisely known of all the physical constants – its value is known<sup>5</sup> within 40 cm/s. A redefinition of the meter has been adopted<sup>6</sup> based on the cesium-atomic-clock frequency standard<sup>7</sup> and a velocity of light in a vacuum of exactly  $2.99792458 \times 10^8$  m/s. Consequently, the meter is now a derived SI unit, defined as the distance traveled *in vacuo* by light in  $1/2.99792458 \times 10^8$  seconds. The velocity, wavelength, and wavenumber of light traveling in the air have slightly different values than they have *in vacuo*. Tables that give corresponding values of  $\bar{v}$  *in vacuo* and in standard air are available.<sup>8</sup>

A plane electromagnetic wave traveling in the  $z$ -direction can, in general, be decomposed into two independent, linearly polarized components. The electric and magnetic fields associated with each of these components are themselves mutually orthogonal and transverse to the direction of propagation. They can be written as  $(E_x, H_y)$  and  $(E_y, H_x)$ . The ratio of the mutually orthogonal  $\mathbf{E}$  and  $\mathbf{H}$  components is called the *impedance*  $Z$  of the medium:

$$\frac{E_x}{H_y} = \frac{-E_y}{H_x} = Z = \sqrt{\frac{\mu_r \mu_0}{\epsilon_r \epsilon_0}} \quad (4.2)$$

Most optical materials have  $\mu_r = 1$ . The negative sign in Equation (4.2) arises because  $(y, x, z)$  is not a right-handed coordinate system. The Poynting vector:

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} \quad (4.3)$$

is a vector that points in the direction of energy propagation of the wave, as shown in Figure 4.1. The local direction of the Poynting vector at a point in a medium is called the *ray* direction. The average magnitude of the Poynting vector is called the *intensity* and is given by:

$$I = \langle |\mathbf{P}| \rangle_{\text{av}} = \frac{|\mathbf{E}|^2}{2Z} \quad (4.4)$$

The factor of two comes from time-averaging the square of the sinusoidally varying electric field.

The wavevector  $\mathbf{k}$  points in the direction perpendicular to the phase front of the wave (the surface of constant phase). In an isotropic medium  $\mathbf{k}$  and  $\mathbf{P}$  are always parallel.

The photon flux corresponding to an electromagnetic wave of average intensity  $I$  and frequency  $\nu$  is:

$$N = \frac{I}{h\nu} = \frac{I\lambda}{hc} \quad (4.5)$$

where  $h$  is Planck's constant,  $6.6 \times 10^{-34}$  J s. For a wave of intensity  $1 \text{ Wm}^2$  and wavelength  $1 \mu\text{m}$  *in vacuo*,  $N = 5.04 \times 10^{18}$  photons  $\text{sec}^{-1} \text{ m}^{-2}$ . Photon energy is sometimes measured in electron volts (eV):  $1 \text{ eV} = 1.60202 \times 10^{-19}$  J. A photon of wavelength  $1 \mu\text{m}$  has an energy of 1.24 eV. It is often important, particularly in the infrared, to know the correspondence between photon and thermal energies. The characteristic thermal energy at absolute temperature  $T$  is  $kT$ . At 300 K,  $kT = 4.14 \times 10^{-21} \text{ J} = 208.6/\text{cm} = 0.026 \text{ eV}$ .

**Table 4.1 Fundamental parameters of electromagnetic radiation and optical media**

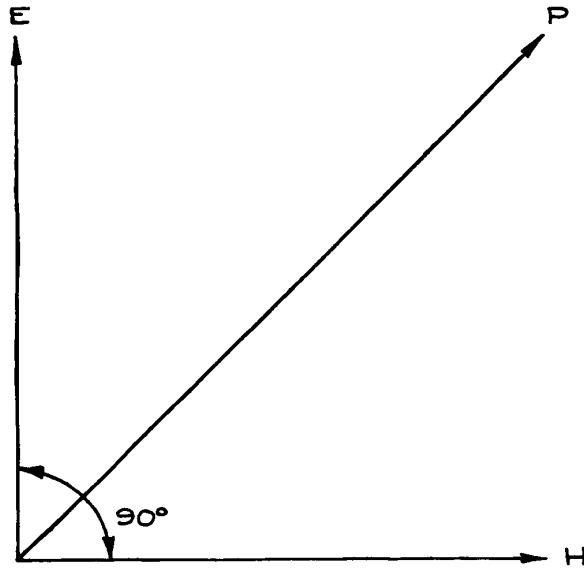
<i>Parameter</i>	<i>Symbol</i>	<i>Value</i>	<i>Units</i>
Velocity of light <i>in vacuo</i>	$c_0 = (\mu_0 \epsilon_0)^{1/2}$	$2.99792458 \times 10^8$	m/sec
Permeability of free space	$\mu_0$	$4\pi \times 10^{-7}$	henry/m
Permittivity of free space	$\epsilon_0$	$8.85416 \times 10^{-12}$	farad/m
Velocity of light in a medium	$c = (\mu_r \mu_0 \epsilon_r \epsilon_0)^{-1/2} = c_0/n$		m/sec
Refractive index	$n = (\mu_r \epsilon_r)^{1/2}$		(dimensionless)
Relative permeability of a medium	$\mu_r$	Usually 1	(dimensionless)
Dielectric constant of a medium	$\epsilon_r$		(dimensionless)
Frequency	$\nu = c/\lambda$		Hz $10^9$ Hz = 1 GHz (gigahertz) $10^{12}$ Hz = 1 THz (terahertz)
Wavelength <i>in vacuo</i>	$\lambda_0 = c_0/\nu^{***}$		m
Wavelength in a medium	$\lambda = c/\nu = \lambda_0/n$		$10^{-3}$ m = 1 mm (millimeter) $10^{-6}$ m = 1 $\mu$ m (micrometer) = 1 $\mu$ (micron) $10^{-9}$ m = 1 nm (nanometer) = 1 m $\mu$ (millimicron) $10^{-12}$ m = 1 pm (picometer) $10^{-10}$ m = 1 Å (Angstrom)
Wavenumber	$\bar{\nu} = 1/\lambda$		/cm (kayser)
Wavevector	$\mathbf{k},  \mathbf{k}  = 2\pi/\lambda$		/m
Photon energy	$E = h\nu$		J $1.60202 \times 10^{19}$ J = 1 eV (electron volt)
Electric field of wave	$\mathbf{E}$		v/m
Magnetic field of wave	$\mathbf{H}$		A/m
Poynting vector	$\mathbf{P} = \mathbf{E} \times \mathbf{H}$		W/m <sup>2</sup>
Intensity	$I = \langle  \mathbf{P} _{av} \rangle =  E ^2/2Z$		W/m <sup>2</sup>
Impedance of medium	$Z = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \left( \frac{\mu_r \mu_0}{\epsilon_r \epsilon_0} \right)^{1/2}$		$\Omega$
Impedance of free space	$Z_0 = (\mu_0/\epsilon_0)^{1/2}$	376.7	$\Omega$

## 4.2 CHARACTERIZATION AND ANALYSIS OF OPTICAL SYSTEMS

Before embarking on a detailed discussion of the properties and uses of passive optical components and systems containing them, it is worthwhile reviewing some of the parameters that are important in characterizing passive optical systems and some of the methods that are useful

in analyzing the way in which light passes through an optical system.

Optical materials have refractive indices that vary with wavelength. This phenomenon is called *dispersion*. It causes a wavelength dependence of the properties of an optical system containing transmissive components. The change of index with wavelength is very gradual, and often negligible, unless the wavelength approaches a region where the material is not transparent. Most materials exhibit *normal dispersion*, where the refractive index



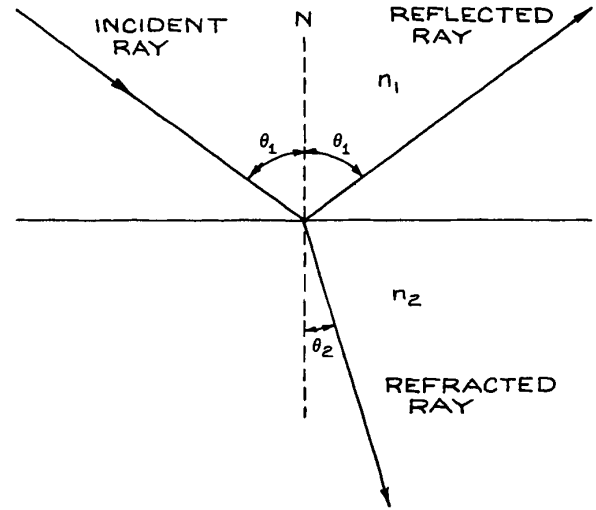
**Figure 4.1** Orientation of the mutually orthogonal electric-field vector **E**, magnetic-field vector **H**, and Poynting vector **P**.

decreases with wavelength. In specific spectral regions they may exhibit *anomalous dispersion*, where the refractive index increases with wavelength over a limited range of wavelengths. A description of the performance of an optical system is often simplified by assuming that the light is *monochromatic* (light that contains only a small spread of wavelength components).

#### 4.2.1 Simple Reflection and Refraction Analysis

The phenomena of reflection and refraction are most easily understood in terms of plane electromagnetic waves – those in which the direction of energy flow (the ray direction) is unique. Other types of wave, such as spherical waves and Gaussian beams, are also important in optical science; however, the part of their wavefront that strikes an optical component can frequently be approximated as a plane wave, so plane-wave considerations of reflection and refraction still hold true.

When light is reflected from a plane mirror, or the planar boundary between two media of different refractive index, the *angle of incidence* is always equal to the *angle of*



**Figure 4.2** Reflection and refraction of a light ray at the boundary between two different isotropic media of refractive indices  $n_1$  and  $n_2$ , respectively. The case shown is for  $n_2 > n_1$ . The incident, reflected, and transmitted rays and the surface normal  $N$  are coplanar.

*reflection*, as shown in Figure 4.2. This is the fundamental *law of reflection*.

When a light ray crosses the boundary between two media of different refractive index, the *angle of refraction*  $\theta_2$ , shown in Figure 4.2, is related to the angle of incidence  $\theta_1$  by *Snell's law*:

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1} \quad (4.6)$$

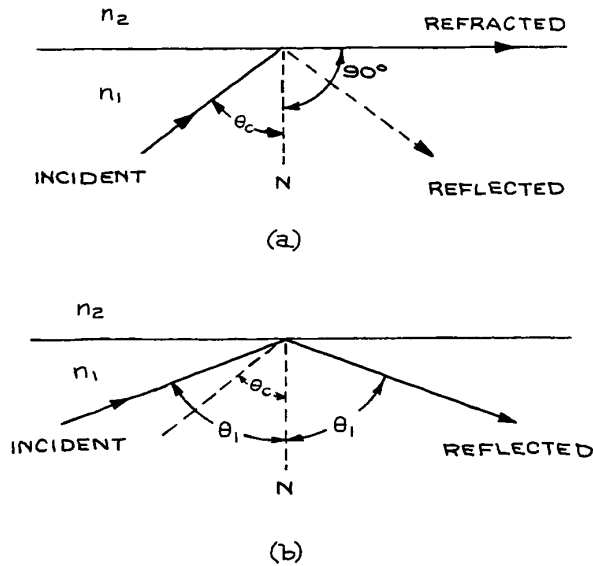
This result is modified if one or both of the media is anisotropic.

If  $n_2 < n_1$  there is a maximum angle of incidence for which there can be a refracted wave, since  $\sin \theta_2$  cannot be greater than unity for a real angle. This is called the *critical angle*  $\theta_c$  and is given by:

$$\sin \theta_c = n_2/n_1 \quad (4.7)$$

as illustrated in Figure 4.3(a)

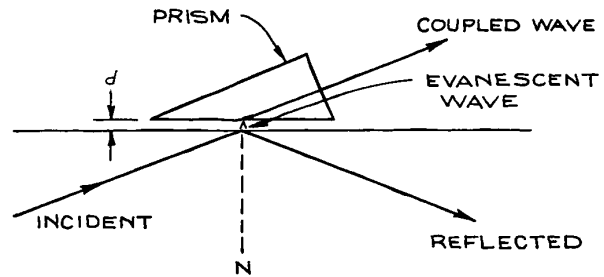
If  $\theta_1$  exceeds  $\theta_c$ , the boundary acts as a very good mirror, as illustrated in Figure 4.3(b). This phenomenon is called *total internal reflection*. Several types of reflecting prisms discussed in Section 4.3.4 operate this way. When total internal reflection occurs, there is no



**Figure 4.3** (a) Critical angle ( $n_1 > n_2$ ); (b) total internal reflection ( $n_1 > n_2$ ).  $N$  is the surface normal.

transmission of energy through the boundary. The fields of the wave do not, however, go abruptly to zero at the boundary. There is an *evanescent wave* on the other side of the boundary, the field amplitudes of which decay exponentially with distance. For this reason, other optical components should not be brought too close to a totally reflecting surface, or energy will be coupled to them via the evanescent wave and the efficiency of total internal reflection will be reduced. With extreme care this effect can be used to produce a variable-reflectivity, totally internally reflecting surface, as can be seen in Figure 4.4. Studies of total internal reflection at interfaces are important in a number of optical measurement techniques, such as attenuated total (internal) reflection<sup>9</sup> and photon scanning tunneling microscopy.<sup>10</sup>

One or both of the media in Figure 4.2 may be anisotropic, like calcite, crystalline quartz, ammonium dihydrogen phosphate (ADP), potassium dihydrogen phosphate (KDP), or tellurium. Then the incident wave in general will split into two components, one of which obeys Snell's law directly (the ordinary wave) and one for which Snell's law is modified (the extraordinary wave). This phenomenon is called *double refraction*.<sup>11-13</sup>



**Figure 4.4** Schema of an evanescent-wave coupler. The amount of intensity reflected or coupled is varied by adjusting the spacing  $d$ ; shown greatly exaggerated in the figure. This spacing is typically on the order of the wavelength. The surface boundary and adjacent prism face must be accurately flat and parallel.

If an optical system contains only planar interfaces, the path of a ray of light through the system can be easily calculated using only the law of reflection and Snell's law. This simple approach neglects diffraction effects, which become significant unless the lateral dimensions (*apertures*) of the system are all much larger than the wavelength (say 10 times larger). The simple behavior of light rays in more complex systems containing nonplanar components, but where diffraction effects are negligible, can be described with the aid of *paraxial-ray analysis*.<sup>14,15</sup> Transmitted and reflected intensities and polarization states cannot be determined by the above methods and are most easily determined by the *method of impedances*.

## 4.2.2 Paraxial-Ray Analysis

A plane wave is characterized by a unique propagation direction given by the wave vector  $\mathbf{k}$ . All fields associated with the wave are, at a given time, equal at all points in infinite planes orthogonal to the propagation direction. In real optical systems ideal plane waves do not exist, as the finite size of the elements of the system restricts the lateral extent of the waves. Nonplanar optical components will cause further deviations of the wave from planarity. Consequently, the wave acquires a ray direction that varies from point to point on the phase front. The behavior of the optical system must be characterized in terms of the deviations its elements cause to the bundle of rays that constitute the propagating, laterally restricted wave. This is most easily done in terms of paraxial rays. In a cylindrically symmetric



optical system (for example, a coaxial system of spherical lenses or mirrors), *paraxial rays* are those rays whose directions of propagation occur at sufficiently small angles  $\theta$  to the symmetry axis of the system that it is possible to replace  $\sin \theta$  or  $\tan \theta$  by  $\theta$ —in other words, paraxial rays obey the small-angle approximation:

$$\sin \theta \simeq \tan \theta \simeq \theta \quad (4.8)$$

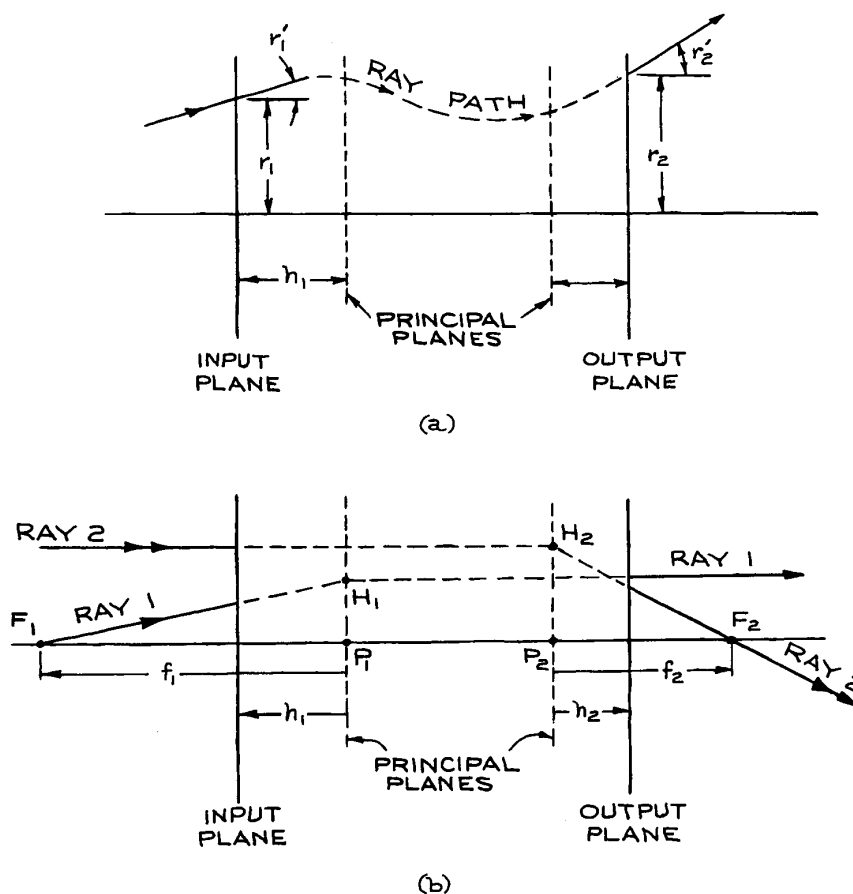
**Matrix Formulation.** In an optical system whose symmetry axis is in the  $z$ -direction, a paraxial ray in a given cross-section ( $z = \text{constant}$ ) is characterized by its

distance  $r$  from the  $z$ -axis and the angle  $r'$  it makes with that axis. Suppose the values of these parameters at two planes of the system (an *input* and an *output* plane) are  $r_1$ ,  $r'_1$  and  $r_2$ ,  $r'_2$ , respectively, as shown in Figure 4.5(a). Then in the paraxial-ray approximation, there is a linear relation between them of the form:

$$\begin{aligned} r_2 &= Ar_1 + Br'_1 \\ r'_2 &= Cr_1 + Dr'_1 \end{aligned} \quad (4.9)$$

or, in matrix notation:

$$\begin{pmatrix} r_2 \\ r'_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} r_1 \\ r'_1 \end{pmatrix} \quad (4.10)$$



**Figure 4.5** (a) Generalized schematic diagram of an optical system, showing a typical ray and its paraxial-ray parameters at the input and output planes; (b) focal points  $F_1$  and  $F_2$  and principal rays of a generalized optical system.

Here:

$$\mathbf{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad (4.11)$$

is called the *ray transfer matrix*. If the media where the input and output ray parameters are measured have the same refractive index, then the determinant of the ray transfer matrix is unity, i.e.,  $AD - BC = 1$ .

Optical systems made of isotropic material are generally reversible – a ray that travels from right to left with input parameters  $r_2, r'_2$  will leave the system with parameters  $r_1, r'_1$ . Thus:

$$\begin{pmatrix} r_1 \\ r'_1 \end{pmatrix} = \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} \begin{pmatrix} r_2 \\ r'_2 \end{pmatrix} \quad (4.12)$$

where the reverse ray transfer matrix is:

$$\begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} \quad (4.13)$$

The ray transfer matrix allows the properties of an optical system to be described in general terms by the location of its *focal points* and *principal planes*, whose location are determined from the elements of the matrix. The significance of these features of the system can be illustrated with the aid of Figure 4.5(b). An input ray that passes through the *first focal point*  $F_1$  (or would pass through this point if it did not first enter the system) emerges traveling parallel to the axis. The intersection point of the extended input and output rays, point  $H_1$  in Figure 4.5(b), defines the location of the *first principal plane*. Conversely, an input ray traveling parallel to the axis will emerge at the output plane and pass through the second focal point  $F_2$  (or appear to have come from this point). The intersection of the extension of these rays, point  $H_2$ , defines the location of the *second principal plane*. Rays 1 and 2 in Figure 4.5(b) are called the *principal rays* of the system. The location of the principal planes allows the corresponding emergent ray paths to be determined, as shown in Figure 4.5(b). The dashed lines in this figure, which permit the geometric construction of the location of output rays 1 and 2, are called *virtual ray paths*. Both  $F_1$  and  $F_2$  lie on the axis of the system. The axis of the system intersects the principal planes at the *principal points*,  $P_1$  and  $P_2$ , in

Figure 4.5(b). The distance,  $f_1$ , from the first principal plane to the first focal point is called the *first focal length*;  $f_2$  is called the *second focal length*.

In many practical situations, the refractive indices of the media to the left of the input plane (the *object space*) and to the right of the output plane (the *image space*) are equal. In this case, several simplifications arise:

$$\begin{aligned} f_1 = f_2 = f &= -\frac{1}{C} \\ h_1 &= \frac{D-1}{C} \\ h_2 &= \frac{A-1}{C} \end{aligned} \quad (4.14)$$

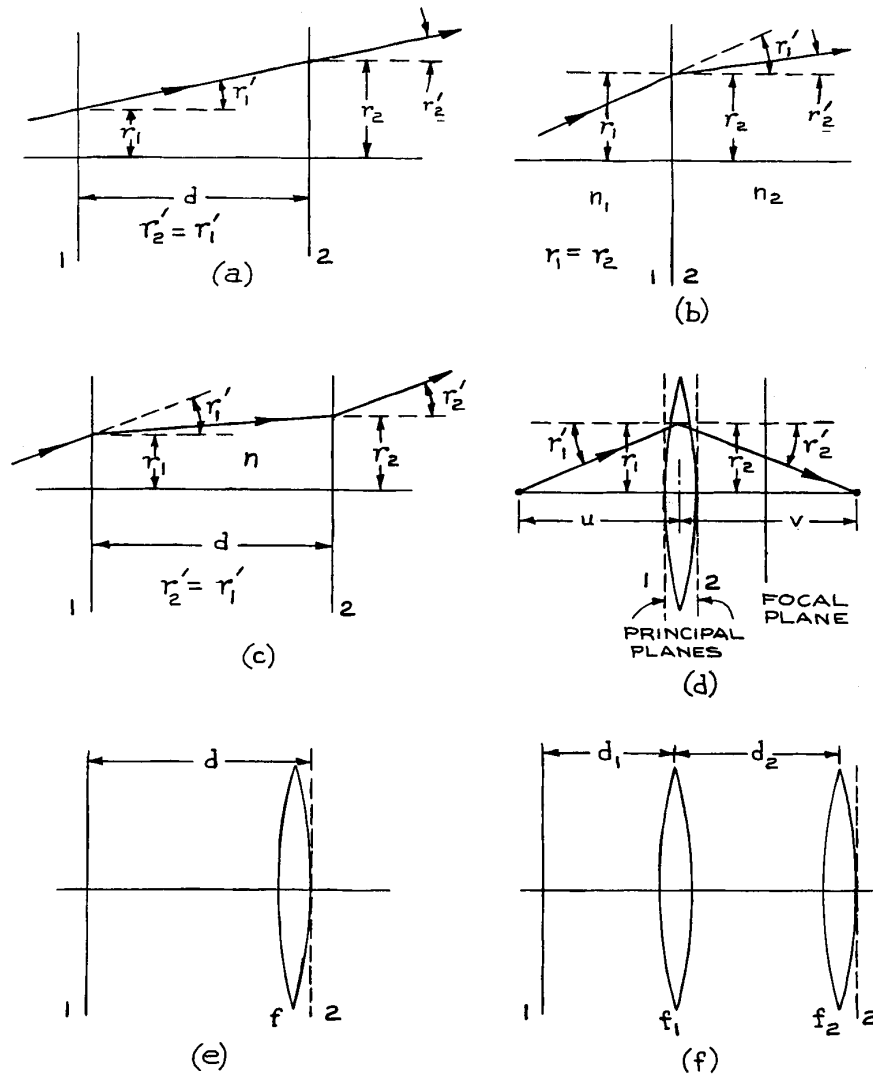
$h_1$  and  $h_2$  are the distances of the input and output planes from the principal planes, measured in the sense shown in Figure 4.5(b).

Thus, if the elements of the transfer matrix are known, the location of the focal points and principal planes is determined. Graphical construction of ray paths through the system using the methods of paraxial *ray tracing* is then straightforward (see Section 4.2.3).

In using the matrix method for optical analysis, a consistent sign convention must be employed. In the present discussion, a ray is assumed to travel in the positive  $z$ -direction from left to right through the system. The distance from the first principal plane to an object is measured positive from right to left – in the negative  $z$ -direction. The distance from the second principal plane to an image is measured positive from left to right – in the positive  $z$  direction. The lateral distance of the ray from the axis is positive in the upward direction, negative in the downward direction. The acute angle between the system-axis direction and the ray, say  $r_1$  in Figure 4.5(a), is positive if a counterclockwise motion is necessary to go from the positive  $z$ -direction to the ray direction. When the ray crosses a spherical interface, the radius of curvature is positive if the interface is convex to the input ray. The use of ray transfer matrices in optical-system analysis can be illustrated with some specific examples.

**(i) Uniform optical medium.** In a uniform optical medium of length  $d$ , no change in ray angle occurs, as illustrated in Figure 4.6(a); so:

$$\begin{aligned} r'_2 &= r'_1 \\ r_2 &= r_1 + dr'_1 \end{aligned} \quad (4.15)$$



**Figure 4.6** Simple optical systems for illustrating the application of ray transfer matrices, with input and output planes marked 1 and 2: (a) uniform optical medium; (b) planar interface between two different media; (c) a parallel-sided slab of refractive index  $n$  bounded on both sides with media of refractive index 1; (d) thin lens ( $r'_2$  is a negative angle; for a thin lens,  $r_2 = r_1$ ); (e) a length of uniform medium plus a thin lens; (f) two thin lenses.

Therefore:

$$\mathbf{M} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \quad (4.16)$$

The focal length of this system is infinite and it has no specific principal planes.

**(ii) Planar interface between two different media.**

At the interface, as shown in Figure 4.6(b), we have

$r_1 = r_2$ , and from Snell's law, using the approximation  $\sin \theta \simeq \theta$ :

$$r'_2 = \frac{n_1}{n_2} r'_1 \quad (4.17)$$

Therefore:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & n_1/n_2 \end{pmatrix} \quad (4.18)$$

**(iii) A parallel-sided slab of refractive index  $n$  bounded on both sides with media of refractive index  $I$  [Figure 4.6(c)].** In this case:

$$\mathbf{M} = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix} \quad (4.19)$$

The principal planes of this system are the boundary faces of the optically dense slab.

**(iv) Thick lens.** The ray transfer matrix of the thick lens shown in Figure 4.7(a) is the product of the three transfer matrices:

$$\begin{aligned} \mathbf{M}_3 \mathbf{M}_2 \mathbf{M}_1 &= (\text{matrix for second spherical interface}) \\ &\quad \times (\text{matrix for medium of length } d) \\ &\quad \times (\text{matrix for first spherical interface}) \end{aligned} \quad (4.20)$$

Note the order of these three matrices;  $\mathbf{M}_1$  comes on the right because it operates first on the column vector that describes the input ray.

At the first spherical surface:

$$n'(r'_1 + \phi_1) = nr = n(r''_2 + \phi_1) \quad (4.21)$$

and, since  $\theta_1 = r_1/R_1$ , this equation can be rewritten as:

$$r''_2 = \frac{nr'_1}{n} + \frac{(n' - n)r_1}{nR_1} \quad (4.22)$$

The transfer matrix at the first spherical surface is:

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 \\ \frac{n' - n}{nR_1} & \frac{n'}{n} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{d_1}{n} & \frac{n'}{n} \end{pmatrix} \quad (4.23)$$

where  $D_1 = (n - n')/R_1$  is called the *power* of the surface. If  $R_1$  is measured in meters, the units of  $D_1$  are *diopeters*.

In the paraxial approximation, all the rays passing through the lens travel the same distance  $d$  in the lens.

Thus:

$$\mathbf{M}_2 = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \quad (4.24)$$

The ray transfer matrix at the second interface is:

$$\mathbf{M}_3 = \begin{pmatrix} 1 & 0 \\ \frac{n - n'}{n'R_2} & \frac{n}{n'} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{D_2}{n'} & \frac{n}{n'} \end{pmatrix} \quad (4.25)$$

which is identical in form to  $\mathbf{M}_1$ . Note that in this case both  $r'_2$  and  $R_2$  are negative. The overall transfer matrix of the thick lens is:

$$\mathbf{M} = \mathbf{M}_3 \mathbf{M}_2 \mathbf{M}_1 = \begin{pmatrix} 1 - \frac{dD_1}{n} & \frac{dn'}{n} \\ \frac{dD_1 D_2}{nn'} - \frac{D_1}{n'} & 1 - \frac{dD_2}{n'} \end{pmatrix} \quad (4.26)$$

If Equations (4.26) and (4.14) are compared, it is clear that the locations of the principal planes of the thick lens are:

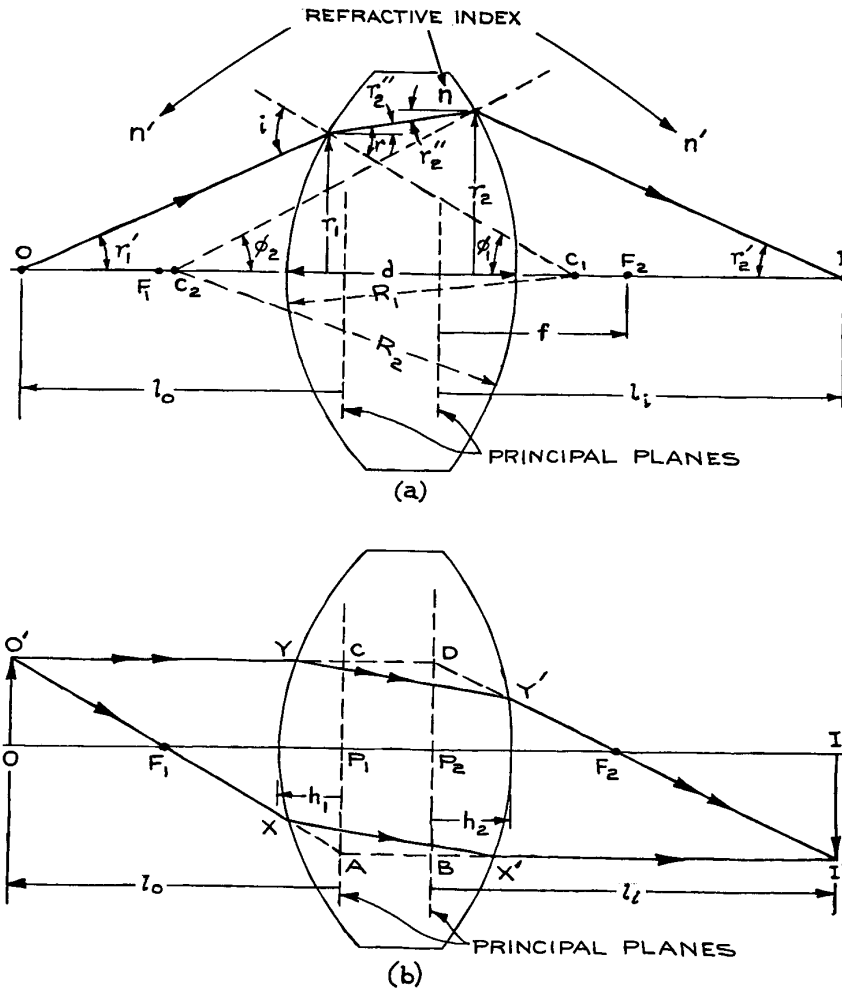
$$h_1 = \frac{d}{\frac{n}{n'} \left( 1 + \frac{D_1}{D_2} - \frac{dD_1}{n} \right)} \quad (4.27)$$

$$h_2 = \frac{d}{\frac{n}{n'} \left( 1 + \frac{D_2}{D_1} - \frac{dD_2}{n} \right)} \quad (4.28)$$

A numerical example will best illustrate the location of the principal planes for a biconvex thick lens. Suppose that:

$$\begin{aligned} n' &= 1(\text{air}), \quad n = 1.5(\text{glass}) \\ R_1 &= -R_2 = 50 \text{ mm}, \\ d &= 10 \text{ mm} \end{aligned} \quad (4.29)$$

In this case,  $D_1 = D_2 = 0.01$  mn and, from Equations (4.27) and (4.28),  $h_1 = h_2 = 3.448$  mm. These principal planes are symmetrically placed inside the lens. Figure 4.7(b) shows how the principal planes can be used to trace the principal-ray paths through a thick lens.



**Figure 4.7** (a) Diagram illustrating ray propagation through a thick lens; (b) diagram showing the use of the principal planes to determine the principal ray paths through the lens. The dashed paths  $XABX'$  and  $YCDY'$  are the virtual-ray paths; the solid lines  $XX'$  and  $YY'$  are the real-ray paths.

From Equation (4.26):

$$r'_2 = -\frac{r_1}{f} + \left(1 - \frac{dD_2}{n'}\right)r'_1 \quad (4.30)$$

where the focal length is:

$$f = \left(\frac{D_1 + D_2}{n'} - \frac{dD_1D_2}{nn'}\right)^{-1} \quad (4.31)$$

If  $l_o$  is the distance from the object  $O$  to the first principal plane, and  $l_i$  the distance from the second principal plane to the image  $I$  in Figure 4.7(b), then from the similar triangles  $OO'F_1$ ,  $P_1AF_1$ , and  $P_2DF_2$ ,  $II'F_2$ ,

$$\begin{aligned} \frac{OO'}{II'} &= \frac{OF_1}{F_1P_1} = \frac{l_o - f_1}{f_1} \\ \frac{OO'}{II'} &= \frac{P_2F_2}{F_2I} = \frac{f_2}{l_i - f_2} \end{aligned} \quad (4.32)$$

If the media on both sides of the lens are the same, then  $f_1 = f_2 = f$  and it immediately follows that:

$$\frac{1}{l_0} + \frac{1}{l_i} = \frac{1}{f} \quad (4.33)$$

This is the fundamental imaging equation.

**(v) Thin lens.** If a lens is sufficiently thin that to a good approximation  $d = 0$ , the transfer matrix is:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ -\frac{D_1 + D_2}{n'} & 1 \end{pmatrix} \quad (4.34)$$

As shown in Figure 4.6(d), the principal planes of such a thin lens are at the lens. The focal length of the thin lens is  $f$ , where:

$$\frac{1}{f} = \frac{D_1 + D_2}{n'} = \left(\frac{n}{n'} - 1\right) \left(\frac{1}{R_1} - \frac{1}{R_2}\right) \quad (4.35)$$

so the transfer matrix can be written very simply as:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}. \quad (4.36)$$

The focal length of the lens depends on the refractive indices of the lens material and of the medium within which it is immersed. In air:

$$\frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{R_2}\right) \quad (4.37)$$

For a biconvex lens,  $R_2$  is negative and:

$$\frac{1}{f} = (n - 1) \left(\frac{1}{|R_1|} + \frac{1}{|R_2|}\right) \quad (4.38)$$

For a biconcave lens:

$$\frac{1}{f} = -(n - 1) \left(\frac{1}{|R_1|} + \frac{1}{|R_2|}\right) \quad (4.39)$$

The focal length of any diverging lens is negative. For a thin lens the object and image distances are measured to a common point. It is common practice to rename the dis-

tances  $l_0$  and  $l_i$  in this case, so that the imaging Equation (4.33) reduces to its familiar form:

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f} \quad (4.40)$$

Then  $u$  is called the object distance (or object *conjugate*) and  $v$  the image distance (or image *conjugate*).

**(vi) A length of uniform medium plus a thin lens [Figure 4.6(e)].** This is a combination of the systems in (i) and (v); its overall transfer matrix is found from Equations (4.16) and (4.36) as:

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & d \\ -1/f & 1 - d/f \end{pmatrix} \end{aligned} \quad (4.41)$$

**(vii) Two thin lenses.** As a final example of the use of ray transfer matrices, consider the combination of two thin lenses shown in Figure 4.6(f). The transfer matrix of this combination is:

$$\begin{aligned} \mathbf{M} &= (\text{matrix of second lens}) \\ &\times (\text{matrix of uniform medium of length } d_2) \\ &\times (\text{matrix of first lens}), \\ &\times (\text{matrix of uniform medium of length } d_1) \end{aligned} \quad (4.42)$$

which can be shown to be:

$$\mathbf{M} = \begin{pmatrix} 1 - \frac{d_2}{f_1} & d_1 + d_2 - \frac{d_1 d_2}{f_1} \\ -\frac{1}{f_1} - \frac{1}{f_2} + \frac{d_2}{f_1 f_2} & 1 - \frac{d_1}{f_1} - \frac{d_2}{f_2} - \frac{d_1}{f_2} + \frac{d_1 d_2}{f_1 f_2} \end{pmatrix} \quad (4.43)$$

The focal length of the combination is:

$$f = \frac{f_1 f_2}{(f_1 + f_2) - d_2} \quad (4.44)$$

The optical system consisting of two thin lenses is the standard system used in analyses of the stability of lens waveguides and optical resonators.<sup>14-16</sup>

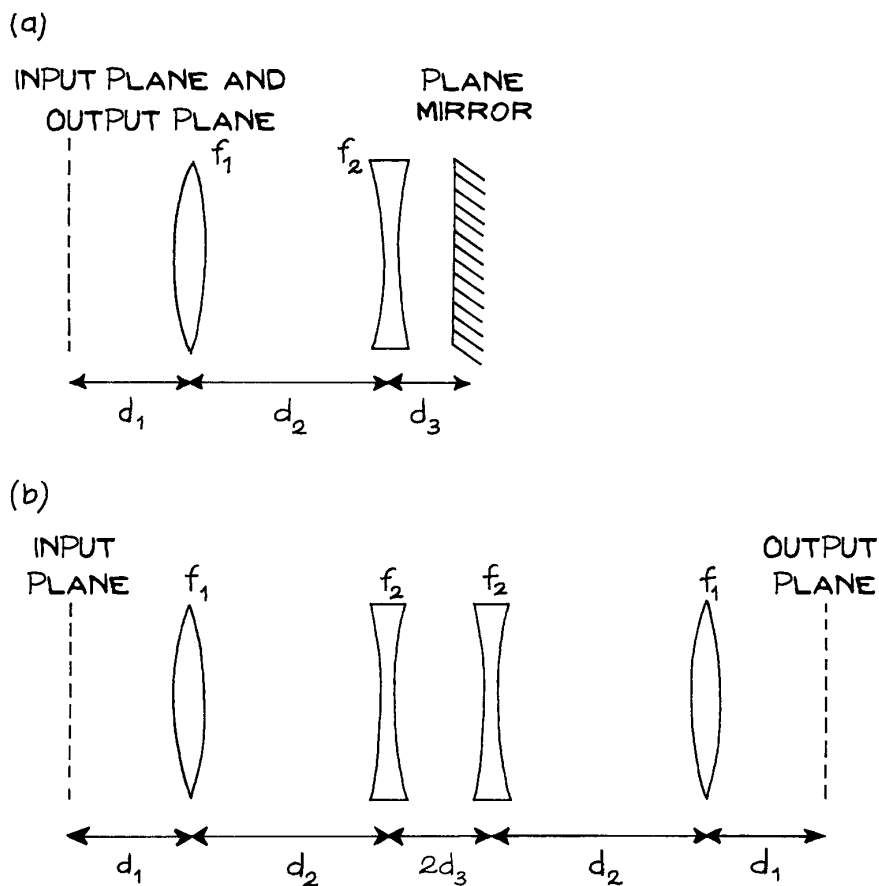
(viii) **Optical systems with plane mirrors.** Optical systems in which a plane mirror is arranged perpendicular to the system axis are most easily analyzed by an *unfolding* technique. For, example if an optical system with ray transfer matrix  $\mathbf{M}$  is terminated by a plane mirror, as shown in Figure 4.8(a), the “unfolded” system is as shown in Figure 4.8(b). The ray parameters at the original input plane, after a ray has passed back through the systems can be found as follows:

$$\begin{pmatrix} r_2 \\ r'_2 \end{pmatrix} = \mathbf{M}_{\text{rev}} \mathbf{M} \begin{pmatrix} r_1 \\ r'_1 \end{pmatrix} \quad (4.45)$$

where  $\mathbf{M}_{\text{rev}}$  is the ray transfer matrix for the optical system in the reverse direction. Equation (4.45) can be also written as

$$\begin{pmatrix} r_2 \\ r'_2 \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{M} \begin{pmatrix} r_1 \\ r'_1 \end{pmatrix} \quad (4.46)$$

(ix) **Spherical mirrors.** The object and image distances from a spherical mirror also obey Equation (4.40), where the focal length of the mirror is  $R/2$ ;  $f$  is positive for a concave mirror, negative for a convex mirror. Positive object and image distances for a mirror are measured



**Figure 4.8** (a) A simple optical system with two lenses and a plane mirror; (b) the equivalent “unfolded” system that can be used for ray transfer matrix analysis.

positive in the normal direction away from its surface. If a negative image distance results from Equation (4.40), this implies a *virtual image* (behind the mirror).

**Ray Tracing.** In designing an optical system, ray tracing is a powerful technique for analyzing the design and assessing its performance. Ray tracing follows the trajectories of light rays through the system. These light rays represent, at each point within the system, the local direction of energy flow, represented mathematically by the Poynting vector:

The trajectory of a light ray is governed by the laws of reflection and refraction. At a mirror surface the angle of reflection is equal to the angle of incidence. At the boundary between two media of different refractive index  $n_1$ ,  $n_2$ , respectively, the angle of incidence,  $\theta_1$ , and refraction,  $\theta_2$ , are related by Snell's Law<sup>b</sup>:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (4.47)$$

In a medium where the refractive index varies from point to point, Equation (4.47) still holds true, but it is better in this case to use the equation of light rays:

$$\frac{d}{ds} \left( n \frac{d\mathbf{r}}{ds} \right) = \text{grad } n \quad (4.48)$$

where  $s$  is the distance measured along the path of the light ray, and  $n(\mathbf{r})$  is the index of refraction at point  $\mathbf{r}$ , measured with respect to the origin.<sup>15</sup>

Ray tracing provides an accurate description of the way light passes through an optical system unless diffraction effects become important. This happens when any apertures in the system become comparable in size to the wavelength, or when two adjacent apertures in the system subtend angles that are comparable to the diffraction angle.

At a surface of diameter  $2a$ , the diffraction angle for light of wavelength  $\lambda$  is:

$$\theta_{\text{diff}} = \frac{1.22\lambda}{2a_1} \quad (4.49)$$

This is the angular position of the first intensity minimum in the ring-shaped diffraction pattern that results when a plane-wave passes through a circular aperture. If a second

aperture of diameter  $2a_2$  is placed a distance  $d$  from the first, the angle subtended is:

$$\theta = \frac{2a_2}{d} \quad (4.50)$$

For diffraction to be negligible requires  $\theta_{\text{diff}} \ll \theta$ , which gives:

$$\frac{\lambda d}{4a_1 a_2} \ll 1 \quad (4.51)$$

often called the Fresnel criterion.

**Paraxial Ray Tracing.** Practical implementation of paraxial-ray analysis in optical-system design can be very conveniently carried out graphically by *paraxial ray tracing*. In ray tracing a few simple rules allow geometrical construction of the principal-ray paths from an object point. (These constructions do not take into account the nonideal behavior, or *aberrations*, of real lenses, which will be discussed later.) The first principal ray from a point on the object passes through (or its projection passes through) the first focal point. From the point where this ray, or its projection, intersects the first principal plane, the output ray is drawn parallel to the axis. The actual ray path between input and output planes can be found in simple cases – for example, the path  $XX'$  in the thick lens shown in Figure 4.7(b). The second principal ray is directed parallel to the axis; from the intersection of this ray, or its projection, with the second principal plane, the output ray passes through (or appears to have come from) the second focal point. The actual ray path between input and output planes can again be found in simple cases – for example, the path  $YY'$  in the thick lens shown in Figure 4.7(b). The intersection of the two principal rays in the image space produces the image point that corresponds to the original point on the object. If only the back projections of the output principal rays appear to intersect, this intersection point lies on a *virtual image*.

In the majority of applications of paraxial ray tracing, a quick analysis of the system is desired. In this case, if all lenses in the system are treated as thin lenses, the position of the principal planes need not be calculated beforehand and ray tracing becomes particularly easy. For each lens, the position of the image of an object is obtained, and this image then becomes the “object” for the next lens, and so

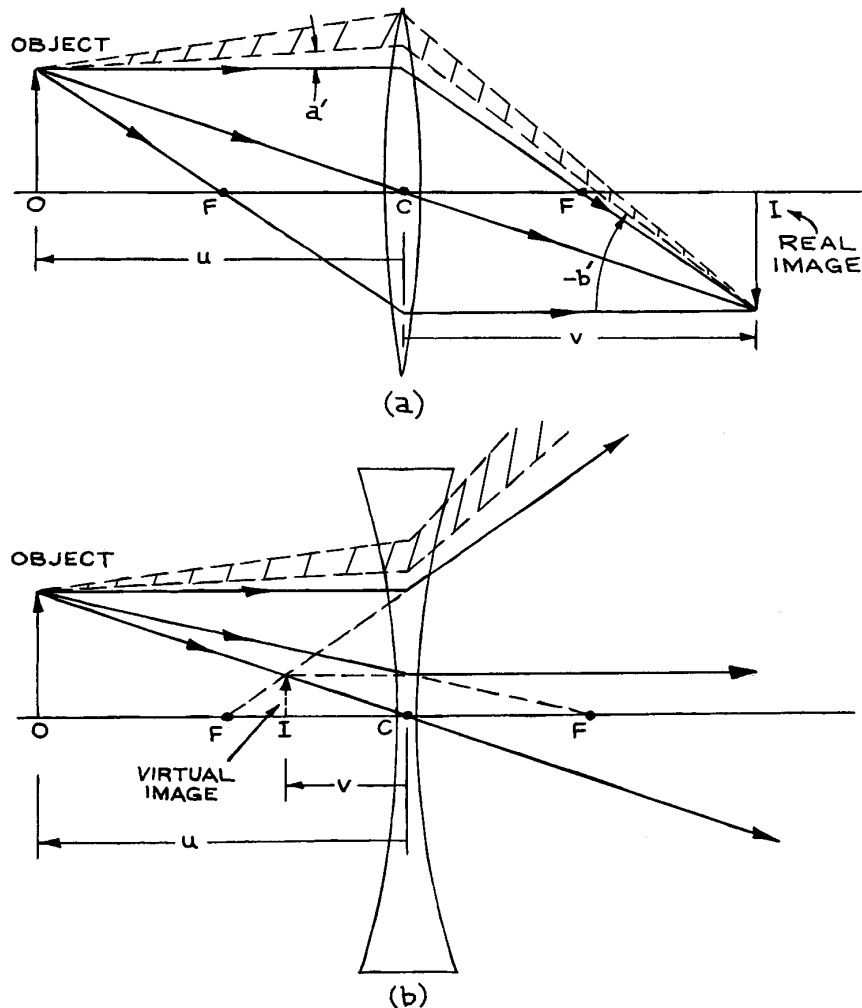


on. For a thin lens, a third principal ray is useful for determining the image location. The ray from a point on the object that passes through the center of the lens is not deviated by the lens.

The use of paraxial ray tracing to determine the size and position of the real image produced by a convex lens, and the virtual image produced by a concave lens are shown in Figure 4.9. Figure 4.9(a) shows a converging lens; the input principal ray parallel to the axis actually passes through the

focal point. Figure 4.9(b) shows a diverging lens; the above ray now emerges from the lens so as to appear to have come from the focal point. More complex systems of lenses can be analyzed the same way. Once the image location has been determined by the use of the principal rays, the path of any group of rays, a ray *pencil*, can be found. See, for example, the cross-hatched ray pencils shown in Figure 4.9.

The use of paraxial ray-tracing rules to analyze spherical-mirror systems is similar to those described above,



**Figure 4.9** Ray-tracing techniques for locating image and ray paths for: (a) a converging thin lens; (b) a diverging thin lens.  $F$  = focal point;  $C$  = center of lens. The principal rays and a general ray pencil are shown in each case.

except that the ray striking the center of the mirror in this case reflects so that the angle of reflection equals the angle of incidence. In all applications of ray tracing, a check on whether this is being done correctly involves examining the way in which a ray is bending at each surface. The refraction (or reflection) relative to the local surface normal should be in the correct direction.

**Imaging and Magnification.** Imaging systems include microscopes, telescopes, periscopes, endoscopes, and cameras. They aim to reproduce at a particular plane in the image space – the *image plane* – an exact replica in terms of relative spatial luminous intensity distribution and spectral content of an object located in the *object plane*. Each point on the object is imaged to a unique point in the image. The relative position and luminous intensity of points in the image plane preserve these qualities from their positions in the object plane. If the image is not a faithful, scaled replica of the object, then the imaging system suffers from *aberrations*.

In paraxial descriptions of an imaging system, often called the *Gaussian* description of the system, the ray transfer matrix from a point on the object to a point on the image can be written as:

$$\begin{pmatrix} r_2 \\ r'_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} r_1 \\ r'_1 \end{pmatrix} \quad (4.52)$$

For any ray angle of  $r'_1$  leaving a point distant  $r_1$  from the axis on the object, the ray in the image plane must pass through a point that is distant  $r_2$  from the axis.

It is clear from Equation (4.52) that for this to be so, the element  $B$  of the ray transfer matrix must be zero. So:

$$r_2 = Ar_1 \quad (4.53)$$

The ratio of image height to object height is the *linear magnification*  $m$  so:

$$m = A = r_2/r_1 \quad (4.54)$$

Note that if the image is inverted,  $m$  is negative. The *angular magnification* of the system is defined as:

$$m' = \left( \frac{r'_2}{r'_1} \right)_{r_1 \rightarrow 0} \quad (4.55)$$

so  $D = m'$ .

If we use the focal length of the imaging system  $f = -1/C$ , then the ray transfer matrix is:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} m & 0 \\ -1/f & m' \end{pmatrix} \quad (4.56)$$

If the media in the space where the object is located (the object space) and the image space are the same then:

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = 1 \quad (4.57)$$

which gives  $mm' = 1$ .

This is also a consequence of the conservation of brightness in an optical system, which will be discussed later.

In Figure 4.9(a) the ratio of the height of the image,  $b$  to the height of the object  $a$  is called the *magnification*  $m$ . In the case of a thin lens:

$$m = \frac{b}{a} = -\frac{v}{u} \quad (4.58)$$

Note that in Figure 4.9(a),  $b$  is negative. In Figure 4.9(b),  $v$  is negative, but  $b$  is positive. The ray transfer matrices  $M$  for a complete description of the imaging systems in Figure 4.9 include the entire system from object O to image I. So, in Figure 4.9(a):

$$\begin{aligned} & \mathbf{M}(\text{matrix for uniform medium of length } u) \\ & \times (\text{matrix for lens}) \\ & \times (\text{Matrix for uniform medium of length } v) \end{aligned} \quad (4.59)$$

**Imaging and Nonimaging Optical Systems.** The techniques of optical-system analysis that we have discussed so far will provide an approximate description of simple systems. In many situations, however, a more detailed analysis is needed. Additional parameters are also introduced to characterize the system. More sophisticated methods of analysis can be illustrated through a discussion of *imaging* and *nonimaging* optical systems.

In any experimental setup that involves the collection of light from a source and its delivery to a light detection system, the properties of the optical system between source and detector should be optimized. This optimization will take different forms depending on the application.

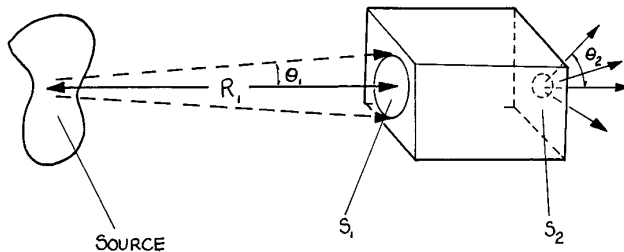
For example, in an experiment in which a weak fluorescent signal from a liquid in a curvette is to be detected, it is generally the aim to maximize the amount of collected fluorescence and deliver this light to the active surface of the detector. This application does not require an *imaging* optical system, although an imaging system is often quite satisfactory. On the other hand, if an image of some object is being delivered to a CCD camera, then an imaging optical system is required.

### 4.2.3 Nonimaging Light Collectors

The properties of a *nonimaging* optical system can be described schematically with the aid of Figure 4.10. Light that enters the front aperture of the system within some angular range, will leave the system through the exit aperture. In such an arrangement, the *brightness* of the light leaving the exit aperture cannot exceed the brightness of the light entering the entrance aperture. The brightness of an emitting object is measured in units of  $W/m^2$  per steradian. In Figure 4.10, if the brightness of the radiation entering the aperture is  $B_1 (W/m^2 sr^{-1})$  then the power entering  $S_1$  from the source is:

$$P_1 = \pi B_1 S_1 \theta_1^2 \quad (4.60)$$

where the angular range of the entering rays is  $\theta_1$ . Assuming that  $R_1^2 \gg S_1$ , this angular range is  $\theta_1 \simeq 1/R_1 \sqrt{S_1/\pi}$ , or



**Figure 4.10** Generalized diagram of a nonimaging system. Light from the source enters the system through the aperture  $S_1$  over an angular range defined by  $\theta_1$  and leaves through aperture  $S_2$  over a range of angles defined by  $\theta_2$ .

$\theta_1^2 \simeq S_1/\pi R_1^2$ . If all the power entering  $S_1$  leaves through  $S_2$ , then the brightness of the emerging radiation is:

$$B_2 \simeq \frac{P_1}{\pi S_2 \theta_2^2} = \frac{\pi B_1 S_1 \theta_1^2}{\pi S_2 \theta_2^2} \quad (4.61)$$

where  $\theta_2$  is the angular spread of emerging rays. So:

$$B_2 = \frac{B_1 S_1 \theta_1^2}{S_2 \theta_2^2} \quad (4.62)$$

and if  $B_2 \leq B_1$ ,

$$\theta_2 \geq \sqrt{\frac{S_2}{S_1}} \theta_1 \quad (4.63)$$

Practical examples of these devices are given in Section 4.3.3.

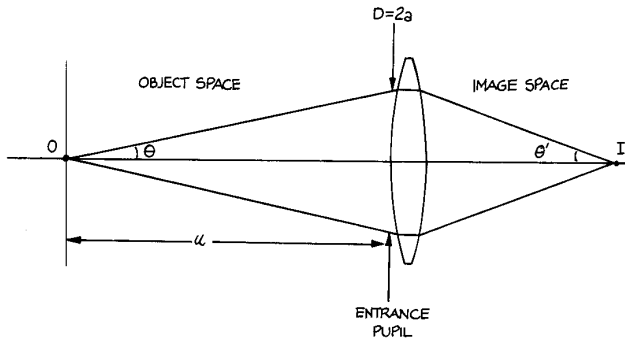
### 4.2.4 Imaging Systems

**Generalized Imaging Systems.** In a generalized imaging system, light from an object, located in the *object plane*, passes through the system until it reaches the *image plane*. In an ideal imaging system all the light rays that leave a point on the object arrive at a single point in the imaging plane. Furthermore, the locations of points in the image plane are such as to preserve the scaling of the image relative to the object. The image may be magnified or demagnified, but it should not be distorted.

A type of imaging called *aplanatic imaging* occurs when rays parallel to the axis produce a sharp image, independent of their distance from the axis. More specifically, if the angle of an input ray with the axis is  $\theta_1$  and the output angle with the axis  $\theta_2$ , then an aplanatic system satisfies the *sine condition*, namely:

$$\frac{\sin \theta_1}{\sin \theta_2} = \text{constant} \quad (4.64)$$

for rays at all distances from the axis that can pass through the system. *Stigmatic imaging* is sharp imaging of a point object to a point image. An object point that is imaged to an image point are together called *conjugate points*.



**Figure 4.11** Single lens optical system to illustrate the concepts of numerical aperture and entrance pupil.

**The Numerical Aperture.** If an optical system is illuminated by a point source on its axis, then the amount of light collected by the system depends on the angle that an effective aperture, called the *entrance pupil*, subtends at the source. This concept will be discussed further in the next section. In Figure 4.11 this angle is defined by:

$$\sin \theta = \frac{D}{2u} \quad (4.65)$$

the *numerical aperture* of the system is:

$$NA = n \sin \theta \quad (4.66)$$

where  $n$  is the refractive index of the object space (which is generally 1, but could be higher – for example in microscopy, when an oil-immersion objective lens is used). Now, a good imaging system obeys the sine condition, which relates the angles subtended by the exit and entrance pupils and the magnification  $m$  by:

$$m = \frac{n \sin \theta}{n' \sin \theta'} \quad (4.67)$$

where the primed quantities refer to the image space. The numerical aperture in the image space is:

$$(NA)' = n' \sin \theta' \quad (4.68)$$

The entrance pupil plays an important geometrical role in determining the amount of light from an object (or light

source) that can enter and pass through an optical system. For example, for an entrance pupil diameter (EPD) of  $D_1$  located a distance  $d_1$  from a point source of light, the fraction of the light emitted by the source that will pass through the system is:

$$\mathcal{F} = \frac{1}{2}(1 - \cos \theta) = \frac{1}{2} \left[ 1 - \left( 1 + \frac{D^2}{4d^2} \right)^{-1/2} \right], \quad (4.69)$$

where  $\theta$  is the angle shown in Figure 4.11. For  $D \ll d$  Equation (4.69) reduces to:

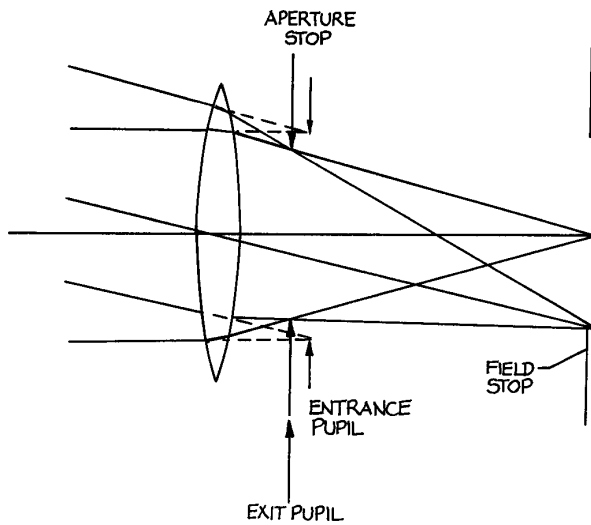
$$\mathcal{F} = \frac{D^2}{16d^2} \quad (4.70)$$

For a lens, the ratio of focal length to EPD is called the *f/number* ( $f/\#$ ) of the lens. An  $f/1$  lens has a ratio of its focal length to aperture (the open lens diameter) equal to unity. For a point source at the focal point of such a lens,  $\mathcal{F} = 0.067$ .

**Apertures, Stops, and Pupils.** Apertures control the ray trajectories that can pass through an optical system. The *aperture stop* is the aperture within the system that limits the angular spread, or diameter, of a cone of rays from an axial point on the object, which can pass through the system. Another aperture in the system may restrict the size or angular extent of an object that the system can image. Awareness of the position and size of these stops in an optical design is crucial in determining the light-gathering power of the system, and its ability to deliver light to an image or detector. The role of the aperture and field stop can be illustrated with Figure 4.12, which shows a single lens being used to image light from a distant object. The circular aperture behind the lens clearly limits the diameter of a bundle of rays from an axial point on the object and is clearly the aperture stop. The second stop, placed in the image plane, clearly controls the maximum angle at which rays from off-axis points on the object can pass through the system. This is the *field stop* in this case.

Whether a given aperture in an optical system is aperture stop or a field stop is not always so easy to determine as was the case in Figure 4.12. Figure 4.13 shows a three-element optical system, which actually constitutes what is

called a Gullstrand ophthalmoscope<sup>c</sup>. The first lens closest to the object, called the *objective* lens, produces a real inverted image, which is then re-inverted by the *erector* lens, to form an image at the focal point of the *eyepiece* lens. Aperture *A* is the aperture stop, and clearly restricts the angular range of rays from an axial point on the object that can pass through the system. Aperture *B* restricts from what height on the object rays can originate, and still pass through the system. This aperture is acting as a field stop.



**Figure 4.12** Single lens imaging system to illustrate the concepts of aperture stop, field stop, and entrance and exit pupils.

The *entrance pupil* is the image of the aperture stop produced by all the elements of the system between the aperture stop and the object. In Figure 4.13 the two elements producing the image are the erector lens and objective lens. The *exit pupil* is the image of the aperture stop produced by all the elements between the aperture stop and the image, which in Figure 4.13 is just the eye lens. The size and location of the entrance and exit pupils determines the light-gathering and light-delivery properties of the system. For example, if the entrance pupil has diameter  $D$ , and lies a distance  $L$  from a point source of radiant power  $P$  (watts) then the power entering the entrance pupil is:

$$P_1 = \frac{P}{2}(1 - \cos \theta), \quad (4.71)$$

where  $\tan \theta = D/2L$ .

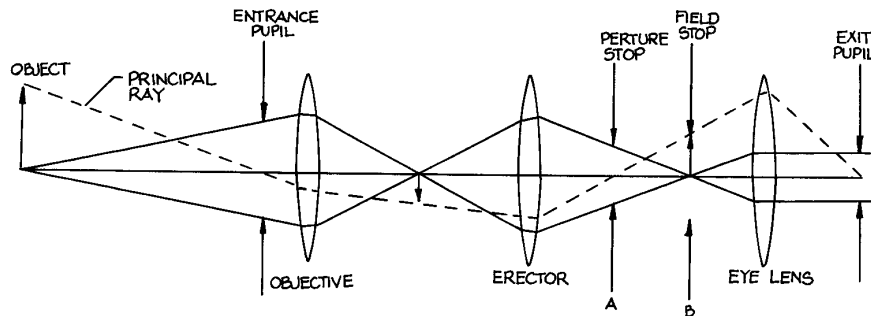
For a point source that is far from the entrance pupil this result becomes:

$$P_1 = \frac{PD_1^2}{16L^2}. \quad (4.72)$$

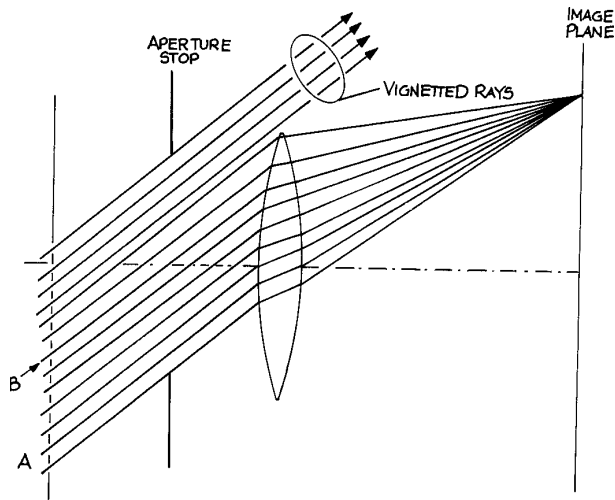
The solid angle subtended by the entrance pupil in this case is  $\Delta\omega = D_1^2/4L^2$ .

The light that leaves the system can be characterized by the total power that leaves the exit pupil, and the angular range of the rays that leave the exit pupil.

Somewhere in an imaging system there is always an *aperture stop*. This may be an actual aperture in an opaque thin screen placed on the system axis, or it may be the clear



**Figure 4.13** Three lens imaging system showing aperture stop, a field stop, and the location of the entrance and exit pupils.



**Figure 4.14** Aperture stop in front of a single lens showing vignetting of a family of rays from a distant off-axis point passing through the lens.

aperture of a lens or mirror in the system. In either case the aperture stop limits the range of rays that can pass through the system. This concept is illustrated in Figure 4.14. A ray of light such as A, that enters the system and just passes the edge of the aperture stop, is called a *marginal ray*. A ray such as B, that enters the system and passes through the center of the aperture stop, is called a *chief ray*. An aperture stop will always allow some light from every point on an object to reach the corresponding point on the image. If the amount of light from the edge of the object (say) that reaches the corresponding point or the object is severely reduced this is referred to as *vignetting*.

In Figure 4.13, a ray from an off-axis point on the object that passes through the center of the aperture stop is called the *principal ray* of the cone of light from the off-axis point.

The ray that enters the system half-way between the highest and lowest rays of an oblique beam is called the *chief ray* of the beam. In the absence of aberrations, the principal ray and chief ray are the same, and both pass through the center of the entrance pupil and aperture stop.

**Vignetting.** In our previous discussion we have simplified the actual performance of a real optical system,

because for oblique rays passing through the system the roles of aperture and field stops become intertwined.

This can be illustrated with Figure 4.15, which shows a simple compound lens in which the apertures of the lenses themselves restrict the rays that can pass through the system. A family of oblique rays no longer fills the aperture stop, because the following lens is acting as a stop. This is called *vignetting*. In imaging systems it does not actually remove points of the object from the image but parts of the image that depend for their illumination on oblique rays (especially at large angles) become less bright. Vignetting is often deliberately allowed to occur in an imaging system so as to remove certain rays from the system that would otherwise suffer severe aberration.

#### 4.2.5 Exact Ray Tracing and Aberrations

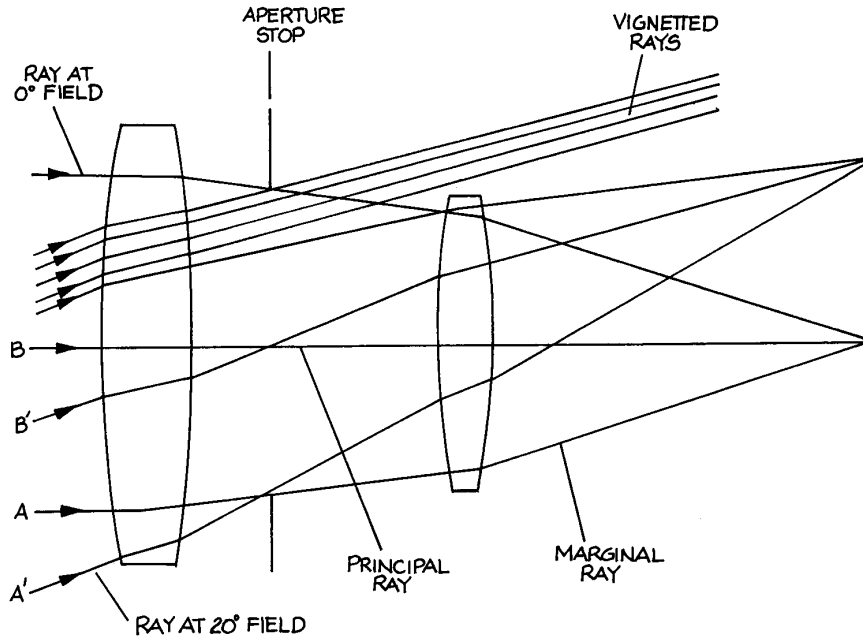
If an imaging system is built with a series of lenses and/or mirrors, then there are inevitably various aberrations. For refraction at a spherical surface this results because, for angles of incidence and refraction  $\theta_1$ ,  $\theta_2$ , respectively, when light travels from a medium with refractive index  $n_1$  to one of refractive index  $n_2$ :

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (4.73)$$

The paraxial approximation, which lacks aberrations, fails because  $\theta_1 \neq \sin \theta_1$ , whereas, in reality,  $\sin \theta = \theta - \theta^3/3! + \theta^5/5! - \theta^7/7!$ . The overall effect of aberrations can be greatly reduced, and some types of aberrations can be eliminated, by using several spherical surfaces, or in some situations by using an appropriate *aspheric* surface, a surface that is not part of a sphere but which is generally describable in terms of some other conic function or power series of the coordinates.

Analysis of imaging systems is most conveniently carried out using exact ray tracing techniques. In exact ray tracing the small angle approximation, Equation (4.8), is not made and Snell's Law is solved exactly for each ray, at each interface.

By injecting many light rays into the system over some defining aperture or range of angles the performance of the system in imaging, focusing, light collection, or light delivery can be evaluated precisely. These days this procedure is carried out with optical design software. The most notable software packages in this regard are Code



**Figure 4.15** Two-lens imaging system showing aperture stop, principal and marginal rays, and vignetted rays.

V<sup>18</sup>, Zemax<sup>19</sup>, Oslo<sup>20</sup>, Solstis<sup>21</sup> and Optikwerk<sup>22</sup>. The wide availability of these software tools, and the essentially generic way in which they are used makes a brief description of how they work and are used worthwhile. Just as an electronic circuit designer would be likely to evaluate a circuit with an analog design tool such as PSpice, or a digital design tool such as Cadence, Magic, or Xilinx, an optical designer can ultimately save time, and avoid mistakes, by numerical simulation of a design.

We will illustrate the use of numerical ray tracing using the set up procedures appropriate to Code V and with a specific example: a double Gauss lens used for imaging.

The various optical elements that make up a double Gauss lens are shown in Figure 4.16. For Code V analysis, the axial position, curvature, material, and aperture size (radius of each surface perpendicular to the axis) must be specified in an *optical table* shown in this case as Table 4.2. Note the entries that are specified in each column.

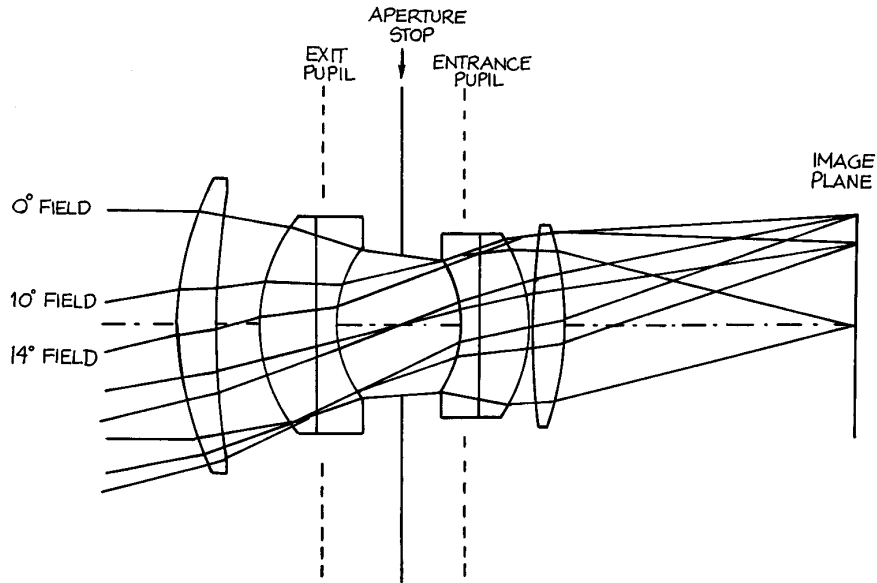
With ray tracing software the various parameters of this lens system can be simultaneously optimized so as to produce the sharpest possible image of an object.

For the double Gauss lens, for example, the curvatures of the various surfaces, their spacings, and the type of glass used for each element can be simultaneously varied. In practice, various constraints must be included in this process: lens elements should not be too thick or too thin, and the glasses chosen must be available. The variable parameters for a glass are its refractive index  $n$  and its *dispersion*  $dn/d\lambda$ , both of which vary with wavelength.

The dispersion is frequently characterized by the Abbe V number, which provides a relative measure of dispersion:

$$V = \frac{n_d - 1}{n_F - n_C} \quad (4.74)$$

where the refractive index  $n$  is specified at the three wavelengths  $d$  (the helium d line at 587.6 nm), F (the hydrogen F line at 486.1 nm), and C (the hydrogen C line at 656.3 nm). Some of the standard wavelengths that are used in optical design are listed in Table 4.3. The different glasses that are available from a manufacturer are generally plotted on a glass chart, which plots  $(n_d - 1)$  against  $V$ .



**Figure 4.16** Double Gauss lens.

Figure 4.17 shows such a chart for glasses available from Schott. Other major suppliers of optical glass are Corning, Hoya, and Ohara.

Two old classifications of glass into *crown* and *flint* glasses can be related to the glass chart. Crown glasses are glasses with a  $V$  value greater than 55 if  $n_d < 1.6$  and  $V > 50$  if  $n_d > 1.6$ . The flint glasses have  $V$  values below these limits. The rare-earth glasses contain rare-earth instead of  $\text{SiO}_2$ , which is the primary constituent of crown and flint glasses.

**Spot Diagram.** If very many rays are launched from a point on the object so as to cover the entrance pupil of the system, the resulting pattern of ray intersections with the image plane is called the *spot diagram*. A tight spot diagram indicates that the effects of several aberrations, notably spherical, astigmatism, coma, and curvature of field have been markedly reduced. It must be remembered that this does not imply that the aberrations themselves have necessarily been eliminated, but that their effects have been reduced in the situation examined. A tighter spot diagram can always be produced by reducing the size of the entrance pupil.<sup>d</sup> Figure 4.18 shows such a

spot diagram for a double Gauss lens being operated with a relatively large entrance pupil.

**Chromatic Aberrations.** Because, in general, the index of refraction increases with decreasing wavelength, the focal length of a singlet lens<sup>e</sup> will, for example, be shorter for blue light than it is for red light. This effect leads to imperfect imaging, so that in white light illumination both axial and off-axis object points will be imaged, not as bright points, but as regions that show blurring from red to blue. For axial object points this blurring is spherically symmetric and results from the image being closer to the lens for blue light than for red light. This is simple *chromatic aberration*. For off-axis object points, the magnification of the system varies with wavelength, leading to what is called *lateral color* or *transverse chromatic aberration*.

Chromatic aberration can be corrected for two colors (generally red and blue) and the imaging of axial points. Such lenses take the form of an achromatic doublet where a doublet lens is generally fabricated from a positive (converging) crown glass element and a negative (diverging) flint glass element.



**Table 4.2 Optical table for code V analysis**

*Double Gauss – U.S. Patent 2, 532, 751*

	<i>RDY</i>	<i>THI</i>	<i>RMD</i>	<i>GLA</i>	<i>CCY</i>	<i>THC</i>	<i>GLC</i>
> OBJ:	INFINITY	INFINITY			100	100	
1:	65.81739	8.746658		BSM24_OHARA	0	100	
2:	179.17158	9.100796			0	0	
3:	36.96542	12.424230		SK1_SCHOTT	0	100	
4:	INFINITY	3.776966		F15_SCHOTT	100	100	
5:	24.34641	15.135377			0	0	
STO:	INFINITY	12.869768			100	0	
7:	-27.68891	3.776966		F15_SCHOTT	0	100	
8:	INFINITY	10.833928		SK16_SCHOTT	100	100	
9:	-37.46023	0.822290			0	0	
10:	156.92645	6.858175		SK16_SCHOTT	0	100	
11:	-73.91040	63.268743			0	PIM	
IMG:	INFINITY	-0.279291			100	0	
SPECIFICATION DATA							
EPD	50.00000						
DIM	MM						
WL	656.30	587.60	486.10				
REF	2						
WTW	1	1	1				
XAN	0.00000	0.00000	0.00000				
YAN	0.00000	10.00000	14.00000				
VUY	0.00000	0.20000	0.40000				
VLY	0.00000	0.30000	0.40000				
REFRACTIVE INDICES							
GLASS CODE		656.30		587.60		486.10	
BSM24_OHARA		1.614254		1.617644		1.625478	
SK1_SCHOTT		1.606991		1.610248		1.617756	
F15_SCHOTT		1.600935		1.605648		1.616951	
SK16_SCHOTT		1.617271		1.620408		1.627559	
SOLVES							
PIM							
NO Pickups defined in system							
INFINITE CONJUGATES							
EFL	99.9866						
EFL	63.2687						
FFL	-17.7398						
FNO	1.9997						
IMG DIS	62.9895						
OAL	84.3452						
PARAXIAL IMAGE							
HT	24.9295						
ANG	14.0000						
ENTRANCE PUPIL							
DIA	50.0000						
THI	68.3242						
EXIT PUPIL							
DIA	58.0886						
THI	-52.8928						

**Table 4.3 Standard wavelengths used in optical design**

Wavelength (nm)	Designation	Source
312.59		mercury
334.15		mercury
365.01	$\ell$	mercury
404.66	h	mercury
435.83	g	mercury
479.99	F'	mercury
486.13	F	hydrogen
546.07	e	mercury
587.56	d	helium
589.29	D	sodium (doublet)
632.80		He-Ne laser
643.85	c'	cadmium
656.27	c	hydrogen
706.52	r	helium
852.11	s	cesium
1013.98	t	mercury
1060.00		neodymium laser
1529.58		mercury
1970.09		mercury
2325.42		mercury

If a compound lens is designed to remove chromatic aberration at three wavelengths it is called an *apochromat* and if correcting for four wavelengths a *superachromat*.

Achromatic doublets, because they have at least three spherical surfaces (or four in an air-spaced achromat) can also better correct for other aberrations than a singlet lens. Such lenses are available in a range of apertures and focal lengths from many manufacturers.

**Geometrical Aberrations.** In a perfect imaging system, all rays from a point on the object would pass through an identical point on the image and there would be a linear relation between the coordinates of points on the image and corresponding points on the object.

For an axially symmetric system we can arbitrarily choose the object point on in Figure 4.19 to be at the  $(x, y)$  point  $(0, h)$  in the object. We take the polar coordinates of a ray from this point to a point  $P$  in the entrance pupil as  $(\rho, \theta)$ . In the general case this ray will intersect the image

plane at point  $I'$  with coordinates in the image plane  $(x', y')$ . The aberrations  $\Delta x'$ ,  $\Delta y'$  are the displacements of  $(x', y')$  from the paraxial image point  $I(x'_0, y'_0)$  for which there is a linear transformation from the coordinates of the point on the object. The lowest-order aberration terms contain terms in either  $h^3$ ,  $h^2\rho$ , or  $h\rho^2$  and are referred to as *third-order aberrations*. Although there are differences in terminology concerning these terms in the literature we can write (following Born and Wolf<sup>11</sup>)

$$\Delta y' = B\rho^3 \cos \theta - Fh\rho^2(2 + \cos Z\theta) + (2C + D)h^2\rho \cos \theta - Eh^3 \quad (4.75)$$

$$\Delta x' = B\rho^3 \sin \theta - Fh\rho^2 \sin 2\theta + Dh^2\rho \sin \theta \quad (4.76)$$

The coefficients  $B$ ,  $C$ ,  $D$ ,  $E$ , and  $F$  characterize the five *primary* or *Seidel* aberrations. Their magnitude in any given optical design can be calculated and visualized with ray tracing optical software.

**Spherical Aberration  $B \neq 0$ .** When all the coefficients in Equation (4.75) and (4.76) are zero except for  $B$ , the remaining effect is spherical aberration. This aberration is strikingly observed in the imaging of an axial point source, which will be imaged to a circular bright region whose radius reveals the extent of the aberration. For an axial point  $h = 0$  and from Equation (4.75) and (4.76),  $\Delta y' = B\rho^3 \cos \theta$ ,  $\Delta x' = B\rho^3 \sin \theta$ , which can be written as:

$$\Delta r = B\rho^3 \quad (4.77)$$

For focusing of parallel light the “best-form” singlet lens for minimum spherical aberration is close to convex-plane (with the convex surface towards the incoming parallel light).

**Coma  $F \neq 0$ .** Coma is the aberration in which the image of an off-axis point varies for rays passing through different regions of the entrance pupil. It is difficult to produce a lens where coma alone can be observed. The spot diagram in Figure 4.18 shows the characteristic “flaring” of the spot pattern, like a comet tail, which gives the aberration its name. Coma can be controlled by varying the curvatures of surfaces in the system.

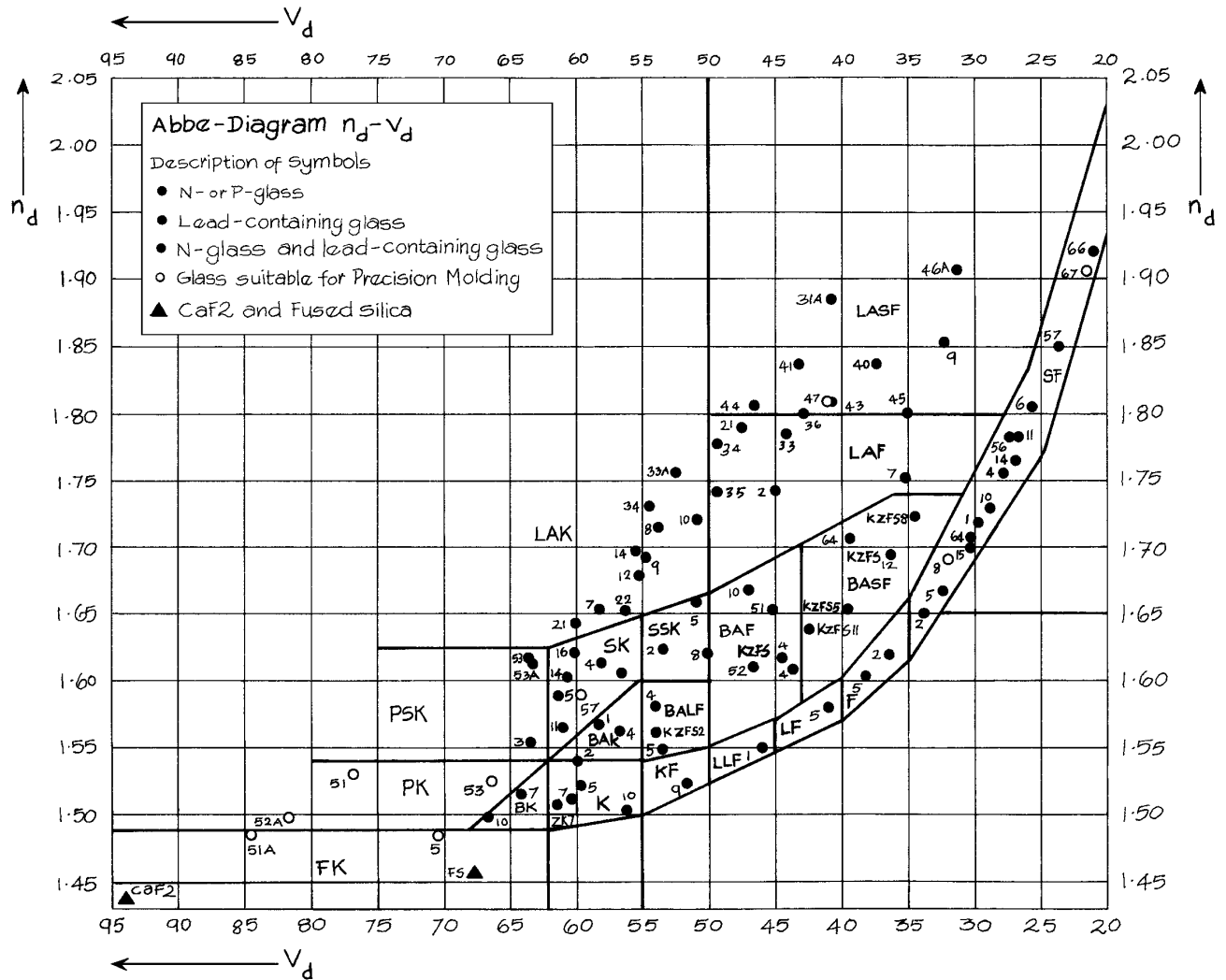
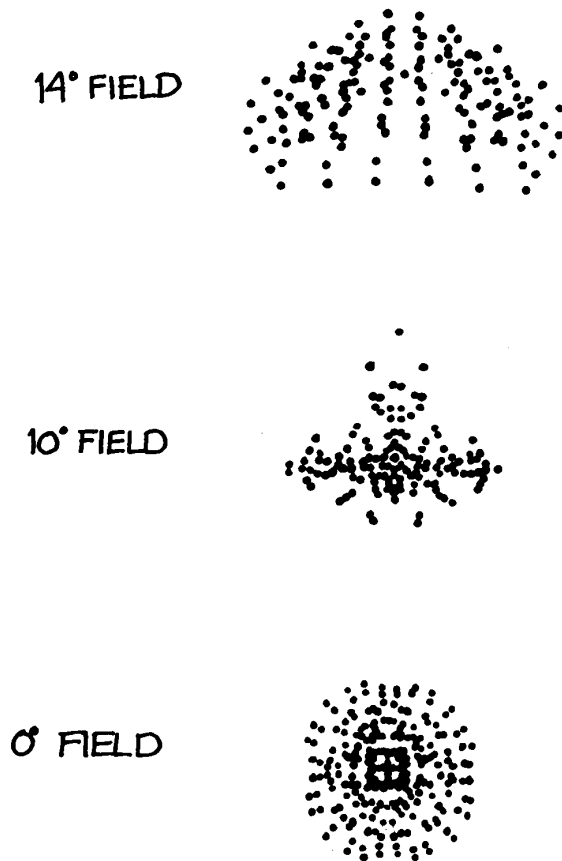


Figure 4.17 The glass chart for glasses available from Schott.

**Astigmatism  $C \neq 0$ .** All the aberrations discussed so far are defects in the imaging of *meridional rays*, that is, rays in the plane containing the axis of the lens and the line from the object point through the center of the lens. The imaging is different for *sagittal rays*, that is, rays in the sagittal plane, which is perpendicular to the meridional plane, as illustrated in Figure 4.20. The resulting aberration is called *astigmatism*. It can be controlled by

lens curvature and refractive-index variations and by the use of apertures to restrict the range of angles and off-axis distances at which rays can traverse the lens.

**Distortion,  $E \neq 0$ .** In distortion, the magnification varies across the image plane. The height of the image above the axis relative to the height of the object varies across the image.



**Figure 4.18** Spot diagrams for a double Gauss lens showing the spot diagrams for families of rays entering the lens at angles of  $0^\circ$ ,  $10^\circ$ , and  $14^\circ$  – the *field angles*.

Each object point appears as a point in the image, and the images of object points on a plane orthogonal to the axis are also on such a plane. The magnification of an object line segment may, however, vary with its distance from the axis. This is called *distortion*. It takes two common forms: *pincushion* and *barrel* distortion, as illustrated in Figure 4.21. Distortion is sensitive to the lens shape and spacing, and to the size and position of apertures in the system – called *stops*.

For further details of aberrations and how to deal with them, we refer the reader to specialized texts on optics

(especially lens design), such as those by Born and Wolf,<sup>11</sup> Ditchburn,<sup>23</sup> Levi,<sup>24</sup> Smith,<sup>25,26</sup> Shannon,<sup>27</sup> and Laikin.<sup>28</sup>

**Curvature of Field,  $D \neq 0$ .** In this aberration, the image of a plane object perpendicular to the axis is sharp over a curved surface in the image space.

**Modulation Transfer Function.** A common way of characterizing the performance of an imaging system is through its *modulation transfer function* or MTF. This refers to the ability of the system to replicate in the image, periodic features in the object. An optical test pattern used to show this consists of a series of white and black bands. These bands will be smeared out in the image to a greater or lesser extent because of aberrations, and ultimately, if geometrical aberrations are absent, by diffraction.

The relative brightness of the object and image will appear schematically as shown in Figure 4.22. We can characterize the *visibility*, *contrast* or *modulation* of the image as:

$$\text{visibility} = \frac{\text{max} - \text{min}}{\text{max} + \text{min}} \quad (4.78)$$

If the visibility is plotted versus the number of lines per millimeter in the object in this case, this shows the modulation transfer function. Figure 4.23 shows an example for the double Gauss lens system shown in Figure 4.16. The MTF is ultimately limited by diffraction. In this case, for a pattern with  $v$  lines per millimeter:

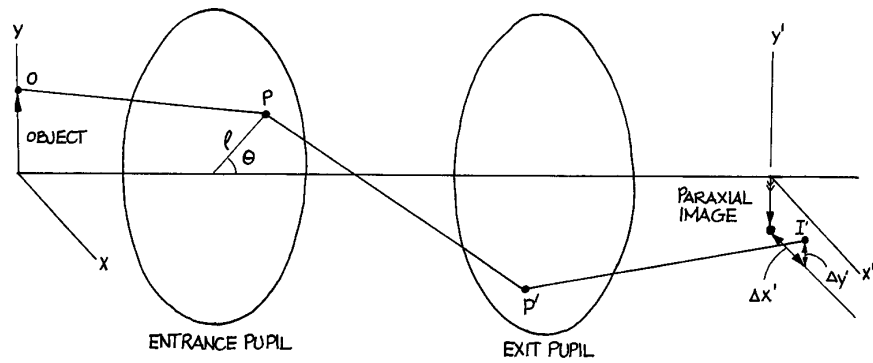
$$\text{MTF}(v) = \frac{1}{\pi} (2\phi - \sin 2\phi) \quad (4.79)$$

where

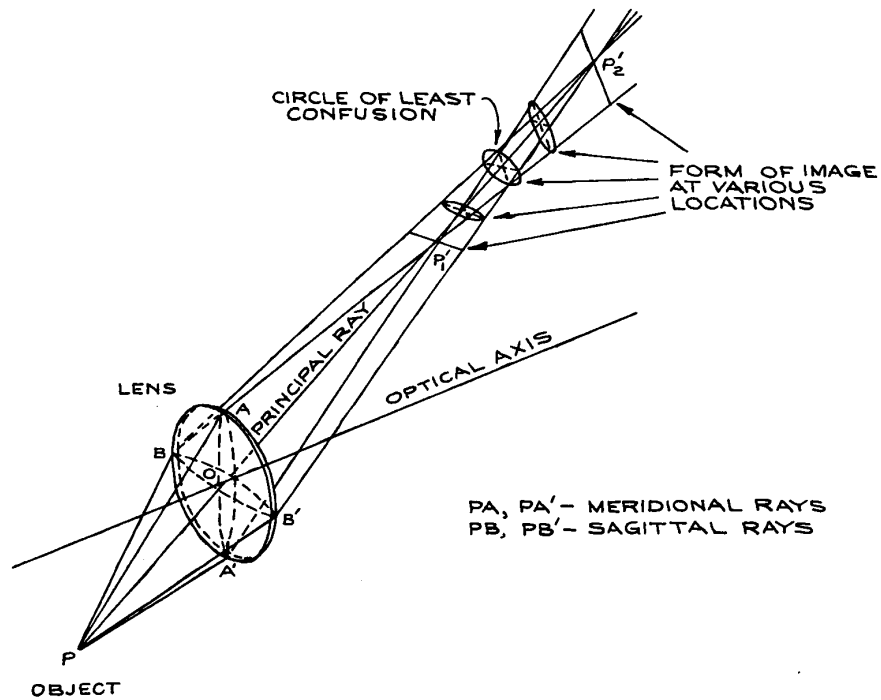
$$\phi = \cos^{-1} \left( \frac{\lambda_0 v}{2\text{NA}} \right) \quad (4.80)$$

In Equation (4.80)  $\lambda_0$  is measured in mm and NA is the numerical aperture of the system.

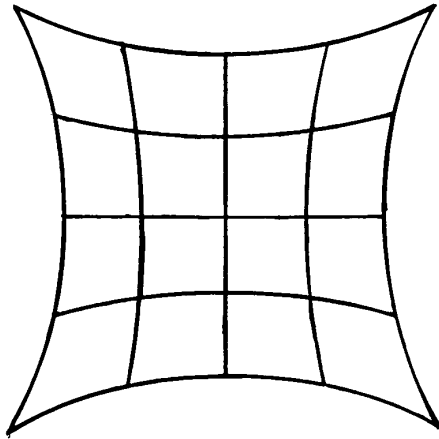
A quantity, closely related to the MTF, for characterizing MTF, and also the focusing of a light beam, or laser beam, is the *Strehl ratio*. The Strehl ratio is the ratio of the illumination at the center of a focused or circular image to the illumination of an equivalent unaberrated image. The interested reader can see a more detailed discussion in Smith.<sup>25</sup>



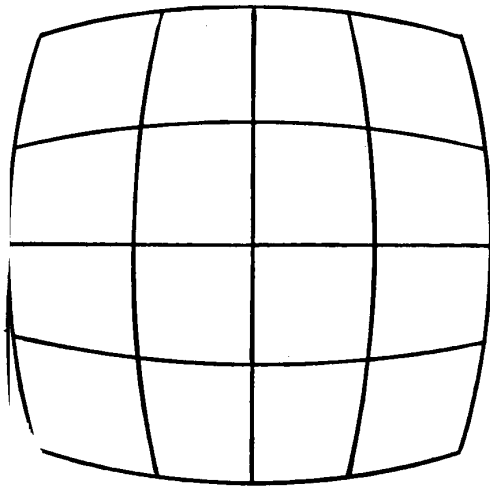
**Figure 4.19** Diagram of a generalized imaging system used to discuss aberrations.



**Figure 4.20** Illustrating astigmatism. Meridional rays  $PA$  and  $PA'$  are imaged at  $P'_1$ , sagittal rays  $PB$  and  $PB'$  are imaged at  $P'_2$ . (After A. C. Hardy and F. H. Perrin, *Principles of Optics*, McGraw-Hill, New York, 1932; by permission of McGraw-Hill Book Company, Inc.)



(a)



(b)

**Figure 4.21** (a) "Pincushion" distortion;  
(b) "barrel" distortion.

### 4.2.6 The Use of Impedances in Optics

The method of impedances is the easiest way to calculate the fraction of incident intensity transmitted and reflected in an optical system. It is also the easiest way to follow the

changes in polarization state that result when light passes through an optical system.

As we have seen in Section 4.1, the impedance of a plane wave traveling in a medium of relative permeability  $\mu_r$ , and dielectric constant  $\epsilon_r$  is:

$$Z = \sqrt{\frac{\mu_r \mu_0}{\epsilon_r \epsilon_0}} = Z_0 \sqrt{\frac{\mu_r}{\epsilon_r}} \quad (4.81)$$

If  $\mu_r = 1$ , as is usually the case for optical media, the impedance can be written as:

$$Z = Z_0/n \quad (4.82)$$

This impedance relates the transverse  $\mathbf{E}$  and  $\mathbf{H}$  fields of the wave:

$$Z = \frac{E_{tr}}{H_{tr}} \quad (4.83)$$

When a plane wave crosses a planar boundary between two different media, the components of both  $\mathbf{E}$  and  $\mathbf{H}$  parallel to the boundary have to be continuous across that boundary. Figure 4.24(a) illustrates a plane wave polarized in the plane of incidence striking a planar boundary between two media of refractive indices  $n_1$  and  $n_2$ . In terms of the magnitudes of the vectors involved:

$$E_i \cos \theta_1 + E_r \cos \theta_1 = E_t \cos \theta_2 \quad (4.84)$$

$$H_i - H_r = H_t \quad (4.85)$$

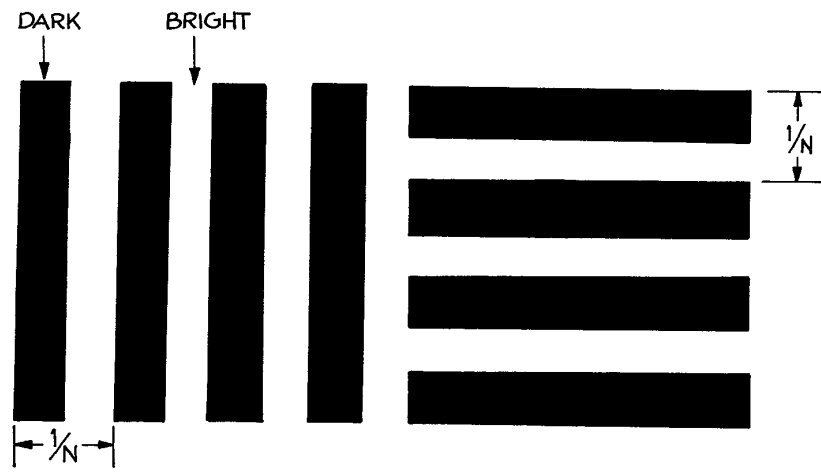
Equation (4.82) can be written as:

$$\frac{E_i}{Z_1} - \frac{E_r}{Z_1} = \frac{E_t}{Z_2} \quad (4.86)$$

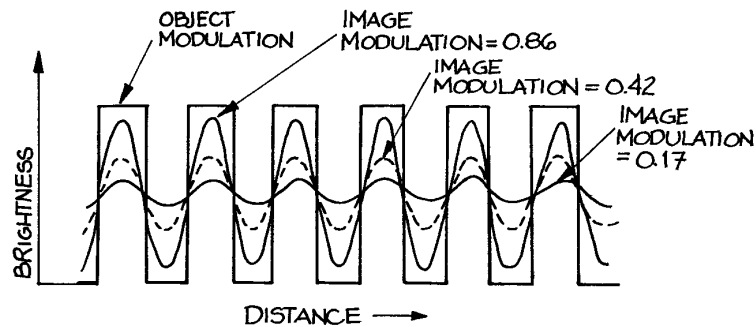
It is easy to eliminate  $E_r$  between Equations (4.84) and (4.86) to give:

$$\rho = \frac{E_r}{E_i} = \frac{Z_2 \cos \theta_2 - Z_1 \cos \theta_1}{Z_2 \cos \theta_2 + Z_1 \cos \theta_1} \quad (4.87)$$

where  $\rho$  is the *reflection coefficient* of the surface. The fraction of the incident energy reflected from the surface



(a)



(b)

**Figure 4.22** (a) An object with sharp contrast between bright and dark bands; (b) the corresponding image brightness variation for differing degrees of modulation.

is called the reflectance,  $R = \rho^2$ . Similarly, the *transmission coefficient* of the boundary is:

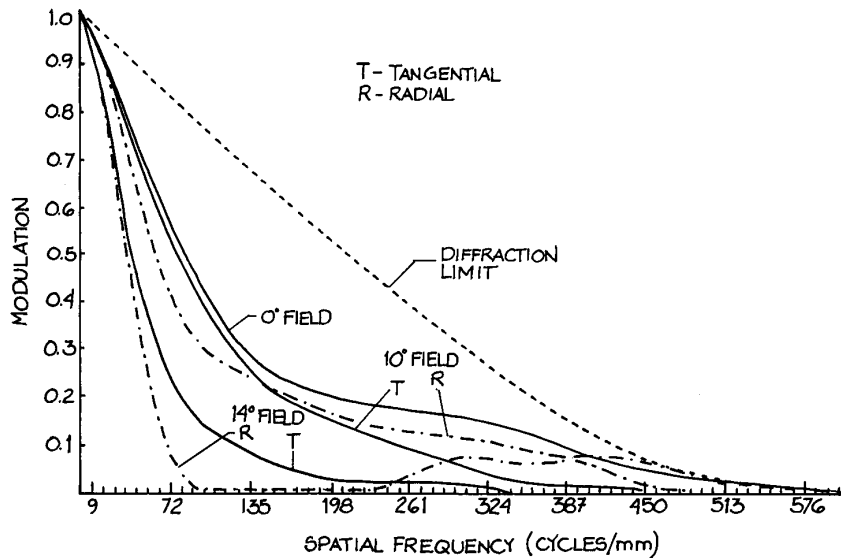
$$\tau = \frac{E_t \cos \theta_2}{E_i \cos \theta_1} = \frac{2Z_2 \cos \theta_2}{Z_2 \cos \theta_2 + Z_1 \cos \theta_1} \quad (4.88)$$

By a similar treatment applied to the geometry shown in Figure 4.24(b), it can be shown that for a plane wave

polarized perpendicular to the plane of incidence:

$$\rho = \frac{E_r}{E_i} = \frac{Z_2 \sec \theta_2 - Z_1 \sec \theta_1}{Z_2 \sec \theta_2 + Z_1 \sec \theta_1} \quad (4.89)$$

$$\tau = \frac{2Z \sec \theta_2}{Z_2 \sec \theta_2 + Z_1 \sec \theta_1} \quad (4.90)$$



**Figure 4.23** Modulation transfer function (MTF) for a double Gauss lens. The MTF for 0°, 10°, and 14° fields in both the tangential and radial directions are shown, and also the diffraction limit of an aberration-free system.

If the effective impedance for a plane wave polarized in the plane of incidence (*P-polarization*)<sup>f</sup> and incident on a boundary at angle  $\theta$  is defined as:

$$Z' = Z \sec \theta \quad (4.91)$$

and for a wave polarized perpendicular to the plane of incidence (*S-polarization*)<sup>g</sup>, often called a TE wave, as:

$$Z' = Z \sec \theta \quad (4.92)$$

then a universal pair of formulae for  $\rho$  and  $\tau$  results:

$$\rho = \frac{Z'_2 - Z'_1}{Z'_1 + Z'_2} \quad (4.93)$$

$$\tau = \frac{2Z'_2}{Z'_1 + Z'_2} \quad (4.94)$$

It will be apparent from an inspection of Figure 4.24 that  $Z'$  is just the ratio of the electric-field component parallel to the boundary and the magnetic-field component parallel to the boundary. For reflection from an ideal mirror,  $Z'_2 = 0$ .

In normal incidence, Equations (4.87) and (4.89) become identical and can be written as:

$$\rho = \frac{Z_2 - Z_1}{Z_2 + Z_1} = \frac{n_1 - n_2}{n_1 + n_2} \quad (4.95)$$

Note that there is a change of phase of  $\pi$  in the reflected field relative to the incident field when  $n_2 > n_1$ .

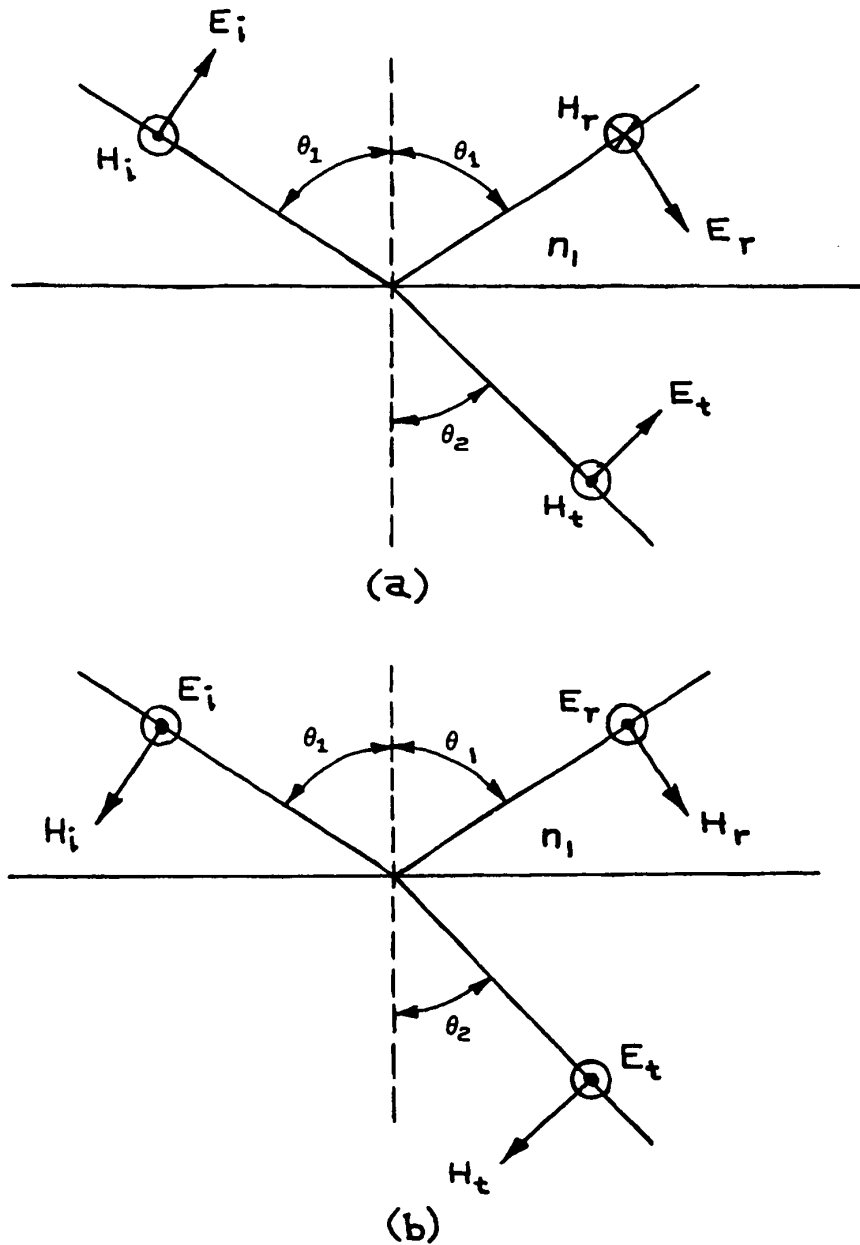
Since intensity  $\propto$  (electric field)<sup>2</sup>, the fraction of the incident energy that is reflected is:

$$R = \rho^2 = \left( \frac{n_1 - n_2}{n_1 + n_2} \right)^2 \quad (4.96)$$

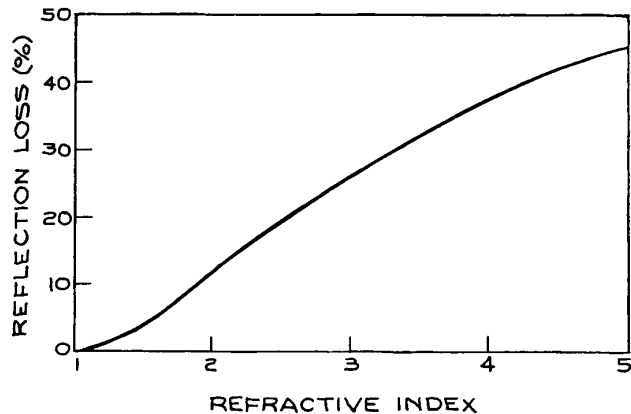
$R$  increases with the index mismatch between the two media, as shown in Figure 4.25. If there is no absorption of energy at the boundary, the fraction of energy transmitted, called the *transmittance*, is

$$T = 1 - R = \frac{4n_1n_2}{(n_1 + n_2)^2} \quad (4.97)$$





**Figure 4.24** Reflection and refraction at a planar boundary between two different dielectric media: (a) wave polarized in the plane of incidence (*P*-polarization); (b) wave polarized perpendicular to the plane of incidence (*S*-polarization).



**Figure 4.25** Reflection loss per surface for normal incidence as a function of refractive index.

Note that because the media on the two sides of the boundary are different,  $T \neq |\tau|^2$ .

**Reflectance for Waves Incident on an Interface at Oblique Angles.** If the wave is not incident normally, it must be decomposed into two linearly polarized components, one polarized in the plane of incidence, and the other polarized perpendicular to the plane of incidence.

For example, consider a plane-polarized wave incident on an air glass ( $n = 1.5$ ) interface at an angle of incidence of  $30^\circ$ , with a polarization state exactly intermediate between the  $S$ -polarization and the  $P$ -polarization. The angle of refraction at the boundary is found from Snell's law:

$$\sin \theta_2 = \frac{\sin 30^\circ}{1.5} \quad (4.98)$$

and so  $\theta_2 = 19.47^\circ$ .

The effective impedance of the  $P$ -component in the air is:

$$Z'_{P1} = 376.7 \cos \theta_1 = 326.23 \Omega \quad (4.99)$$

and in the glass:

$$Z'_{P2} = \frac{376.7}{1.5} \cos \theta_2 = 236.77 \Omega \quad (4.100)$$

Thus, from Equation (4.87), the reflection coefficient for the  $P$ -component is:

$$\rho_P = \frac{236.77 - 326.33}{236.77 + 326.23} = -0.159 \quad (4.101)$$

The fraction of the intensity associated with the  $P$ -component that is reflected is  $\rho_P^2 = 0.0253$ .

For the  $S$ -component of the input wave:

$$Z'_{S1} = \frac{376.7}{\cos \theta_1} = 434.98 \Omega \quad (4.102)$$

$$Z'_{S2} = \frac{376.7}{1.5 \cos \theta_2} = 266.37 \Omega \quad (4.103)$$

$$\rho_S = \frac{266.37 - 434.98}{266.37 + 434.98} = -0.240$$

The fraction of the intensity associated with the  $S$ -polarization component that is reflected is  $\rho_S^2 = 0.0578$ . Since the input wave contains equal amounts of  $S$ - and  $P$ -polarization, the overall reflectance in this case is:

$$R = \langle \rho^2 \rangle_{av} = 0.0416 \simeq 4\% \quad (4.104)$$

Note that the reflected wave now contains more  $S$ -polarization than  $P$ , so the polarization state of the reflected wave has been changed – a phenomenon that will be discussed further in Section 4.3.6.

**Brewster's Angle.** Returning to Equation (4.87), it might be asked whether the reflectance is ever zero. It is clear that  $\rho$  will be zero if:

$$n_1 \cos \theta_2 = n_2 \cos \theta_1 \quad (4.105)$$

which, from Snell's law [Equation (4.6)], gives:

$$\cos \theta_1 = \frac{n_1}{n_2} \sqrt{1 - \left(\frac{n_1}{n_2}\right)^2 \sin^2 \theta_1} \quad (4.106)$$

giving the solution:

$$\theta_1 = \theta_B = \arcsin \sqrt{\frac{n_2^2}{n_1^2 + n_2^2}} = \arctan \frac{n_2}{n_1} \quad (4.107)$$

The angle  $\theta_B$  is called *Brewster's angle*. A wave polarized in the plane of incidence and incident on a boundary at

this angle is totally transmitted. This allows the design of low-reflection-loss windows in laser systems, as will be seen in Section 4.6.2. If Equation (4.89) is inspected carefully, it will be seen that there is no angle of incidence that yields zero reflection for a wave polarized perpendicular to the plane of incidence.

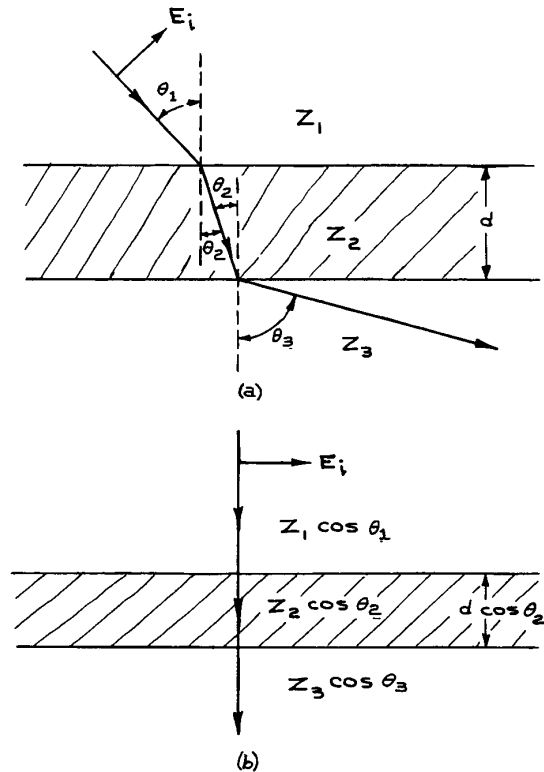
**Transformation of Impedance through Multilayer Optical Systems.** The impedance concept allows the reflection and transmission characteristics of multilayer optical systems to be evaluated very simply. If the incident light is incoherent (a concept discussed in more detail later in Section 4.5.1), then the overall transmission of a multilayer structure is just the product of the transmittances of its various interfaces. For example, an air–glass interface transmits about 96% of the light in normal incidence. The transmittance of a parallel-sided slab is  $0.96 \times 0.96 = 92\%$ . This simple result ignores the possibility of interference effects between reflected and transmitted waves at the two faces of the slab. If the faces of the slab are very flat and parallel, and if the light is coherent, such effects cannot be ignored. In this case, the method of transformed impedances is useful.

Consider the three-layer structure shown in Figure 4.26(a). The path of a ray of light through the structure is shown. The angles  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$  can be calculated from Snell's law. As an example, consider a wave polarized in the plane of incidence. The effective impedances of media 1, 2, and 3 are

$$\begin{aligned} Z'_1 &= Z_1 \cos \theta_1 = \frac{Z_0 \cos \theta_1}{n_1} \\ Z'_2 &= Z_2 \cos \theta_2 = \frac{Z_0 \cos \theta_2}{n_2} \\ Z'_3 &= Z_3 \cos \theta_3 = \frac{Z_0 \cos \theta_3}{n_3} \end{aligned} \quad (4.108)$$

It can be shown that the reflection coefficient of the structure is exactly the same as for the equivalent structure in Figure 4.26(b) in normal incidence, where the effective thickness of layer 2 is now:

$$d' = d \cos \theta_2 \quad (4.109)$$



**Figure 4.26** (a) Wave polarized in the plane of incidence passing through a dielectric slab of thickness  $d$  and impedance  $Z_2$  separating two semi-infinite media of impedances  $Z_1$  and  $Z_3$ , respectively; (b) equivalent structure for normal incidence.

The reflection coefficient of the structure can be calculated from its equivalent structure using the transformed impedance concept.<sup>15,29</sup>

The transformed impedance of medium 3 at the boundary between media 1 and 2 is:

$$Z''_3 = Z'_2 \left( \frac{Z'_3 \cos k_2 d' + i Z'_2 \sin k_2 d'}{Z'_2 \cos k_2 d' + i Z'_3 \sin k_2 d'} \right) \quad (4.110)$$

where  $k_2 = 2\pi/\lambda_2 = 2\pi n_2/\lambda_0$ . The reflection coefficient of the whole structure is now just:

$$\rho = \frac{Z''_3 - Z'_1}{Z''_3 + Z'_1} \quad (4.111)$$

the reflectance of the whole structure is  $R = |\rho|^2$ , and its transmittance is:

$$T = 1 - R \quad (4.112)$$

In a structure with more layers, the transformed impedance formula [Equation (4.108)] can be used sequentially, starting at the last optical surface and working back to the first. More examples of the use of the technique of transformation of impedance will be given in Sections 4.3.1 and 4.3.7.

**Optical System Design Software.** As well as the ray tracing optical design programs mentioned earlier, such as Code V, there is a broad range of other software available commercially for optical system design including: laser system design (GLAD, PARAXIA, ASC, and LASFIT); atmospheric optical properties and sensors (ONTAR); optical system solid modeling (LightTools, TracePro); projection systems, fiber optics, and complex sources (ASAP); beam propagation for integrated and fiber optics (BeamProp); vector diffraction grating analysis software and simulation (GSOLVER); optical thin-film and coating design, material analysis, ellipsometry, spectrophotometry and thin-film metrology; [Film Wizard, FilmSpectrum, TFCalc, Thin Film Design Package (Thin Film Center)]; and analysis of laser beam propagation (FRESNEL).

## 4.2.7 Gaussian Beams

*Gaussian beams* are propagating-wave solutions of Maxwell's equations that are restricted in lateral extent even in free space. They do not need any beam-confining reflective planes, as do the confined electromagnetic-field modes of waveguides.<sup>29</sup> Lasers frequently emit narrow beams of light that are very close in character to ideal Gaussian beams.

The field components of a plane transverse electromagnetic wave of angular frequency  $\omega$  propagating in the  $z$ -direction are of the form:

$$V = V_0 e^{i(\omega t - kz)} \quad (4.113)$$

where  $V_0$  is a constant. A Gaussian beam is of the form:

$$V = \Psi(x, y, z) e^{i(\omega t - kz)} = U(x, y, z) e^{i\omega t} \quad (4.114)$$

where, for example, for a particular value of  $z$ ,  $\Psi(x, y, z)$  gives the spatial variation of the fields in the  $xy$  plane.

$\overline{\Psi} * \overline{\Psi}$  gives the relative intensity distribution in the plane; for a Gaussian beam this gives a localized intensity pattern. The various Gaussian-beam solutions of Maxwell's equations are denoted as  $TEM_{mn}$  modes; a detailed discussion of their properties is given by Kogelnik and Li.<sup>16</sup> The output beam from an ideal laser is generally a TEM mode or a combination of TEM modes. Many laser systems operate in the fundamental Gaussian mode, denoted  $TEM_{00}$ . This is the only mode that will be considered in detail here.

For the  $TEM_{00}$  mode,  $\Psi(x, y, z)$  has the form:

$$\Psi(x, y, z) = \exp \left\{ -i \left[ P(z) + \frac{kr^2}{2q(z)} \right] \right\} \quad (4.115)$$

where  $P(z)$  is a *phase factor*,  $q(z)$  is called the *beam parameter*,  $k$  equals  $2\pi/\lambda$ , and  $r^2$  equals  $x^2 + y^2$ . The beam parameter,  $q(z)$ , is usually written in terms of the phase-front curvature of the beam  $R(z)$ , and its *spot size*,  $w(z)$ , as:

$$\frac{1}{q} = \frac{1}{R} - \frac{i\lambda}{\pi w^2} \quad (4.116)$$

$q$  and  $P$  obey the following relations:

$$\frac{dq}{dz} = 1 \quad (4.117)$$

$$\frac{dP}{dz} = -\frac{i}{q} \quad (4.118)$$

For a particular value of  $z$ , the intensity variation in the  $xy$  plane (the mode pattern) is, from Equation (4.115):

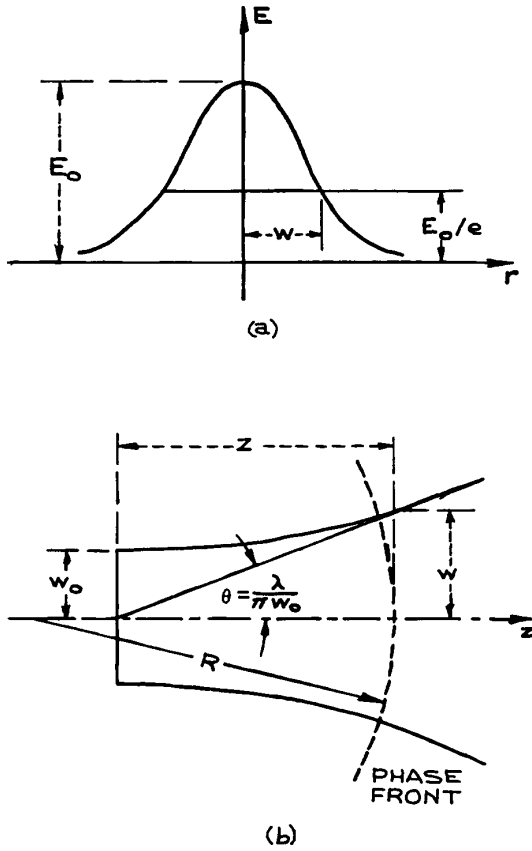
$$\Psi^* \Psi = \exp \left[ \frac{-ikr^2}{2} \left( \frac{1}{q(z)} - \frac{1}{q^*(z)} \right) \right] \quad (4.119)$$

which, from Equation (4.116) gives:

$$\overline{\Psi} \overline{\Psi}^* = e^{-2r^2/w^2} \quad (4.120)$$

Thus,  $w(z)$  is the distance from the axis of the beam ( $x = y = 0$ ) where the intensity has fallen to  $1/e^2$  of its axial value and the fields to  $1/e$  of their axial magnitude. The radial distribution of both intensity and field strength of the  $TEM_{00}$  mode is Gaussian, as shown in Figure 4.27. Clearly, from Equation (4.117):

$$q = q_0 + z \quad (4.121)$$



**Figure 4.27** (a) Radial amplitude variation of the TEM<sub>00</sub> Gaussian beam; (b) contour of a Gaussian beam.

where  $q_0$  is the value of the beam parameter at  $z = 0$ . From Equation (4.118):

$$\begin{aligned} P(z) &= -i \ln q + \text{constant} \\ &= -i \ln(q_0 + z) + \text{constant} \end{aligned} \quad (4.122)$$

Writing the constant in Equation (4.122) as  $\theta + i \ln q_0$ , the full spatial variation of the Gaussian beam is:

$$\begin{aligned} U &= \exp \left\{ -i \left[ kz - i \ln \left( 1 + \frac{z}{q_0} \right) + \theta \right. \right. \\ &\quad \left. \left. + \frac{kr^2}{2} \left( \frac{1}{R(z)} - \frac{i\lambda}{\pi w^2} \right) \right] \right\} \end{aligned} \quad (4.123)$$

The phase angle  $\theta$  is usually set equal to zero. In the plane  $z = 0$ :

$$U = \exp \left\{ -i \frac{kr^2}{2} \left( \frac{1}{R(0)} - \frac{i\lambda}{\pi w(0)^2} \right) \right\} \quad (4.124)$$

The surface of constant phase at this point is defined by the equation:

$$\frac{kr^2}{2R(0)} = \text{constant} \quad (4.125)$$

If the constant is taken to be zero and  $R(0)$  infinite, then  $r$  becomes indeterminate and the surface of constant phase is the plane  $z = 0$ . In this case,  $w(0)$  can have its minimum value anywhere. This minimum value is called the *minimum spot size*  $w_0$ , and the plane  $z = 0$  is called the *beam waist*. At the beam waist:

$$\frac{1}{q_0} = \frac{-i\lambda}{\pi w_0^2} \quad (4.126)$$

Using Equations (4.116) and (4.120), for any arbitrary value of  $z$ :

$$w^2(z) = w_0^2 \left[ 1 + \left( \frac{\lambda z}{\pi w_0^2} \right)^2 \right] \quad (4.127)$$

The radius of curvature of the phase front at this point is:

$$R(z) = z \left[ 1 + \left( \frac{\pi w_0^2}{\lambda z} \right)^2 \right] \quad (4.128)$$

From Equation (4.127) it can be seen that the Gaussian beam expands in both the positive and negative  $z$ -directions from its beam waist along a hyperbola that has asymptotes inclined to the axis at an angle:

$$\theta_{\text{beam}} = \arctan \frac{\lambda}{\pi w_0} \quad (4.129)$$

as illustrated in Figure 4.27(b). The surfaces of constant phase of the Gaussian beam are, in reality, parabolic. For  $r_2 \ll z_2$  (which is generally true, except close to the beam waist), they are spherical surfaces with the radius of curvature  $R(z)$ . Although they will not be discussed further here, the higher-order Gaussian beams, denoted TEM <sub>$m$</sub> , are also characterized by spot sizes and radii of curvature

that are identical to those of the TEM<sub>00</sub> mode and obey Equations (4.127) and (4.128).

**Focusing a Laser beam with a Lens.** A lens can be used to focus a laser beam to a small spot, or systems of lenses may be used to expand the beam and recollimate it (i.e., minimize the beam divergence). In such an application a thin lens will not alter the transverse intensity pattern of the beam at the lens, but it will alter its radius of curvature. Far enough from the beam waist, the radius of curvature of a Gaussian beam behaves exactly as a true spherical wave, since, for  $z \gg \pi\omega_0^2/\lambda$ , Equation (4.128) becomes:

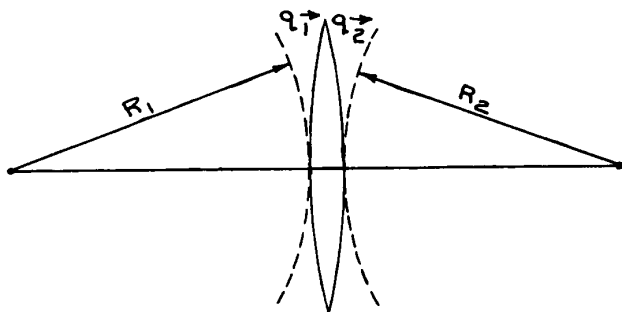
$$R(z) = z \quad (4.130)$$

Now, when a spherical wave of radius  $R_1$  strikes a thin lens, the object distance is clearly also  $R_1$  – the distance to the point of origin of the wave. Therefore, the radius of curvature  $R_2$  immediately after passage through the lens must obey:

$$\frac{1}{R_2} = \frac{1}{R_1} - \frac{1}{f} \quad (4.131)$$

as shown in Figure 4.28. Thus, if  $w$  is unchanged at the lens, the beam parameter after passage through the lens obeys

$$\frac{1}{q_2} = \frac{1}{q_1} - \frac{1}{f} \quad (4.132)$$



**Figure 4.28** Transformation of a Gaussian beam by a thin lens.

It is straightforward to use this result in conjunction with Equations (4.116) and (4.120) to give the minimum spot size  $w_f$  of a TEM<sub>00</sub> Gaussian beam focused by a lens as:

$$w_f = \frac{f\lambda}{\pi w_1} \left[ \left( 1 - \frac{f}{R_1} \right)^2 + \left( \frac{\lambda f}{\pi w_1^2} \right)^2 \right]^{1/2} \quad (4.133)$$

where  $w_1$  and  $R_1$  are the laser-beam spot size and radius of curvature at the input face of the lens. If the lens is placed very close to the waist of the beam being focused, Equation (4.133) reduces to:

$$\omega f = f\theta_B \quad (4.134)$$

where  $\theta_B$  is the beam divergence at the input face of the lens. If  $\theta_B$  is a small angle, then  $w_f$  is located almost at the focal point of the lens – in reality very slightly closer to the lens. It is worthwhile comparing the result given by Equation (4.134) with the result obtained for a plane wave being focused by a lens. In the latter case, the lens diameter  $D$  is the factor that limits the lateral extent of the wave being focused. Diffraction theory<sup>11,23</sup> shows that, in this case, 84% of the energy is focused into a region of diameter:

$$S = 2.44\lambda f/D \quad (4.135)$$

This is the diameter of the focused Airy disk diffraction pattern of the lens aperture.

It should be noted that the spot size to which a laser beam can be focused cannot be reduced without limit merely by reducing the focal length of the focusing lens. In the limit the lens becomes a sphere and, of course, is then no longer a thin lens. In practice, to focus a laser beam to a small spot, the value of  $\theta_B$  should be first reduced by expanding the beam and then recollimating it – generally with a Galilean telescope, as will be seen in Section 4.3.3. To prevent diffraction effects, the lens aperture should be larger than the spot size at the lens:  $D = 2.8w$  is a common size used. In this case, if the focusing lens is placed at the beam waist of the collimated beam, the focal-spot diameter is:

$$2w_f = \frac{5.6\lambda f}{\pi D} = \frac{1.78\lambda f}{D} \quad (4.136)$$

In practice, it is difficult to manufacture spherical lenses with very small values of  $f/D$  (called the  $f/\#$  or  $f/\text{number}$ )

and at the same time achieve the diffraction-limited performance predicted by this equation. Commercially available spherical lenses<sup>30</sup> achieve values of  $2w_f$  of about  $10\lambda$ . Smaller spot sizes are possible with aspheric lenses.

**The Beam Parameter  $M^2$ .** Real laser beams are often characterized by their  $M^2$  parameter or *focusability parameter*. Because of deviations from perfect Gaussian behavior, when a real laser beam is focused by an ideal lens placed at the beam waist the focused spotsize is not  $w_f = f\theta_B$ , as predicted by Equation (4.134), but is instead:

$$w_{fr} = M^2 f \theta_B \quad (4.137)$$

For asymmetrical laser beams there will be two beam parameters  $M_x^2$  and  $M_y^2$  for the two lateral coordinates of the laser beam. The  $M^2$  parameter cannot be measured from a single measurement of the laser beam. A series of measurements on a focused beam are made, such as the position of the focus relative to the focal length of the lens, the new beam divergence angle after the focus, the spotsize(s) at the focal point of the lens, and the new minimum spotsize(s). Commercial instruments that measure  $M^2$  are available from Coherent, Photon, Inc., and Spiricon. Here, and wherever suppliers are mentioned in the text, the list is intended to be representative, but not necessarily exhaustive.

## 4.3 OPTICAL COMPONENTS

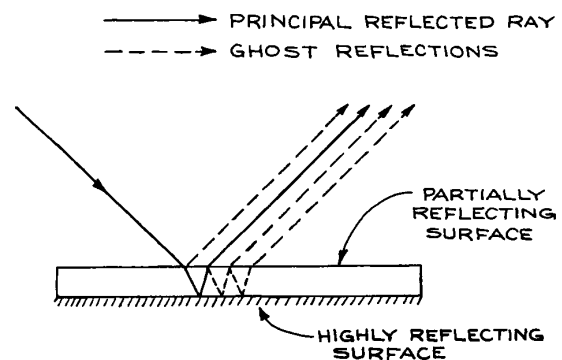
### 4.3.1 Mirrors

When light passes from one medium to another of different refractive index, there is some reflection, so the interface acts as a partially reflecting mirror. By applying an appropriate single-layer or multilayer coating to the interface between the two media, the reflection can be controlled so that the reflectance has any desired value between 0 and 1. Both flat and spherical mirrors made in this way are available – there are numerous suppliers.<sup>31</sup> If no transmitted light is required, high-reflectance mirrors can be made from metal-coated substrates or from metals themselves. Mirrors that reflect and transmit roughly equal amounts of incident light are often referred to as *beamsplitters*.

**Flat Mirrors.** Flat mirrors are used to deviate the path of light rays without any focusing. These mirrors can have their reflective surface on the front face of any suitable substrate, or on the rear face of a transparent substrate. Front-surface, totally reflecting mirrors have the advantage of producing no unwanted additional or *ghost* reflections; however, their reflective surface is exposed. Rear-surface mirrors produce ghost reflections, as illustrated in Figure 4.29, unless their front surface is antireflection coated; but the reflective surface is protected. Most household mirrors are made this way. The cost of flat mirrors depends on their size and on the degree of flatness required.

Mirrors for high-precision applications – for example, in visible lasers – are normally specified to be flat between  $\lambda/10$  and  $\lambda/20$  for visible light. This degree of flatness is not required when the mirror is merely for light collection and redirection, for example to reflect light onto the surface of a detector. Mirrors for this sort of application are routinely flat to within a few wavelengths of visible light. Excellent mirrors for this purpose can be made from “float” plate glass,<sup>32</sup> which is flat to 1 or 2 wavelengths per inch.

New superpolishing techniques have allowed the production of *super mirrors*, which have surface roughness down to 0.1nm or below and loss (absorption plus scattering) below  $10^{-6}$  per reflection. Super mirrors allow the construction of optical resonators with very high Q values,



**Figure 4.29** Production of ghost reflections by a rear-surface mirror whose front surface is not antireflection coated.

up to 1 million, for applications such as cavity ring down spectroscopy and high finesse interferometry (see Section 4.7.4). Such mirrors are available from Los Gatos Research, Newport Corporation, and the Optical Corporation. More discussion of the issues involved in fabricating such mirrors is available in references 34–37.

**Spherical Mirrors.** Spherical mirrors are widely used in laser construction, where the radii of curvature are typically rather long – frequently 20 m radius or more. They can be used whenever light must be collected and focused; however, spherical mirrors are only good for focusing nearly parallel beams of light that strike the mirror close to normal. In common with spherical lenses, spherical mirrors suffer from various imaging defects called aberrations, which have been discussed in Section 4.2.3. A parallel beam of light that strikes a spherical mirror far from normal is not focused to a small spot. When used this way the mirror is *astigmatic*, that is, off-axis rays are focused at different distances in the horizontal and vertical planes. This leads to blurring of the image, or at best the focusing of a point source into a line image.

A useful practical way to distinguish between a flat mirror and one with a large radius of curvature is to use the mirror to view a sharp object in grazing incidence. A curved mirror surface will reveal itself through the blurring of the image, while a flat mirror will give a sharp image.

**Paraboloidal and Ellipsoidal Mirrors.** *Paraboloidal mirrors* will produce a parallel beam of light when a point source is placed at the focus of the paraboloidal surface. Thus, these mirrors are very useful in projection systems. They are usually made of polished metal, although versions using Pyrex substrates are also available. An important application of parabolic mirrors is in the off-axis focusing of laser beams, using off-axis mirrors in which the axis of the paraboloid does not pass through the mirror. When they are used in this way there is complete access to the focal region without any shadowing, as shown in Figure 4.30. Spherical mirrors, on the other hand, do not focus well if they are used in this way. Off-axis paraboloidal mirrors, as well as metal axial paraboloidal mirrors, are available from Edmund Optics, II-VI Infrared, Kugler, Lightwave Enterprises, Melles Griot, Newport, Opti-Forms, OptoSigma, Space Optics Research Lab, and Thor Labs.

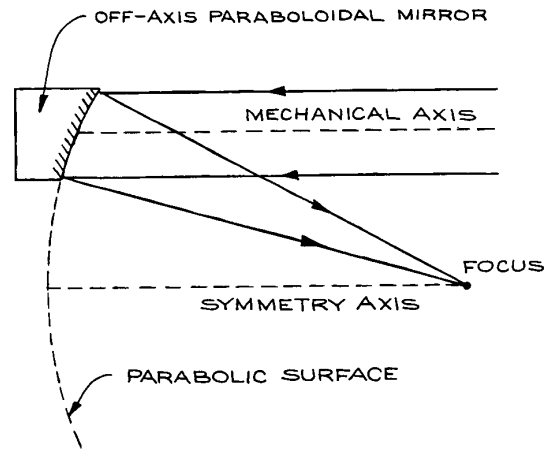


Figure 4.30 Off-axis paraboloidal mirror.

*Ellipsoidal mirrors* are also used for light collecting. Light that passes through one focus of the ellipsoid will, after reflection, also pass through the other. These mirrors, like all-metal paraboloidal mirrors, are generally made by electroforming. The surface finish obtained in this way is quite good, although not so good as can be obtained by optical polishing. Rhodium-plated electroformed ellipsoidal mirrors are available from Edmund Optics, Melles Griot, Opti-Forms, and Spectrum Scientific.

**Dielectric Coatings.** Several different kinds of single-layer and multilayer dielectric coatings are available commercially on the surface of optical components. These coatings can either reduce the reflectance or enhance it in some spectral region.

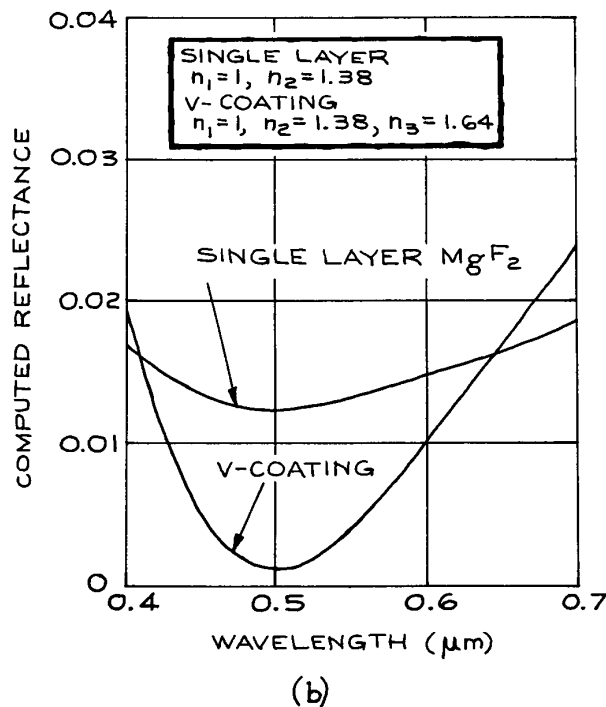
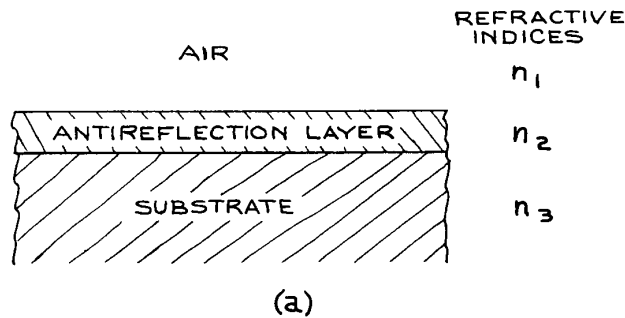
(i) **Single-layer antireflection coatings.** The mode of operation of a single-layer antireflection (AR) coating can be conveniently illustrated by the method of transformed impedances. Suppose medium 3 in Figure 4.31(a) is the surface to be AR coated. Apply to the surface a coating with an effective thickness:

$$d' = \lambda_2/4 \quad (4.138)$$

where  $\lambda_2 = c_0/vn_2$ . The actual thickness of layer 2 is:

$$d = d'/\cos\theta_2 \quad (4.139)$$





**Figure 4.31** (a) Geometry of single-layer antireflection coating; (b) reflectances of single- and double-layer antireflection coatings on a substrate of refractive index 1.64. (From *Handbook of Lasers*, R.J. Pressley (Ed.), CRC Press, Cleveland, 1971; by permission of CRC Press, Inc.)

The transformed impedance of medium 3 at the first interface is found by substitution in Equation (4.110):

$$Z''_3 = Z'_2/Z'_3 \quad (4.140)$$

To reduce the reflection coefficient to zero, we need  $Z''_3 = Z'_1$ ; so the effective impedance of the antireflection layer must be:

$$Z'_2 = \sqrt{Z'_1 Z'_3} \quad (4.141)$$

Thus, to eliminate reflection of waves linearly polarized in the plane of incidence and incident at angle  $\theta_1$ ,  $n_2$  must be chosen to satisfy:

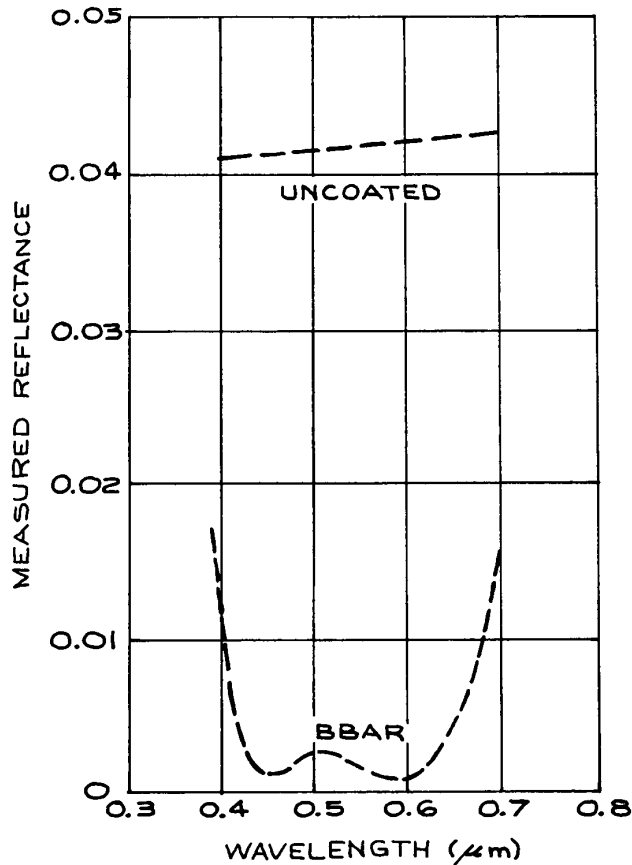
$$\frac{\cos \theta_2}{n_2} = \sqrt{\frac{\cos \theta_1 \cos \theta_3}{n_1 n_3}} \quad (4.142)$$

For use in normal incidence,  $n_2 = \sqrt{n_1 n_3}$  and  $d = \lambda/4$ . In normal incidence, an AR coating works for any incident polarization. To minimize reflection at an air-flint-glass ( $n = 1.7$ ) interface, we would need:

$$n_2 = \sqrt{1.7} = 1.3 \quad (4.143)$$

Magnesium fluoride with a refractive index of 1.38 and cryolite (sodium aluminum fluoride) with a refractive index of 1.36 come closest to meeting this requirement in the visible region of the spectrum. Optical components such as camera lenses are usually coated with one of these materials to minimize reflection at 550 nm. The slightly greater Fresnel reflection that results in the blue and red gives rise to the characteristic purple color of these components in reflected light often called “blooming”.

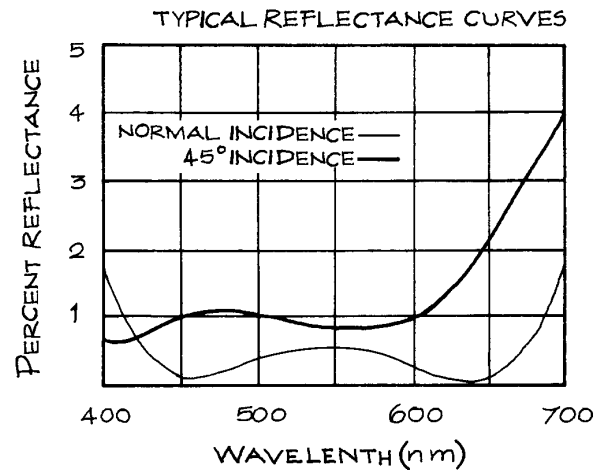
**(ii) Multilayer antireflection coatings.** The minimum reflectance of a single-layer antireflection-coated substrate is  $(n_2^2 - n_1 n_3)/(n_2^2 + n_1 n_3)^2$ . Available robust optical coating materials such as  $\text{MgF}_2$ , however, do not have a sufficiently high refractive index to reduce the reflectance to zero with a single layer. Two-layer coatings, often called V-coatings because of the shape of their transmission characteristic, reduce reflection better than a single layer, as shown in Figure 4.31(b). Multilayer coatings can be used to reduce reflectance over a broader wavelength region than a V-coating, as shown in Figure 4.32. Such broadband coatings are usually of proprietary design: they are available commercially from many sources, as are coatings of other types.<sup>31</sup> Some companies offer both optical components and coatings, while others specialize in coatings and will coat customers' own materials. Melles Griot offer a HEBBAR (high efficiency broad-band antireflection)



**Figure 4.32** Reflectance of a typical broad-band antireflection coating (BBAR) showing considerable reduction in reflectance below the uncoated substrate. (From *Handbook of Lasers*, R.J. Pressley (Ed.), CRC Press, Cleveland, 1971; by permission of CRC Press, Inc.)

coating that provides very low reflectance over a broad bandwidth, and is relatively insensitive to angle of incidence, as shown in Figure 4.33.

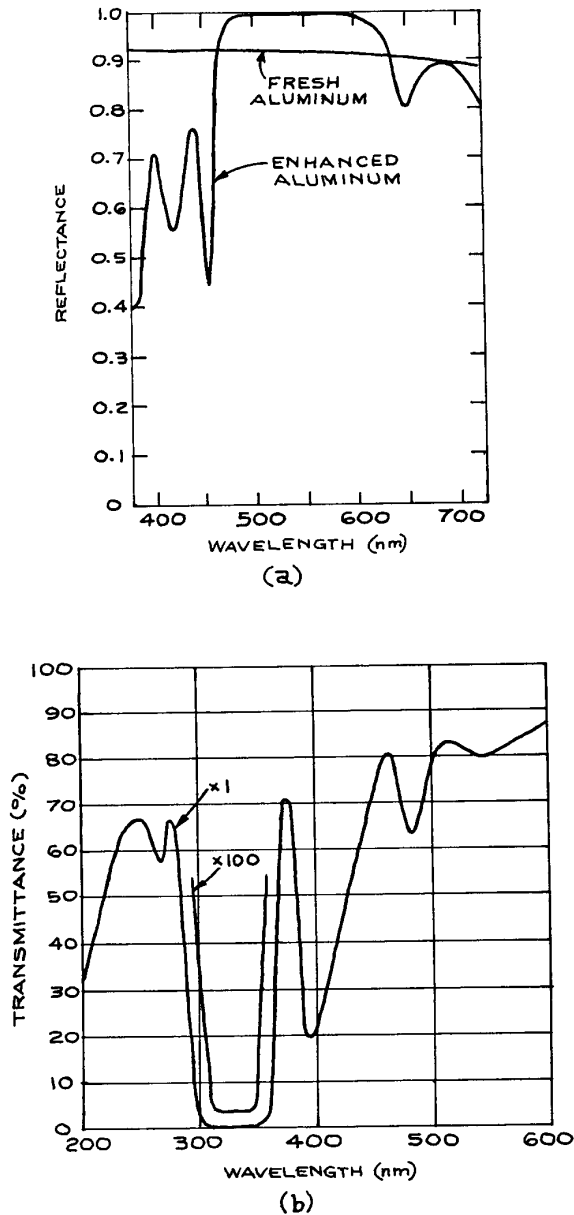
**(iii) High-reflectance coatings.** The usually high reflectance of a metal surface can be further enhanced, as shown in Figure 4.34(a), by a multiple dielectric coating consisting of an even number of layers, each a quarter wavelength thick, and of alternately high and low refractive index, with a low-refractive-index material next to the metal. A very high reflectance over a narrow wavelength range can be



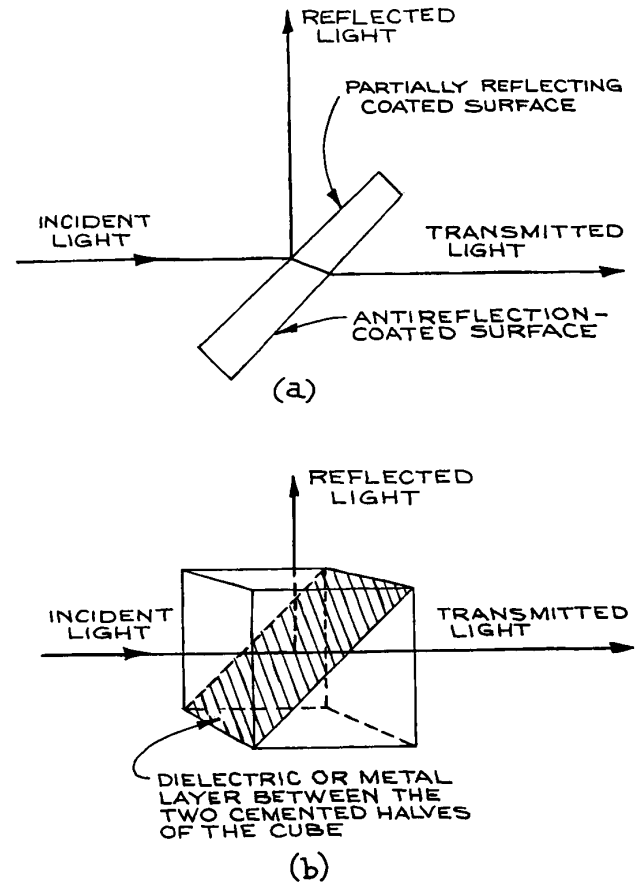
**Figure 4.33** The reflectance spectrum of an HEBBAR high reflectance broad-band coating.

achieved, as shown in Figure 4.34(b), by using a dielectric substrate, such as fused quartz, and an odd number of quarter-wavelength layers of alternately high and low refractive index with a high-index layer on the substrate. Broad-band high-reflectance coatings are also available where the thickness of the layers varies around an average value of a quarter wavelength.

**Beamsplitters.** Beamsplitters are semitransparent mirrors that both reflect and transmit light over a range of wavelengths. A good beamsplitter has a multilayer dielectric coating on a substrate that is slightly wedge-shaped to eliminate interference effects, and antireflection-coated on its back surface to minimize ghost images. The ratio of reflectance to transmittance of a beamsplitter depends on the polarization state of the light. The performance is usually specified for light linearly polarized in the plane of incidence (*P*-polarization) or orthogonal to the plane of incidence (*S*-polarization). Cube beamsplitters are pairs of identical right-angle prisms cemented together on their hypotenuse faces. Before cementing, a metal or dielectric semireflecting layer is placed on one of the hypotenuse faces. Antireflection-coated cube prisms have virtually no ghost image problems and are more rigid than plate type beamsplitters. The operation of both types of beamsplitter is illustrated in Figure 4.35.



**Figure 4.34** (a) Reflectance of a plain aluminum mirror and an aluminum mirror with four dielectric overlayers. (From *Handbook of Lasers*, R.J. Pressley (Ed.), CRC Press, Cleveland, 1971; by permission of CRC Press); (b) typical transmittance characteristic of a narrow-band, maximum-reflectance, multilayer dielectric coating on a dielectric substrate.



**Figure 4.35** (a) Conventional planar beamsplitter (wedge angle greatly exaggerated); (b) cube beamsplitter.

**Pellicles.** Pellicles are beam-splitting mirrors made of a high-tensile-strength polymer stretched over a flat metal frame. The polymer film can be coated to modify the reflection–transmission characteristics of the film. The polymer generally used, nitrocellulose, transmits in the visible and near infrared to about  $2\mu\text{m}$ . These devices have some advantages over conventional coated-glass or quartz beamsplitters; the thinness of the polymer film virtually eliminates spherical and chromatic aberrations (see Section 4.3.3) when diverging or converging light passes through them, and ghost-image problems are virtually eliminated. They will, however, produce some wavefront distortion, typically about two waves per inch,

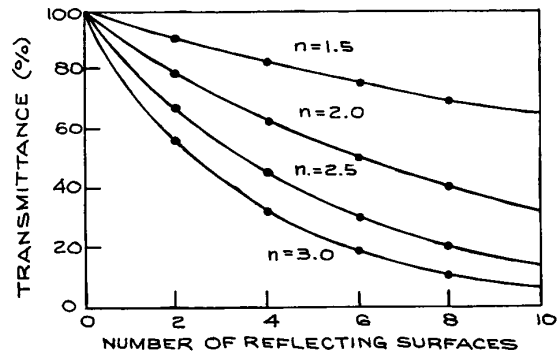
and are not suitable for precision applications. Pellicles are available from several suppliers<sup>31</sup>, such as Edmund Optics, Melles Griot, National Photocolor, and Newport.

### 4.3.2 Windows

An optical window serves as a barrier between two media, for example, as an observation window on a vacuum system or on a liquid or gas cell, or as a Brewster window on a laser. When light passes through a window which separates two media, there is, in general, some reflection from the window and a change in the state of polarization of both the reflected and the transmitted light. If the window material is not perfectly transparent at the wavelength of interest, there is also absorption of light in the window, which at high light intensities will cause the window to heat. This will cause optical distortion of the transmitted wave, and at worst – in high-power laser applications – damage the window on its surface, internally, or both. Additionally, if the two surfaces of a window are both very flat (roughly speaking, one wave per centimeter or better) and close to parallel, the window will act as an *etalon* (see Sections 4.3.7 and 4.7.4) and exhibit distinct variations in transmission with wavelength. To circumvent this difficulty, which is a particular nuisance in experiments using lasers, most precision flat optical windows are constructed as a slight wedge, with an angle usually about 30'.

The details of the reflection, refraction, and change of polarization state that occur when light strikes a window surface are dealt with in detail in Sections 4.2.3 and 4.3.6. These considerations also apply when light strikes a lens or prism. The reflection at such surfaces can be reduced by a single-layer or multilayer dielectric antireflection coating. In optical systems, such as multielement camera lenses, where light crosses many such surfaces, antireflection coatings are very desirable. Otherwise a severe reduction in transmitted light intensity will result, as illustrated in Figure 4.36.

Optical cells, or cuvettes, are transparent containers, generally made of glass or fused quartz, which are used in absorption and fluorescence spectrophotometers, in light-scattering and turbidity measurements, and in many other specialized experiments. These cells are usually rectangular or cylindrical in shape and are available in many sizes and configurations, including provision for liquid circulation. Suppliers include Harrick Scientific, Hellma,



**Figure 4.36** Transmittance of a multielement optical system as a function of the refractive index of the elements.

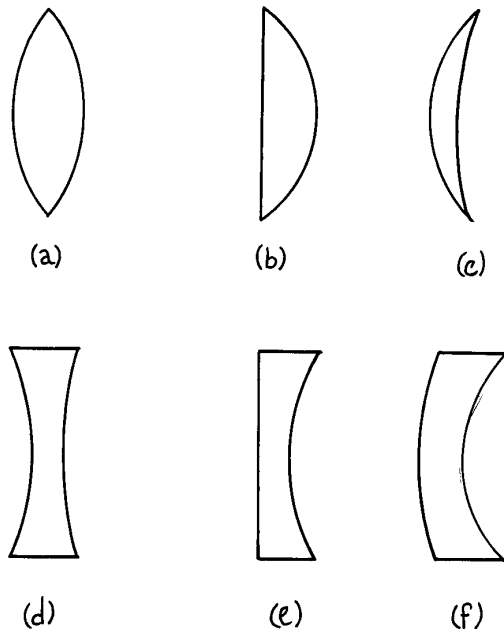
International Crystal Laboratories, Newport/Oriel, NSG Precision Cells, Optiglas, Spectrocell, and Starna.

### 4.3.3 Lenses and Lens Systems

**Singlet Lenses.** *Singlet lenses* are the simplest and most readily available lenses for nonexacting applications. They come in six basic formats, shown in Figure 4.37. The lens is specified by the material of which it is made, the curvatures  $R_1$  and  $R_2$  of its two faces, the thickness  $d$  of the lens at its mid point, and its aperture diameter  $D$ , which also indirectly specifies its thickness at the edge. The lens has an effective focal length (EFL),  $f$  measured from its principal planes, and front and back focal lengths, which specify the distances of the focal points from the front and back surfaces of the lens, as shown in Figure 4.38. A *biconvex lens* has two generally spherical surfaces. In the standard sign convention the first of these has a positive radius and the second a negative radius. A symmetrical biconvex lens has  $R_1 = -R_2$ . A *plano-convex* lens has one flat face, while a *convex meniscus* lens has a first surface of positive radius and a second surface of a larger positive radius. Biconvex, plano-convex, and convex meniscus lenses all have a positive EFL. *Biconcave*, *plano-concave*, and *concave meniscus lenses* have surface curvatures, opposite in sign to their convex counterparts, and all have a negative EFL.

All singlet lenses suffer from spherical and chromatic aberration. They are inexpensive, but suffer from significant

amounts of aberration, so are not a good choice in imaging situations. For focusing parallel light (infinite object conjugate), or for producing a parallel beam of light, plano-convex lenses or convex meniscus lenses that are close to

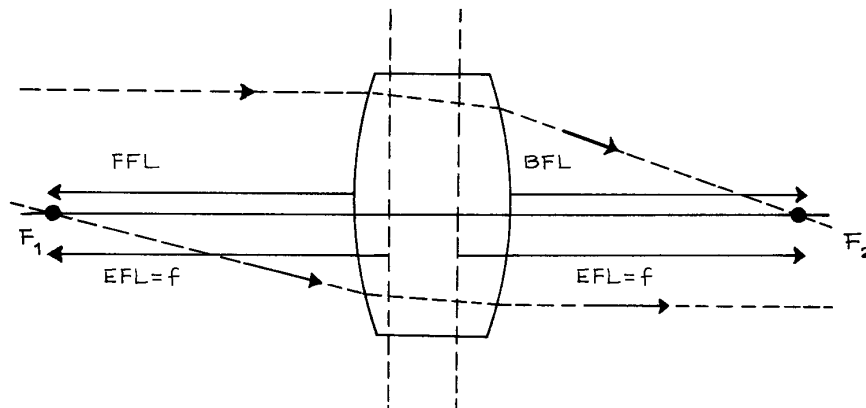


**Figure 4.37** The six basic singlet lens types: (a) biconvex; (b) plano-convex; (c) convex meniscus; (d) biconcave; (e) plano-concave; (f) concave meniscus.

plano-convex exhibit the least spherical aberration for a specified EFL. For best performance in these applications the parallel light should pass through the more strongly curved of the two surfaces. For imaging at different conjugate ratios the shape of singlet that produces the least spherical aberration is called the *best form* lens for that application. For example, for equal object and image distances (conjugate ratio = 1) the best form singlet is a symmetrical biconvex lens. Singlet lenses can provide satisfactory performance provided they are used with relatively large  $f$ /numbers. If singlet lenses are used with lasers then there is no chromatic aberration.

**Doublets.** If a positive focal length lens of one type of glass is combined with a weaker negative lens made from a different type of glass, then the lens combination can exhibit reduced spherical and chromatic aberration. *Doublets* can be either air-spaced or cemented. They are widely available from many suppliers, such as CVI/Melles Griot, Newport Corporation, OptoSigma, Rolyn, and Thorlabs. For laser applications, even without the concern for chromatic aberration, doublets will provide superior focusing performance to singlets.

**Camera lenses.** *Compound lenses* involve more than two spherical surfaces and can be designed to have reduced aberrations. Camera lenses are multi-element compound lenses especially designed for imaging applications.

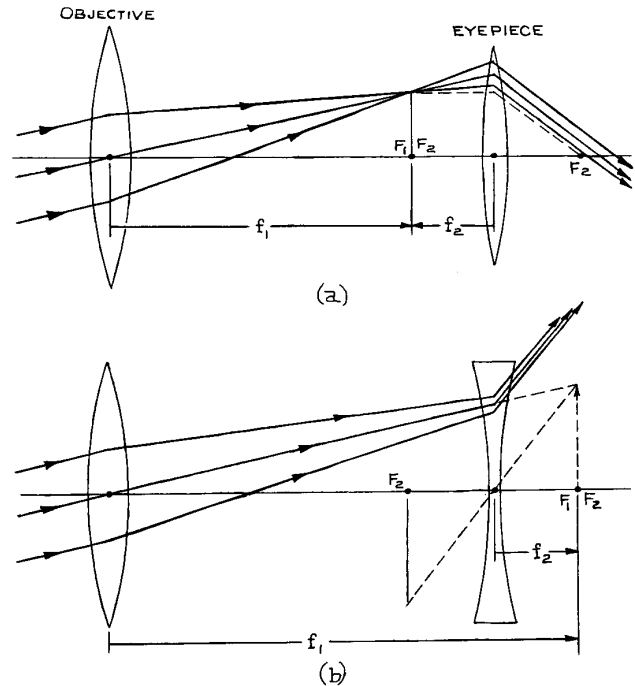


**Figure 4.38** Diagram showing the effective focal length (EFL), the front (FFL) focal length, and the back focal length (BFL) of a thick lens.

A wide variety of high-performance camera lenses is available commercially from companies such as Canon, Fujinon, Kodak, Leitz, Minolta, Nikon, Olympus, Sigma, and Zeiss. The choice of camera lenses for a specific application is generally guided by two parameters: (i) the *field of view* (FOV), which can be specified by the apex angle of the cone of light that can be collected by the lens and (ii) the aperture of the lens. Both these parameters influence the light-gathering power of the lens.

*Telecentric lenses* are special compound lenses used in applications such as machine vision and lithography where object or image movement does not produce a change in magnification.<sup>33–38</sup> For a lens with object space telecentricity, a change in the object distance will not result in the image changing size. For such a lens the aperture stop is located in the back focal plane and the entrance pupil is at infinity. All the principal rays in object space are parallel to the axis. For a lens with image space telecentricity, moving the image plane to focus or defocus the system will not change the size of the image. In this case the aperture stop is located in the front focal plane and the exit pupil is at infinity. In the image space all the principal rays are parallel to the axis. If a lens is telecentric for both the object and image spaces it is called *doubly telecentric*. The primary disadvantage of telecentric lenses is that they have a very limited field of view: object and image have to lie within a cylindrical region whose radius in object and image space is limited by the size of the lens elements. Telecentric lenses are available from Computer Optics, Edmund Optics, Mekoptik, Navitar, Opto-Engineering, and Sill Optics.

**Simple and Galilean Telescopes.** The operation of a *simple telescope* is illustrated in Figure 4.39(a). The objective is usually an *achromat* (a lens in which chromatic aberration – the variation of focal length with wavelength – has been minimized). It has a long focal length  $f_1$  and produces a real image of a distant object, which can then be examined by the eyepiece lens of focal length  $f_2$ . This design, which produces an inverted image, is often called an *astronomical telescope*. If desired, the final image can be erected with a third lens, in which case the device is called a *terrestrial telescope*. The eyepiece of the telescope can be a singlet lens, but composite eyepiece designs such as the Ramsden and the Kellner<sup>9</sup> are also common. Most



**Figure 4.39** Ray paths through telescopes in *normal* adjustment (object and final image at infinity): (a) astronomical telescope; (b) Galilean telescope.

large astronomical telescopes use spherical mirrors as objectives, and various configurations are used, such as the Newtonian, Cassegrain, and Schmidt.<sup>11,23–25</sup> Since most telescope development has been for astronomical applications, it is outside the scope of this book to give a detailed discussion; for further details the reader should consult Born and Wolf,<sup>11</sup> Levi,<sup>24</sup> Meinel,<sup>38</sup> or Brouwer and Walther.<sup>39</sup>

Small astronomical and terrestrial telescopes are widely available at relatively low cost. Two principal manufacturers are Celestron<sup>40</sup> and Meade.<sup>41</sup>

The *Galilean telescope* illustrated in Figure 4.39(b), first constructed by Galileo in 1609, is the earliest telescope of which there exists definite knowledge. It produces no real intermediate image, but the final image is erect. In *normal adjustment*, the focal point of the diverging eyepiece is outside the telescope and coincident with that of the objective. In the simple telescope the two focal points also

coincide, but are between the two. The magnification  $M$  produced by either type of telescope can be written as:

$$M = -f_1/f_2 \quad (4.144)$$

Since the simple telescope has two positive (converging) lenses, its overall magnification is negative, which indicates that the final image is inverted.

Simple and Galilean telescopes have practical uses in the laboratory that are distinct from their traditional use for observing distant objects.

**Laser-Beam Expanders and Spatial Filters.** Laser beams can be expanded and recollimated (or focused and recollimated) with simple or Galilean telescope arrangements, as illustrated in Figure 4.40. For optimum recollimation the spacing of the two lenses should be adjustable, as fine adjustment about the spacing  $f_2 + f_1$  will be necessary for recollimating a Gaussian beam. In this application the Galilean telescope has the advantage that the laser beam is not brought to an intermediate focus inside the beam expander. Gas breakdown at such an internal focus can

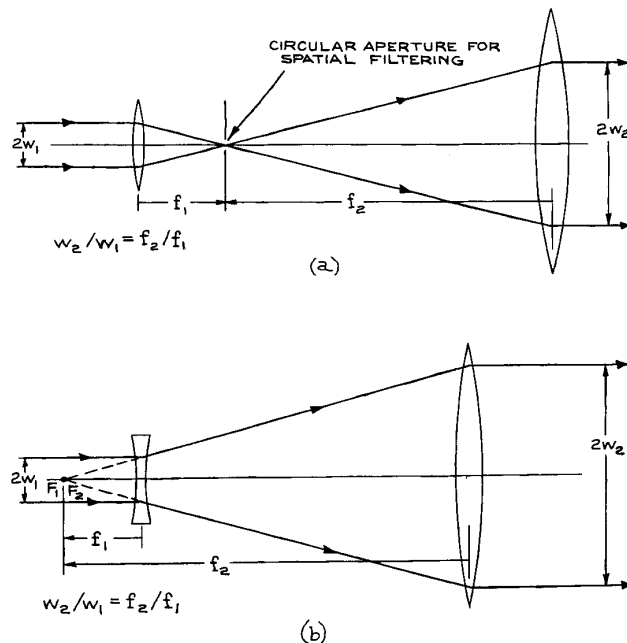
occur with high-power laser beams, although this problem can be solved in simple-telescope beam expanders by evacuating the telescope. Since laser beams are highly monochromatic, beam expanders need not be constructed from achromatic lenses. Attempts should, however, be made to minimize spherical aberration and beam distortion. It is best to use precision antireflection-coated lenses if possible. Laser-beam-expanding telescopes that are very well corrected for spherical aberration are available from Janos, Melles Griot, Newport/Oriel, Rolyn, Sigma Koki, Space Optics Research Labs, and Special Optics, among others.<sup>31</sup> Infrared laser-beam expanders usually have ZnSe or germanium lenses. The use of biconvex and biconcave lenses distributes the focusing power of the lenses over their surfaces and minimizes spherical aberration. Better cancellation of spherical aberration is possible in Galilean beam expanders than with simple telescopes.

If a small circular aperture is placed at the common intermediate focus of a simple-telescope beam expander, the device becomes a spatial filter as well. Although ideally the output beam from a laser emitting a TEM<sub>00</sub> mode has a Gaussian radial intensity profile, in practice the radial profile may have some irregular structure. Such beam irregularities may be produced in the laser or by passage of the beam through some medium. If such a beam is focused through a small enough aperture and then recollimated, the irregular structure on the radial intensity profile can be removed and a smooth profile restored, as illustrated in Figure 4.41. This is called *spatial filtering*. The minimum aperture diameter that should be used is

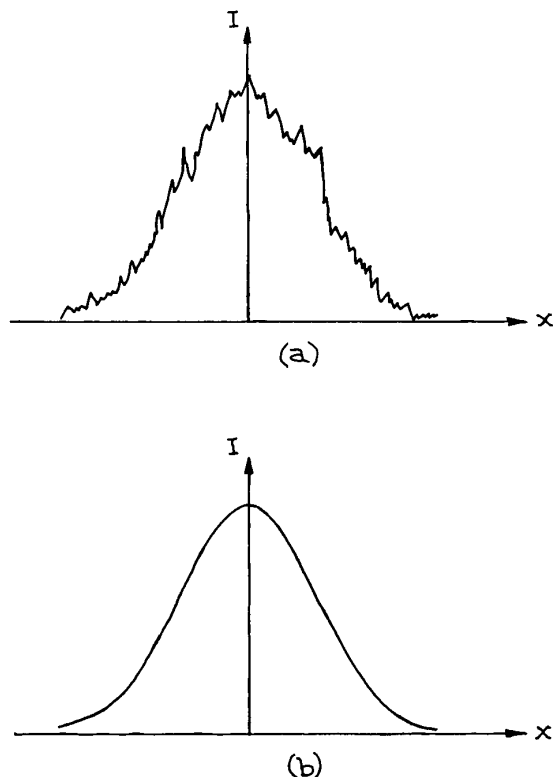
$$D_{\min} = \frac{2f_1\lambda}{\pi w_1} \quad (4.145)$$

where  $f_1$  is the focal length of the focusing lens and  $w_1$  is the spot size at that lens. Typical aperture sizes used in commercial spatial filters for visible lasers are on the order of 10  $\mu\text{m}$ . This aperture must be accurately and symmetrically positioned at the focal point, so it is usually mounted on an adjustable XY translation stage.

**Lens Aberrations.** As discussed in Section (4.2.5), real lenses suffer from various forms of aberration, which lead the image of an object to be an imperfect reconstruction of it. If a particular aberration is judged to be detrimental in a



**Figure 4.40** Laser-beam expanders: (a) focusing type with spatial filter; (b) Galilean type.



**Figure 4.41** Intensity variation  $I(x)$  along a diameter of a  $TEM_{00}$  laser beam possessing spatial noise structure: (a) before and (b) after spatial filtering.

particular experimental situation, then a special lens combination can often be obtained that minimizes that aberration – occasionally at the expense of worsening others.

**(i) Chromatic aberrations.** Because the refractive index of the material of a lens varies with wavelength, so does its focal length. Therefore rays of different wavelengths from an object form images in different locations. Chromatic aberration is almost eliminated with the use of an *achromatic doublet*. This is a pair of lenses, usually consisting of a positive crown-glass lens and a negative flint-glass lens, cemented together. These two lenses cancel each other's chromatic aberrations exactly, at specific wavelengths in the blue and red, and almost exactly in the region in between. Good camera lenses are well corrected for chromatic aberrations.

Achromatic doublets are available from numerous suppliers, including Edmund Optics.

**(ii) Spherical aberration.** This aberration can be minimized in a single lens by distributing the curvature between its two surfaces without altering the focal length. So, for example, a biconvex lens will produce less spherical aberration than a plano-convex one of the same focal length. Spherical aberration can be minimized in an achromatic doublet with appropriate choice of curvatures for the constituent lens elements.

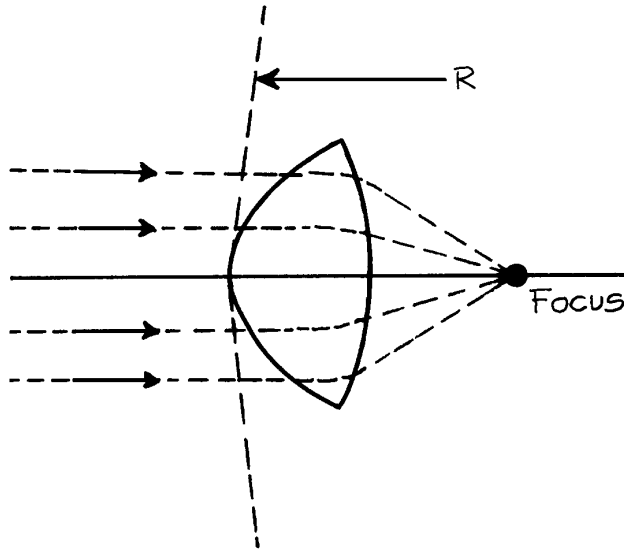
**(iii) Coma.** When an object point is off the axis of a lens, its image is produced in different lateral positions by different zones of the lens. Coma can be controlled by an appropriate choice of lens curvatures.

Other forms of aberration also occur, such as *field curvature*, in which an object plane orthogonal to the lens axis is imaged as a curved surface.

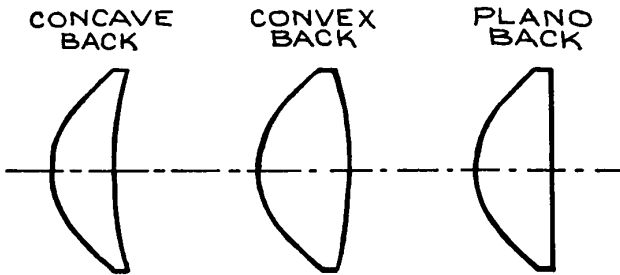
**(iv) Astigmatism.** This aberration can be controlled by lens curvature and index variations, and by the use of apertures to restrict the range of angles and off-axis distances at which rays can traverse the lens.

**Aspheric Lenses.** If one face of a singlet lens is manufactured as a nonspherical (or *aspherical*), surface then at a specified design wavelength all spherical aberration can be eliminated. With such a lens it is possible to obtain a much smaller focal length than with a conventional spherical lens of the same diameter, without increasing the spherical aberration of the lens. An ideal aspheric lens exactly cancels the spherical aberration that would otherwise be present in the optical system. Such lenses can produce the smallest focal spots when used with collimated laser light. The collimated beam should enter the aspherical surface and the focus will be on the opposite side, as shown in Figure 4.42. The second face of an aspherical lens is often plane, or weakly spherical as shown in Figure 4.43. If optimum performance at a particular wavelength and focal length is required, then a custom asphere can be manufactured, but this can be expensive, unless a large number of such lenses are required. Aspheres are often made by moulding, so once the mould has been manufactured additional lenses are less expensive.





**Figure 4.42** Focusing light with an aspherical lens.



**Figure 4.43** Cross-sections of three common types of aspheric lens.

The standard formula that describes an aspheric surface is:

$$z = \left(\frac{1}{R}\right) \frac{r^2}{1 + \sqrt{1 - (1 + K)\frac{r^2}{R^2}}} + A_4 r^4 + A_6 r^6 + A_8 r^8 + A_{10} r^{10} + \dots \quad (4.146)$$

where  $z$  is the depth (or sag) of the surface,  $r$  is the lateral position,  $K$  is the conic constant of the surface, and  $R$  is the

effective curvature at the vertex of the surface, as shown in Figure 4.42. The conic constant  $K$  has the following values for different surface profiles:  $K > 0$  ellipse,  $K = 0$  sphere,  $-1 < K < 0$  ellipse,  $K = -1$  parabola, and  $K < -1$  hyperbola. Most aspherical designs use only the even order terms, as shown in Equation (4.146), but the odd terms could also be included.

Aspheric lenses can collect and focus light rays over a much larger solid angle than conventional lenses and can be used at  $f$ -numbers as low as 0.6. Aspheric lenses save space and energy. They can be used as collection lenses very close to small optical sources, such as high-pressure mercury or xenon arc lamps, and for collecting radiation over a large solid angle and focusing it onto a detector element. Such lenses are often referred to as *condensers*. Condensers are widely used in microscopes, where they focus light from the microscope light source onto a specimen. In such applications the aspherical surface should be on the light source side. Because aspheric lenses are generally designed only to minimize spherical aberrations, they will contribute to chromatic aberration, coma, distortion, astigmatism, and curvature of field. Consequently, an aspheric lens should not be used under circumstances where these aberrations will be compounded by transmission through additional components of the optical system. Glass aspheric lenses are available in a range of diameters and focal lengths from several suppliers,<sup>31</sup> such as Fresnel Technologies, Melles Griot, Moritex, OptoSigma, and Thorlabs. Aspheric lenses in exotic materials can be obtained as custom items.

**Graded Index Lenses.** Small lenses with a convenient cylindrical shape can be made from *graded index* material, which has a refractive index that varies from point to point inside the material. A commonly encountered axi-symmetric index variation is Gaussian, or approximately so. The refractive index as a function of distance  $r$  from the axis of symmetry is:

$$n = n_0 e^{-r^2/2\sigma^2} \quad (4.147)$$

For distances from the axis that satisfy  $r \ll \sigma$ , Equation (4.147) can be written as

$$n = n_0 \left(1 - \frac{r^2}{2\sigma^2}\right) = n_0 - \frac{1}{2} n_2 r^2 \quad (4.148)$$

where  $n_2 = n_0/\sigma^2$ . This is a quadratic index profile. A cylindrical piece of such transparent material with flat ends has an overall ray transfer matrix from air, through a length  $d$  of the graded medium and back into air:

$$\begin{aligned} \begin{pmatrix} A & B \\ C & D \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & n_0 \end{pmatrix} \begin{pmatrix} \cos(\frac{d}{\sigma}) & \sigma \sin(\frac{d}{\sigma}) \\ -\frac{1}{\sigma} \sin(\frac{d}{\sigma}) & \cos(\frac{d}{\sigma}) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/n_0 \end{pmatrix} \\ &= \begin{pmatrix} \cos(\frac{d}{\sigma}) & \frac{\sigma}{n_0} \sin(\frac{d}{\sigma}) \\ -\frac{n_0}{\sigma} \sin(\frac{d}{\sigma}) & \cos(\frac{d}{\sigma}) \end{pmatrix} \end{aligned} \quad (4.149)$$

The focal length of this GRIN lens is:

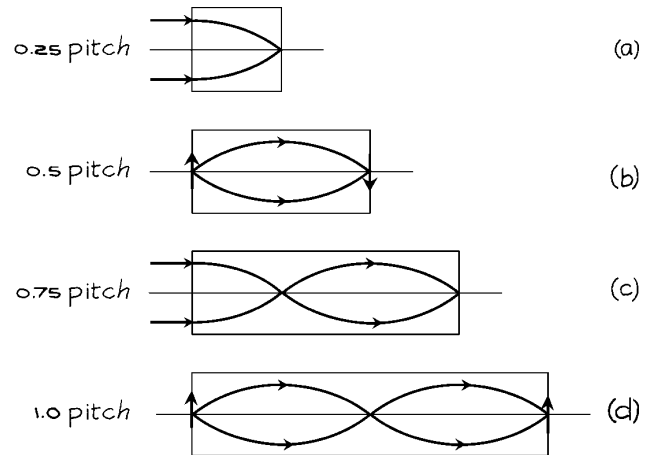
$$f = \frac{\sigma}{n_0 \sin\left(\frac{d}{\sigma}\right)} \quad (4.150)$$

which can also be written as:

$$f = \frac{1}{\sqrt{n_0 n_2} \sin(d \sqrt{n_2/n_0})} \quad (4.151)$$

Rays that enter a graded index lens follow sinusoidal paths until they reach the back surface of the lens. The oscillation pattern inside the lens describes its *pitch*. Figure 4.44 shows schematically how the pitch is defined. A 1/4 pitch graded index lens, for example, will take input parallel light and bring it to a focus at the exit face of the lens. Graded index lenses are very convenient to use in small optical arrangements because their flat ends make them easy to couple to other structures, such as semiconductor lasers or optical fibers. They are available commercially as *GRIN* or *SELFOC* lenses from suppliers such as LINOS Photonics, Melles Griot, NSG, Red Optronics, Schott, and Thorlabs. These lenses should not be used in precision imaging applications because they generally suffer from relatively severe aberrations, especially chromatic aberration.

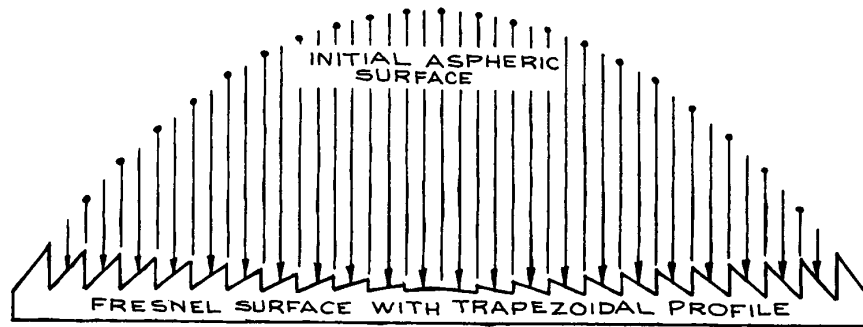
**Fresnel Lenses.** A *Fresnel lens* is an aspheric lens whose surface is broken up into many concentric annular rings. Each ring refracts incident rays to a common focus, so that a very large-aperture, small  $f$ -number, thin aspheric lens results. Figure 4.45 illustrates the construction and focusing characteristics of such a lens. These lenses are



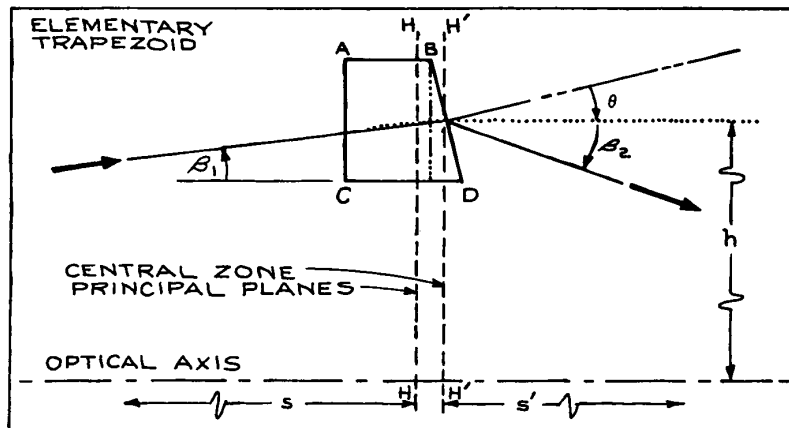
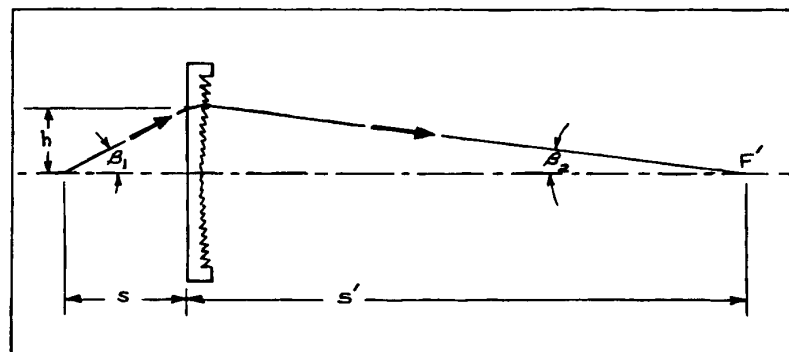
**Figure 4.44** Schematic operation of graded index lenses: (a) a quarter pitch lens, which can be used to either focus parallel light or collimate light from a point source on one face of the lens; (b) a half pitch lens, which inverts an image at its front face; (c) a three quarter pitch lens; (d) a whole pitch lens.

generally manufactured from precision molded acrylic. Because the refractive index of acrylic varies little with wavelength, from 1.51 at 410 nm to 1.488 at 700 nm, these lenses can be very free of chromatic and spherical aberration throughout the visible spectrum. Fresnel lenses should be used so that they focus to the plane side of the lens, and their surfaces should not be handled. They are inexpensive and available from very many suppliers, including Edmund Scientific, Fresnel Technologies, Lexitek, Newport/Oriel, and Wavelength Optics. They should not be considered for use in applications where a precision image or diffraction-limited operation is required. Fresnel lenses also scatter much more light than conventional lenses.

Fresnel lenses are an example of a more general class of optics called diffractive optics. These are structured optical elements that use diffraction to redirect light, split a light beam into many separate beams, focus light, produce diffuse light, or correct distorted wavefronts. They are available from Holographix, MEMS Optical, Optocraft, RPC Photonics, Silios Technologies, and Stocker Yale.



(a)



(b)

**Figure 4.45** (a) A Fresnel lens, made up of trapezoidal concentric sections (it replaces a bulky aspheric lens); (b) illustration of how the focusing characteristics of a Fresnel lens result from refraction at the individual surface grooves. (Courtesy of Melles Griot, Inc.)

**Cylindrical Lenses.** *Cylindrical lenses* are planar on one side and cylindrical on the other. In planes perpendicular to the cylinder axis they have focusing properties identical to a spherical lens, but they do not focus at all in planes containing the cylinder axis. Cylindrical lenses focus an extended source into a line, and so are very useful for imaging sources onto monochromator slits, although perfect matching of  $f$ /numbers is not possible in this way. These lenses are also widely used for focusing the output of solid-state and nitrogen lasers into a line image in the dye cell of dye lasers (see Section 4.6.3). Cylindrical lenses are available from Bond Optics, Esco, Infrared Optical Products, Melles Griot, Newport/Oriel, OFR (now Thorlabs), Optimax, and Rolyon.

**Optical Concentrators.** *Optical concentrators* are nonimaging light collectors that are useful in delivering light to a detector. They effectively increase the collection area of the detector, although they affect the angles of light rays.

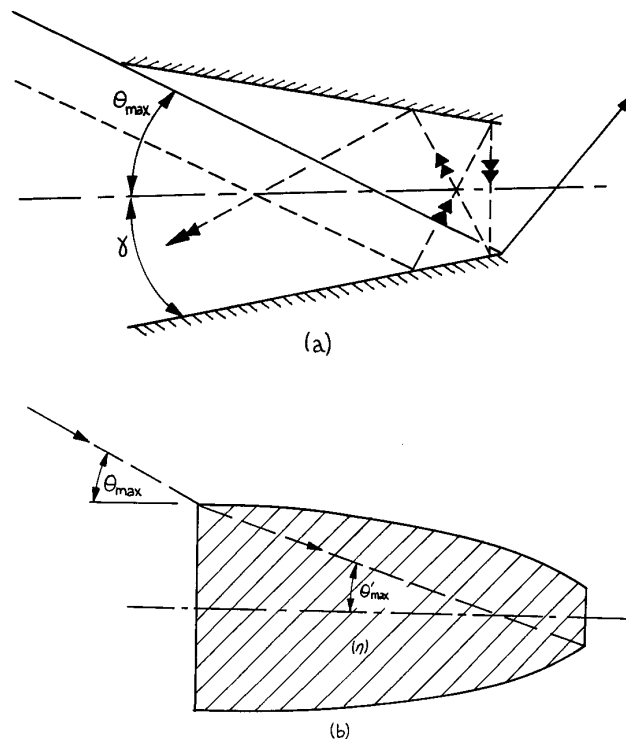
If light enters an optical system of entrance aperture diameter  $2a$  over a range of angles  $2\theta$  and leaves through an aperture of diameter  $2a'$  over a range of angles  $2\theta'$  then in the paraxial approximation:

$$a\theta = a'\theta' \quad (4.152)$$

which follows from the brightness theorem. If the refractive indices of the input and output media are  $n, n'$ , respectively, then this relation becomes:

$$na\theta = n'a'\theta' \quad (4.153)$$

The quantity  $n^2a^2\theta^2$ , which is unchanged in going through the system, is called the *étendue* of the system. Figure 4.46 old shows two simple concentrator designs, a linear conical concentrator, and a compound parabolic concentrator (CPC), often called a “Winston” Cone. These devices are hollow or solid cones, which reflect entering rays one or more times before they leave the exit aperture. The range of light angles leaving such a device is greater than the range of light angle entering, by the linear ratio between the entrance and exit apertures. Such a nonimaging light collector could be used to collect light from a diffusely



**Figure 4.46** Simple optical concentrator designs: (a) linear cone; (b) compound parabolic concentrator.

emitting source and direct it onto a small area photoelector; however, the detector’s active surface must be placed close to the exit aperture because of the larger range of angles of the emerging light rays.

Nonimaging light collectors are useful for collecting light whose spatial distribution may be fluctuating, yet whose overall power may remain relatively constant, for example, in the case of a laser beam that has passed through a fluctuating medium.

The *concentration ratio* is:<sup>42,43</sup>

$$C = a/a' = \frac{n' \sin \theta'}{n \sin \theta} \quad (4.154)$$

In 2-D the maximum concentration ratio is:

$$C_{\max} = \left( \frac{n'}{n \sin \theta} \right) \quad (4.155)$$

and in 3-D for an axisymmetric concentrator:

$$C_{\max} = \left( \frac{n'}{n \sin \theta} \right)^2 \quad (4.156)$$

The CPC can be fabricated from a solid transparent material with refractive index  $n$ . For total internal reflection to occur for all internal rays that enter within a range of angles  $\theta$ :

$$\sin \theta' \leq 1 - \left( \frac{2}{n^2} \right) \quad (4.157)$$

or

$$\sin \theta \leq n - \frac{2}{n} \quad (4.158)$$

To be useful as a material for a CPC it is clear that  $n > \sqrt{2}$ .

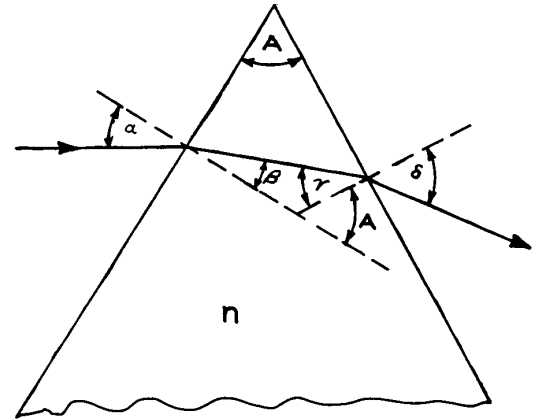
Since the rays leaving a concentrator can cover a solid angle up to  $2\pi$  sr, when the maximum concentration ratio is used, a detector used in conjunction with the concentrator would need to have its photosensitive surface directly at the exit aperture.

Simple concentrators can be made by direct machining of plastic material such as lexan (refractive index  $\sim 1.585$ ) using a numerically controlled lathe, followed by polishing. Alternatively, a stainless-steel mandrel could be machined to the required shape, and then a mold made by electroplating over the stainless steel. After the hollow mold is made, then plastic concentrators can be fabricated using transparent epoxy resin placed into the mold and allowed to harden.

### 4.3.4 Prisms

Prisms are used for the dispersion of light into its various wavelength components and for altering the direction of beams of light. Prisms are generally available in most materials that transmit ultraviolet, visible, and infrared light. High-refractive-index semiconductor materials, such as silicon, germanium, and gallium arsenide, however, are rarely used for prisms. Consult Section 4.4 for a list of suppliers of optical materials.

The deviation and dispersion of a ray of light passing through a simple prism can be described with the aid of Figure 4.47. The various angles  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  satisfy Snell's law



**Figure 4.47** Passage of a ray of light through a prism.

irrespective of the polarization state of the input beam, provided the prism is made of an isotropic material. It is easy to show that:

$$\sin \delta = \sin A (n^2 - \sin^2 \alpha)^{1/2} - \cos A \sin \alpha \quad (4.159)$$

and:

$$D = \alpha + \delta - A \quad (4.160)$$

where  $A$  is the apex angle shown in Figure 4.47. The exit ray will not take the path shown if  $\gamma$  is greater than the critical angle.

The dispersion of the prism is defined as:

$$\frac{d\delta}{d\lambda} = \frac{d\delta}{dn} \frac{dn}{d\lambda} \quad (4.161)$$

Substituting from Equation (4.159), this becomes:

$$d\delta = \frac{\sin A}{\cos \beta \cos \delta} \frac{dn}{d\lambda} \quad (4.162)$$

When the prism is used in the position of minimum deviation:

$$\beta = \gamma = \frac{1}{2}A \quad (4.163)$$

and:

$$\alpha = \delta = \frac{1}{2}(D_{\min} + A) \quad (4.164)$$

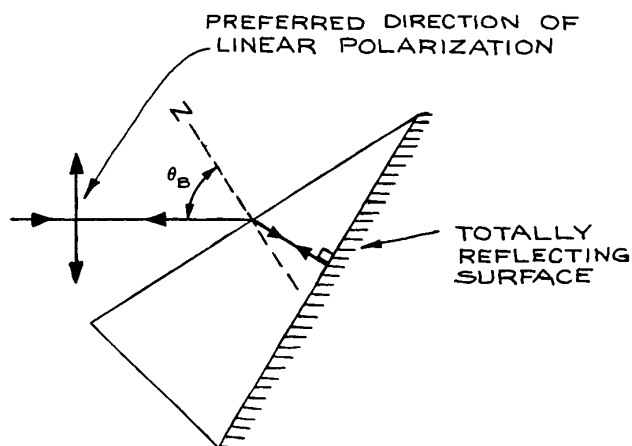
By the sine rule, Equation (4.162) can be written in the form:

$$\frac{d\delta}{d\lambda} = \frac{t}{W} \frac{dn}{d\lambda} \quad (4.165)$$

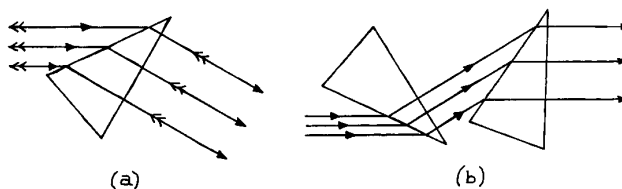
where  $t$  is the distance traveled by the ray through the prism and  $W$  is the dimension shown in Figure 4.47.

A simple, yet important, application of a prism as a dispersing element is in the separation of a laser beam containing several well-spaced wavelengths into its constituent parts. Situations where this might be necessary include isolation of the 488.8 nm line in the output of an argon ion laser oscillating simultaneously at 476.5, 488.8, and 514.5 nm and other lines, or in nonlinear generation where a fundamental and second harmonic must be separated. If the spatial width of the laser beam is known, then an equation such as (4.165) will determine at what distance from the prism the different wavelengths are spatially separated from each other. The advantages of prisms in this application are that they can introduce very little scattered light, produce no troublesome ghost images, and can be cut so that they operate with  $\alpha = \delta = \theta_B$ , thereby eliminating reflection losses at their surfaces. Brewster-angle prisms are used in this way inside the cavity of a laser to select oscillation at a particular frequency. A modified prism called a reflecting *Littrow prism* is frequently used in this way, as shown in Figure 4.48. A Littrow prism is designed so that for a particular wavelength the refracted ray on entering the prism travels normal to the exit face. Thus, if the exit face is reflectively coated, for the specified wavelength the incident light is returned along its original path. Generally, a dispersing optical element used in *Littrow* has this retroreflective characteristic at a particular wavelength.

If a beam of parallel monochromatic light passes through a prism, unless the prism is used in the symmetrical minimum-deviation position, the width of the beam will be increased or decreased in one dimension. This effect is illustrated in Figure 4.49. By the use of two identical prisms in an inverted configuration, the compression or expansion can be accomplished without an angular change in beam direction. Prism beam expanders employing this principle are used in some commercial dye lasers for expanding a small-diameter, intracavity laser beam onto a diffraction grating to achieve greater laser line nar-



**Figure 4.48** Reflecting prism used in Littrow at Brewster's angle. Only light of a specific wavelength will be refracted at the entrance face of the prism and then reflected back on itself at the coated reflecting face.



**Figure 4.49** (a) One-dimensional beam expansion (or compression) with a prism; (b) one-dimensional beam expansion without beam deflection using two oppositely oriented, similar prisms.

rowing. Such prism pairs can be used, in conjunction with lenses, to produce *anamorphic* laser beam expanders. These devices are important for the circularization and collimation of the beams from semiconductor lasers. These lasers generally emit an elliptical-shaped laser beam whose beam divergence angle is different in two orthogonal planes containing the beam axis.

*Triangular prisms* with a wide range of shapes and vertex angles are available. The most common types are the equilateral prism, the right-angle prism, and isosceles and Littrow prisms designed for Brewster-angle operation. Such prisms are readily available in glass and fused silica

from many suppliers.<sup>31</sup> Triangular and other prisms also find wide use as reflectors, utilizing total internal reflection inside the prism or a reflective coating on appropriate faces. Some specific reflective applications of prisms are worthy of special note.

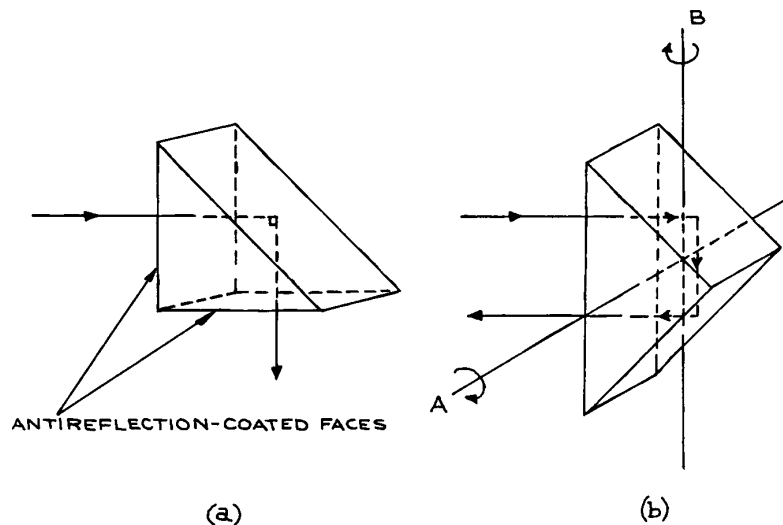
**Right-Angle Prisms.** Right-angle prisms can be used to deviate a beam of light through  $90^\circ$ , as shown in Figure 4.50(a). Antireflection coatings on the faces shown are desirable in exacting applications. A right-angle prism can also be used to reflect a beam back parallel to its original path, as shown in Figure 4.50(b). The retro-reflected beam is shifted laterally. The prism will operate in this way provided the incident beam is in the plane of the prism cross-section. The retroreflection effect is thus independent of rotation about the axis A in Figure 4.50(b). If the angle of incidence differs significantly from zero, the roof faces of the prism need to be coated; otherwise total internal reflection at these surfaces may not result. If such a prism is rotated about the axis B, only in a single orientation does retroreflection result. Roof prisms are available

from numerous suppliers, including Ealing, Edmund Scientific, Melles Griot, Newport/Oriel, Opto-Sigma, and Rolyn.

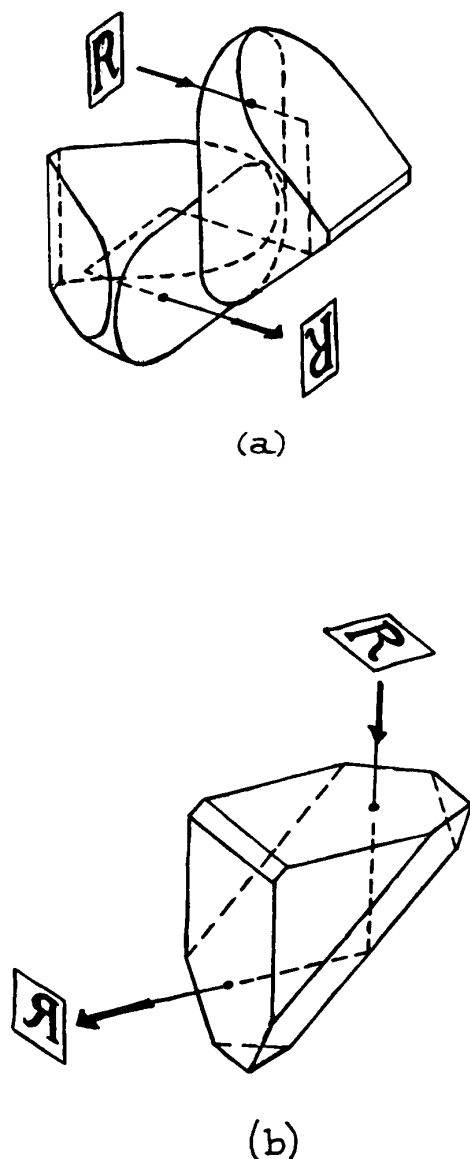
*Porro prisms* are roof prisms with rounded corners on the hypotenuse face and beveled edges. They are widely used in pairs for image erection in telescopes and binoculars, as shown in Figure 4.51(a). They are available from Anchor Optics, Edmund Optics, and Sunny Precision Optics. *Amici prisms* are right-angle prisms where the hypotenuse face has been replaced by a  $90^\circ$  roof. They are available from Anchor Optics, Edmund Optics, Red Optronics, Starna, and TrustOptics. An image viewed through an Amici prism is left-to-right reversed (reverted) and top-to-bottom inverted, as shown in Figure 4.51(b).

**Dove, Penta, Polygon, Rhomboid, and Wedge Prisms.** These prisms are illustrated in Figure 4.52.

*Dove prisms* are used to rotate the image in optical systems; rotation of the prism at a given angular rate causes the image to rotate at twice this rate. Dove prisms have a length-to-aperture ratio of about five, so they must



**Figure 4.50** (a) Right-angle prism used for  $90^\circ$  beam deflection; (b) right-angle prism used for retroreflection. Incident and reflected ray are parallel only if the incident beam is in the plane of the prism cross-section; however, in this orientation, retroreflection is independent of orientation about the axis A within a large angular range.



**Figure 4.51** (a) Two Porro prisms used as an image-erecting element; (b) Amici prism. (Courtesy of Melles Griot, Inc.)

be used with reasonably well-collimated light. They are available from Anchor Optics, Edmund Optics, Esco Optics, Melles Griot, Newport, Red Optronics, Rolyn, TrustOptics, and Sunny Precision Optics.

*Penta prisms* deviate a ray of light by  $90^\circ$  without inversion (turning upside down) or reversion (right-left reversal) of the image. This  $90^\circ$  deviation applies to all rays incident on the useful aperture of the prism, irrespective of their angle of incidence. Penta prisms do not operate by total internal reflection, so their reflective surfaces are coated. These prisms are useful for  $90^\circ$  deviation when vibrations or other effects prevent their alignment being well controlled. They are available from Anchor Optics, Edmund Optics, Melles Griot, OptoSigma, Red Optronics, Rolyn, Sunny Precision Optics, and TrustOptics.

*Polygon prisms*, usually octagonal, are available from Rolyn and are used for high-speed light-ray deviation, for example, in high-speed rotary prism cameras.

*Rhomboid prisms* are used for lateral deviation of a light ray. When used in imaging applications, there is no change of orientation of the image. Rhomboid prisms are available from Edmund Optics, JML Optical, Precision Glass and Optics, and Rolyn.

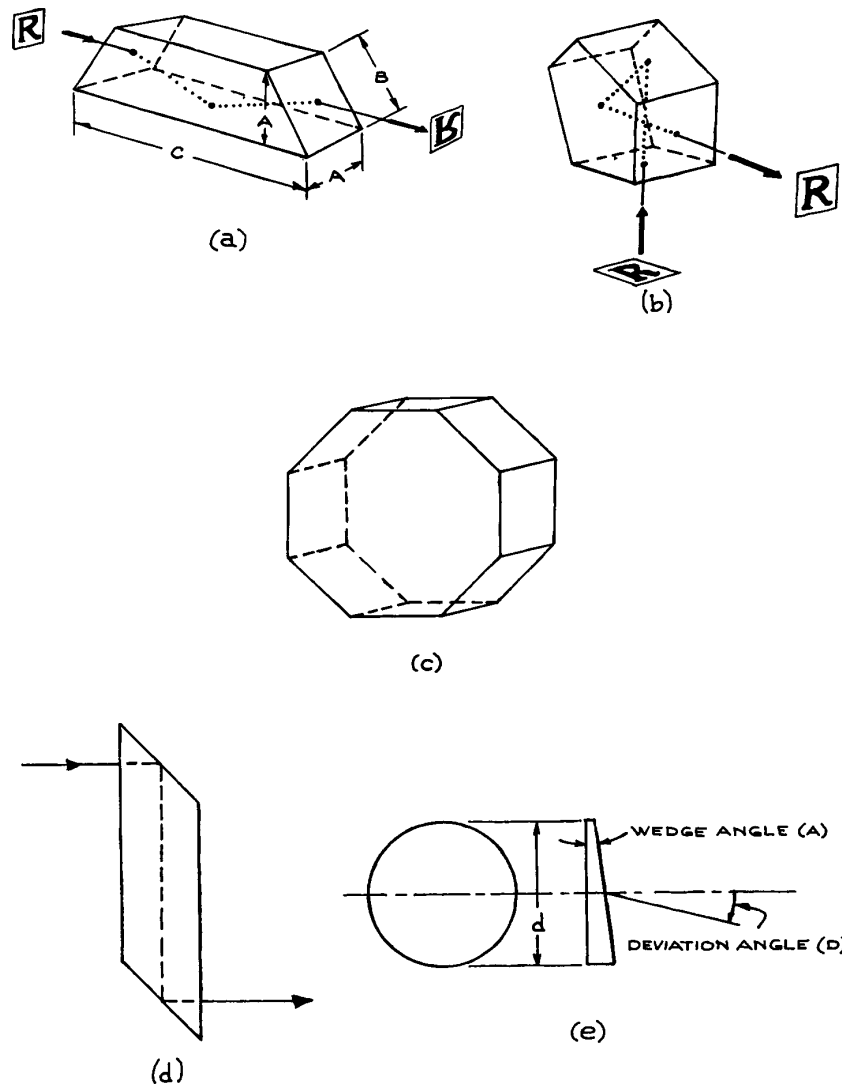
*Wedge prisms* are used in beam steering. The minimum angular deviation  $D$  of a ray passing through a thin wedge prism of apex angle  $A$  is:

$$D = \arcsin(n \sin A) - A \simeq (n - 1)A \quad (4.166)$$

where  $n$  is the refractive index of the prism material. A combination of two identical wedge prisms can be used to steer a light ray in any direction lying within a cone of semivertical angle  $2D$  about the original ray direction. Wedge prisms are available from Anchor Optics, Edmund Optics, EKSPLA, Melles Griot, Precision Glass and Optics, Red Optronics, Rolyn, and Sunny Precision Optics.

**Corner-Cube Prisms (Retroreflectors).** *Corner-cube prisms* are exactly what their name implies, prisms with the shape of a corner of a cube, cut off orthogonal to one of its triad (body-diagonal) axes. The front face of the resultant prism is usually polished into a circle as shown in Figure 4.53. As a result of three total internal reflections, these prisms reflect an incident light ray back parallel to its original direction, no matter what the angle of incidence is. The reflected ray is shifted laterally by an amount that depends on the angle of incidence and the point of entry of the incident ray on the front surface of the prism. These prisms are invaluable in experiments where a light beam

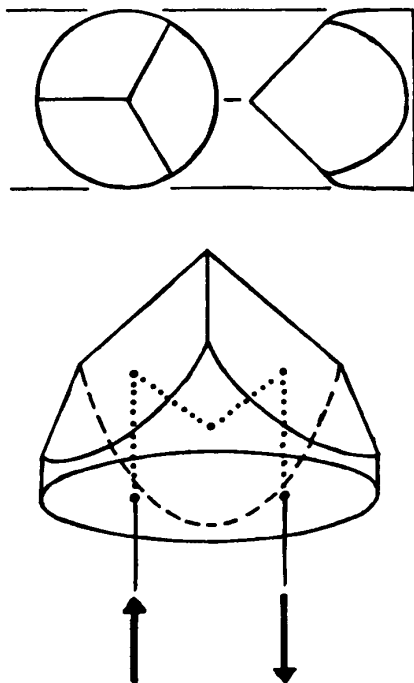




**Figure 4.52** (a) Dove prism; (b) penta prism; (c) octagonal prism; (d) rhomboid prism; (e) wedge prism, [(a), (b), and (e) courtesy of Melles Griot, Inc.]

must be reflected back to its point of origin from some (usually distant) point – for example, in long-path absorption measurements through the atmosphere with a laser beam. They are available commercially from Edmund Optics, Knight Optical, Melles Griot, Newport, Precision Glass and Optics, Precision Optical, Red Optronics, Rolyn, and Sunny Precision Optics. The angular deviation of the

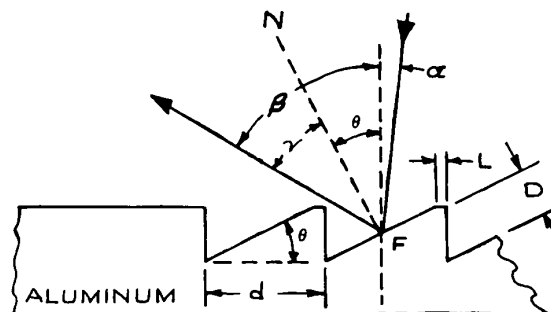
reflected ray from its original direction can be held to less than a half second of arc in the best prisms. For infrared applications beyond the transmission of fused silica, corner-cube prisms are not readily available, but retroreflectors can be made by using three square mirrors butted together to form a hollow cube corner. Precision hollow corner cubes of this type are available from PLX.



**Figure 4.53** Corner-cube prism.  
(Courtesy of Melles Griot, Inc.)

### 4.3.5 Diffraction Gratings

A *diffraction grating* is a planar or curved optical surface covered with straight, parallel, equally spaced grooves. They are manufactured in three principal ways: by direct ruling on a substrate (called a *master grating*), by replication from a master grating, and holographically. Such gratings can be made for use in transmission, but are more commonly used in reflection. Light incident on the grooved face of a reflection grating is diffracted by the grooves; the angular intensity distribution of the diffracted light depends on the wavelength and angle of incidence of the incident light and on the spacing of the grooves. This can be illustrated with reference to the plane grating shown in Figure 4.54. The grooves are usually produced in an aluminum or gold layer that has been evaporated onto a flat optical substrate. For high-intensity laser applications the substrate should have high thermal conductivity. Master gratings ruled on metal are available for this purpose



**Figure 4.54** Blazed groove profile. The thickness of the aluminum is generally equal to the blaze wavelength, and the depth of the groove is about half the thickness. The dimensions shown are approximately those for a grating ruled with 1200 grooves per millimeter, blazed at 750 nm.  $D$  = depth of ruling;  $d$  = groove spacing;  $L$  = unruled land;  $\alpha$  = angle of incidence;  $\beta$  = angle of diffraction;  $\gamma$  = angle of reflection;  $\theta$  = blaze angle. The line  $N$  is normal to the groove face  $F$ . (From D. Richardson, "Diffraction Gratings," in *Applied Optics and Optical Engineering*, Vol. 5, R. Kingslake (Ed.), Academic Press, New York, 1969; by permission of Academic Press).

from Diffraction Products, Jobin-Yvon, and Richardson Grating Laboratory (though Newport). If the light diffracted at angle  $\beta$  from a given groove differs in phase from light diffracted from the adjacent groove by an integral multiple  $m$  of  $2\pi$ , a maximum in diffracted intensity will be observed. This condition can be expressed as:

$$m\lambda = d(\sin \alpha \pm \sin \beta) \quad (4.167)$$

The plus sign applies if the incident and the diffracted ray lie on the same side of the normal to the grating surface;  $m$  is called the order of the diffraction. For  $m = 0$ ,  $\alpha = \beta$  and the grating acts like a mirror. In any other order, the diffraction maxima of different wavelengths lie at different angles. The actual distribution of diffracted intensity among the various orders depends on the profile of the grating grooves. If the grooves are planar, cut at an angle  $\theta$  to the plane of the grating, maximum diffracted intensity into a particular order results if the angle of diffraction and the angle of reflection are the same. In this case:

$$\beta - \theta = \gamma \quad (4.168)$$

$\theta$  is called the *blaze angle*. The wavelength for which the angle of reflection from the groove face and the diffraction angle are the same is called the *blaze wavelength*,  $\lambda_B$ . It is common to use a grating in a Littrow configuration so that the diffracted ray lies in the same direction as the incident ray. In this case:

$$m\lambda = 2d \sin \beta \quad (4.169)$$

The wavelength at which the light reflects normally from the grating groove satisfies:

$$m\lambda_B = 2d \sin \theta \quad (4.170)$$

Thus, for example, a grating could be specified to be blazed at 600 nm in the first order; it would, of course, be simultaneously blazed in the second order for 300 nm, the third order for 200 nm, and so on. Gratings are also available with rectangular shaped (laminar profile) or sinusoidal groove profiles. Laminar profile grooves have low second-order efficiency, which can be advantageous in vacuum ultraviolet applications: second-order rejection filters are not available at such short wavelengths. Sinusoidal gratings will operate over a broad spectral range, but their diffraction efficiency into a particular order is at best 33%. There are many suppliers of both plane and concave gratings, particularly Diffraction Products, Digital Optics, Gentec, Jobin-Yvon, Newport, Optometrics, Spectrogon, Spectrum Scientific, and TVC Jarrell-Ash.

**Resolving Power.** The *resolving power* of a grating is a measure of its ability to separate two closely spaced wavelengths. The resolving power depends both on the dispersion and the size of the grating. For a fixed angle of incidence, the *dispersion* is defined as:

$$\left(\frac{d\beta}{d\lambda}\right)_\alpha = \frac{m}{d \cos \beta} \quad (4.171)$$

Thus, the dispersion can be increased by increasing the number of lines per millimeter ( $1/d$ ), by operating in a high order, and by using a large angle of incidence (grazing incidence). The incidence, however, must not be so flat as to make the projection of the groove width perpendicular to the incoming light smaller than the wavelength of the light.

The resolving power of a grating is  $\Delta\lambda/\lambda$ , where  $\lambda$  and  $\lambda + \Delta\lambda$  are two closely spaced wavelengths that are just resolved by the grating. The limiting resolution depends on the pro-

jected width of the grating perpendicular to the diffracted beam. This width is  $Nd \cos \beta$ , where  $N$  is the total number of lines in the grating. Diffraction theory predicts that the angular resolution of an aperture of this size is  $\Delta\phi \simeq \lambda/(Nd \cos \beta)$ . The angle between the two closely spaced wavelengths, from the dispersion relation, is:

$$\Delta\beta = \frac{m\Delta\lambda}{d \cos \beta} \quad (4.172)$$

In the limit of resolution,  $\Delta\phi = \Delta\beta$ , which gives:

$$\lambda/\Delta\lambda = mN \quad (4.173)$$

Thus, from Equation (4.167):

$$\frac{\lambda}{\Delta\lambda} = \frac{Nd(\sin \alpha \pm \sin \beta)}{\lambda} \quad (4.174)$$

and so it is clear that large gratings used at high angles give the highest resolving power. The resolving powers available from 1 cm wide gratings range up to about  $5 \times 10^5$ .

**Concave gratings.** have long been used in short wavelength applications because they provide diffractive and focusing functions simultaneously. The development of holographic concave gratings has made these devices widely available. They can be used to correct for aberrations which occur when they are used in spectrometers (see Section 4.7), and to allow compact versions of such instruments to be produced.

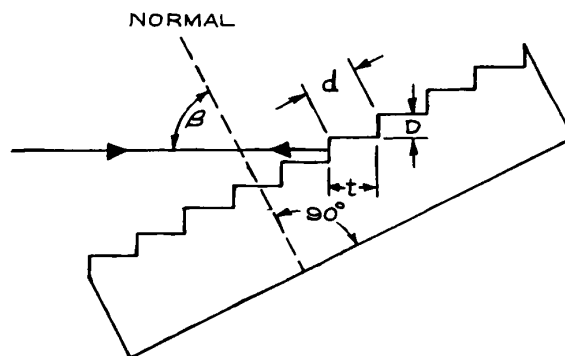
**Efficiency.** The efficiency of a grating is a measure of its ability to diffract a given wavelength into a particular order of the diffraction pattern. Blazing of the grating is the main means for obtaining efficiency in a particular order; without it, the diffracted energy is distributed over many orders. Gratings are available with efficiencies of 95% or more at their blaze wavelength, so they are virtually as efficient as a mirror, yet retain wavelength selectivity.

**Defects in Diffraction Gratings.** In principle, if an ideal grating is illuminated with a plane monochromatic beam of light, diffracted maxima occur only at angles that satisfy the diffraction equation (4.157). In practice, however, this is not so. Gratings manufactured by ruling a metal-coated substrate with a ruling engine exhibit undesirable additional maxima.

Periodic errors in the spacing of the ruled grooves produce *Rowland ghosts*. These are spurious intensity maxima, usually symmetrically placed with respect to an expected maximum and usually lying close to it. The strongest Rowland ghosts from a modern ruled grating will be less than 0.1% of the expected diffraction peaks. *Lyman ghosts* occur at large angular separations from their parent maximum, usually at positions corresponding to a simple fraction of the wavelength of the parent maximum – 4/9 or 5/9, for example. They are also associated with slow periodic errors in the ruling process. Lyman-ghost intensities from modern gratings are exceedingly weak (0.001% of the parent or less). Ghosts can be a problem when used for the detection of weak emissions in the presence of a strong laser signal, the ghosts of which can (and have been) mistakenly identified as other real spectral lines. If there is any suspicion of this, the weak signal should be checked for its degree of correlation with the laser signal – a linear correlation would be strong evidence for a ghost. The development of unruled holographic gratings, which are essentially perfect, plus the improvement of ruled gratings made by interferometrically controlled ruling engines, has considerably reduced the problem of ghosts.

There are other diffraction-grating defects. *Satellites* are misplaced spectral lines, which can be numerous, occurring very close to the parent, usually so close that they can only be discerned under conditions of high resolution: they arise from small local variations in groove spacing. *Scattering* gives rise to an apparent weak continuum over all diffraction angles when a grating is illuminated with an intense monochromatic source such as a laser. Scattering can arise from microscopic dust particles or nongroove-like, random defects on the diffraction-grating surface. In practice it does not follow that a diffraction grating that shows apparent blemishes, such as broad, shaded bands, will perform defectively. One should never attempt to remove such apparent visual blemishes by cleaning or polishing the grating.

**Specialized Diffraction Gratings.** Concave gratings are frequently used in vacuum-ultraviolet spectrometers and spectrographs, as they combine the functions of both dispersing element and focusing optics. Thus, fewer reflective surfaces are required – a highly desirable feature of an instrument used at short wavelengths, where reflectances of all materials decrease markedly.



**Figure 4.55** Echelle grating used in Littrow.

Gratings for use at long wavelengths, above 50  $\mu\text{m}$ , use relatively few grooves per millimeter. Such gratings are generally ruled directly on metal and are frequently called *echelettes* (little ladders) because of the shape of their grooves.

*Echelle* gratings are special gratings designed to give very high dispersion, up to  $10^6$  for ultraviolet wavelengths, when operated in a very high order. They have the surface profile illustrated in Figure 4.55, where the dimension  $D$  is much larger than in a conventional grating, and can range up to several micrometers. Echelles are widely used as the wavelength-tuning element in pulsed dye lasers because of their high dispersion and efficient operation. Other uses of diffraction gratings, such as in the production of Moiré fringes and in interferometers, will not be discussed here. These subjects have been dealt with by Girard and Jacquinet.<sup>44</sup>

**Practical Considerations in Using Diffraction Gratings.** The main considerations in specifying a diffraction grating are the wavelength region where maximum efficiency is required (which will specify the blaze angle), and the number of grooves per millimeter and the size of the grating (which will determine the resolving power). Most gratings are rectangular or circular, the latter being primarily for laser-cavity wavelength selection. Diffraction gratings should be mounted with the care due all precision components. Their surfaces should *never* be touched and should be protected from dust. It is important to note that if a grating is illuminated with a given wavelength, say 300 nm,

then diffraction maxima will occur in the same positions as would be found for 600 and 900 nm. This difficulty can be avoided by using appropriate filters to prevent unwanted wavelengths from passing through the system.

### 4.3.6 Polarizers

**Polarized Light.** If the electric vector of an electromagnetic wave always points in the same direction as the wave propagates through a medium, then the wave is said to be *linearly polarized*. The *direction of linear polarization* is defined as the direction of the electric displacement vector  $\mathbf{D}$ , where:

$$\mathbf{D} = \epsilon_r \epsilon_0 \mathbf{E} \quad (4.175)$$

Except in anisotropic media, where  $\epsilon_r$  is a tensor,  $\mathbf{D}$  and  $\mathbf{E}$  are parallel and the direction of linear polarization can be taken as the direction of  $\mathbf{E}$ . If a combination of two linearly polarized plane waves of the same frequency, but having different phases, magnitudes, and polarization directions, is propagating in the  $z$ -direction, the resultant light is said to be *elliptically polarized*. Such a pair of waves, in general, have resultant electric fields in the  $x$ - and  $y$ -directions that can be written as:

$$E_x = E_1 \cos \omega t \quad (4.176)$$

$$E_y = E_2 \cos(\omega t + \phi) \quad (4.177)$$

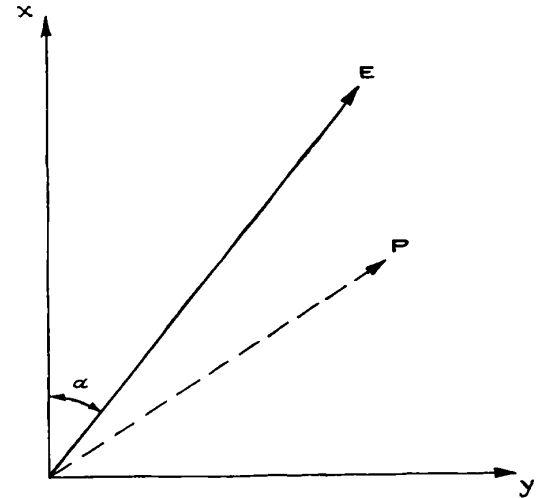
where  $\phi$  is the phase difference between these two resultant field components. These are the parametric equations of an ellipse. If  $\phi = \pm\pi/2$  and  $E_1 = E_2 = E_0$ , then:

$$E_x^2 + E_y^2 = E_0^2 \quad (4.178)$$

which is the equation of a circle. This represents *circularly polarized* light. Then the instantaneous angle that the total electric field vector makes with the  $x$ -axis, as illustrated in Figure 4.56, is:

$$\alpha = \arctan \frac{E_y}{E_x} = \arctan \mp \tan \omega t = \mp \omega t \quad (4.179)$$

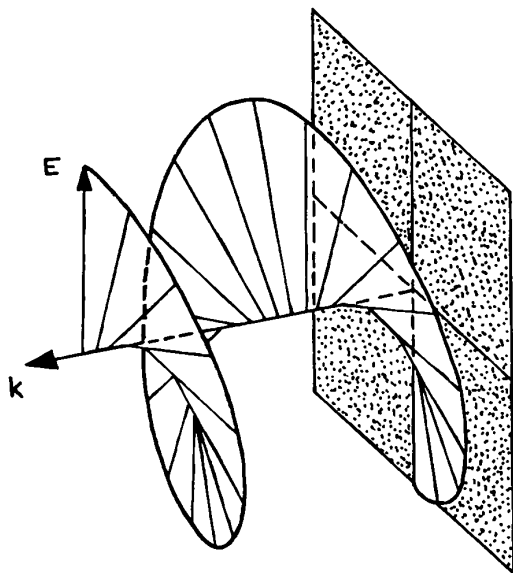
For  $\phi = \pi/2$ , the resultant electric vector rotates counter-clockwise viewed in the direction of propagation – this is *right-hand circularly polarized light*. If  $\phi = -\pi/2$  the rotation is clockwise – this is *left-hand circularly polarized*



**Figure 4.56** Instantaneous direction of the electric vector of an electromagnetic wave.

*light*. The motion of the total electric vector as it propagates is illustrated in Figure 4.57. If  $\phi = 0$  or  $\pi$  we have *linearly polarized light*. Just as circularly polarized light can be viewed as a superposition of two linearly polarized waves with orthogonal polarizations, a linearly polarized wave can be regarded as a superposition of left- and right-hand circularly polarized waves. If an electromagnetic wave consists of a superposition of many independent linearly polarized waves of independent phase, amplitude, and polarization direction, it is said to be *unpolarized*.

In anisotropic media – media with lower than cubic symmetry – there is at least one direction, and at most two directions, along which light can propagate with no change in its state of polarization, independent of its state of polarization. This direction is called the *optic axis*. *Uniaxial crystals* have one such axis; *biaxial crystals* have two. When a wave does not propagate along the optic axis in such crystals, it is split into two polarized components with orthogonal linear polarizations. In uniaxial crystals these two components are called the *ordinary* and *extraordinary* waves. They travel with different phase velocities, characterized by two different refractive indices,  $n_o$  and  $n_e(\theta)$ , respectively, where  $\theta$  is the angle between the  $\mathbf{k}$  vector of the wave and the optic axis. This phenomenon is referred to as *birefringence*.



**Figure 4.57** Left-hand circularly polarized light. (From E. Wahlstrom, *Optical Crystallography*, 3rd edn., John Wiley & Sons, Inc., New York, 1960; by permission of John Wiley & Sons, Inc.)

**D** and **E** are not necessarily parallel in an anisotropic medium; however, **D** and **H** are orthogonal to the wave vector of any propagating wave. Consequently the Poynting vector of the wave does not, in general, lie in the same direction as the wave vector. The direction of the Poynting vector is the direction of energy flow – the ray direction. When a plane wave crosses the boundary between an isotropic and an anisotropic medium, the path of the ray will not, in general, satisfy Snell's law. The angles of refraction of the ordinary and extraordinary rays will be different. This phenomenon is called *double refraction*. In uniaxial crystals, however, the ordinary ray direction at a boundary does satisfy Snell's law. For further details of these and other optical characteristics of anisotropic media, the reader should consult Born and Wolf,<sup>11</sup> Wahlstrom,<sup>12</sup> and Davis.<sup>15</sup>

If linearly polarized light propagates into a birefringent material, unless its polarization direction matches the allowed direction of the ordinary or extraordinary wave, it will be split into two components polarized in these two

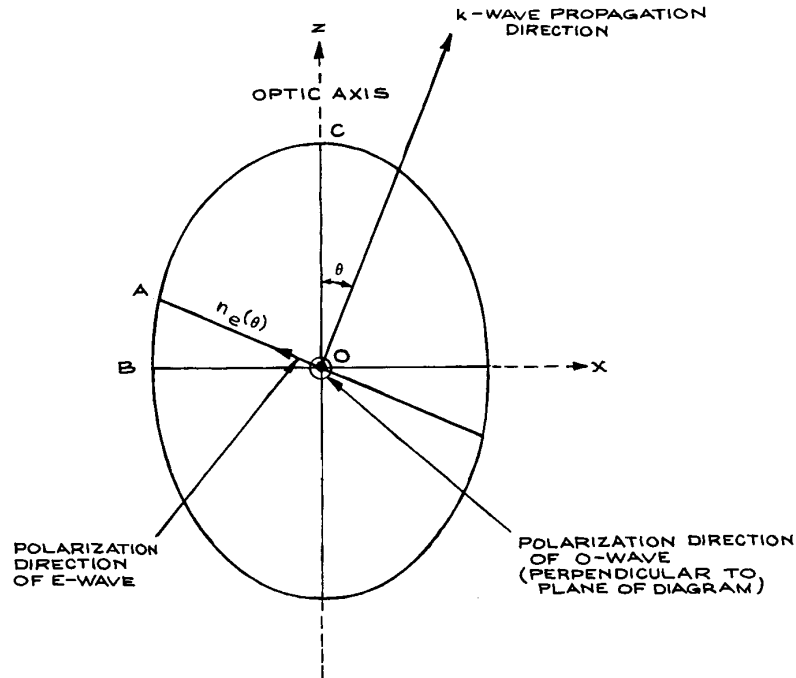
allowed directions. These two components propagate at different velocities and experience different phase changes on passing through the crystal. For light of free-space wavelength  $\lambda_0$  passing through a uniaxial crystal of length  $L$ , the birefringent phase shift is:

$$\Delta\phi = \frac{2\pi L}{\lambda_0} [n_e(\theta) - n_o] \quad (4.180)$$

In uniaxial crystals the allowed polarization directions are perpendicular to the optic axis (for the ordinary wave) and in the plane containing the propagation direction and the optic axis (for the extraordinary wave), as shown in Figure 4.58. If the input wave is polarized at an angle  $\beta$  to the ordinary polarization direction and  $\Delta\phi = (2n + 1)\pi$ , the output wave will remain linearly polarized, but its direction of polarization will have been rotated by an angle of  $2\beta$ . This rotation is always toward the ordinary polarization direction, so a round-trip pass through the material does not affect the polarization state. It is usual to make  $\beta = 45^\circ$ , in which case the crystal rotates the plane of polarization by  $90^\circ$ . Such a device is called a  $(2n + 1)$ th-order *half-wave plate*. On the other hand, if  $\beta = 45^\circ$  and  $\Delta\phi = (2n + 1)\pi/2$ , the ordinary and extraordinary waves recombine to form circularly polarized light. Such a device is called a  $(2n + 1)$ th-order *quarter-wave plate*. If  $\beta$  is not  $45^\circ$ , a quarter-wave plate will convert linearly polarized light into elliptically polarized light.

**Polarization Changes on Reflection.** As shown in Section 4.2.4, if a linearly polarized wave reflects from a dielectric surface (or mirror), its state of linear polarization may be changed. For example, if a vertically polarized wave traveling horizontally striking a mirror at  $45^\circ$  is polarized in the plane of incidence, then after reflection it will be traveling vertically and will be horizontally polarized. A reflection off a second mirror at  $45^\circ$ , so oriented that the plane of incidence is perpendicular to the polarization, will produce a horizontally traveling, horizontally polarized wave. This exemplifies how successive reflections can be used to rotate the plane of polarization of a linearly polarized wave, and is illustrated in Figure 4.59.

If circular or elliptically polarized light reflects from a dielectric surface or mirror, its polarization state will, in general, be changed. As an example, consider reflection at a metal mirror. From Equations (4.87) and (4.89), for the



**Figure 4.58** Cross-section of an ellipsoidal figure called the indicatrix, which allows determination of the refractive indices and permitted polarization directions of a wave traveling in a uniaxial crystal. The equation of the ellipse shown is  $(x^2/n_o^2) + (z^2/n_e^2) = 1$ , where  $OB = n_o$  and  $OC = n_e$ ,  $OA = n(\theta)$  is the effective extraordinary refractive index for a wave traveling at angle  $\theta$  to the optic axis. The ordinary refractive index for this wave is still  $n_o$  and is independent of  $\theta$ .

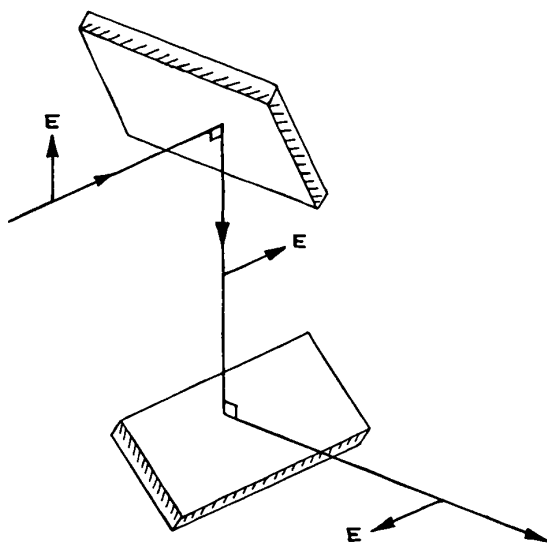
in-plane and perpendicular polarized components of the incident wave, respectively, we have:

$$\begin{aligned} \rho_{\parallel} &= -1 \\ \rho_{\perp} &= -1 \end{aligned} \quad (4.181)$$

These reflection coefficients ensure that the tangential component of the electric field goes to zero at the conducting surface of the mirror. As illustrated in Figure 4.60(a) and (b), this leads to a conversion of left- to right-hand circularly polarized light on reflection. The change of polarization state on reflection from a dielectric interface depends on the angle of incidence and whether  $n_2 > n_1$ . A special case, illustrated in Figure 4.60(c), shows left-hand circularly polarized light being converted to linearly

polarized light on reflection from an interface placed at Brewster's angle. Such changes of polarization state must be taken into account when designing an optical system to work with polarized light. The change in polarization of a beam of light as it passes through a series of optical components can be calculated very conveniently by the use of Jones or Mueller calculus, which use matrix methods to describe the state of polarization of the beam and its interaction with each component. Details of these techniques are given by Shurcliff.<sup>13</sup>

**Linear Polarizers.** A *linear polarizer* changes unpolarized light to linearly polarized light or changes polarized light to a desired linear polarization.

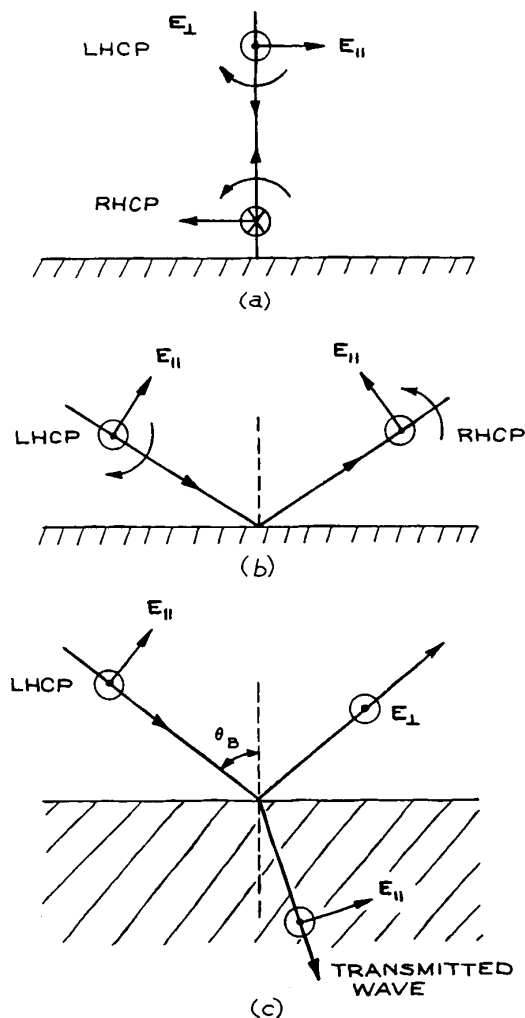


**Figure 4.59** Rotation of the plane of polarization of a linearly polarized beam by successive reflection at two mirrors.

The simplest linear polarizers are made from dichroic materials – materials that transmit one polarization, either ordinary or extraordinary, and strongly absorb the other. Modern dichroic linear polarizers, the commonest of which is Polaroid, are made of polymer films in which long-chain molecules with appropriate absorbing side groups are oriented by stretching. The stretched film is then sandwiched between glass or plastic sheets. These polarizers are inexpensive, but cannot be used to transmit high intensities because they absorb all polarizations to some degree. They cannot be fabricated to very high optical quality, and generally only transmit about 50% of light already linearly polarized for maximum transmission. They do, however, work when the incident light strikes them at any angle up to grazing incidence. The extinction coefficient  $K$  that can be achieved with crossed polarizers is defined as:

$$K = \log_{10} \frac{T_0}{T_{90}} \quad (4.182)$$

where  $T_0$  and  $T_{90}$  are the transmittances of parallel and crossed polarizers, respectively.  $K$  is a useful quantity for specifying a linear polarizer – the larger its value, the



**Figure 4.60** Changes of polarization state on reflection: (a) and (b) conversion of left-hand circularly polarized light (LHCP) to right-hand circularly polarized light (RHCP) on reflection at a perfect metal mirror; (c) conversion of LHCP to linear polarization on reflection at Brewster's angle.

better the polarizer. For dichroic polarizers, values of  $K$  up to about  $10^4$  can be obtained. Polarizing beamsplitting cubes provide linearly polarized light of relatively high purity (98% or better) and can handle higher light intensities. Corning manufacture a sheet polarizer called Polaroid made from glass containing billions of tiny silver



crystals. This material provides high throughput in the red to near-infrared region, and extinction up to  $10^4$ .

Higher quality, but more expensive, linear polarizers can be made by using the phenomenon of double refraction in transparent birefringent materials such as calcite, crystalline quartz, or magnesium fluoride. Figure 4.61 shows schematically the way in which various polarizing prisms of this type operate. Some of these polarizers generate two orthogonally polarized output beams separated by an angle, while others, for example the Glan–Taylor, reject one polarization state by total internal reflection at a boundary. Extinction coefficients as high as  $10^6$  and good optical quality can be obtained from such polarizers. The acceptance angle of these polarizers is generally quite small, although it can range up to about  $38^\circ$  with an Ahrens polarizer. Suppliers of polarizing prisms include Argyle International, Gooch and Housego, Karl Lambrecht Corporation, II-VI Infrared, Inrad, Lambda Research Optics, Meadowlark Optics, Melles Griot, Newport/Oriel, OptoSigma, and Precision Optical.

Infrared polarizers (beyond about  $7\ \mu\text{m}$ ) are often made in the form of very many fine, parallel, closely spaced metal wires. Waves with their electric vector perpendicular to the wires are transmitted; the parallel polarization is reflected. Such polarizers are available from Coherent/Molelectron, Specac, and Thorlabs.

Linear polarizers for any wavelength where a transparent window material is available can be made using a stack of plates placed at Brewster's angle. Unpolarized light passing through one such plate becomes slightly polarized, since the component of incident light polarized in the plane of incidence is completely transmitted, whereas only about 90% of the energy associated with the orthogonal polarization is transmitted. A pile of 25 plates gives a high degree of polarization. Stacked-plate polarizers are rather cumbersome, but they are most useful for polarizing high-energy laser beams where prism polarizers would suffer optical damage. They are available commercially from Inrad and II-VI Infrared (using ZnSe). For visible and near-infrared light they can be easily made using a stack of microscope slides.

**Retardation Plates.** *Quarter-wave plates* are generally used for converting linear to circularly polarized light. They should be used with the optic-axis direction at  $45^\circ$

to the incident linear polarization direction. They are available commercially in two forms: multiple- and single-order. Multiple-order plates produce a phase change (retardation) between the ordinary and extraordinary waves of  $(2n + 1)\pi/2$ . This phase difference is very temperature-sensitive – a thickness change of one wavelength will produce a substantial change in retardation. Single order plates are very thin, their thickness being:

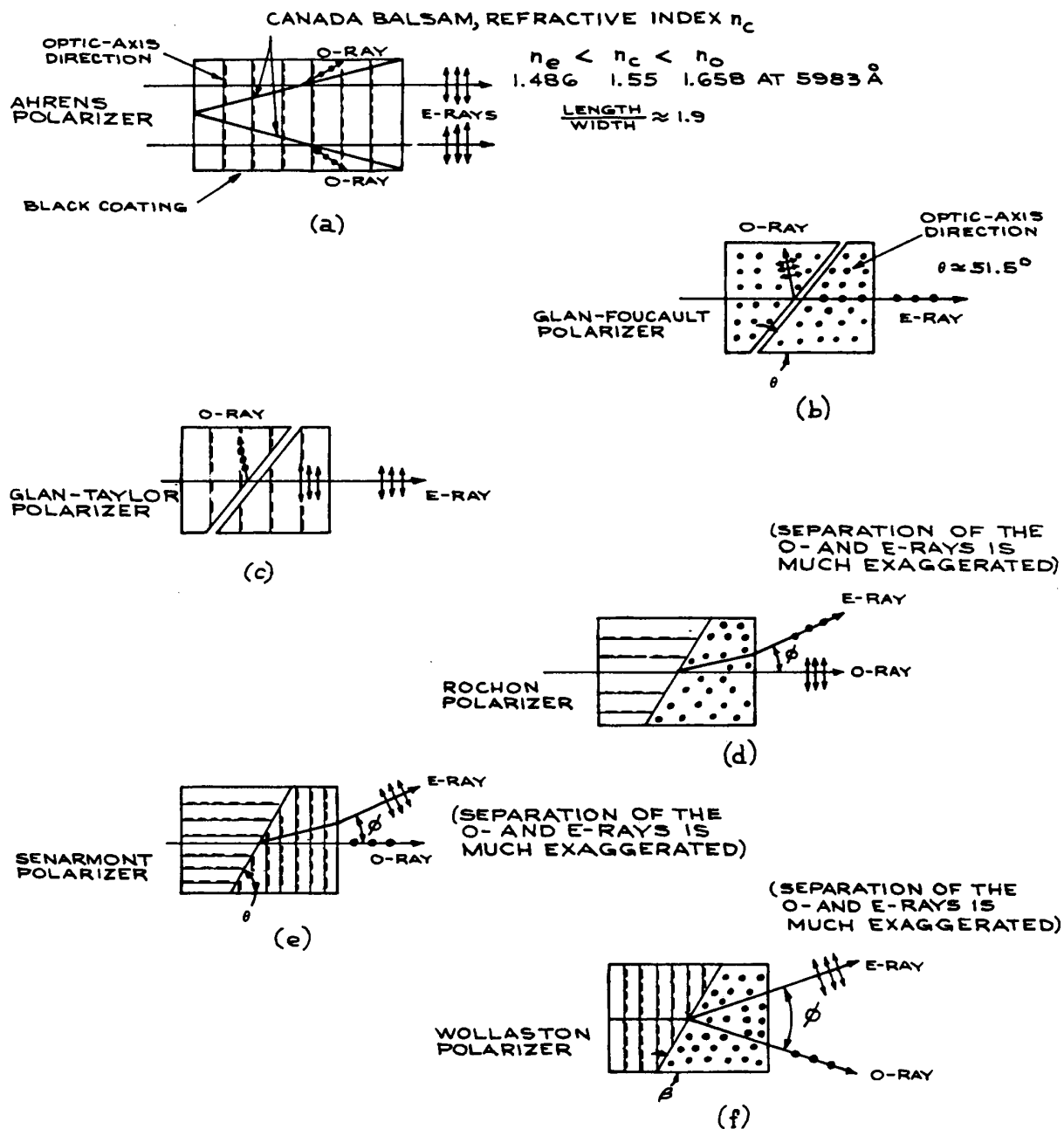
$$L = \frac{\lambda_0}{4(n_e - n_o)} \quad (4.183)$$

For example, with calcite, which has  $n_o = 1.658$ ,  $n_e = 1.486$ , and with  $\lambda_0 = 500\ \text{nm}$  this thickness is only  $726.7\ \text{nm}$ . Consequently, most single-order plates are made by stacking a  $(2n + 1)$ th-order quarter-wave plate on top of a  $2n$ th-order plate whose optic axis is orthogonal to that of the first plate. The net retardation is just  $\pi/2$  and is very much less temperature-sensitive.

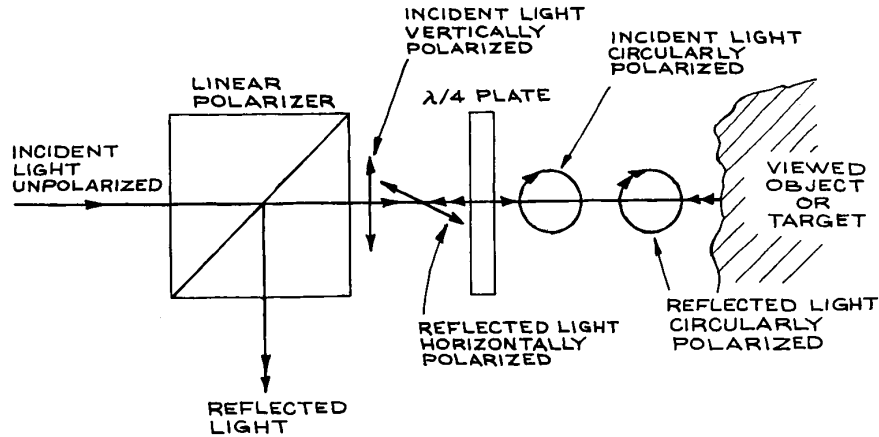
Direct single-order quarter-wave plates for low-intensity applications can be made from mica, which cleaves naturally in thin slices. The production of a plate for a particular wavelength is a trial-and-error procedure.<sup>12,45</sup> Quarter-wave plates are invaluable, in combination with a linear polarizer, for reducing reflected glare and for reducing back reflections into an optical system, as shown in Figure 4.62.

*Half-wave plates* are generally used for rotating the plane of polarization of linearly polarized radiation. Multiple- and zero-order plates are available; as in the case of quarter-wave plates, the operating wavelength must be specified in buying one. This operating wavelength can be tuned to some extent by tilting the retardation plate. Half-wave plates cannot be used to make an optical isolator, as the polarization change that occurs on one pass through the plate is reversed on the return path.

Retardation plates only produce their specified retardation at a particular wavelength, and are usually specified for normal, or almost normal incidence. Retardation plates are available from several suppliers, including Karl Lambrecht Corporation, Optics for Research, Inrad, II-VI Infrared, Esco, Melles Griot, Newport, OptoSigma and Thorlabs. Continuously adjustable retardation plates, called Babinet-Soleil compensators, are available from Karl Lambrecht Corporation, Melles Griot, II-VI Infrared, New Focus, Continental Optical, and OFR (now Thorlabs), among others.<sup>31</sup>



**Figure 4.61** Construction of various polarizers using birefringent crystals: (a) Ahrens polarizer – three calcite prisms cemented together with Canada balsam, whose refractive index is intermediate between  $n_o$  and  $n_e$  for calcite. The O-ray suffers total internal reflection at the cement and is absorbed in the black coating on the sides; (b) Glan-Foucault polarizer – two calcite prisms separated with an air gap. The O-ray is totally internally reflected and either absorbed in the sides or transmitted



**Figure 4.62** An optical isolator constructed from a linear polarizer and a quarter-wave plate.

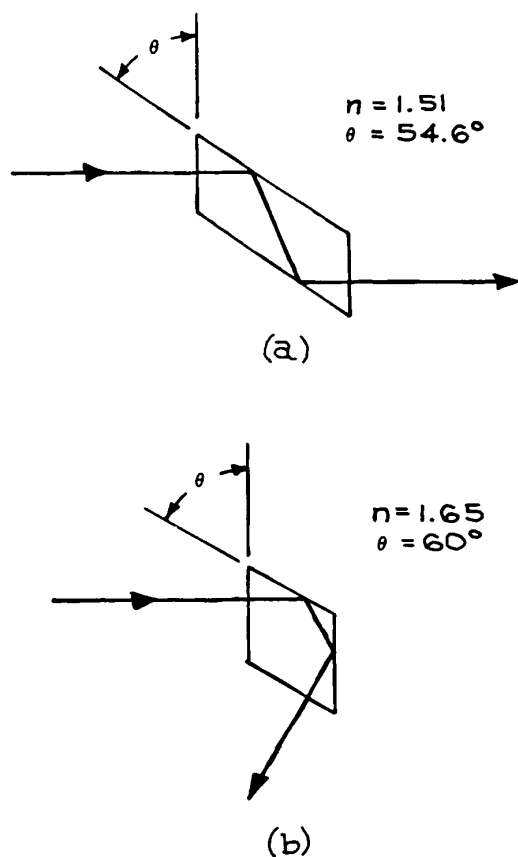
**Retardation Rhombs.** Because, in general, there is a different phase shift on total internal reflection for waves polarized in and perpendicular to the plane of incidence, a retarder that is virtually achromatic can be made by the use of two internal reflections in a rhomboidal prism of appropriate apex angle and refractive index. Two examples are shown in Figure 4.63. The Fresnel rhomb, for example, uses glass of index 1.51, and with an apex angle of approximately  $54.6^\circ$  produces a retardation of exactly  $90^\circ$ . The optimum angle must, however, be determined by trial and error, and the desired retardation will only be obtained for a specific angle of incidence. Fresnel rhombs are available from II-VI Infrared, Karl Lambrecht (KLC), and Newport/Oriel.

### 4.3.7 Optical Isolators

*Optical isolators* are devices that allow light to pass through in one direction, but not in the reverse direction. These devices contain two linear polarizers and between them a medium (generally a special glass), which exhibits the Faraday effect. The medium is placed in an axial magnetic field provided by a permanent magnet or magnets. A Faraday-active medium rotates the plane of an input linearly polarized wave, and the direction of rotation depends on the relative direction of light propagation and the magnetic field direction. In an isolator, the optical path length through the Faraday material and the magnetic field strength are selected to provide a  $45^\circ$  rotation of the input linear polarization, which then lines up with the exit

**Caption for Figure 4.61 (contd.)** if the sides are polished. A Glan–Thompson polarizer is similar; (c) Glan–Taylor polarizer – similar to a Glan–Foucault polarizer except for the optic-axis orientation, which ensures that the transmitted E-ray passes through the air gap nearly at Brewster’s angle; consequently, transmission losses are much reduced from those of a Glan–Foucault. For high-power laser applications, the side faces can be polished at Brewster’s angle; (d) Rochon polarizer – two calcite prisms with orthogonal optic axes cemented together with Canada balsam.  $\phi$  depends on the interface angle. The intensities of the transmitted O- and E-rays are different. Other birefringent materials can be used; in the case of quartz, for example, the E-ray would exit below the O-ray in the diagram shown here; (e) Senarmont polarizer – two calcite prisms cemented together with Canada balsam. The angle  $\phi$  depends on  $\theta$ . This type of polarizer is less commonly used than the Rochon polarizer; (f) Wollaston polarizer – two calcite prisms cemented together with Canada balsam. The O- and E-ray transmitted intensities are different. The beam separation angle  $\phi$  depends on the interface angle  $\theta$ .

polarizer set at  $45^\circ$  to the input polarizer, as shown in Figure 4.64. Light passing in the reverse direction has its plane of linear polarization also rotated by  $45^\circ$ , but in a sense that makes it orthogonal to the next polarizer so



**Figure 4.63** Rhomb retarders: (a) Fresnel; (b) Mooney.

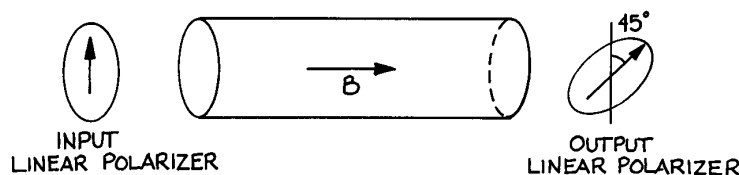
negligible light emerges. Commercially available Faraday isolators typically provide 30 dB isolation and must be selected for the specific wavelength desired. Both free space and optical-fiber-based devices are available from companies such as AC Photonics, Conoptics, Electro-Optics Technology, Isowave, Leysop, Linos Photonics, Namiki, New Focus, Newport, and OFR (now Thorlabs).

Optical isolators are particularly useful in certain precision experiments using lasers, to avoid *feedback instability*. If some of the light from a laser re-enters the laser, the amplitude and phase of the laser will fluctuate, and this can sometimes be very marked.

### 4.3.8 Filters

*Filters* are used to select a particular wavelength region from light containing a broader range of wavelengths than is desired. Thus, for example, they allow red light to be obtained from a white light source, or they allow the isolation of a particular sharp line from a lamp or laser in the presence of other sharp lines or continuum emission. Although filters do not have high resolving power (see Section 4.3.5), they have much higher optical efficiency at their operating wavelength than higher-resolving-power, wavelength-selective instruments, such as prism or grating monochromators. That is, they have a high ratio of transmitted flux to input flux in the wavelength region desired. Filters of several types are available for different applications.

**Color Filters.** A *color filter* selectively transmits a particular spectral region while absorbing or reflecting others. These simple filters are glasses or plastics containing absorbing materials such as metal ions or dyes, which have characteristic transmission spectra. Such color filters are available from Chance-Pilkington,



**Figure 4.64** Optical isolator that uses two linear polarizers and a Faraday material producing a  $45^\circ$  polarization rotation.

Corning, CVI, Hoya, Kodak (Wratten filters), Kopp Glass, Melles-Griot (now CVI-Melles-Griot), Newport, Newport Thin Film Laboratory, Omega Optical, Praezisions Glas & Optik, Schott, Rolyn, Rosco, and Sterling Precision Optical. Transmission curves for these filters are available from the manufacturers, and in tabulations of physical and chemical data.<sup>7,46</sup>

Care must be taken when color filters are being used to block a strong light signal, such as a laser. These filters may fluoresce, so although the primary incident light is blocked, a weak, longer-wavelength fluorescence can occur. We have seen this phenomenon quite strongly when using orange plastic to block a blue/green laser.

**Band-Pass Filters.** A *band-pass filter* is usually a *Fabry–Perot etalon* of small thickness. Although etalons will be discussed in further detail in Section 4.7.4, a brief discussion is in order here.

In its simplest form, the Fabry–Perot etalon consists of a plane, parallel-sided slab of optical material of refractive index  $n$  and thickness  $d$  with air on both sides. In normal incidence the device has maximum transmission for wavelengths for which the thickness of the device is an integral number of half wavelengths. In this case, from Equation (4.110):

$$Z''_3 = Z'_3 = Z'_1 \quad (4.184)$$

and there is zero reflection. For incidence at angle  $\theta$ , regardless of the polarization state of the light, there is maximum transmission for wavelengths (*in vacuo*) that satisfy:

$$m\lambda_0 = 2d \cos \theta_2 \quad (4.185)$$

where  $\theta_2$  is the angle of refraction in the etalon.

A Fabry–Perot etalon is a “comb” filter. (Figure 4.159 shows an example of a transmission characteristic of such a device.) If the thickness of the etalon is small, the transmission peaks are broad. In a typical band-pass filter operated as an etalon, all the transmission peaks but the desired one can be suppressed by additional absorbing or multilayer reflective layers. The spectral width of the transmission peak of the filter can be reduced by stacking individual etalon filters in series. A single etalon filter might consist, for example, of two multilayer reflective stacks separated by a half-wavelength-thick layer, as

shown in Figure 4.65(a), while a two-etalon filter would have two half-wavelength-thick layers bounded by multilayer reflective stacks. Commercially available band-pass filters for use in the visible usually have a passband ranging from below 1 nm to 50 nm (full width at half maximum transmission). The narrowest band-pass filters are frequently called “spike filters.” Filters to transmit particular wavelengths, such as 632.8 nm (He-Ne laser), 514.5 and 488 nm (argon ion laser), 404.7, 435.8, 546.1, 577.0, and 671.6 nm (mercury lamp), 589.3 nm (sodium lamp), and other wavelengths, are available as standard items.<sup>31</sup> Filters at nonstandard wavelengths can be fabricated as custom items. Bandpass filters for use in the ultraviolet usually have lower transmittance, typically 20%, than filters in the visible, whose transmittance usually ranges from 40% to 70%, being lowest for the narrowest passband. Infrared band-pass filters are also readily obtained, at least out to about 10  $\mu\text{m}$ , and have good peak transmittance. The typical available band-pass for a filter whose peak transmission is  $\lambda_{\text{peak}}$  can usually be estimated as  $\lambda_{\text{peak}}/50$ , although narrower band-pass filters can be obtained at higher cost. Band-pass filters are available from many suppliers, including Andover Optical, LOT-Oriel, Melles Griot, Newport, Omega Optical, Spectrogon, Spectrum Thin Films, and Sterling Precision Optical.

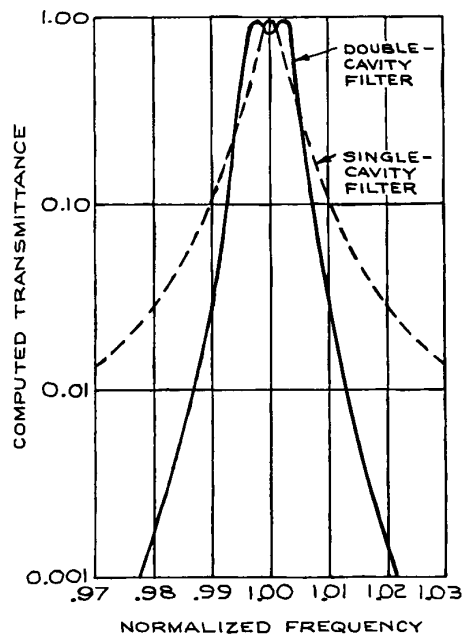
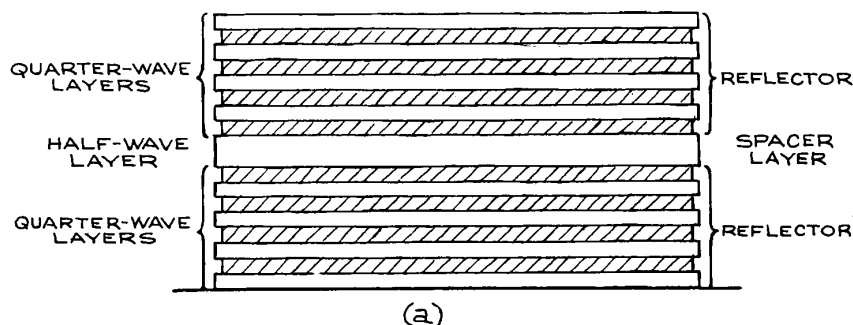
Band-pass filters are usually designed for use in normal incidence. Their peak of transmission can be moved to shorter wavelengths by tilting the filter, although the sharpness of the transmission peak will be degraded by this procedure. If a filter designed for peak wavelength  $\lambda_0$  is tilted by an angle  $\theta$ , its transmission maximum will shift to a wavelength:

$$\lambda_s = \lambda_0 \sqrt{1 - \frac{\sin^2 \theta}{n^2}} \quad (4.186)$$

where  $n$  is the refractive index of the half-wavelength-thick layer of the filter. For small tilt angles:

$$\lambda_0 - \lambda_s = \frac{\lambda_0 \sin^2 \theta}{2n^2} \quad (4.187)$$

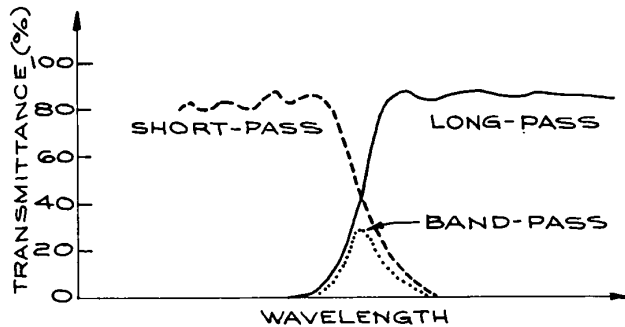
Thus, a band-pass filter whose peak-transmission wavelength is somewhat longer than desired can be optimized by tilting.



**Figure 4.65** (a) Construction of an all-dielectric single-cavity Fabry-Perot interference filter; (b) theoretical transmittance characteristics of single- and double-cavity filters. (From *Handbook of Lasers*, R. J. Pressley (Ed.), CRC Press, Cleveland, 1971; by permission of CRC Press, Inc.)

**Long- and Short-Wavelength-Pass Filters.** A *long-wavelength-pass filter* is one that transmits a broad spectral region beyond a particular cutoff wavelength,  $\lambda_{\min}$ . A *short-wavelength-pass filter* transmits in a broad spectral region below its cutoff wavelength  $\lambda_{\max}$ . A com-

bination of a long-wavelength-pass and a short-wavelength-pass filter can be used to produce a band-pass filter, as shown in Figure 4.66. Although long- and short-wavelength-pass filters usually involve multilayer dielectric stacks, the inherent absorption and transmission characteristics of materials can be utilized. Semiconductors exhibit



**Figure 4.66** Schematic transmission of long- and short-pass optical filters showing the band-pass transmission characteristics of the two in combination.

fairly sharp long-wavelength-pass behavior beginning at the band gap energy. For example, germanium is a long-pass filter with  $\lambda_{\min} = 1.8 \mu\text{m}$ .

**Holographic Notch Filters.** These filters are special diffractive structures that reflect a narrow range of wavelengths ( $\sim 5 \text{ nm}$ ) very strongly, and consequently they reduce transmitted light substantially (40–60 dB) over the same narrow range. These filters are ideal for rejection of laser light in experiments where a weaker light signal – fluorescence or Raman scattering – is being studied, and some scattering or reflection of the excitation laser light is occurring. These filters are available from Andor Technology, Del Mar Ventures, Kaiser Optical Systems, MK Photonics, and Semrock.

**Rugate Filters.** Rugate filters are notch interference filters with high rejection of a narrow band of wavelengths. In contrast to conventional multilayer dielectric interference filters, in which the refractive index variation from layer to layer is a square wave, *rugate filters* use a sinusoidally variable refractive index throughout the oxide film layers. They can provide a single notch band without harmonics, as shown in Figure 4.67. A rugate filter can look almost completely colorless and transparent, and yet reject laser light at a specified wavelength very efficiently. They are available from Advanced Technology Coatings, Barr Associates, Edmund Optics, Gist Optics, and Rugate Technologies.

**Reststrahlen Filters.** When ionic crystals are irradiated in the infrared, they reflect strongly when their absorption coefficient is high, and their refractive index changes sharply. This high reflection results from resonance between the applied infrared frequency and the natural vibrational frequency of ions in the crystal lattice. The characteristic reflected light from different crystals, termed *Reststrahlen*, allows specific broad spectral regions to be isolated by reflection from the appropriate crystal. Some examples of Reststrahlen filters are given in Figure 4.68.

**Christiansen Filters.** In the far infrared, the alkali halides exhibit anomalous dispersion: at certain wavelengths their refractive indices pass through unity. (Normal dispersion involves an increase of refractive index with wavelength.) Thus a powdered alkali halide, which would generally attenuate a transmitted wave severely because of scattering, will transmit well at wavelengths where its refractive index is the same as that of air. Filters using such an alkali halide powder held between parallel plates are called Christiansen filters and provide sharp transmission peaks at certain wavelengths listed in Table 4.4.

**Neutral-Density Filters.** *Neutral-density (ND) filters* are designed to attenuate light uniformly over some broad spectral region. The ideal neutral-density filter should have a transmittance that is independent of wavelength. These filters are usually made by depositing a thin metal layer on a transparent substrate. The layer is kept sufficiently thin that some light is transmitted through it. The optical density  $D$  of a neutral-density filter is defined as:

$$D = \log_{10}(1/T) \quad (4.188)$$

where  $T$  is the transmittance of the filter.  $T$  is controlled both by reflection from and by absorption in the metal film. If such filters are stacked in series, the optical density of the combination is the sum of the optical densities of the individual filters, provided the filters are positioned so that multiple reflection effects do not occur between them in the direction of interest.

Neutral-density filters are used for calibrating optical detectors and for attenuating strong light signals falling

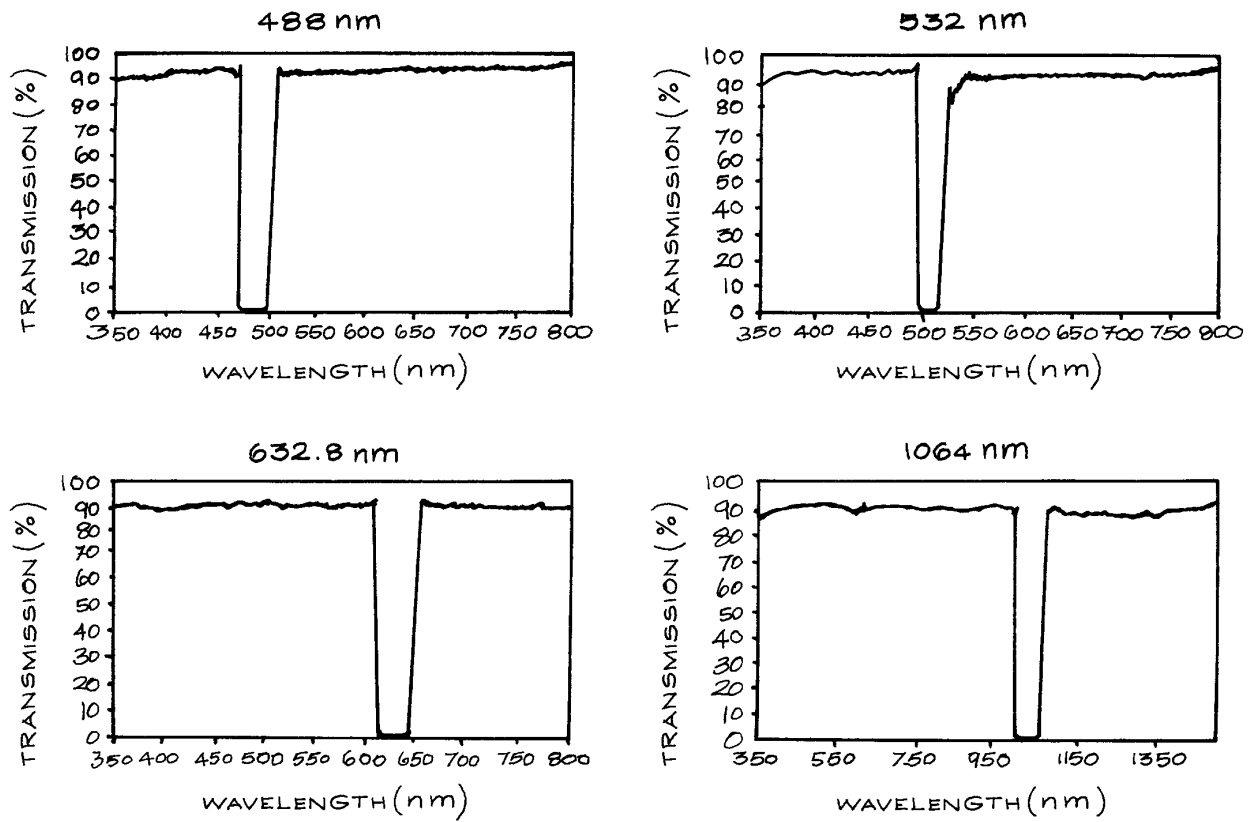


Figure 4.67 Transmission characteristics of some rugate filters.

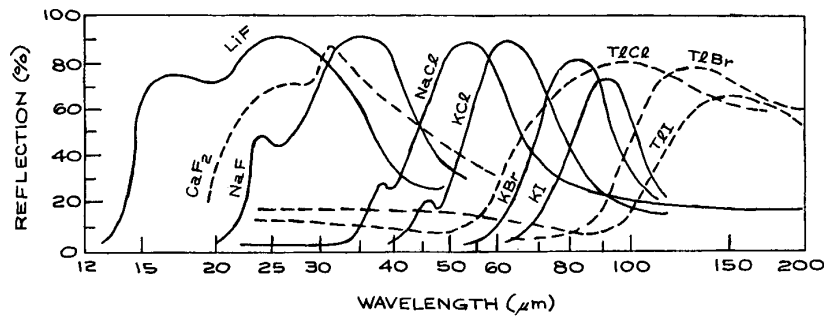


Figure 4.68 Reststrahlen filters. (Courtesy of Harshaw Chemical Co.)



**Table 4.4 Christiansen filters**

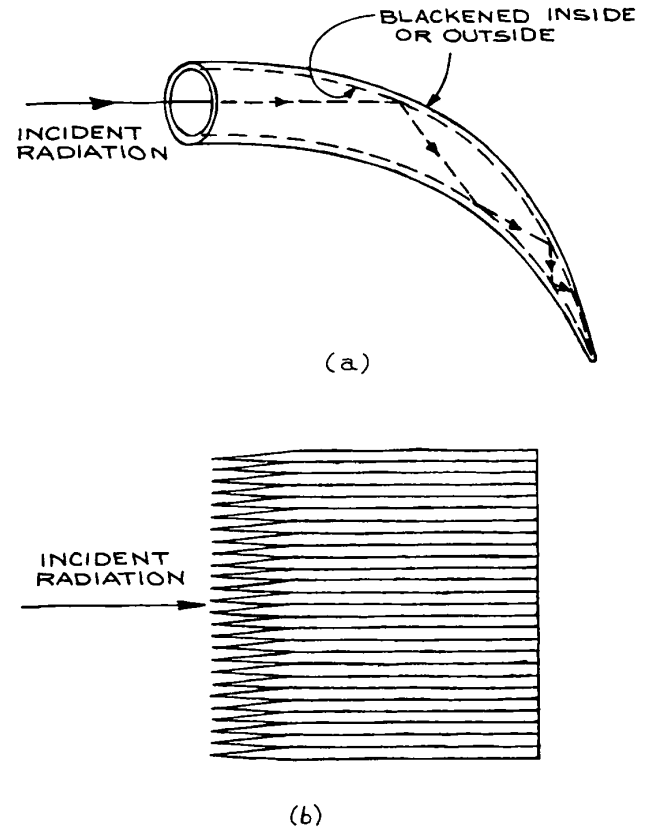
<i>Crystal</i>	<i>Wavelength of Maximum Transmission (<math>\mu\text{m}</math>)</i>
LiF	11.2
NaCl	32
NaBr	37
KCl	37
RbCl	45
NaI	49
CsCl	50
KBr	52
CsBr	60
TlBr	64
KI	64
RbBr	65
RbI	73
TlI	90

on detectors to ensure that they respond linearly. Because neutral-density filters usually reflect and transmit light, they can be used as beamsplitters and beamcombiners for any desired intensity ratio. There are many suppliers.<sup>31</sup> Variable ND filters are available from HB Optical, Melles Griot, Nova Phase, Reynard, and Singh-Ray.

**Light Traps.** If a beam of light must be totally absorbed (for example, in an application where any reflected light from a surface could interfere with some underlying weak emission), a light trap can be constructed. The two types illustrated in Figure 4.69 work well: the Wood's horn, in which a beam of light entering the trap is gradually attenuated by a series of reflections at an absorbing surface made in the form of a curved cone (usually made of glass); and a stack of razor blades, which absorbs incident light very efficiently.

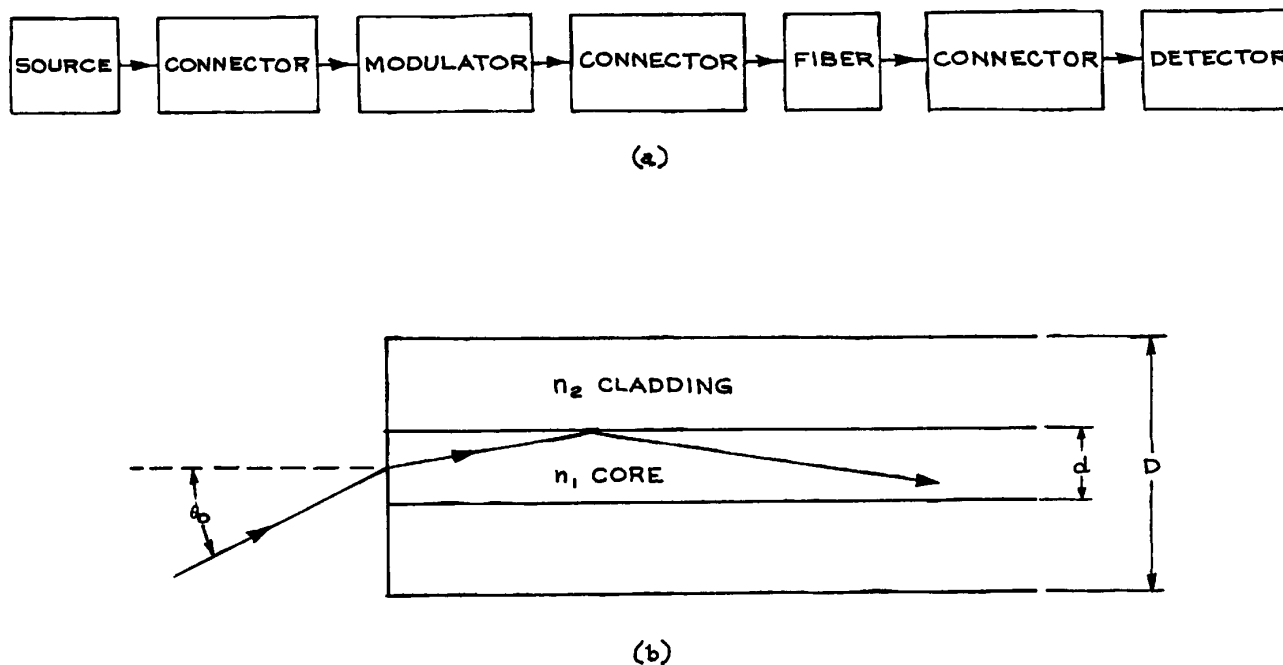
### 4.3.9 Fiber Optics

As is well known, the use of optical fibers has become widespread in the telecommunications field. Consequently, there are numerous suppliers of the various components and subsystems involved.<sup>31</sup> The use of optical fibers in scientific research can be a very valuable technique, and



**Figure 4.69** Light traps: (a) Wood's horn; (b) stacked razor blades.

one that is not particularly complicated. For example, in experiments where radio frequency (rf) interference, pickup, and ground loops are a problem, an experimental signal can be used to modulate a small light-emitting diode or laser. The modulated optical signal is then passed along an optical fiber to the observation location, where the original electrical signal can be recovered with a photodiode. Complete analog transmission systems of this kind are available from A.A. Lab Systems, Alcatel, Analog Modules, Force, JDS Uniphase, Luxlink, and Metrotek Industries. A schematic diagram of the various components of such a system is shown in Figure 4.70(a). The source may be supplied fabricated directly onto a fiber. The modulator is frequently unnecessary, as the source itself can be directly amplitude-modulated. The detector may also be



**Figure 4.70** (a) Fundamental components of a fiber-optic data link; (b) meridional section through a step-index optical fiber. The core refractive index is  $n_1$ ; the cladding refractive index is  $n_2$ . The core and cladding diameters are  $d$  and  $D$ , respectively. A ray entering the fiber will be totally internally reflected provided its angle of incidence is less than  $\theta_0$ , where  $\sin \theta_0 = \text{NA}$ .

supplied fabricated directly onto the end of the fiber. A few specific points are worthy of note for the experimentalist who may wish to utilize this technology.

For certain laboratory applications the construction of a fiber system without any connectors is straightforward. Light from the source is focused into the end of the fiber with a microscope objective. Light emerging from the other end of the fiber is focused onto the detector in a similar way. A convenient range of microscope objectives for this purpose is available from Newport, who also supply a wide range of components for holding and positioning fibers. The choice of lens focal length and placement is governed by the *numerical aperture* (NA) of the fiber. The meaning of this parameter can be understood with reference to Figure 4.70(b), which shows a meridional section through a so-called step-index fiber. In this composite fiber the cylindrical core has refractive index  $n_1$ , and the surrounding cladding has index  $n_2$ , where for total internal

reflection of rays to occur in the core,  $n_1 > n_2$ . Light entering the fiber at angles  $\leq \theta_0$  will totally internally reflect inside the core. The NA is:

$$\sin \theta_0 = \sqrt{n_1^2 - n_2^2} \quad (4.189)$$

Commercially available fibers typically have a smoothly varying radial index profile, but the NA is still the appropriate parameter for determining the acceptance angle. Two specific types of fiber are in most common use: single-mode and multimode. Single-mode fibers typically have core diameters on the order of  $10 \mu\text{m}$  and cladding diameters of  $125 \mu\text{m}$ . They require very precise connectors and are only needed in specialized experiments, for example in high data rate optical communications, where the ability of the fiber to support only a single propagating mode is important. Multimode fibers have larger core diameters, from  $50 \mu\text{m}$  to above  $1 \text{mm}$ , and cladding diameters somewhat larger than

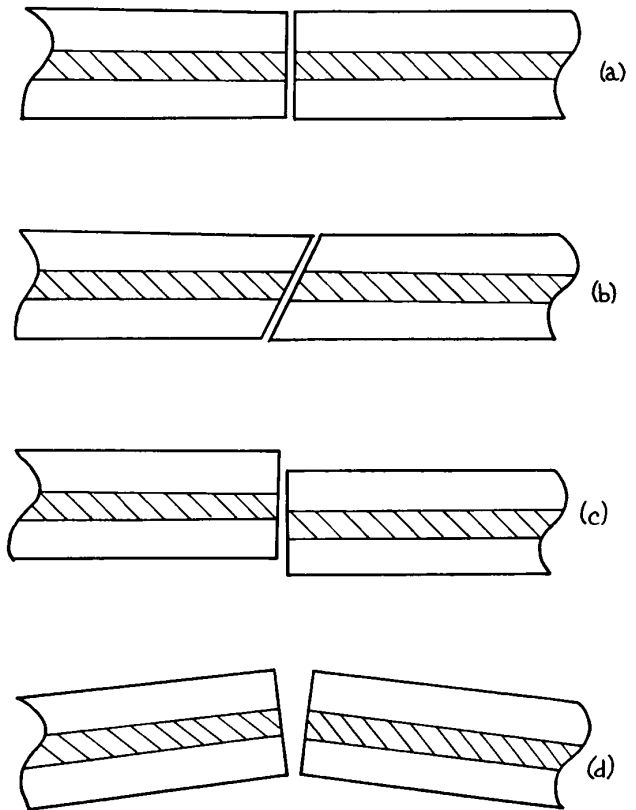
their respective core diameters. These fibers are easy to use: the larger sizes are frequently used to channel light for illumination into positions that are difficult to access, or to collect light from one location and channel it to a detector somewhere else. Large fibers that are suitable for these purposes are available from AFL Telecommunications, Berkshire Photonics, Edmund Optics, Fiberguide Industries, Newport, Polymicro Technologies, and 3M Speciality Optical Fibers.

Fibers can be cut to provide a flat end face with the aid of specialist cleaving tools of varying degrees of precision and complexity, available from suppliers such as Newport and York. In simple experimental setups an adequate cleaved flat face can be obtained in the following way:

- (1) Remove the outer plastic protective coating from the cladding in the region that is to be cut by using a fiber-stripping tool or by immersing this part of the fiber in methylene chloride (the active ingredient in most proprietary paint and varnish strippers). To clean a fiber that has been stripped of its cladding use Kim-wipes or lens tissue moistened with pure isopropyl alcohol. Do not wipe the bare fiber with dry tissue.
- (2) Lightly scratch the cladding with a cleaving tool (slightly more precise than a conventional glass cutter).
- (3) Fasten one half of the fiber to a flat surface with masking tape, and pull the other half of the fiber axially away, keeping the fiber flat on the surface.

This procedure usually gives a flat enough face for use with a microscope objective. If fibers are to be cut and fitted with connectors, they must be cut and polished according to the instructions supplied with the connector.

**Fiber Optic Connectors.** If two identical fibers are to be joined together, this can be done with mechanical coaxial connectors, with a mechanical splice, or the fibers can be fused together with a fusion splicer. In all cases the flat, cleaved faces of the two fibers to be joined must be brought together so that the two faces are in close proximity, are parallel, and the fibers are coaxial. Proper alignments for orthogonal and angle cleaves are shown in Figure 4.71(a) and (b), respectively. Any lateral or tilt offset of the fibers when they are joined will reduce coupling efficiency. Precision mechanical optical fiber connectors are able to couple two fibers together with



**Figure 4.71** Alignment of optical fibers for optical connection: (a) correct alignment for perpendicular-cleaved fibers; (b) correct alignment for angle-cleaved fibers; (c) lateral misalignment; (d) angle misalignment.

losses in the range 0.1–0.2 dB. Sometimes the connector contains a small amount of index-matching gel that is squeezed in the gap between the fibers.

Mechanical fiber splices are an inexpensive way to join fibers. They generally have a sleeve-like construction into which the two fibers are pushed and then glued or clamped in place. Better yet, a *fusion splicer* can be used to fuse fibers together; this is, however, an expensive device. A good fusion splice can have a loss below 0.1 dB. After two fibers are fused together, a *splice protector* should be included if the joined fiber is to be protected from failure of the splice.

The use of optical fiber devices is much simplified if these devices are purchased with appropriate fiber optic connectors already attached. A convenient way to obtain

a connectorized fiber where access to one free end is required is to buy a fiber *jumper*, which is the optical equivalent of patch cord or coaxial cable. By cutting the jumper cable, a length of fiber with a connector on one end is easily obtained. Fiber jumpers with various kinds of connectors, different fiber optic cables, and of various lengths are readily available (Cableorganizer.com, Data-max Technologies, Fiberall, Fiber.com, Fiber Optic Cable Shop, L-Com, Optica, Stonewall Cable, and Unicom). Fiber optic connectors come in many types for both single mode and multimode fibers: FC, LC, SMA, ST, SC, MT-RJ, MU, 50/125, D4, E2000, 62.5/125, 9/125, and Opto-clips. For special applications, angle-cleaved connectors are available, which reduce the inevitable small reflection that occurs at a fiber connector. A comprehensive discussion of practical fiber optics can be found in Goff.<sup>47</sup>

*Polarization-maintaining (PM) fiber*, will preserve the polarization state of launched light (to about 1% purity). Special connectors and splicing techniques are needed for PM fiber, as the azimuthal orientation of the two fibers is now important.

#### 4.3.10 Precision Mechanical Movement Systems

It is frequently necessary to be able to position components in an optical system with high precision. The type of stage with which this is done depends on the scale of movement, the accuracy, repeatability resolution, and freedom from backlash that is required. There are a large number of suppliers of stages, including translation, rotation, tilting, multi-axis positioning and combinations of these including Aerotech, Daedal, Melles Griot, Newport, New Focus, Optosigma, Thor Labs, and Velmex. Both manual and motor-driven stages are available. Although precision mechanical stages can be built in the laboratory, they are widely available commercially. It is not likely that an experimentalist will save money by building rather than buying a precision stage.

The accuracy of adjustment is the difference between the actual and desired positions. Repeatability refers to the ability to reproduce a desired position after a movement of a stage to a different position. In a fair comparison of stages this should involve significant, comparable movements away from, and then back to a desired position.

Resolution is the smallest motion that a stage can make, which for a screw-driven device will be limited by factors such as the number of threads per inch in the lead screw, and friction. A precision micrometer might, for example, have a resolution of about 1/10 of the smallest scale graduation around the micrometer head. Another factor of importance is creep. After a stage has been set to a desired position does it stay there, or does removal of stress from a positioning screw cause it to relax. Absence of creep means that a mount, once positioned, holds its position for a long period. It is our experience that stainless-steel mounts are less susceptible to creep than aluminum mounts.

Resolution alone as a specification does not mean much without additional specification of accuracy, repeatability, and stability. The long-term stability of a mount is also influenced by the thermal stability and expansion of the materials from which the stage is made. The conventional materials of construction are aluminum alloy and stainless steel.

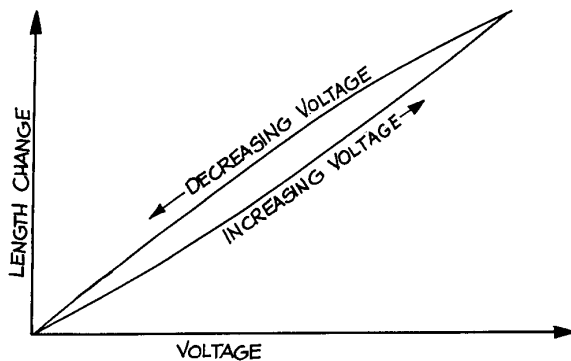
Aluminum has a higher coefficient of thermal expansion ( $24 \times 10^{-6}/\text{K}$ ) compared with stainless steel ( $10\text{--}16 \times 10^{-6}/\text{K}$ ), but aluminum has a higher thermal conductivity and reaches thermal equilibrium more quickly, which minimizes thermally induced strain. Aluminum takes a poor thread and usually brass threaded inserts must be used for brass or stainless-steel positioning screws.

The thermal stability of an optical mount can be improved by using mounting assemblies constructed with two dissimilar metals whose thermal expansions offset each other. Commercially available mounts of the highest precision may include this technology. The manufacturer's technical specifications should be consulted.

It is convenient to divide the movement scales available into three broad categories.

- (1) *Medium* – the scale of movement provided by manual micrometer drives or precision lead screws. Typical micrometer drives provide up to 50 mm travel with accuracy on the order of 1  $\mu\text{m}$ . Some commercial drives claim a resolution below 0.1  $\mu\text{m}$ , but, in practice, such claims must be reviewed with skepticism, since the backlash, creep and repeatability of all mechanical, screw-based, stages is likely to be worse than this.

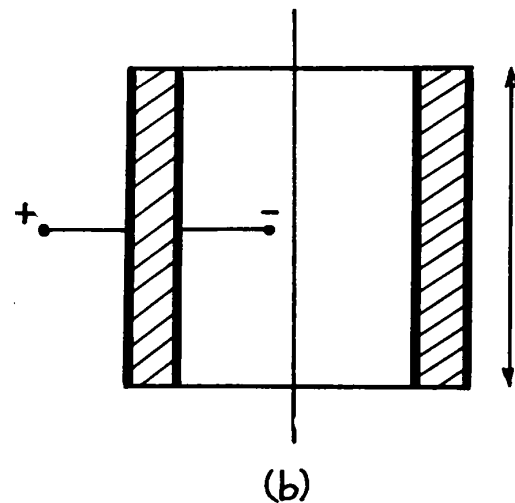
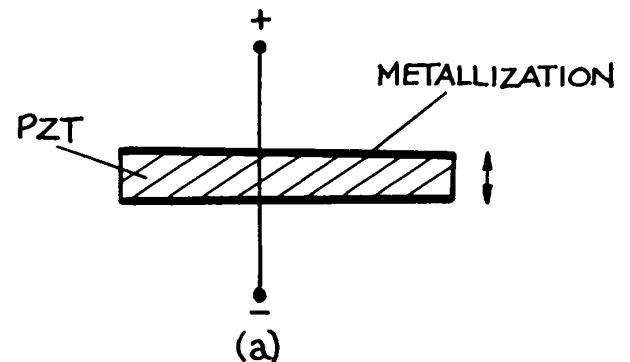
- (2) *Fine* – The scales of movement, precision, and repeatability provided by motor-driven stages, which may be driven either by a stepping motor or dc motor. For repeatability, some form of encoding of the position is needed, or feedback based on the characteristics of the state. Position encoding is usually done with either a shaft encoder, which tells the drive motor on the lead screw exactly what its angular position is, or with a linear encoder. A linear encoder gives a more precise, repeatable measure of the actual location of the object being moved. Precision drives incorporating encoders are available from griffin Motion, Newport, Physik Instrumente (PI), Quater, and Thorlabs. Precise motor driven stages are also available from Aerotech, J.A. Noll, Daedal, Melles-Griot, Newport, and New Focus.
- (3) *Nanoscale* – positioning down to 1 nm resolution, with varying degrees of repeatability. Positioning on this scale requires the use of piezoelectric or electrostatic transducers. Piezoelectric transducers generally use a ceramic material such as lead zirconate titanate (PZT), which changes its shape in an applied electric field. Typical PZT transducers provide mechanical motion on the order of a few hundreds of nanometers for applied voltages on the order of 1 kV. They provide resolution of 10 nm or less, but do exhibit hysteresis, as shown in Figure 4.72, unless some absolute position encoding scheme is used. Individual PZT transducers are generally available either in the form of tubes or disks, as shown in Figure 4.73. Larger motions can be



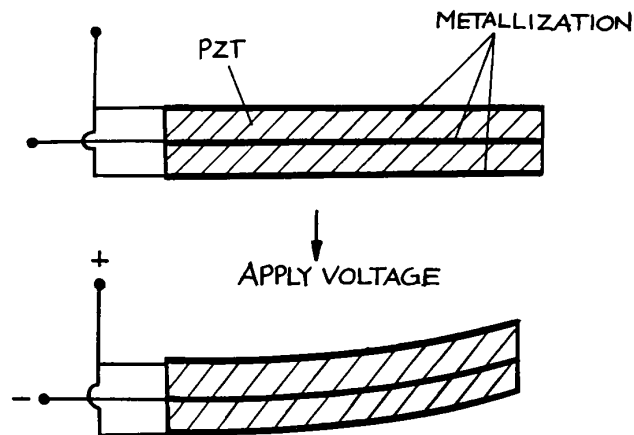
**Figure 4.72** Hysteresis in the length change of a piezoelectric element with applied voltage.

obtained by stacking transducers in either a bimorph or polymorph arrangement. In a bimorph, two slabs of PZT are driven so that one slab extends while the other gets shorter, which causes the bimorph to bend, as shown in Figure 4.74.

In a slab transducer, voltage is applied across the narrow dimension of the slab, which will then change its thickness with applied voltage. In a cylindrical transducer voltage is applied between the inner and outer metallized cylindrical surface of the tube, which then changes its length in the



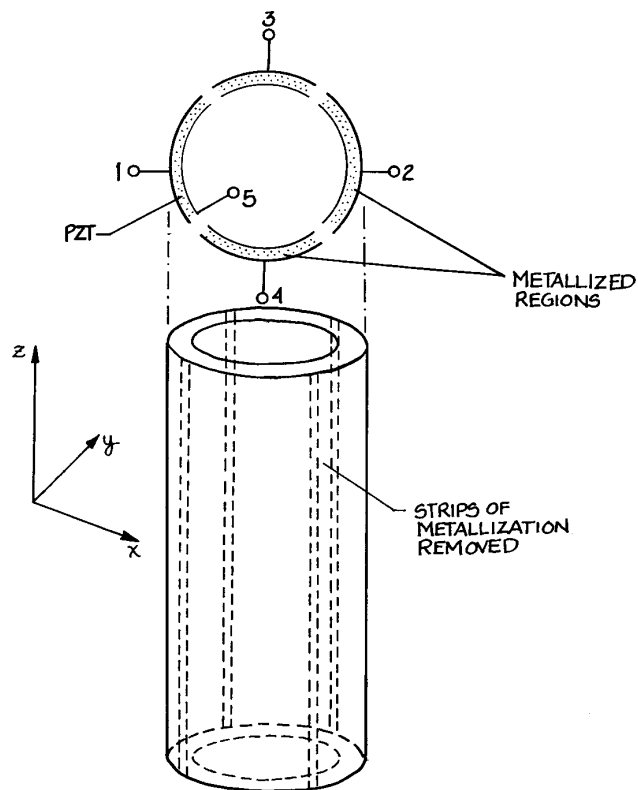
**Figure 4.73** Simple piezoelectric elements: (a) slab; (b) cylinder.



**Figure 4.74** Operation of a bimorph piezoelectric element in a parallel configuration.

axial direction. Nanoscale precision piezoelectric drives are available commercially from Attocube, Mad City Labs, Melles Griot, Queensgate Instruments, Polytec, Physik Instrumente, and Thorlabs, among others. Bimorphs are available from FDK Products and Morgan Electro Ceramics.

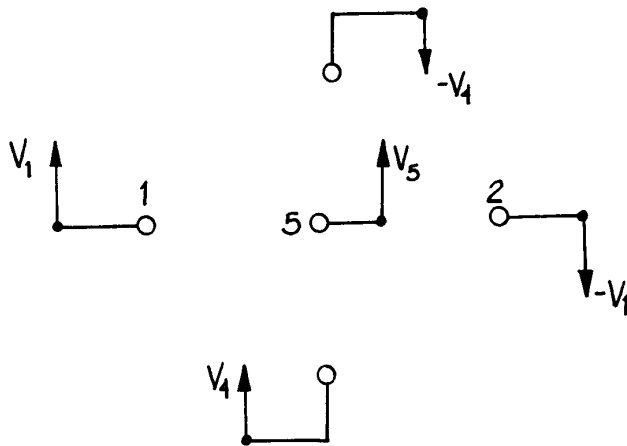
Small scale ( $\sim 10\ \mu\text{m}$ ) precision motion in each of three orthogonal axes can be accomplished inexpensively with a modified piezotube. A typical piezotube is on the order of 30 mm long by 10 mm, and is generally supplied with a metal coating on its inner and outer cylindrical surfaces. The outer metallization should be removed in four segments approximately 1 mm wide spaced  $90^\circ$  apart on the outer cylindrical surface, as shown in Figure 4.75. The four segments are then driven by a “bridge” amplifier current, in which opposite pairs of electrodes are driven in a push-pull configuration, as shown schematically in Figure 4.76. Driving electrodes 1 and 2 produces an  $x$ -directed tilt of the tube, electrodes 3 and 4 produce a  $y$ -directed tilt of the tube. An additional amplifier drives electrode 5 for  $z$ -directed extension of the tube. Because the interelectrode voltages used need to be as large as 1000 V, special op-amps are required. Appropriate amplifiers are available from Apex Microtechnology Corporation. Piezo tubes are available from Argillon, APC International, Boston Piezo-Optics, Omega Piezo Technologies, Physik Instrumente (PI), and Sensor Technology.



**Figure 4.75** A cylindrical piezoelectric element with outer quadrant electrodes. The element can be driven to provide axial motion as well as tilt in two orthogonal axes.

#### 4.3.11 Devices for Positional and Orientational Adjustment of Optical Components

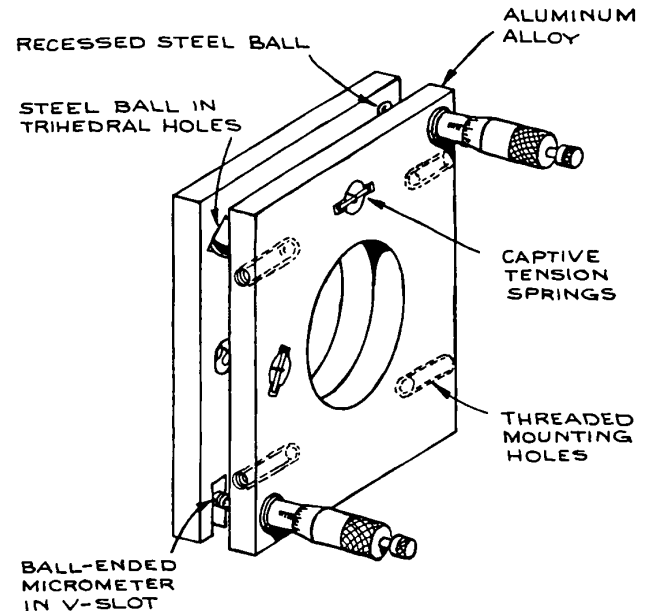
For positioning a given component, the maximum number of possible adjustments is six: translation along three mutually orthogonal axes and rotation about these (or other) axes. Mounts can be made that provide all six of these adjustments simultaneously. Usually, however, they are restricted and have no more degrees of freedom than are necessary. Thus, for example, a translation stage is usually designed for motion in only one dimension; independent motion along other axes can then be obtained by stacking one-dimensional translation stages. Orientation



**Figure 4.76** Voltage drive scheme for a piezoelectric element such as the one shown in Figure (4.69). Electrodes 1–4 are the quadrant electrodes, Electrode 5 is the inner electrode.

stages are usually designed for rotation about one axis (polar rotation) or two orthogonal axes (mirror mounts or tilt tables). Once again, such mounts can be stacked to provide additional degrees of freedom. It is often unnecessary for angular adjustment to take place about axes that correspond to translation directions.

The ideal optical mount provides independent adjustment in each of its degrees of freedom without any interaction with other potential degrees of freedom. It should be sturdy, insensitive to extraneous vibrational disturbances, (such as air or structure-borne acoustic vibrations), and free of backlash during adjustment. Generally it should be designed on kinematic principles (see Section 1.7.1). The effect of temperature variations on an optical mount can also be reduced, if desired, by constructing the mount from low-coefficient-of-expansion material, such as Invar, or by using thermal-expansion compensation techniques. Very high rigidity in an optical mount generally implies massive construction. When weight is a limitation, honey-combed material or material appropriately machined to reduce excess weight can be used. Most commercially available mounts are not excessively massive, but represent a compromise between size, weight, and rigidity. Generally, mounts do not need to be made any more rigid than mechanical considerations involving Young's modulus and likely applied forces on the mount would dictate.



**Figure 4.77** Kinematic double-hinge mirror mount.

The design of optical mounts on kinematic principles involves constraining the mount just enough so that, once adjusted, it has no degrees of freedom. Some examples will illustrate how kinematic design is applied to optical mounts.

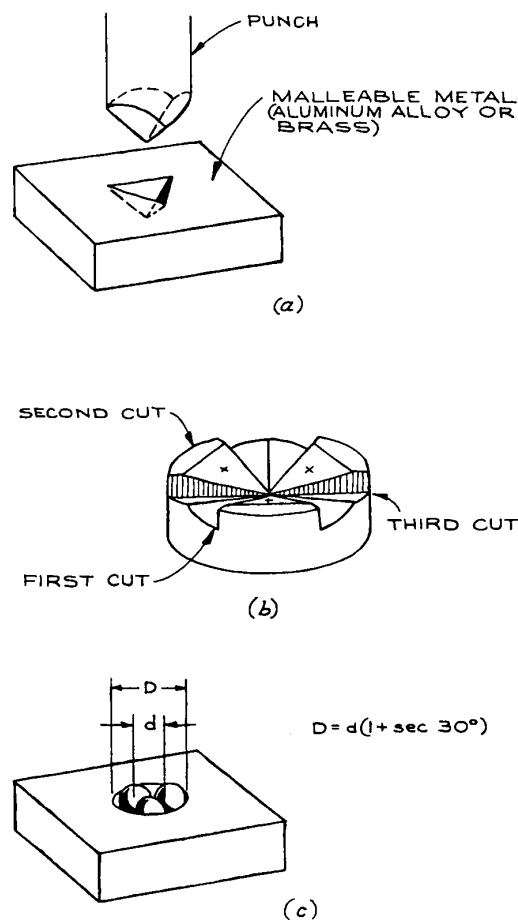
**Two-Axis Rotators.** Within this category the commonest devices are adjustable mirror or grating mounts and tilt tables used for orienting components such as prisms. Devices that adjust orientation near the horizontal are usually called tilt adjusters. Mirror holders, on the other hand, involve adjustment about a vertical or near-vertical axis. A basic device of either kind involves a fixed stage equipped with three spherical or hemispherical ball bearings, which locate, respectively, in a V-groove, in a conical hole, and on a plane surface machined on a movable stage. A typical device is shown in Figure 4.77. The movable stage makes contact with the three balls on the fixed stage at just six points – sufficient to prevent motion in its six degrees of freedom – and is held in place with tension springs between the fixed and movable stages. Adjustment is provided by making two of the locating balls the ends of fine-thread screws or micrometer heads. Suitable micrometer heads for this

purpose are manufactured by Brown and Sharpe, Mitutoyo, Moore and Wright, Shardlow, and Starrett, and are available from most wholesale machine-tool suppliers. The ball of the micrometer should be very accurately centered on the spindle. The adjustment of the mount shown in Figure 4.77 is about two almost orthogonal axes. There is some translation of the center of the movable stage during angular adjustment, which can be avoided by the use of a true gimbal mount.

A few of the constructional details of the mount in Figure 4.77 are worthy of note. Stainless steel, brass, and aluminum are all suitable materials of construction; an attempt to reduce the thermal expansion of the mount by machining it from a low-expansion material such as Invar may not meet with success. Extensive machining of low-expansion-coefficient alloys generally increases the coefficient, unless they are carefully annealed afterward. Putting the component in boiling water and letting it gradually cool will lead to partial annealing. True kinematic design requires ball location in a trihedral hole, in a V-slot, and on a plane surface (or in three V-slots, but this no longer permits angular adjustment about orthogonal axes) (see also Section 1.7.1). If the mount is made of aluminum, the locating hole, slot, and plane on the movable stage should be made of stainless steel, or some other hard material, shrunk-fit into the main aluminum body. A true trihedral slot can be made with a punch, as shown in Figure 4.78(a), or it can be machined with a 45° milling cutter as shown in Figure 4.78(b). A conical hole is not truly kinematic; a compromise design is shown in Figure 4.79.

Although motion of the mount in Figure 4.77 results from direct micrometer adjustment, various commercial mounts incorporate a reduction mechanism to increase the sensitivity of the mount. Typical schemes that are used involve adjustment of a wedge-shaped surface by a moving ball, as shown in Figure 4.80, or the use of a differential screw drive. The two or more movable stages of an adjustable mount can be held together with captive tension springs, as in Figure 4.77. Designs involving only small angular adjustment can utilize the bending of a thin section of material, as shown in Figure 4.81. The design of such mounts, called *flexures*, is discussed in detail in Section 1.7.6.

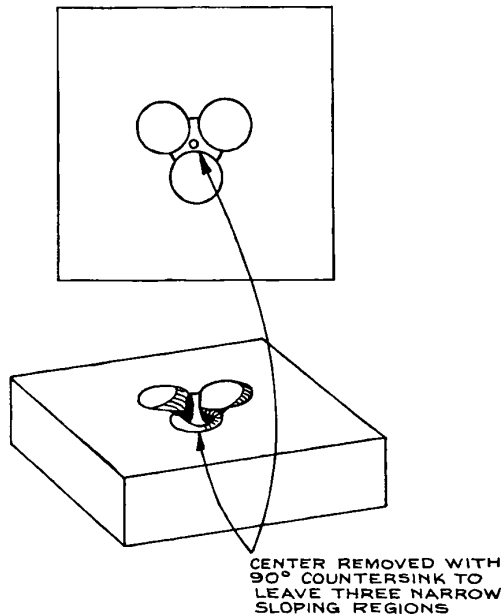
The basic principle of a gimbal mount for two-axis adjustment is shown in Figure 4.82. True orthogonal motion about two axes can be obtained in such a design without



**Figure 4.78** Methods for making trihedral locating holes for kinematic design; (a) trihedral steel punch; (b) trihedral hollow machined with 45° milling tool; (c) steel balls fitting tightly in a hole.

translation of the center of the adjusted component. A kinematically designed gimbal mount is, however, much more complicated in construction than the mount shown in Figure 4.82. Precision gimbal mounts should be purchased rather than built. Very many companies manufacture adequate devices with a variety of designs, and their prices are lower than the cost of duplicating such devices in the laboratory. Suppliers of gimbal mounts include Aerotech, Daedal, New Focus, Newport, J. A. Noll, Standa, and Thorlabs. Most of these manufacturers also manufacture good nongimbal mounts similar to the one shown in Figure 4.77.





**Figure 4.79** Compromise design for kinematic trihedral hole.

**Translation Stages.** Translation stages are generally designed to provide linear motion in a single dimension and can be stacked to provide additional degrees of freedom. Mounts for centering optical components, particularly lenses, incorporate motion in two orthogonal (or near-orthogonal) axes in a single mount. The simplest translation stages usually involve some sort of dovetail groove arrangement, as shown in Figure 4.83(a). Greater precision can be obtained with ball bearings captured between V-grooves with an arrangement to prevent adjacent balls from touching, as shown in Figure 4.83(b), or with crossed roller bearings, as shown in Figure 4.83(c). A simple ball-bearing groove design is shown in Figure 4.84, although variants are possible. Suppliers of precision, roller bearing translators include Aerotech, Daedal, Melles Griot, Newport, OptoSigma, and Thorlabs. Vertical translators using a fine-thread screw drive are available from Newport. Stages can be built to provide short-range, highly stable, frictionless precision movement by using flexures, as shown in Figure 4.85. Typical travel of a flexure stage is on the order of 10% to 15% of the dimensions of the stage.

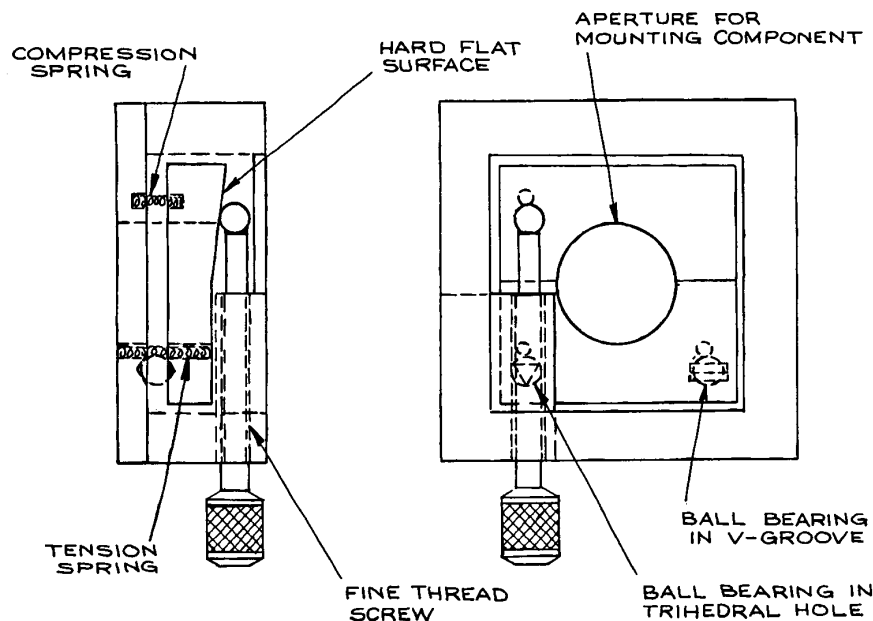
Less precise, but generally adequate, nonball-bearing translation stages are available from Newport and Velmex. Mounts for centering optical components such as lenses are available from Aerotech, Ealing, and Newport, among others. A simple kinematic design for a centering mount is illustrated in Figure 4.86.

**Rotators and Other Mounts.** Rotators are used, for example, for orienting polarizers, retardation plates, prisms, and crystals. They provide rotation about a single axis. Precision devices using ball bearings are available from Aerotech, New Focus, Newport, OptoSigma, and Thorlabs, among various other suppliers.

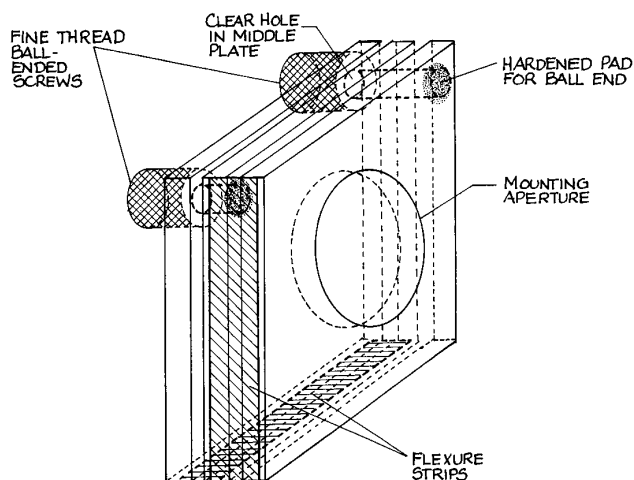
There are numerous more complicated optical mounts than those described in the previous sections, which provide in a single mount any combination of degrees of freedom desired. Motor-driven and computer-controllable versions of all these mounts are generally available. For details, the catalogs of the manufacturers previously mentioned should be consulted.

**Optical Benches and Components.** For optical experiments that involve the use of one or more optical components in conjunction with a source and/or a detector in some sort of linear optical arrangement, construction using an optical bench and components is very convenient, particularly in breadboarding applications. The commonest such bench design is based on the equilateral triangular bar developed by Zeiss and typically machined from lengths of stabilized cast iron. Aluminum benches of this type are also available, but are less rigid. Components are mounted on the bench in rod-mounted holders that fit into carriers that locate on the triangular bench. A cross-sectional diagram of a typical bench and carrier is shown in Figure 4.87.

Coarse adjustment along the length of the bench is provided by moving the carrier manually. On many benches this is accomplished with a rack-and-pinion arrangement. A wide range of mounts and accessories are available, providing for vertical, lateral, and longitudinal adjustment of mounted-component positions. Holders for filters, for centering lenses, and for holding and rotating prisms and polarizers are readily obtained. Precision gimbal mounts, translators, and rotators can be readily incorporated into the structure. Examples of devices that lend themselves well



**Figure 4.80** Cutaway views of a kinematic, single-axis rotator incorporating a reduction mechanism involving a ball bearing on an angled surface. Two rotators of this kind could be incorporated in a single mount to give two-axis gimbal adjustment.



**Figure 4.81** Flexure type double-hinge mirror design.

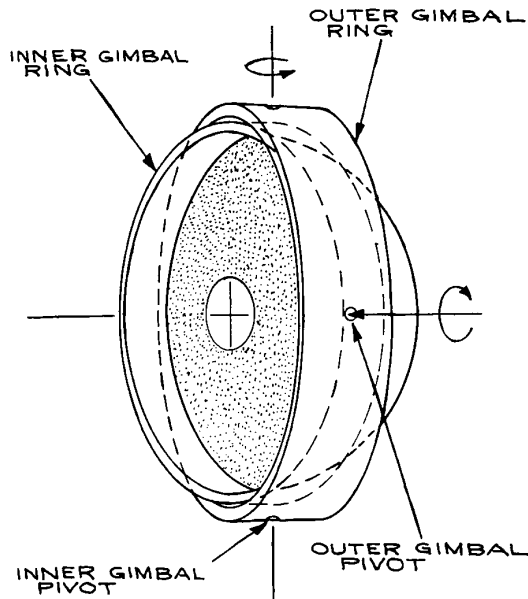
to optical-bench construction are telescopes, collimators, laser-beam expanders and spatial filters, optical isolators, and electro-optic light modulators. Optical benches and accessories are available from Edmund Optics, Lasico, Linos Photonics, Melles Griot, and Newport. Optical benches are no longer as commonly used or available as they once were because of the wide availability of optical tables and mounts, and the additional flexibility provided by their use.

**Piezoelectric Transducers.** Most PZTs are made from ferroelectric ceramics, such as lead zirconate titanate or barium titanate, but polymers such as polyvinylidene fluoride (PVF) are also used. Simple ceramic transducers are usually supplied as flat discs with a metalized layer on each flat face, or as short hollow cylinders, in which case the metalized electrodes are deposited on the inner and outer cylindrical surfaces. The flat discs change their thickness with applied voltage, the cylindrical types change their axial length. Typical length changes are on the order of  $10^{-9}$  m/V. Therefore, these devices are particularly suitable for very

fine (submicron) positioning. Much larger motions can be obtained with bimorph PZTs. Bimorphs can be stacked to provide macroscopic motions for modest applied voltages. Excursions of several millimeters can be obtained for voltages of 1 kV. Even larger motions can be obtained with Inchworm transducers, available from EXFO-Burleigh and Sensor Technology. Basic ceramic transducer elements are available from EDO Electro-Ceramic Products, Physik Instrumente (PI), Piezo Systems, CTS Piezoelectric Prod-

ucts, Polytec, and Queensgate Instruments. PVF transducers are available from Atochem North America (formerly Pennwalt). Complete transducer assemblies are available from EXFO-Burleigh, Newport/Klinger, and Polytec Optronics.

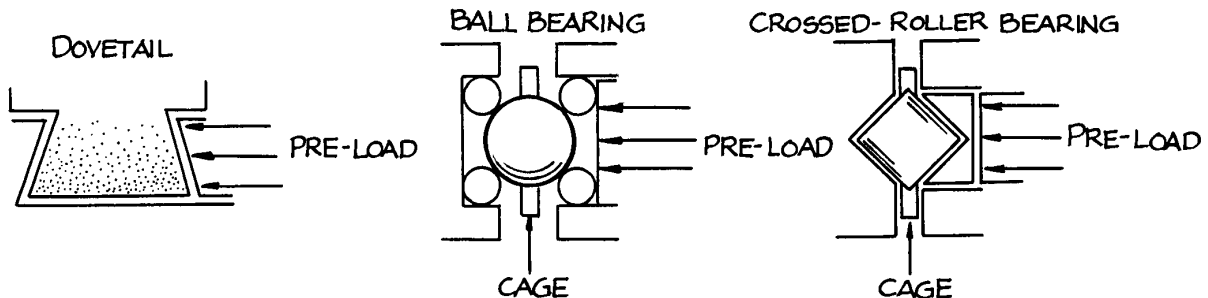
**Voice Coil Actuators.** Voice coil actuators are a convenient way to generate linear movement in the range between micrometers and tens of millimeters, with zero hysteresis, and up to audio frequencies. They are essentially the same as the mechanism in a loud speaker that moves the diaphragm to generate sound. They are direct drive linear actuators that incorporate magnets and an electrical coil. The voice coil itself is a noncommutated, two terminal device. It has linear control characteristics, zero hysteresis, zero backlash, and infinite position sensitivity. It has short mechanical and electrical time constants and a high output power-to-weight ratio. A voice coil is generally operated in a control loop where a control signal is provided from a position readout device. Voice coil actuators are available from Airex, BEI Kimco, Equipment Solutions, H2W Technologies, Innotics, and Orlin.



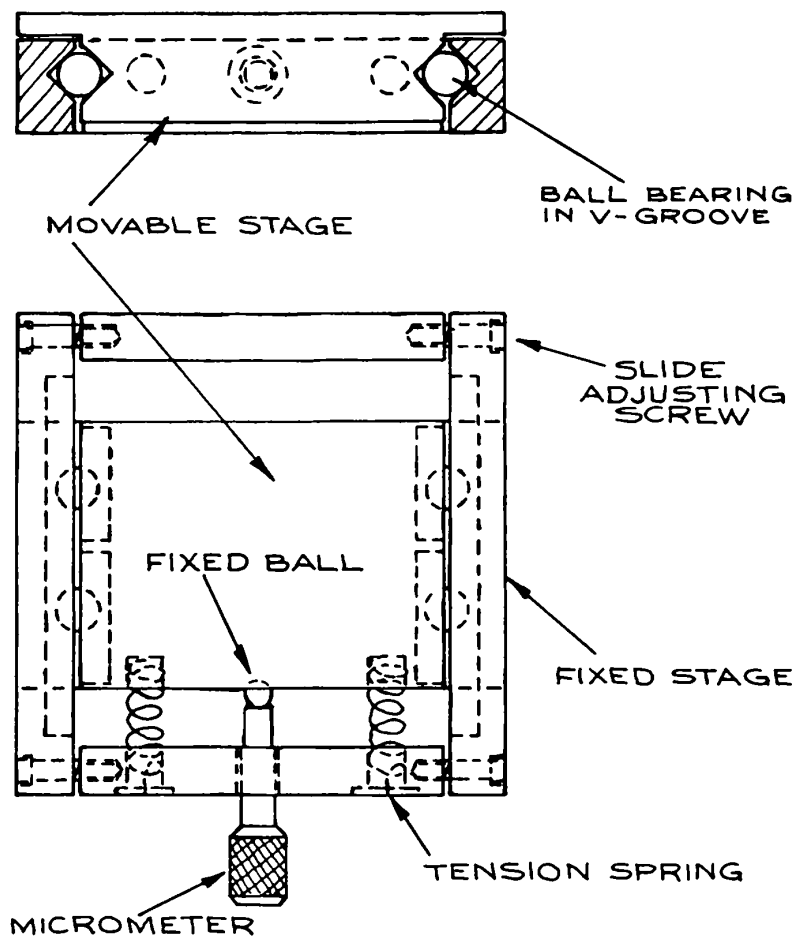
**Figure 4.82** Schematic diagram of a true two-axis gimbal mount. The center of the mounted component does not translate during rotation.

#### 4.3.12 Optical Tables and Vibration Isolation

A commercially made optical table provides the perfect base for construction of an optical system. Such tables have tops made of granite, steel, invar, superinvar (very expensive), or ferromagnetic or nonmagnetic stainless steel. When vibrational isolation of the table top is required, the top is mounted on pneumatic legs. Such an arrangement effectively isolates the table top from floor



**Figure 4.83** Groove mounting arrangements for linear translation stages: (a) dovetail V-groove; (b) ball bearing; (c) crossed roller bearing.

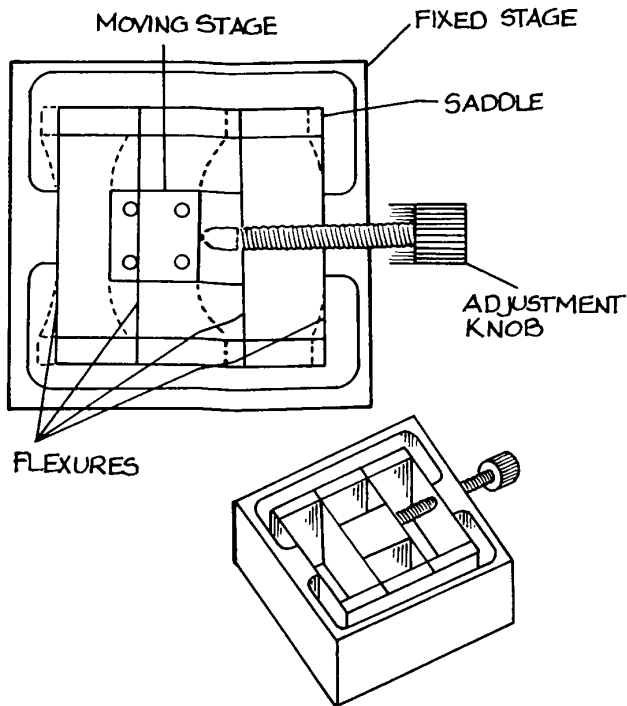


**Figure 4.84** Section drawings of a simple 1-D translation stage.

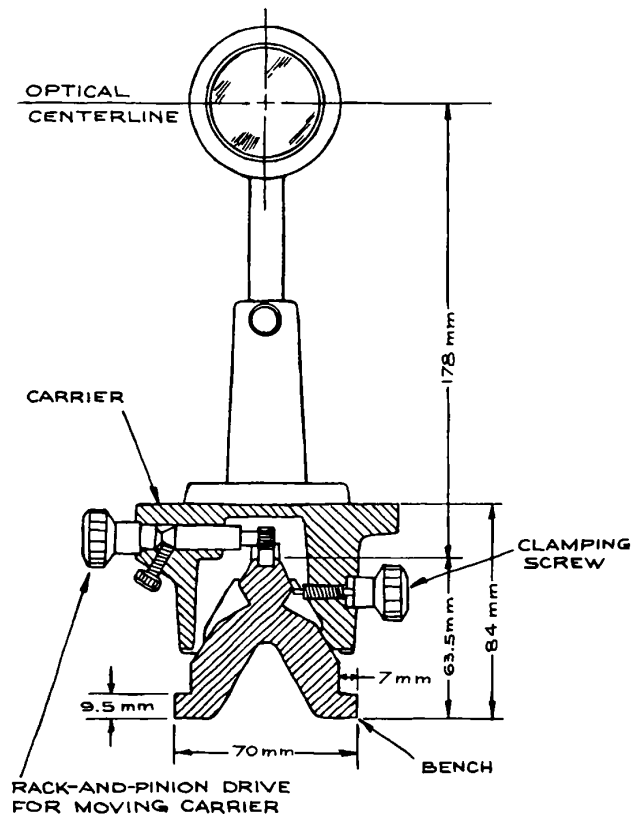
vibrations above a few Hertz. To minimize acoustically driven vibrations of the table top itself, the better metal ones employ a laminated construction with a corrugated or honeycomb metal cellular core bonded with epoxy to the top and bottom table surfaces.

Granite table tops are very much less intrinsically resonant than metal tops and represent the ultimate in dimensional stability and freedom from undesirable vibrational modes. Complete pneumatically isolated optical table systems, nonisolated tables, and table tops alone are available from Kinetic Systems, Melles Griot, Newport, Standa, Thorlabs, and TMC. In selecting a table, points of comparison to look for include the following:

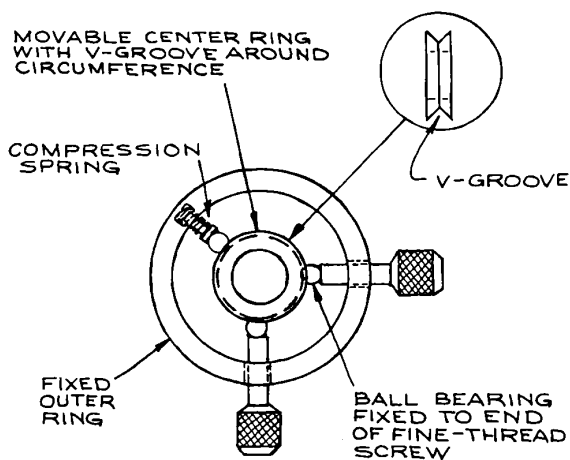
- (1) The overall flatness of the table top (flatness within  $50\ \mu\text{m}$  is good, although greater flatness can be obtained)
- (2) The low-frequency vibration isolation characteristics of the pneumatic legs
- (3) The rigidity of the table top (how much it deforms with an applied load at the center or end)
- (4) The isolation from both vertical and horizontal vibrations
- (5) The frequencies and damping of the intrinsic vibrational modes
- (6) Are the threaded holes in the table top sealed so that liquid spills do not disappear into the interior honeycombs of the table?



**Figure 4.85** Linear translation stage using flexure mounts.



**Figure 4.87** Typical dimensions of a Zeiss triangular optical bench with carrier and mounted optical component. (Courtesy of Ealing Corporation).



**Figure 4.86** A simple precision centering stage.

Other criteria, such as cost and table weight, may be important. Typical sizes of readily available tops range up to 5 × 12 ft (larger tables are also available) and in thickness from 8 in. to 2 ft. The thicker the table, the greater its rigidity and stability. Most commercial tables are available with 1/4-20 threaded holes spaced on 1 in. centers over the surface of the table. Better designs use blind holes of this type so that liquids spilt on the table do not vanish into its interior. Ferromagnetic table tops permit the use of magnetic hold-down bases for mounting optical components. Such mounts are available from numerous optical suppliers. It is cheaper, and just as satisfactory, to buy magnetic bases designed for machine-tool use from a machine-tool supply company.

Steel and granite surface plates that can be used as small, nonisolated optical tables are available from machine-tool supply companies. If it is desired to construct a vibration-isolated table in the laboratory, a 2 in. thick aluminum plate resting on underinflated automobile-tire inner tubes works rather well, although the plate will have resonant frequencies. Very flat aluminum plate, called *jig plate*, can be obtained for this application at a cost not much greater than ordinary aluminum plate. Just mounting a heavy plate on dense polyurethane foam also works surprisingly well. In fact, the vibrational isolation of a small optical arrangement built on an aluminum plate and mounted on a commercial optical table is increased by having a layer of thick foam between plate and table top.

Even the best-constructed precision optical arrangement on a pneumatic isolation table is subject to airborne acoustic disturbances. A very satisfactory solution to this problem is to surround the system with a wooden particle-board box lined with acoustic absorber. Excellent absorber material, a foam-lead-foam-lead-foam sandwich called Hushcloth DS, is available from American Acoustical Products.

#### 4.3.13 Alignment of Optical Systems

Small helium-neon lasers are now so widely available and inexpensive that they must be regarded as the universal tool for the alignment of optical systems. Complex systems of lenses, mirrors, beamsplitters, and so on can be readily aligned with the aid of such a laser: even visible-opaque infrared components can be aligned in this way by the use of reflections from their surfaces. Fabry-Perot interferometers, perhaps the optical instruments most difficult to align, are easily aligned with the aid of a laser. The laser beam is shone orthogonally through the first mirror of the interferometer (easily accomplished by observing part of the laser beam reflect back on itself), and the second interferometer mirror is adjusted until the resultant multiple-spot pattern between the mirrors coalesces into one spot.

#### 4.3.14 Mounting Optical Components

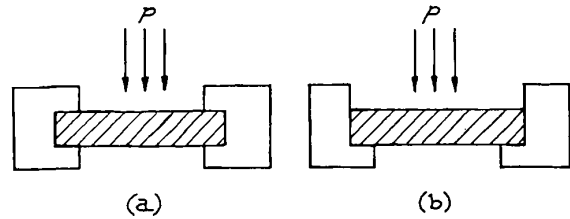
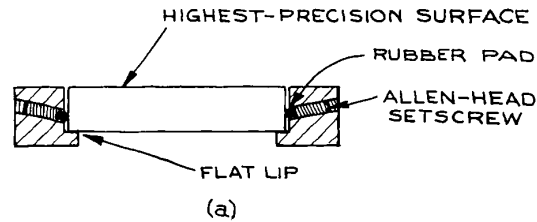
A very wide range of optical devices and systems can be constructed using commercially available mounts designed to hold prisms, lenses, windows, mirrors, light sources, and detectors. Some commercially made optical

mounts merely hold the component without allowing precision adjustment of its position. Many of the rodmounted lens, mirror, and prism holders designed for use with standard optical benches fall into this category, although they generally have some degree of coarse adjustability; however, a wide range of commercially available optical mounts allow precision adjustment of the mounted component.

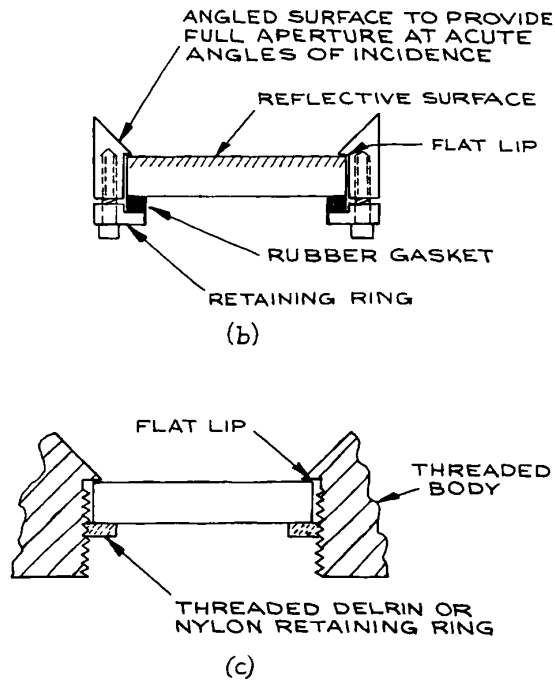
The construction of a complete optical system from commercial mounts can be very expensive, while providing in many cases unnecessary, and consequently undesirable, degrees of freedom in the adjustment and orientation of the various components of the system. Although extra adjustability is often desirable when an optical system is in the “breadboard” stage, in the final version of the system it reduces overall stability. An adjustable mount can never provide the stability of a fixed mount of comparable size and mass. Stability in the mounting of optical components is particularly important in the construction of precision optical systems for use, for example, in interferometry and holography.

A custom-made optical mount must allow the insertion and stable retention of the optical component without damage to the component, which is usually made of some fragile (glassy or single-crystal) material. If possible, the precision surface or surfaces of the component should not contact any part of the mount. The component should be clamped securely against a rigid surface of the mount. It should not be clamped between two rigid metal parts: thermal expansion or contraction of the metal may crack the component or loosen it. The clamping force should be applied by an intermediary rubber pad or ring. Some examples of how circular mirrors or windows can be mounted in this way are shown in Figure 4.88. If a precision polished surface of the component must contact part of the mount (for example, when a precision lens, window, mirror, or prism is forming a vacuum-tight seal), the metal surface should be finely machined: the edges of the mounting surface must be free of burrs. Ideally, the metal with which the component is in contact should have a lower hardness than the component. A thin piece of paper placed between the component and metal will protect the surface of the component without significantly detracting from the rigidity of the mounting.

Generally speaking, unless an optical component has one of a few standard shapes and sizes, it will not fit



**Figure 4.89** Window mounting arrangements: (a) clamped or fixed edge (maximum stress at edge); (b) unclamped or freely supported edge (maximum stress at center).



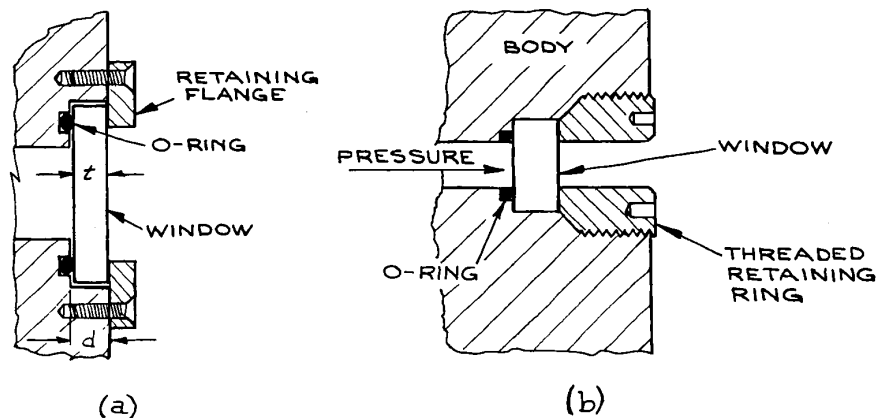
**Figure 4.88** Ways for mounting delicate optical components: (a) no contact between precision surface of component and mount; (b) precision surface in minimal contact with mount and held in place with retaining ring; (c) precision surface in minimal contact with bottom of recess in mount and held in place with threaded plastic ring.

directly in a commercial mount and a custom adaptor will need to be made.

It is often necessary to mount a precision optical component (usually a flat window or mirror, but occasionally a lens or spherical mirror) so that it provides a vacuum or pressure-tight seal. There are two important questions to

be asked when this is done: can the component, of a certain clear aperture and thickness, withstand the pressure differential to which it will be subjected, and will it be distorted optically either by this pressure differential or the clamping forces holding it in place?

There are various ways of mounting a circular window so that it makes a vacuum- or pressure-tight seal. Demountable seals are generally made by the use of rubber O-rings (see Section 3.5.2) for vacuum, or flat rubber gaskets for pressure. More-or-less permanent seals are made with the use of epoxy resin, by running a molten bead of silver chloride around the circumference of the window where it sits on a metal flange, or in certain cases by directly soldering or fusing the window in place. Some crystalline materials, such as germanium, can be soldered to metal; many others can be soldered with indium if they, and the part to which they are to be attached, are appropriately treated with gold paint beforehand.<sup>48</sup> If a window is fused in place, its coefficient of expansion must be well matched to that of the part to which it is fixed. Suitably matched glass, fused silica, and sapphire windows on metal are available. Either the glass and metal are themselves matched, or an intermediate graded seal is used (see Section 2.2.4). To decide what thickness-to-unsupported-diameter ratio is appropriate for a given pressure differential, it must be determined whether the method of mounting corresponds to Figure 4.89(a) or (b). In either case, the maximum stress in the window must be kept below the modulus of rupture  $Y$  of the window by an appropriate safety factor  $S_f$ . A safety factor of four is generally adequate for determining a safe window thickness, but a larger factor may be necessary to avoid



**Figure 4.90** Mounting arrangements for optical windows in vacuum and pressure applications: (a) simple vacuum window; (b) high-pressure window.

deformation of the window that would lead to distortion of transmitted wavefronts. The desired thickness-to-free-diameter ratio for a pressure differential  $p$  is found from the formula:

$$\frac{t}{D} = \sqrt{\frac{KpS_f}{4Y}} \quad (4.190)$$

where  $K$  is a constant whose value can be taken as 0.75 for a clamped edge and 1.125 for a free edge (see also Section 3.6.3).

A window mounted with an O-ring in a groove, as shown in Figure 4.90, may not provide an adequate seal at moderately high pressures, say above  $7\text{MPa}^h$ , unless the clamping force of the window on the ring is great enough to prevent the pressure differential deforming the ring in its grooves and creating a leak. A better design for high-pressure use, up to about  $50\text{MPa}$ , is shown in Figure 4.90(b): the clamping force and internal pressure act in concert to compress the ring into position. Clamping an optical-quality window or an O-ring seal of the sort shown in Figure 4.90(a) may cause slight bending of the window.

If an optical window has been fixed “permanently” in place with epoxy resin, it may be possible to recover the window, if it is made of a robust material, by heating the seal to a few hundred degrees or by soaking it in methylene chloride or a similar commercial epoxy remover.<sup>49</sup>

If optical windows are to be used in pressure cells where there is a pressure differential of more than 500 bars ( $\approx 7000$  psi), O-rings are no longer adequate and special window-mounting techniques for very high pressures must be used.<sup>50–54</sup> Optical cells for use at pressures up to 3 kbar are available from ISS, Inc.

### 4.3.15 Cleaning Optical Components

It is always desirable for the components of an optical system to be clean. At the very least, particles of dust on mirrors, lenses, and windows increase scattering; organic films lead to unnecessary absorption, possible distortion of transmitted wavefronts, and – in experiments with ultraviolet light sources – possible unwanted fluorescence. In optical systems employing lasers, dust and films will frequently lead to visible, undesirable interference and diffraction effects. At worst, dirty optical components exposed to high-intensity laser beams can be permanently damaged, as absorbing dust and films can be burnt into the surface of the component.

The method used for cleaning optical components, such as mirrors, lenses, windows, prisms, and crystals, will depend on the nature of both the optical component and the type of contaminant to be removed. For components made of hard materials, such as borosilicate glass, sapphire, quartz, or fused silica, gross amounts of contamination can



be removed by cleaning in an ultrasonic tank. The tank should be filled with water and the components to be cleaned placed in a suitable pure solvent, such as acetone or trichloroethylene, contained in a clean glass or plastic beaker. If trichloroethylene is used, the operation should be performed inside a fume hood because there is evidence that this solvent is a human carcinogen. The ultrasonic cleaning is accomplished by placing the beaker in the tank of water. Dust can be removed from the surface of hard materials, including most modern hard dielectric reflecting and antireflecting coatings, by brushing with a soft camels-hair brush. Suitable brushes are available from most photographic stores. Lens tissue and solvent can be used to clean contaminants from both hard and medium-hard surfaces, provided the correct method is used. The category of medium-hard materials includes materials such as arsenic trisulphide, silicon, germanium, gallium arsenide, KRS-5, and tellurium. Information about the relative hardness of these and other materials is contained in Table 4.5. Dust should first be removed from the component by wiping, very gently, in one direction only, with a folded piece of dry lens tissue. The tissue should be folded so that the fingers do not come near any part of the tissue that will contact the surface. Better still, if a cylinder of dry nitrogen is available, dust can be blown from the surface of the component. Because the gas can itself contain particles, it should be passed through a suitable filter, such as fine sintered glass, before use. Clean gas can be used in this way to remove dust from the surface of even very delicate components, such as aluminum and gold-coated mirrors, and diffraction gratings. The use of gas jets for dust removal requires caution: delicate structures have been damaged by the force of the gas. Commercial "dust-off" sprays are available that use a liquified propellant gas that vaporizes on leaving the spray-can nozzle. These sprays are not recommended for cleaning delicate or expensive components. They can exert considerable force at close range and can also deposit unvaporized propellant on the component, or condense water vapor into the surface of a hygroscopic material.

There are two good methods for cleaning hard and medium-hard components with lens tissue and solvent. The lens tissue should be folded as indicated previously, and a few drops of solvent placed on the tissue in the area to be used for the cleaning. The tissue should be wiped just

once in a single direction across the component and then discarded. Fingers should be kept away from the solvent-impregnated part of the tissue, and it may be desirable to hold the component to be cleaned with a piece of dry tissue to avoid the transmission of finger grease to its surface. Hemostats, available from surgical supply stores, are very convenient for holding tissue during cleaning operations. To clean a single, fairly flat optical surface, such as a dielectric coated mirror or window, place a single lens tissue flat on the surface. Put one drop of clean solvent (spectroscopically pure grades are best) on the tissue over the component. Then draw the tissue across the surface of the component so that the dry portion of the tissue follows the solvent-soaked part and dries off the surface.

Several suitable solvents exist for these cleaning operations, such as acetone, trichloroethylene, and methyl, ethyl, and isopropyl alcohols. Diethyl ether is a very good solvent for cleaning laser mirrors, but its extreme flammability makes its use undesirable.

If there is any doubt as to the ability of an optical component to withstand a particular solvent, the manufacturer should be consulted. This applies particularly to specially coated components used in ultraviolet and some infrared systems. Some delicate or soft components cannot be cleaned at all without running the risk of damaging them permanently, although it is usually safe to remove dust from them by blowing with clean, dust-free gas. It is worth noting that diffraction gratings frequently look as though they have surface blemishes. These blemishes frequently look much worse than they actually are and do not noticeably affect performance. Never try to clean diffraction gratings. It is also very difficult to clean aluminum- or gold-coated mirrors that are not overcoated with a hard protective layer such as silicon monoxide. Copper mirrors are soft and should be cleaned with extreme care.

Most optical components used in the visible and near ultraviolet regions of the spectrum are hard; soft materials one often used in infrared systems. In this region several soft crystalline materials such as sodium chloride, potassium chloride, cesium bromide, and cesium iodide are commonly used. These materials can be cleaned, essentially by repolishing their surfaces. For sodium and potassium chloride windows this is particularly straightforward. Fold a piece of soft, clean muslin several times to form a glass pad. Place the pad on a flat surface – a piece of glass

Table 4.5 Characteristics of optical materials

Material	Useful Transmission Range ( $\geq 10\%$ transmission) in 2 mm Thickness	Index of Refraction [wavelength ( $\mu\text{m}$ ) in parentheses]	Knoop Hardness	Melting Point ( $^{\circ}\text{C}$ )
LiF	0.104–7	1.60(0.125), 1.34(4.3)	100	870
MgF <sub>2</sub>	0.126–9.7	$n_o = 1.3777$ , $n_e = 1.38950(0.589)^f$	415	1396
CaF <sub>2</sub>	0.125–12	1.47635(0.2288), 1.30756(9.724)	158	1360
BaF <sub>2</sub>	0.1345–15	1.51217(0.3652), 1.39636(10.346)	65	1280
Sapphire(Al <sub>2</sub> O <sub>3</sub> )	0.15–6.3	$n_o = 1.8336(0.26520)^f$ , $n_e = 1.5864(5.577)^f$ , $n_e$ slightly less than $n_o$	1525–2000 <sup>c</sup>	2040 $\pm$ 10
Fused silica (SiO <sub>2</sub> )	0.165–4 <sup>d</sup>	1.54715(0.20254), 1.40601(3.5)	615	1600
Pyrex 7740	0.3–2.7	1.474(0.589), $\approx 1.5(2.2)$	$\approx 600$	820 <sup>g</sup>
Vycor 7913	0.26–2.7	1.458(0.589)	—	1200
As <sub>2</sub> S <sub>3</sub>	0.6–0.13	2.84(1.0), 2.4(8)	109	300
RIR 2	$\approx 0.4$ –4.7	1.75(2.2)	$\approx 600$	$\approx 900$
RIR 20	$\approx 0.4$ –5.5	1.82(2.2)	542	760
NaF	0.13–12	1.393(0.185), 0.24(24)	60	980
RIR 12	$\approx 0.4$ –5.7	1.62(2.2)	594	$\approx 900$
MgO	0.25–8.5	1.71(2.0)	692	2800
Acrylic	0.340–1.6	1.5066(0.4101), 1.4892(0.6563)	—	Distorts at 72
Silver chloride (AgCl)	0.4–32	2.134(0.43), 1.90149(20.5)	9.5	455
Silver bromide (AgBr)	0.45–42	2.313(0.496), 2.2318(0.671)	$\geq 9.5$	432
Kel-F	0.34–3.8	—	—	—
Diamond (type IIA)	0.23–200+	2.7151(0.2265), 2.4237(0.5461)	5700–10400 <sup>c</sup>	
NaCl	0.21–25	1.89332(0.185), 1.3403(22.3)	18	803
KBr	0.205–25	1.55995(0.538), 1.46324(25.14)	7	730
KCl	0.18–30	1.78373(0.19), 1.3632(23)	—	776
CsCl	0.19– $\approx 30$	1.8226(0.226), 1.6440(0.538)	—	646
CsBr	0.21–50	1.75118(0.365), 2.55990(39.22)	19.5	636
KI	0.25–40	2.0548(0.248), 1.6381(1.083)	5	723
CsI	0.235–60	1.98704(0.297), 1.61925(53.12)	—	621
SrTiO <sub>3</sub>	0.4–7.4	2.23(2.2), 2.19(4.3)	620	2080
SrF <sub>2</sub>	0.13–14	1.438(0.538)	130	1450
Rutile (TiO <sub>2</sub> )	0.4–7	$n_o = 2.5(1.0)$ , $n_e = 2.7(1.0)^f$	880	1825
Thallium bromide (TlBr)	0.45–45	2.652(0.436), 2.3(0.75)	12	460
Thallium bromiodide (KRS-5)	0.56–60	2.62758(0.577), 2.21721(39.38)	40	414.5
Thallium chloro bromide (KRS-6)	0.4–32	2.3367(0.589), 2.0752(24)	39	423.5
ZnSe	0.5–22	2.40(10.6)	150	—
Irtran 2 (ZnS)	0.6–15.6	2.26(2.2), 2.25(4.3)	354	800
Si	1.1–15 <sup>e</sup>	3.42(5.0)	1150	1420
Ge	1.85–30 <sup>e</sup>	4.025(4.0), 4.002(12.0)	692	936
GaAs	1–15	3.5(1.0), 3.135(10.6)	750	1238
CdTe	0.9–16	2.83(1.0), 2.67(10.6)	45	1045
Te	3.8–8+	$n_o = 6.37(4.3)$ , $n_e = 4.93(4.3)^f$	—	450
CaCO <sub>3</sub>	0.25–3	$n_o = 1.90284(0.200)$ , $n_e = 1.57796(0.198)^f$ $n_o = 1.62099(2.172)$ , $n_e = 1.47392(3.324)$	—	—
			135	894.4 <sup>b</sup>

Table 4.5 (contd.)

<i>Material</i>	<i>Thermal-Expansion Coefficient (<math>10^{-6}/K</math>)</i>	<i>Solubility in Water [g/(100 g), 20°C]</i>	<i>Soluble in</i>	<i>Comments</i>
LiF	9	0.27	HF	Scratches easily
MgF <sub>2</sub>	16	$7.6 \times 10^{-3}$	HNO <sub>3</sub>	—
CaF <sub>2</sub>	25	$1.1 \times 10^{-3}$	NH <sub>4</sub> salts	Not resistant to thermal or mechanical shock
BaF <sub>2</sub>	26	0.12	NH <sub>4</sub> Cl	Slightly hygroscopic, sensitive to thermal shock
Al <sub>2</sub> O <sub>3</sub>	6.66, <sup>a</sup> 5.0 <sup>b</sup>	$9.8 \times 10^{-5}$	NH <sub>4</sub> salts	Very resistant to chemical attack, excellent material
SiO <sub>2</sub>	0.55	0.00	HF	Excellent material
Pyrex	3.25	0.00	HF, hot H <sub>2</sub> PO <sub>4</sub>	Excellent mechanical, optical properties
Vycor	0.8	0.00	HF, hot H <sub>2</sub> PO <sub>4</sub>	Excellent mechanical, optical properties
As <sub>2</sub> S <sub>3</sub>	26	$5 \times 10^{-5}$	Alcohol	Nonhygroscopic
RIR 2	8.3	0.00	1% HNO <sub>3</sub>	Good mechanical, optical properties
RIR 20	9.6	—	—	Good mechanical, optical properties
NaF	36	4.2	HF	Lowest ref. index of all known crystals
RIR 12	8.3	—	—	Good mechanical and optical properties
MgO	43	$6.2 \times 10^{-4}$	NH <sub>4</sub> salts	Nonhygroscopic; surface scum forms if stored in air
Acrylic	110–140	0.00	Methylene chloride	Easily scratched, available in large sheets
AgCl	30	$1.5 \times 10^{-4}$	NH <sub>4</sub> OH, KCN	Corrosive, nonhygroscopic, cold-flows
AgBr	—	$12 \times 10^{-6}$	KCN	Cold-flows
Kel-F	—	—	—	Soft, easily scratched
Diamond	0.8	0.00	—	Hardest material known; thermal conductivity $6 \times$ Cu at room temp., chemically inert
NaCl	44	36	H <sub>2</sub> O, glycerine	Corrosive, hygroscopic
KBr	—	65.2	H <sub>2</sub> O, glycerine	Hygroscopic
KCl	—	34.35	H <sub>2</sub> O, glycerine	Hygroscopic
CsCl	—	186	Alcohol	Very hygroscopic
CsBr	48	124	H <sub>2</sub> O	Soft, easily scratched, hygroscopic
KI	—	144.5	Alcohol	Soft, easily scratched, hygroscopic, sensitive to thermal shock
CsI	50	160	Alcohol	Soft, easily scratched, very hygroscopic
SrTiO <sub>3</sub>	9.4	—	—	Refractive index $\approx \sqrt{5}$
SrF <sub>2</sub>	—	$1.17 \times 10^{-2}$	Hot HCl	Slightly sensitive to thermal shock
TiO <sub>2</sub>	9	0.00	H <sub>2</sub> SO <sub>4</sub>	Nonhygroscopic, nontoxic
TlBr	—	0.0476	Alcohol	Flows under pressure, toxic
KRS-5	51	<0.0476	HNO <sub>3</sub> , aqua regia	Cold-flows, nonhygroscopic, toxic
KRS-6	60	0.32	HNO <sub>2</sub> , aqua regia	Cold-flows, nonhygroscopic, toxic
ZnSe	8.5	0.001	—	Very good infrared materials, also transparent to some visible light
ZnS	—	$6.5 \times 10^{-5}$	HNO <sub>3</sub> , H <sub>2</sub> SO <sub>4</sub>	—
Si	4.2	0.00	HF + HNO <sub>3</sub>	Resistant to corrosion, must be highly polished
Ge	5.5	0.00	Hot H <sub>2</sub> SO <sub>4</sub> , aqua regia	to reduce scattering losses at surface

Table 4.5 (contd.)

Material	Thermal-Expansion Coefficient ( $10^{-6}/K$ )	Solubility in Water [ $g/(100\text{ g}), 20^{\circ}C$ ]	Soluble in	Comments
GaAs	5.7	0.00	—	Very good high-power infrared-laser window material
CdTe	4.5	—	—	—
Te	16.8	0.00	H <sub>2</sub> SO <sub>4</sub> , HNO <sub>3</sub>	Poisonous, soft, easily scratched
CaCO <sub>3</sub>	—	$1.4 \times 10^{-3}$	Acids, NH <sub>4</sub> Cl	—
	—	—	—	—

<sup>a</sup> Parallel to c-axis.

<sup>b</sup> Perpendicular to c-axis.

<sup>c</sup> Depends on crystal orientation.

<sup>d</sup> Depends on grade.

<sup>e</sup> Long-wavelength limit depends on purity of material.

<sup>f</sup> Birefringent.

<sup>g</sup> Softening temperature.

is best. Stretch the pad tight on the surface and fix it securely at the edges with adhesive tape. Dampen an area of the pad with a suitable solvent such as trichloroethylene or alcohol. Place the window to be polished on the solvent-soaked area, and work it back and forth in a figure-eight motion. Gradually work the disc onto the dry portion of the pad. This operation can be repeated as many times as necessary to restore the surface of the window. Do not attempt to clean windows of this kind, which are hygroscopic, by wiping them with solvent-soaked tissue. The evaporation of solvent from the surface will condense water vapor onto it and cause damage. The above recommended cleaning procedure is quite satisfactory for cleaning even laser windows. Small residual surface blemishes and slightly foggy areas do not detract significantly from the transmission of such windows in the middle infrared and beyond – which in any case is the only spectral region where they should be used. For removing slightly larger blemishes and scratches from such soft windows, very fine aluminum oxide powder (available from Adolf Meller) can be mixed with the cleaning solvent on the pad in the first stages of polishing. Perhaps the best way of all for cleaning hard and medium-hard optical components is to vapor-degrease them. This is a very general procedure for cleaning precision components. Place some trichloroethylene or isopropyl

alcohol in a large beaker (500 ml or larger) to a depth of about 1 cm. Suspend the components to be cleaned in the top of the beaker so that their critical surfaces face downward. An improvised wire rack is useful for accomplishing this. Cover the top of the beaker tightly with aluminum foil. Place the beaker on an electric hot plate, and heat the solvent until it boils. The solvent vapor thus produced is very pure. It condenses on, and drips off, the suspended item being cleaned, effectively removing grease and dirt. If large items are being cleaned, it may be necessary to blow air on the aluminum foil with a small fan. At the end of several minutes turn off the heat and remove the clean items while they are still warm. This cleaning procedure is recommended for components that must be extremely clean, such as laser Brewster windows.

## 4.4 OPTICAL MATERIALS

The choice of materials for the various components of an optical system such as windows, prisms, lenses, mirrors, and filters is governed by several factors:

- (1) Wavelength range to be covered
- (2) Environment and handling that components must withstand

- (3) Refractive-index considerations
- (4) Intensity of radiation to be transmitted or reflected
- (5) Cost.

#### 4.4.1 Materials for Windows, Lenses, and Prisms

For the purposes of classification, the characteristics of various materials for use in transmissive applications will be considered in three spectral regions:

- (1) Ultraviolet, 100–400 nm
- (2) Visible and near infrared, 400 nm–2  $\mu\text{m}$
- (3) Middle and far infrared, 2–1000  $\mu\text{m}$ .

There are, of course, many materials that can be used in part of two or even three of these regions. Table 4.5 summarizes the essential characteristics of all the common, and most of the rarely used, materials in these three regions. The useful transmission range given for each material is only a guide; the transmission at the ends of this range can be increased and the useful wavelength range of the component extended by using thinner pieces of material – if this is possible. The refractive index of all these materials varies with wavelength, as is illustrated for several materials in Figure 4.91. The Knoop hardness<sup>53</sup> is a static measure of material hardness based on the size of impression made in the material with a pyramidal diamond indenter under specific conditions of loading, time, and temperature. Roughly speaking, a material with a Knoop hardness above about 60 is hard enough to withstand the cleaning procedures described in Section 4.3.15.

Transmission curves for many of the useful optical materials summarized in Table 4.5 are collected together in Figures 4.92 – 4.101. Most of these materials are available from several suppliers. Some of the materials listed in Table 4.5 are worthy of brief extra comment.

**Ultraviolet Transmissive Materials.** The following are suitable for use in the UV region.

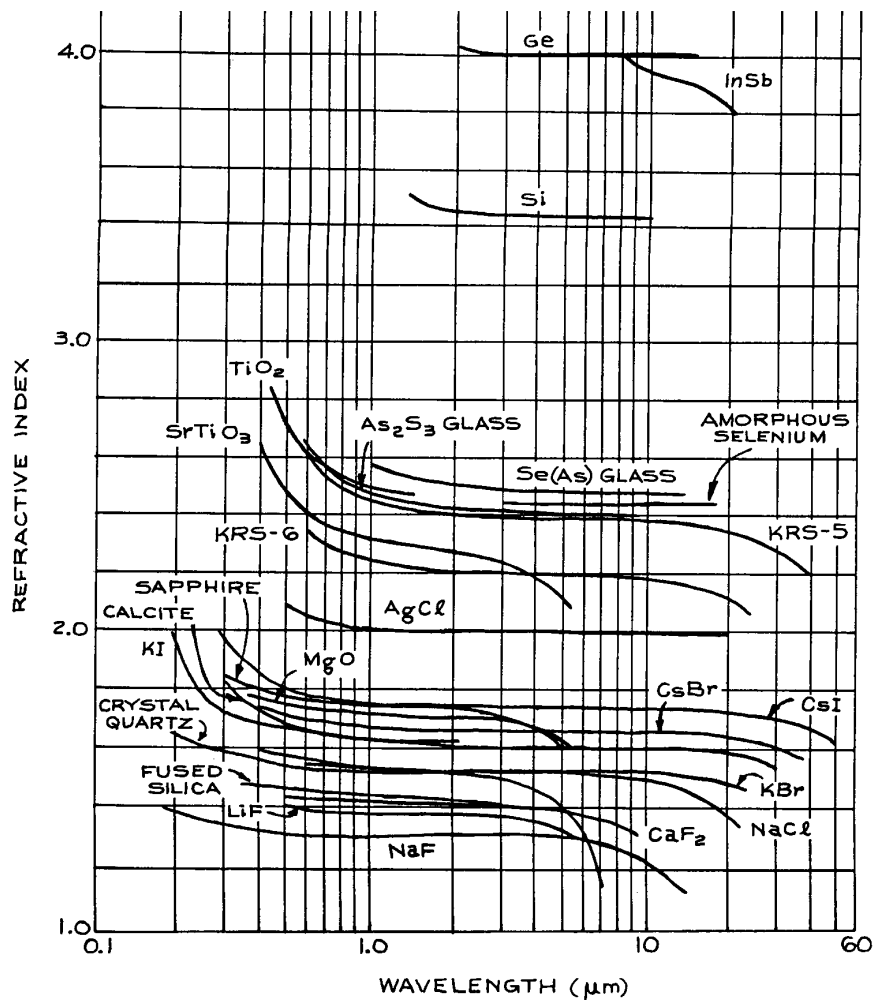
**(i) Lithium fluoride.** Lithium fluoride (LiF, Figure 4.92) has useful transmittance further into the vacuum ultraviolet than any other common crystal. The transmission of vacuum-ultraviolet quality crystals is more than 50% at 121.6 nm for 2 mm thickness, and thinner crystals have useful

transmission down to 104 nm. The short-wavelength transmission of the material deteriorates on exposure to atmospheric moisture or high-energy radiation. Moisture does not affect the infrared transmission, which extends to 7  $\mu\text{m}$ . Several grades of LiF are available. Vacuum-ultraviolet-grade material is soft, but visible-grade material is hard. LiF can be used as a vacuum-tight window material by sealing with either silver chloride or a suitable epoxy.

**(ii) Magnesium fluoride.** Vacuum-ultraviolet-grade magnesium fluoride ( $\text{MgF}_2$ , Figure 4.92) transmits farther into the ultraviolet than any common material except LiF. The transmission at 121.6 nm is 35% or more for a 2 mm thickness.  $\text{MgF}_2$  is recommended for use as an ultraviolet-transmissive component in space work, as it is only slightly affected by ionizing radiation.  $\text{MgF}_2$  is birefringent and is used for making polarizing components in the ultraviolet. Irtran 1 is a polycrystalline form of magnesium fluoride manufactured by Kodak, which is not suitable for ultraviolet or visible applications. Its useful transmission extends from 500 nm to 9  $\mu\text{m}$ .

**(iii) Calcium fluoride.** Calcium fluoride ( $\text{CaF}_2$ , Figure 4.92) is an excellent, hard material that can be used for optical components from 125 nm to beyond 10  $\mu\text{m}$ . Vacuum-ultraviolet-grade material transmits more than 50% at 125.7 nm for a 2 mm path length. Calcium fluoride is not significantly affected by atmospheric moisture at ambient temperature. Calcium fluoride lenses are available from Argus, Janos Technology, Linos Photonics, Meller Optics, Melles Griot, Newport, Optovac, Newport/Oriel and Precision Optical. Irtran 3 is a polycrystalline form of  $\text{CaF}_2$ , which is not suitable for ultraviolet or visible applications, but transmits in the infrared to 11.5  $\mu\text{m}$ .

**(iv) Barium fluoride.** Although barium fluoride ( $\text{BaF}_2$ , Figure 4.93) is slightly more water soluble than calcium or magnesium fluoride, it is more resistant than either of these to ionizing radiation. It is a good general-purpose optical material from ultraviolet to infrared. It has excellent transmission to beyond 10  $\mu\text{m}$ , and windows that are not too thick (< 2 mm) can be used in  $\text{CO}_2$  laser applications, except at very high energy densities. Barium fluoride lenses are available from Infrared Optical Products and International Scientific Products.

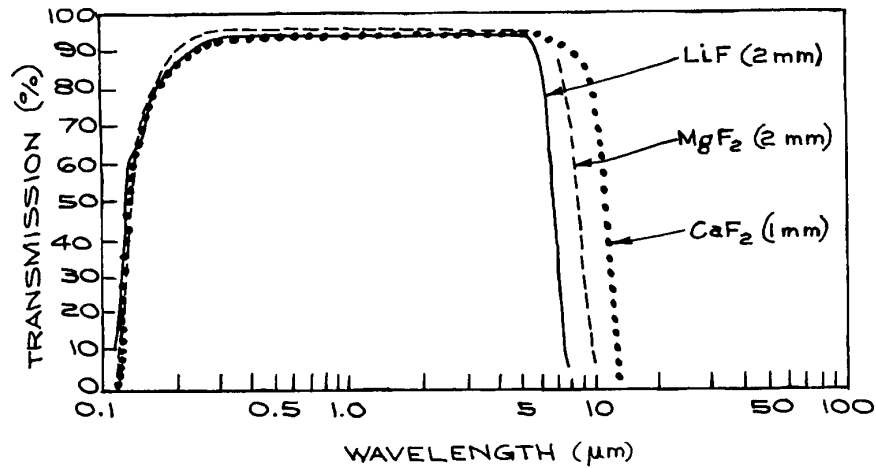


**Figure 4.91** The refractive indices of various optical materials. [Adapted, with permission, from W. C. Wolfe and S. S. Ballard, *Optical Materials, Films and Filters for Infrared Instrumentation*, Proceedings of the IRE, **47**, 1540-1546, 1959. © 1959 IRE (now IEEE).]

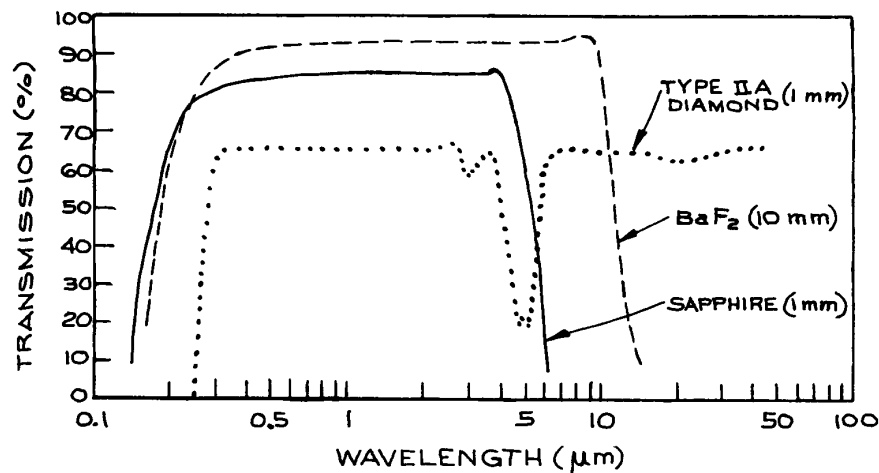
(v) **Synthetic sapphire.** Synthetic sapphire (corundum,  $\text{Al}_2\text{O}_3$ ; Figure 4.93) is probably the finest optical material available in applications from 300 nm to 4  $\mu\text{m}$ . It is very hard, strong, and resistant to moisture and chemical attack (even HF below 300 °C). The useful transmissive range extends from 150 nm to 6  $\mu\text{m}$ , the transmission of a 1 mm thickness being 21% and 34%, respectively, at these two

wavelengths. Sapphire optical components are available from many suppliers.<sup>31</sup>

Sapphire is also resistant to ionizing radiation, has high thermal conductivity, and can be very accurately fabricated in a variety of forms. It has low dispersion, however, and is not very useful as a prism material. Chromium-doped  $\text{Al}_2\text{O}_3$  (ruby) is used in laser applications.



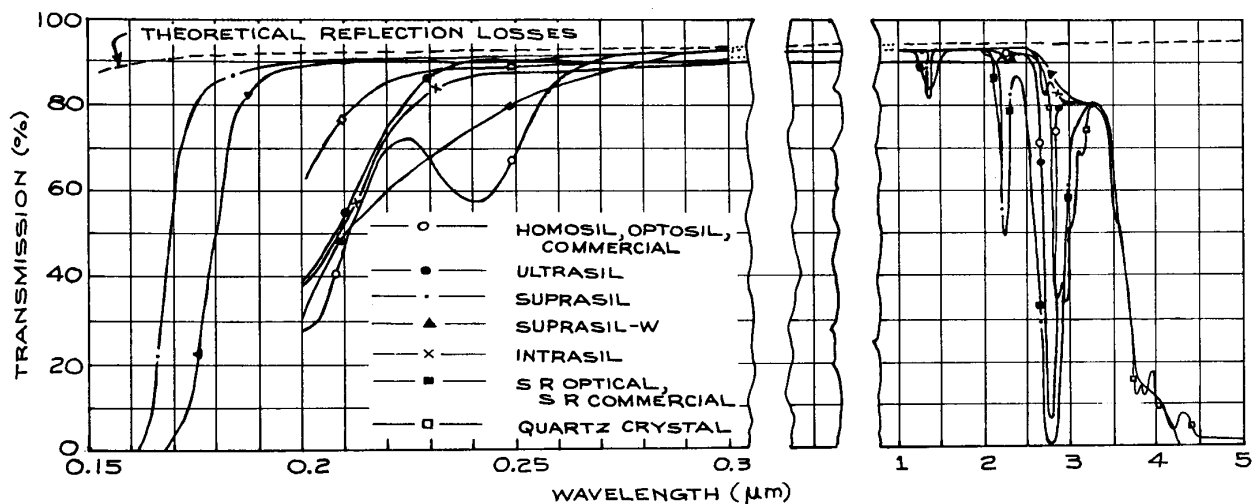
**Figure 4.92** Transmission curves of lithium fluoride, magnesium fluoride, and calcium fluoride windows of specified thicknesses.



**Figure 4.93** Transmission curves of sapphire, barium fluoride, and type-IIA diamond windows of specified thicknesses.

(vi) **Fused silica.** Fused silica ( $\text{SiO}_2$ , Figure 4.94) is the amorphous form of crystalline quartz. It is almost as good an optical material as sapphire and is much cheaper. Its useful transmission range extends from about 170 nm to about 4.5  $\mu\text{m}$ . Special ultraviolet-transmitting grades are manufactured (such as Spectrosil and Suprasil), as well as infrared grades (such as Infrasil) from which undesirable absorption features at 1.38, 2.22, and 2.72  $\mu\text{m}$ , due to

residual OH radicals, have been removed. Fused silica has a very low coefficient of thermal expansion,  $5 \times 10^{-7}/\text{K}$  in the temperature range from 20 to 900  $^\circ\text{C}$ , and is very useful as a spacer material in applications such as Fabry–Perot etalons and laser cavities. Fused-silica optical components such as windows, lenses, prisms, and etalons are readily available. Crystalline quartz is birefringent and is widely used in the manufacture of



**Figure 4.94** Transmission characteristics of various grades of fused silica; all measurements made on 10 mm thick materials. (Courtesy of Heraeus-Amersiol, Inc.)

polarizing optics, particularly quarter-wave and half-wave plates.

Very pure fused silica is also the material of which most optical fibers are made. In fiber optic applications the index gradient inside the fiber is generally produced by germania ( $\text{GeO}_2$ ) doping. The pure silica used in optical fibers is most transparent in the infrared, generally near  $1.55 \mu\text{m}$ .

**(vii) Diamond.** Windows made of type-IIA diamond (Figure 4.93) have a transmittance that extends from 230 nm to beyond  $200 \mu\text{m}$ , although there is some absorption between 2.5 and  $6.0 \mu\text{m}$ . The high refractive index of diamond,  $n = 2.4$ , over a very wide band, leads to a reflection loss of about 34% for two surfaces. Diamond windows are extremely expensive, (\$10,000–20,000 for a 1 cm diameter, 1 mm thick window); however, some properties of diamond are unique: its thermal conductivity is six times that of pure copper at  $20^\circ\text{C}$ , and it is the hardest material known, very resistant to chemical attack, with a low thermal-expansion coefficient and high resistance to radiation damage.

**Visible- and Near-Infrared-Transmissive Materials.** the materials so far considered as ultraviolet-transmitting are also excellent for use in the

visible and near infrared. There are, however, several materials that transmit well in the visible and near infrared, but are not suitable for ultraviolet use.

**(i) Glasses.** There is an extremely large number of types of glass used in the manufacture of optical components. They are available from many manufacturers, such as Chance, Corning, Hoya, O'Hara, and Schott.

Two old classifications of glass into *crown* and *flint* glasses can be related to the glass chart. Crown glasses are glasses with a  $V$  value greater than 55 if  $n_d < 1.6$  and  $V > 50$  if  $n_d > 1.6$ . The flint glasses have  $V$  values below these limits. The rare earth glasses contain rare earths instead of  $\text{SiO}_2$ , which is the primary constituent of crown and flint glasses. Glasses containing lanthanum contain the symbol La.

A few common glasses are widely used in the manufacture of lenses, windows, prisms, and other components. These include the borosilicate glasses Pyrex, BK7/A, and crown (BSC), as well as Vycor, which is 96% fused quartz. Transmission curves for Pyrex No. 7740 and Vycor No. 7913 are shown in Figure 4.95. These glasses are not very useful above  $2.5 \mu\text{m}$ , their transmission at  $2.7 \mu\text{m}$  is down to 20% for a 10 mm thickness. In the spectral region between 350 nm and  $2.5 \mu\text{m}$ , however, they are excellent



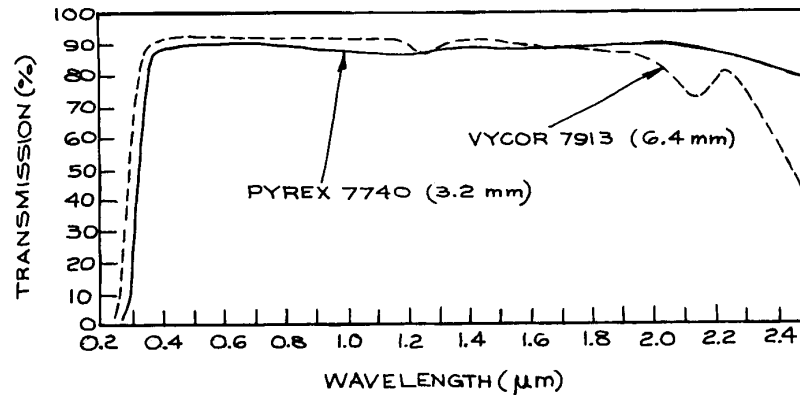


Figure 4.95 Transmission curves for Pyrex 7740 and Vycor 7913.

transmissive materials. They are inexpensive, hard, and chemically resistant, and can be fabricated and polished to high precision. Special glasses for use further into the infrared are available from Hoya and Corning. In window applications these glasses do not appear to offer any particular advantages over more desirable materials such as infrared-grade fused silica or sapphire; however, in specialized applications, such as infrared lens design, their higher refractive indices are useful.

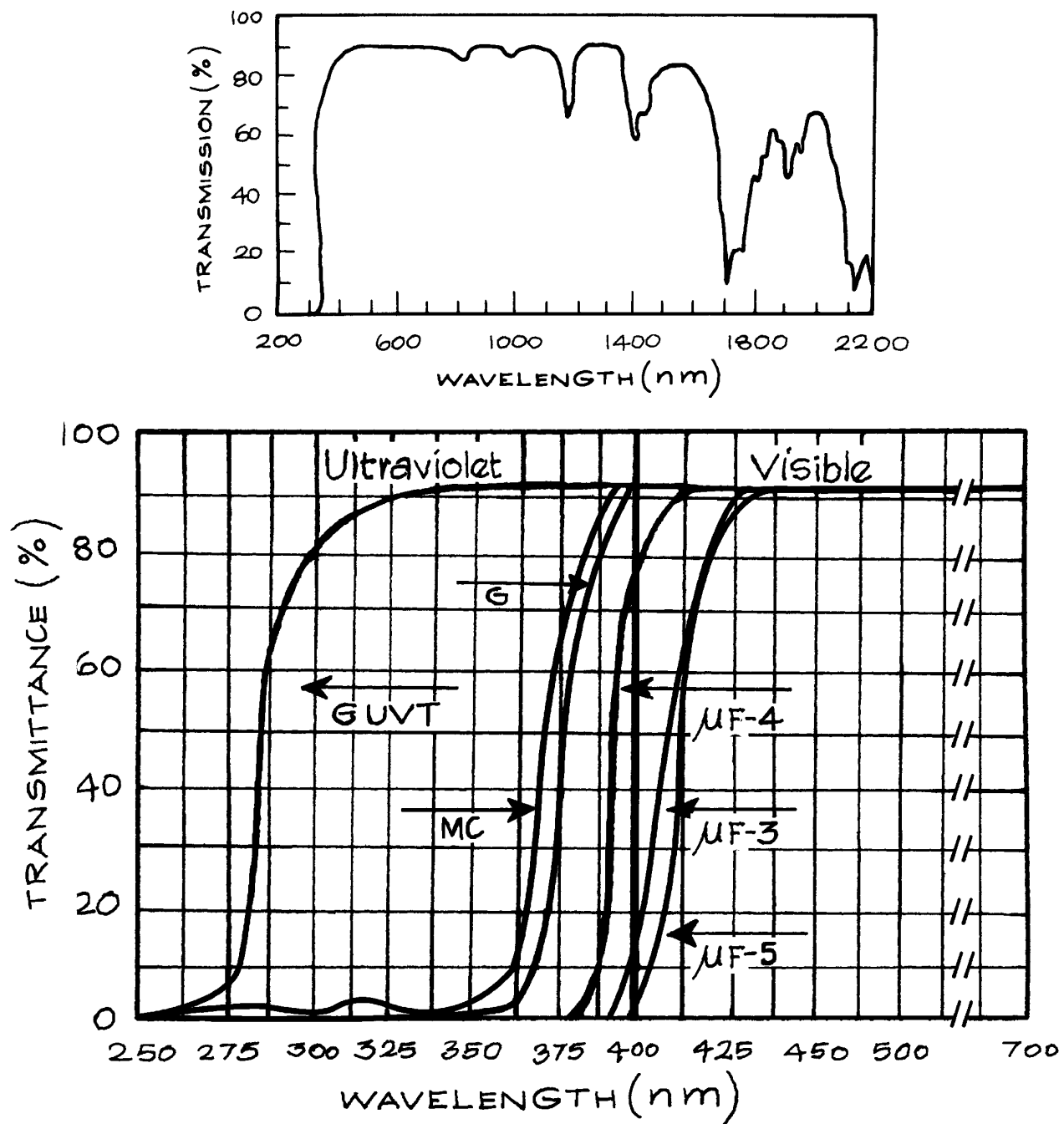
For special applications, many colored filter glasses can be obtained. Glasses that transmit visible but not infrared, or vice versa, or that transmit some near ultraviolet, but no visible, and many other combinations are available. The reader should consult the catalogs published by the various manufacturers and suppliers such as Chance-Pilkington, Corning, CVI, Hoya, Kopp Glass, Melles-Griot (now CVI-Melles-Griot), NPræzisions Glas & Optik, Schott, and Rolyn.

Special glasses whose expansion coefficients match selected metals can be made because the composition of glass is continuously variable. These glasses are useful for sealing to the corresponding metals in order to make windows on vacuum chambers. Corning glass No. 7056 is often used in this way, as it can be sealed to the alloy Kovar. Vacuum-chamber windows of quartz or sapphire are also available,<sup>54</sup> but their construction is complex, and consequently they are costly. Various grades of flint glass are available that have high refractive indices and are therefore useful for prisms. For a tabulation of the types

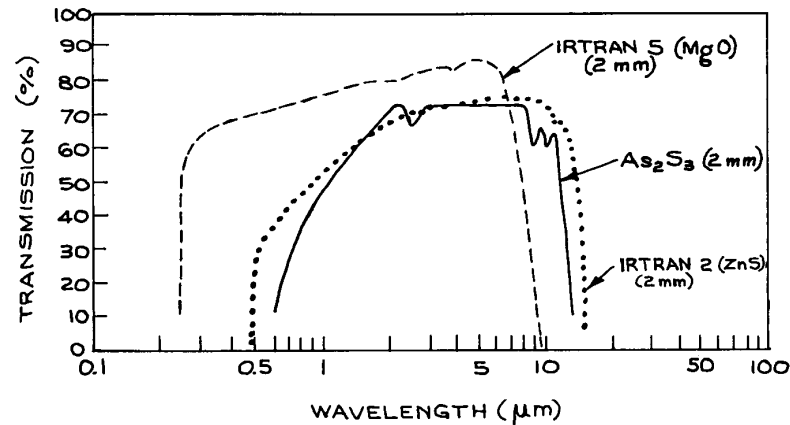
available the reader is referred to Kaye and Laby.<sup>7</sup> The refractive indices available range up to 1.93 for Chance-Pilkington Double Extra Dense Flint glass No. 927210, but this and other high-refractive-index glasses suffer long-term damage such as darkening when exposed to the atmosphere.

**(ii) Plastics.** In certain noncritical applications, such as observation windows, transparent plastics such as polymethylmethacrylate (Acrylic, Plexiglas, Lucite, Perspex) or polychlorotrifluoroethylene (Kel-F) are cheap and satisfactory. They cannot be easily obtained in high optical quality (with very good surface flatness, for example) and are easily scratched. The useful transmission range of Acrylic windows 1 cm thick runs from about 340 nm to 1.6  $\mu\text{m}$ . A transmission-versus-wavelength curve for this material is shown in Figure 4.96. One important application of Acrylic is in the manufacture of Fresnel lenses (see Section 4.3.3). Kel-F (poly-chloro-trifluoroethylene) is useful up to about 3.8  $\mu\text{m}$ , although it shows some decrease in transmission near 3  $\mu\text{m}$ . It is similar to polytetrafluoroethylene (Teflon) and is very resistant to a wide range of chemicals – even gaseous fluorine.

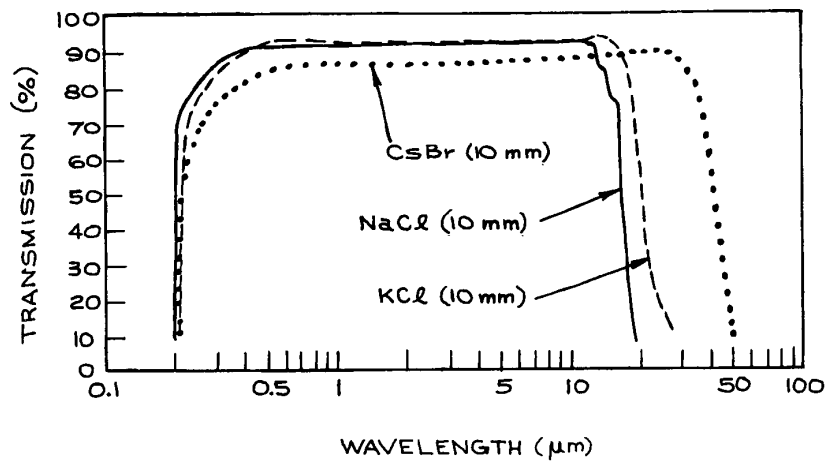
**(iii) Arsenic trisulfide.** Arsenic trisulfide ( $\text{As}_2\text{S}_3$ , Figure 4.97) is a red glass that transmits well from 800 nm to 10  $\mu\text{m}$ . Its usefulness stems from its relatively low price, ease of fabrication, nontoxicity, and resistance to moisture. Although primarily an infrared-transmissive material, it



**Figure 4.96** (a) Transmission curve for 3.175 mm thick Acrylic. (Courtesy of Melles Griot, Inc.); (b) Transmission curve of various grades of acrylic available from Altumax International.



**Figure 4.97** Transmission curves for Irtran 2 (ZnS), Irtran 5 (MgO), and  $\text{As}_2\text{S}_3$  windows of specified thicknesses.



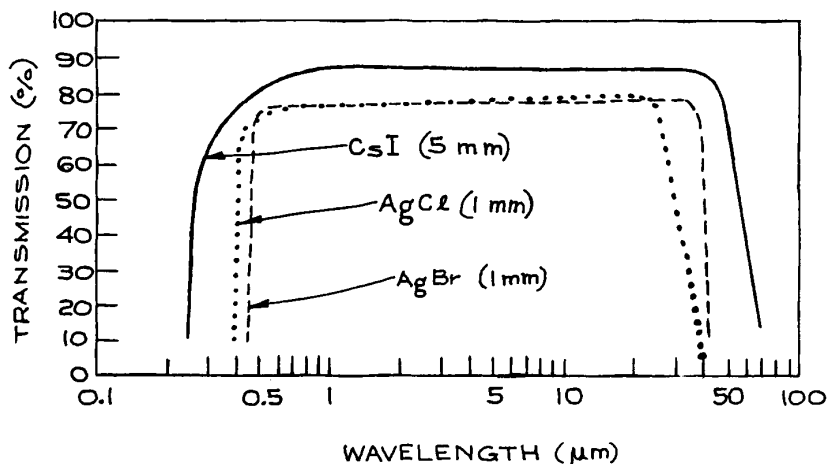
**Figure 4.98** Transmission curves for cesium bromide, potassium chloride, and sodium chloride windows of specified thickness.

is also transmissive in the red, which facilitates the alignment of optical systems containing it.  $\text{As}_2\text{S}_3$  is relatively soft and will cold-flow over a long period of time. A wide range of  $\text{As}_2\text{S}_3$  lenses are available from Amorphous Materials, OFR (Thorlabs), Recon Optical, and REFLEX Analytical.

#### **Middle- and Far-Infrared-Transmissive Materials.**

The following are useful well into the IR region.

(i) **Sodium chloride.** Sodium chloride (NaCl, Figure 4.98) is one of the most widely used materials for infrared-transmissive windows. Although it is soft and hygroscopic, it can be repolished by simple techniques and maintained exposed to the atmosphere for long periods without damage simply by maintaining its temperature higher than ambient. Two simple ways of doing this are to mount a small tungsten-filament bulb near the window, or to run a heated wire around the periphery of small



**Figure 4.99** Transmission curves for cesium iodide, silver bromide, and silver chloride windows of specified thicknesses.

windows. High-precision sodium chloride windows are in widespread use as Brewster windows in high-energy, pulsed CO<sub>2</sub> laser systems.

**(ii) Potassium chloride.** Potassium chloride (KCl, Figure 4.98) is very similar to sodium chloride. It is also an excellent material for use as a Brewster window in pulsed CO<sub>2</sub> lasers, as its transmission at 10.6 μm is slightly greater than that of sodium chloride.

**(iii) Cesium bromide.** Cesium bromide (CsBr, Figure 4.98) is soft and extremely soluble in water, but has useful transmission to beyond 40 μm. It will suffer surface damage if the relative humidity exceeds 35%.

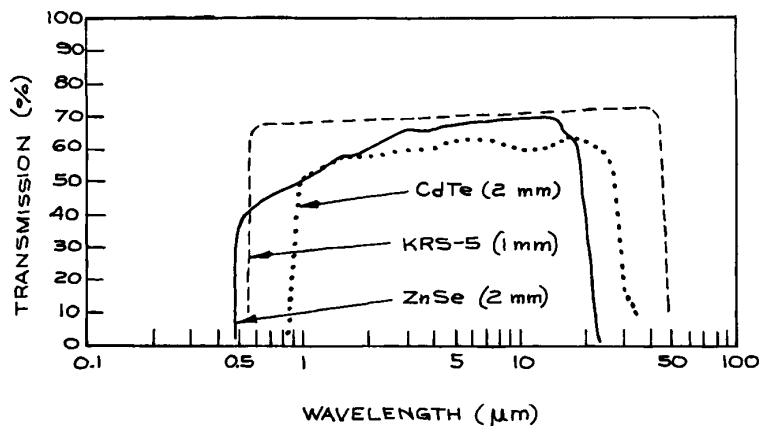
**(iv) Cesium iodide.** Cesium iodide (CsI, Figure 4.99) is used for both infrared prisms and windows; it is similar to CsBr, but its useful transmission extends beyond 60 μm.

**(v) Thallium bromoiodide (KRS-5).** Thallium bromoiodide (Figure 4.100), widely known as KRS-5, is an important material because of its wide transmission range in the infrared – from 600 nm to beyond 40 μm for a 5-mm thickness – and its low solubility in water. KRS-5 will survive atmospheric exposure for long periods of time and can even be used in liquid cells in direct contact with aqueous solutions. Its refractive index is high: reflection

losses at 30 μm are 30% and it has a tendency to cold-flow. KRS-5 lenses are available from Almaz Optics, Infrared Optical Products, ISP Optics, and Janos Technology.

**(vi) Zinc selenide.** Chemical-vapor-deposited zinc selenide (ZnSe, Figure 4.100) is used widely in infrared-laser applications. It is transmissive from 500 nm to 22 μm. Zinc selenide is hard enough that it can be cleaned easily. It is resistant to atmospheric moisture and can be fabricated into precision windows and lenses. Its transmissive qualities in the visible (it is orange-yellow in color) makes the alignment of infrared systems using it much more convenient than in systems using germanium, silicon, or gallium arsenide. Irtran 4 is a polycrystalline form of zinc selenide originally developed by Kodak, which is available from Harrick Scientific and Perkin Elmer.

**(vii) Plastic films.** Several polymer materials are sold commercially in film form, including polyethylene (Polythene, Polyphane, Poly-Fresh, Dinethene, etc.), polyvinylidene chloride copolymer (Saran Wrap), polyethylene terephthalate (Mylar, Melinex, Scotchpar), and polycarbonate (Lexan). These plastic materials can be used to wrap hygroscopic crystalline optical components to protect them from atmospheric moisture. If the film is wrapped tightly, it will not substantially affect the optical quality of the wrapped components in the infrared. Care



**Figure 4.100** Transmission curves for zinc selenide, cadmium telluride, and KRS-5 windows of specified thicknesses.

should be taken to use a polymer film that transmits the wavelength for which the protected component is to be used. For example, several commercial plastic food wraps can be used to wrap NaCl or KCl windows for use in CO<sub>2</sub>-laser applications. Not all will prove satisfactory, however, and a particular plastic film should be tested for transmission before use. A note of caution: plastic films contain plasticizers, the vapor from which may be a problem.

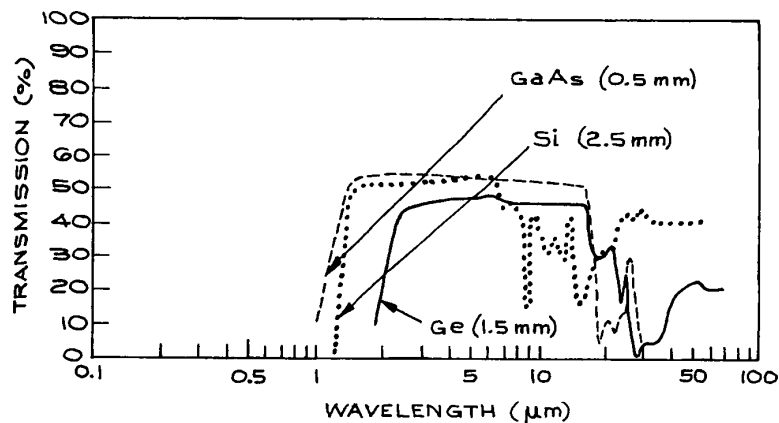
**(viii) Plastics and other materials for far-infrared use.**

Plastics are widely used for windows, beamsplitters, and light guides in the far infrared. A Mylar film is widely used as the beam-splitting element in far-infrared Michelson interferometers, and polyethylene, polystyrene, nylon, and Teflon are used as windows in far-infrared spectrometers and lasers. Black polyethylene excludes visible and near-infrared radiation, but transmits far-infrared. Lenses and stacked-sheet polarizers for far-infrared use can be made from polyethylene. Several crystalline and semiconducting materials, although they absorb in the middle infrared, begin to transmit again in the far infrared. A 1 mm thick quartz window has 70% transmission at 100 μm and more than 90% transmission from 300 to 1000 μm. The alkali and alkaline-earth halides such as NaF, LiF, KBr, KCl, NaCl, CaF<sub>2</sub>, SrF<sub>2</sub>, and BaF<sub>2</sub> show increasing transmission for wavelengths above a critical value that varies from one crystal to another but is in the 200–400 μm range. Further details of the specialized area of

far-infrared materials and instrumentation are available in the literature.<sup>55–57</sup>

**Semiconductor Materials.** Several of the most valuable infrared optical materials are semiconductors. In very pure form these materials should be transmissive to all infrared radiation with wavelengths longer than that corresponding to the band gap. In practice the long-wavelength transmission will be governed by the presence of impurities – a fact that is put to good use in the construction of infrared detectors. Semiconductors make very convenient “cold mirrors,” as they reflect visible radiation quite well but do not transmit it.

**(i) Silicon.** Silicon (Si, Figure 4.101) is a hard, chemically resistant material that transmits wavelengths beyond about 1.1 μm. It does, however, show some absorption near 9 μm, so it is not suitable for CO<sub>2</sub>-laser windows. It has a high refractive index (the reflection loss for two surfaces is 46% at 10 μm), but can be antireflection-coated. Because of its high refractive index it should always be highly polished to reduce surface scattering. Silicon has a high thermal conductivity and a low thermal expansion coefficient, and it is resistant to mechanical and thermal shock. Silicon mirrors can therefore handle substantial infrared-laser power densities. Silicon lenses are available from II-VI Infrared, ISP Optics, Lambda Research Optics, Laser Research Optics, Meller Optics, and V.A. Optical.



**Figure 4.101** Transmission curves for the semiconductor materials germanium, gallium arsenide, and silicon.

**(ii) Germanium.** Germanium (Ge, Figure 4.101) transmits beyond about 1.85  $\mu\text{m}$ . Very pure samples can be transparent into the microwave region. It is somewhat brittle, but is still one of the most widely used window and lens materials in infrared-laser applications. It is chemically inert and can be fabricated to high precision. Germanium has good thermal conductivity and a low coefficient of thermal expansion; it can be soldered to metal. In use, its temperature should be kept below 40  $^{\circ}\text{C}$ , as it exhibits thermal runaway – its absorption increases with increasing temperature. A wide range of germanium lenses are available from Lambda Research Optical, Laser Research Optics, Meller Optics, Umicore, and V.A. Optical.

**(iii) Gallium arsenide.** Gallium arsenide (GaAs, Figure 4.101) transmits from 1 to 15  $\mu\text{m}$ . It is a better, although more expensive, material than germanium, particularly in CO- and CO<sub>2</sub>-laser applications. It is hard and chemically inert, and maintains a very good surface finish. It can handle large infrared power densities because it does not exhibit thermal runaway until it reaches 250  $^{\circ}\text{C}$ . It is also used for manufacturing infrared-laser electro-optic modulators. Gallium arsenide lenses are available from Infrared Optical Products, ISP Optics, Lambda Research Optics, and REFLEX Analytical.

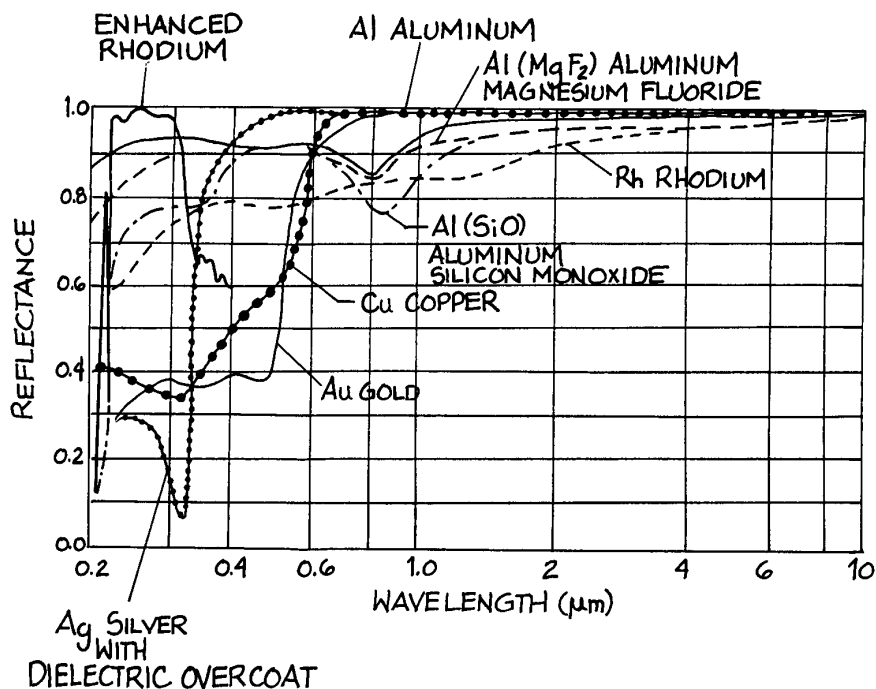
**(iv) Cadmium telluride.** Cadmium telluride (CdTe, Figure 4.100) is another excellent material for use between 1

and 15  $\mu\text{m}$ . It is quite hard, chemically inert, and takes a good surface finish. It is also used for manufacturing infrared-laser electro-optic modulators. Irtran 6 is a polycrystalline form of cadmium telluride available from Kodak; it transmits to 31  $\mu\text{m}$ . Cadmium telluride lenses are available from ISP Optics, and REFLEX Analytical.

**(v) Tellurium.** Tellurium is a soft material that is transparent from 3.3  $\mu\text{m}$  to beyond 11  $\mu\text{m}$ . It is not affected by water. Its properties are highly anisotropic; it can be used for second harmonic generation with CO<sub>2</sub> laser radiation. Tellurium should not be handled, as it is toxic and can be absorbed through the skin.

#### 4.4.2 Materials for Mirrors and Diffraction Gratings

**Metal Mirrors.** The best mirrors for general broadband use have pure metallic layers, vacuum-deposited or electrolytically deposited on glass, fused-silica, or metal substrates. The best metals for use in such reflective coatings are aluminum, silver, gold, and rhodium. Solid metal mirrors with highly polished surfaces made of metals such as stainless steel, copper, zirconium-copper, and molybdenum are also excellent, particularly in the infrared. The reflectance of a good evaporated metal coating exceeds that of the bulk metal. Flat solid-metal mirrors can be made

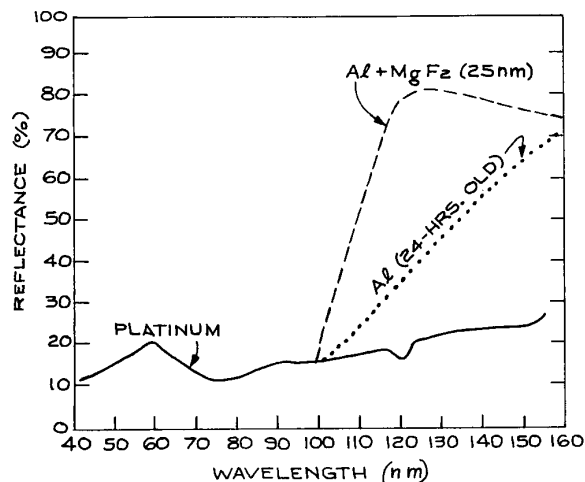


**Figure 4.102** Reflectance of freshly deposited films of aluminum, copper, gold, rhodium, and coated silver, aluminum and rhodium as a function of wavelength from 0.2 to 10  $\mu\text{m}$ .

fairly cheaply by machining the mirror blank, surface-grinding it, and finally polishing the surface. If precision polishing facilities are not available, there are numerous optical component suppliers who will undertake the polishing of such surface-ground blanks to flatnesses as good as  $\lambda/10$  in the visible. Finished flat and spherical solid-metal mirrors are available from II-VI Infrared, Opti-Forms, Space Optics Research Labs (SORL), and Spawr. Solid-metal mirrors are most valuable for handling high-power infrared laser beams when metal-coated mirrors cannot withstand the power dissipation in the slightly absorbing mirror surface. Water-cooled copper mirrors for handling very intense infrared laser beams are available from Spawr and Umicore Laser Optics Ltd.

The spectral reflectances in normal incidence of several common metal coatings are compared in Figure 4.102. Several points are worthy of note. Gold and copper are not good for use as reflectors in the visible region. Unpro-

tected gold, silver, copper, and aluminum are very soft. Aluminum rapidly acquires a protective layer of oxide after deposition; this significantly reduces the reflectance in the vacuum ultraviolet and contributes to increased scattering throughout the spectrum. Aluminum and gold are frequently supplied commercially with a thin protective dielectric layer, which increases their resistance to abrasion without significantly affecting their reflectance, and also protects the aluminum from oxidation. The protective layer is usually a  $\lambda/2$  layer (at 550 nm) of SiO for aluminum mirrors used in the visible region of the spectrum. Aluminum coated with a thin layer of MgF<sub>2</sub> can be used as a reflector in the vacuum ultraviolet, although the reflectance is substantially reduced below about 100 nm, as shown in Figure 4.103. Many optical instruments operating in the vacuum ultraviolet, including some spectrographs, have at least one reflective component. Above 100 nm, coated aluminum is almost always used for this



**Figure 4.103** Normal-incidence reflectance of platinum, aluminum with a 25 nm thick overcoat of MgF<sub>2</sub>, and unprotected aluminum 24 hours after film deposition.

reflector. Below 100 nm, platinum and indium have superior reflectance to aluminum. At very short wavelengths, below about 40 nm, the reflectance of platinum in normal incidence is down to only 10%, as shown in Figure 4.103. Therefore, reflective components in this wavelength region are generally used at grazing incidence, as the reflectance under these conditions can remain high. For example, at 0.832 nm the reflectance of an aluminum film at a grazing angle of 0.5° is above 90%.

Freshly evaporated silver has the highest reflectance of any metal in the visible; however, it tarnishes so rapidly that it is rarely used except on internal reflection surfaces. In this case the external surface is protected with a layer of Inconel or copper and a coat of paint.

**Multilayer Dielectric Coatings.** Extremely high-reflectance mirrors (up to 99.99%) can be made by using multilayer dielectric films involving alternate high- and low-refractive-index layers, ranging around  $\lambda/4$  thick, deposited on glass, metal, or semiconductor substrates. The layers are made from materials that are transparent in the wavelength region where high reflectance is required. Both narrow-band and broad-band reflective-coated mirrors are available commercially. Figure

4.34(a) shows an example of each. Multilayer or single-layer dielectric-coated mirrors having almost any desired reflectance at wavelengths from 150 nm to 20  $\mu\text{m}$  are also commercially available.<sup>31</sup>

**Substrates for Mirrors.** The main factors influencing the choice of substrate for a mirror are: (1) the dimensional stability required; (2) thermal dissipation; (3) mechanical considerations such as size and weight; and (4) cost. Glass is an excellent substrate for most totally reflective mirror applications throughout the spectrum, except where high thermal dissipation is necessary. Glass is inexpensive and strong, and takes a surface finish as good as  $\lambda/200$  in the visible. Fused silica and certain ceramic materials such as Zerodur (available from Schott Glass) are superior, but more expensive, alternatives to glass when greater dimensional stability and a lower coefficient of thermal expansion (CTE) are required. Fused silica has the lowest CTE,  $\sim 5 \times 10^{-7}/\text{K}$  of any readily available material. The ceramic Zerodur has a CTE that can be 10 times smaller. The alloy Superinvar has a CTE similar to quartz, but generally over a restricted temperature range. Machining or polishing of low CTE materials can produce stresses in them and increase their expansion coefficients. Metal mirrors are frequently used when a very high light flux must be reflected, but even the small absorption loss in the reflecting surface necessitates cooling of the substrate. Metals are, however, not so dimensionally stable as glasses or ceramics, and their use as mirrors should be avoided in precision applications, particularly in the visible or ultraviolet.

Partially reflecting, partially transmitting mirrors are used in many optical systems, such as interferometers and lasers. In this application the substrate for the mirror has to transmit in the spectral region being handled. The substrate material should possess the usually desirable properties of hardness and dimensional stability, plus the ability to be coated. For example, germanium, silicon, and gallium arsenide are frequently used as partially transmitting mirror substrates in the near infrared. Only the least expensive partially reflecting mirrors use thin metal coatings. Such coatings are absorbing, and multilayer, dielectric-coated reflective surfaces are much to be preferred.

There are numerous suppliers of totally and partially reflecting mirrors.<sup>31</sup> The experimentalist need only



determine the requirements and specify the mirror appropriately.

Mirrors that reflect visible radiation and transmit infrared (*cold mirrors*) or transmit visible radiation and reflect infrared (*hot mirrors*) are available as standard items from several suppliers.<sup>31</sup>

**Diffraction Gratings.** Diffraction gratings are generally ruled on glass, which may then be coated with aluminum or gold for visible or infrared reflective use, respectively. Many commercial gratings are replicas made from a ruled master, although high-quality holographic gratings are available from several suppliers.<sup>31</sup> For high-power infrared-laser applications, gold-coated replica gratings are unsatisfactory and master gratings ruled on copper should be used. Such gratings are available from Diffraction Products, Jobin-Yvon, and Richardson Grating Laboratory (though Newport).

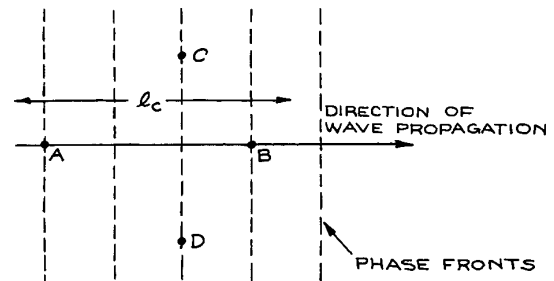
## 4.5 OPTICAL SOURCES

Optical sources fall into two categories: *incoherent sources* such as mercury or xenon arc, tungsten filament, and sodium lamps; and *coherent sources* (lasers). There are many different types of laser, of which some exhibit a high degree of coherence and others are not significantly more coherent than a line source, such as a low-pressure mercury lamp.

Incoherent sources fall into two broad categories: *line sources* and *continuum sources*. Line sources emit most of their radiation at discrete wavelengths, which correspond to strong spectral emission features of the excited atom or ion that is the emitting species in the source. *Continuum sources* emit over some broad spectral region, although their radiant intensity varies with wavelength.

### 4.5.1 Coherence

There are two basic coherence properties of an optical source. One is a measure of its relative spectral purity. The other depends on the degree to which the wavefronts coming from the source are uniphase or of fixed spatial phase variation. A wavefront is said to be *uniphase* if it has the same phase at all points on the wavefront. The Gaussian TEM<sub>00</sub> mode emitted by many lasers has this property.



**Figure 4.104** Temporal and spatial coherence.

The two types of coherence are illustrated by Figure 4.104. If the phases of the electromagnetic field at two points, *A* and *B*, longitudinally separated in the direction of propagation, have a fixed relationship, then the wave is said to be *temporally coherent* for times corresponding to the distance from *A* to *B*. The existence of such a fixed phase relationship could be demonstrated by combining waves extracted at points separated like *A* and *B* and producing an interference pattern. The Michelson interferometer (see Section 4.7.6) demonstrates the existence of such a fixed phase relationship. The maximum separation of *A* and *B* for which a fixed phase relationship exists is called the *coherence length* of the source  $l_c$ . The coherence length of a source is related to its *coherence time*  $\tau_c$  by:

$$l_c = c\tau_c \quad (4.191)$$

Conceptually  $l_c$  is the average length of the uninterrupted wave trains from the source between random phase interruptions. Fourier theory demonstrates that the longer in time a sinusoidal wave train is observed, the narrower is its frequency spread. Thus the *spectral width* of a source can be related to its coherence time by writing:

$$\Delta\nu \simeq 1/\tau_c \quad (4.192)$$

The most coherent conventional lamp sources are stabilized low-pressure mercury lamps, which can have  $\tau_c$  ranging up to about 10 ns. Lasers, however, can have  $\tau_c$  ranging up to at least 1 ms.

*Spatial coherence* is a measure of phase relationships in the wavefront transverse to the direction of wave propagation. In Figure 4.104, if a fixed phase relationship exists between points *C* and *D* in the wavefront, the wave is said to be spatially coherent over this region. The *area of*

*coherence* is a region of the wavefront within which all points have fixed phase relationships. Spatial coherence can be demonstrated by placing pinholes at different locations in the wavefront and observing interference fringes. Temporally incoherent sources can exhibit spatial coherence. Small sources (so-called point sources) fall into this category. The light from a star can be spatially coherent.

Extended, temporally incoherent sources have a low degree of spatial coherence because light coming from different parts of the source has different phases. Lasers emitting a TEM<sub>00</sub> (fundamental) mode have very good spatial coherence over their whole beam diameter.

#### 4.5.2 Radiometry: Units and Definitions

*Radiometry* deals with the measurement of amounts of light. In radiometric terms the characteristics of a light source can be specified in several ways.

*Radiant power*,  $W$ , measured in watts, is the total amount of energy emitted by a light source per second. The spectral variation of radiant power can be specified in terms of the radiant power density per unit wavelength interval,  $W_\lambda$ . Clearly:

$$W = \int_0^{\infty} W_\lambda d\lambda \quad (4.193)$$

If a light source emits radiation only for some specific duration – which may be quite short in the case of a flash-lamp – it is more useful to specify the source in terms of its *radiant energy output*,  $Q_e$ , measured in joules. If the source emits radiation for a time  $T$ , we can write:

$$Q_e = \int_0^T W(t) dt \quad (4.194)$$

The amount of power emitted by a source in a particular direction per unit solid angle is called *radiant intensity*  $I_e$ , and is measured in watts per steradian. In general:

$$W = I_e(\omega) d\omega \quad (4.195)$$

where the integral is taken over a closed surface surrounding the source. If  $I_e$  is the same in all directions, the source is said to be an isotropic radiator. At a distance  $r$  from such

a source, if  $r$  is much greater than the dimensions of the source, the radiant flux crossing a small area  $\Delta S$  is:

$$\phi_e = \frac{I_e \Delta S}{r^2} \quad (4.196)$$

The *irradiance* at this point, measured in W/m<sup>2</sup>, is:

$$E_e = \frac{I_e}{r^2} \quad (4.197)$$

which is equal to the average value of the Poynting vector measured at the point. The radiant flux emitted per unit area of a surface (whether this be emitting or merely reflecting and scattering radiation) is called the *radiant emittance*  $M_e$ , and it is measured in W/m<sup>2</sup>. For an extended source, the radiant flux emitted per unit solid angle per unit area of the source is called its *radiance* or *brightness*  $L_e$ :

$$L_e = \frac{\delta I_e}{\delta S_n} \quad (4.198)$$

where the area  $\delta S_n$  is the projection of the surface element of the source in the direction being considered. When the light emitted from a source or scattered from a surface has a radiance that is independent of viewing angle, the source or scatterer is called a perfectly diffuse or *Lambertian radiator*. Clearly, for such a source, the radiant intensity at an angle  $\theta$  to the normal to the surface is:

$$I_e(\theta) = I_e(0) \cos \theta \quad (4.199)$$

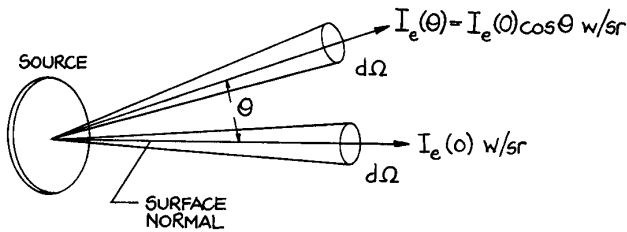
as illustrated in Figure 4.105.

The total flux emitted per unit area of such a surface is its radiant emittance, which in this case is:

$$M_e = \pi I_e(0) \quad (4.200)$$

Illuminated diffusing surfaces made of finely ground glass or finely powered magnesium oxide will behave as Lambertian radiators.

For plane waves, since all the energy in the wave is transported in the same direction, the concepts of radiant intensity and emittance are not useful. It is customary to specify the radiant flux crossing unit area normal to the direction of propagation, and call this the *intensity*  $I$  of the plane wave. Because lasers emit radiation into an extremely small solid angle, they have very high radiant intensity, and it is once again more usual to refer simply to the *intensity* of



**Figure 4.105** Radiant intensity characteristics of a Lambertian radiator.

the laser beam at a point as the energy flux per second per unit area. The total power output of a laser is:

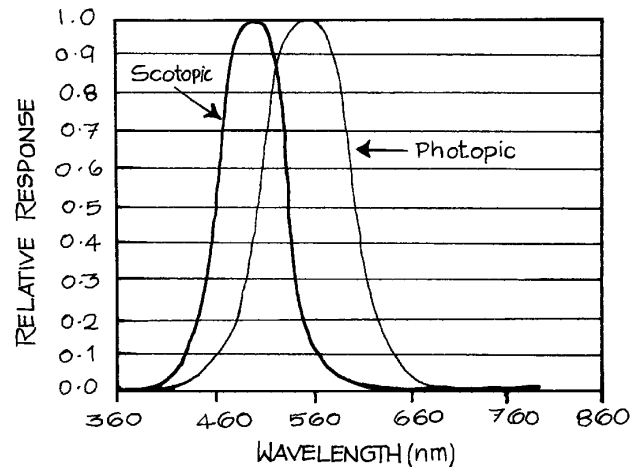
$$W = \int_{\text{beam}} I dS \quad (4.201)$$

### 4.5.3 Photometry

*Photometry* is the measurement of light that is in, or close to, the spectral region where the human eye responds. It is related to radiometry, but its units are weighted by the response of the human eye. The human eye provides a nonlinear and wavelength-dependent subjective impression of radiometric quantities. The response function of the human eye extends roughly from 360 to 700 nm, with a peak at 555 nm for the photopic (light-adapted) eye, where the *cones* are the dominant receptors. The dark adapted human eye exhibits the *scotopic response*, which comes from the light- (but not color-) sensitive *rods* in the periphery of the retina. The photopic and scotopic response are shown in Figure 4.106. Measures in photometry take into account the so-called photopic spectral luminous efficiency function  $V(\lambda)$ . Thus, for example, in physiological photometry the *luminous flux*  $F$  is related to radiant flux  $\phi_e(\lambda)$  by:

$$F = K \int_0^{\infty} V(\lambda) \phi_e(\lambda) d\lambda \quad (4.202)$$

where  $K$  is a constant. When  $F$  is measured in lumens and  $\phi_e(\lambda)$  in watts,  $K = 683 \text{ lumen/W}$ . In practice, the integral in equation (4.202) need only be evaluated between 360 nm and 830 nm, as  $V(\lambda)$  is zero outside this range.



**Figure 4.106** The photopic and scotopic spectral response of the human eye.

Other photometric quantities that may be encountered in specifications of light sources are as follows:

- (1) The *luminous intensity*  $I_v$ , measured as candela (Cd), where:

$$1 \text{ candela} = 1 \text{ lumen str}$$

This is the photometric equivalent of the radiometric quantity radiant intensity. The formal scientific specification of the candela is that it is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of wavelength 550 nm and that has a radiant intensity in that direction of 1/683 watt per steradian.

- (2) The *illuminance*,  $E_v$ , measured in lumen/m<sup>2</sup>  $\equiv$  lux; lumen/cm<sup>2</sup>  $\equiv$  phot; or lumen/ft<sup>2</sup>  $\equiv$  footcandle.
- (3) The *luminance* or *brightness*,  $L_v$ , measured in candela/m<sup>2</sup>  $\equiv$  nit; candela/cm<sup>2</sup>  $\equiv$  stilb; candela/ $\pi$  ft<sup>2</sup>  $\equiv$  footlambert; candela/ $\pi$  m<sup>2</sup>  $\equiv$  apostilb; or candela/ $\pi$  cm<sup>2</sup> lambert.
- (4) The *luminous energy*,  $Q_v$ , measured in lumen seconds  $\equiv$  talbot. The correspondences between radiometric and photometric quantities are summarized in Table 4.6.

For a Lambertian source the luminance is independent of the observation direction. Photometric description of the characteristics of light sources should be avoided in strict scientific work, but some catalogs of light sources use

**Table 4.6 Correspondence between radiometric and photometric quantities**

<i>Radiometric Quantity</i>	<i>Radiometric Unit</i>	<i>Photometric Quantity</i>	<i>Photometric Unit</i>
Radiant power	watt (W)	Power 66	lumen (lm)
Radiant emittance	W/m <sup>2</sup>	Illuminance	lumen/m <sup>2</sup> = lux (lx)
Radiant intensity	W/sr	Luminous intensity	lumen/sr = candela (cd)
Radiance (Brightness)	W/m <sup>2</sup> /sr <sup>1</sup>	Luminance (Brightness)	lm/m <sup>2</sup> /sr <sup>1</sup> = cd/m <sup>2</sup> = nit
Radiant Energy	Joule	Luminous Energy	lm s ≡ talbot

photometric units to describe lamp performance. For further details of photometry, and other concepts, such as color in physiological optics, the reader should consult Levi<sup>24</sup> or Fry.<sup>58</sup>

Calibration standards and devices for radiometric and photometric calibration of light sources are available from Gamma Scientific, International Light, Optronic Laboratories, and UDT Instruments (formerly Graseby).

#### 4.5.4 Line Sources

Line sources are used as wavelength standards for calibrating spectrometers and interferometers; as sources in atomic absorption spectrometers; in interferometric arrangements for testing optical components, such as Twyman–Green interferometers (Section 4.7.6); in a few special cases for optically pumping solid-state and gas lasers; and for illumination.

The emission lines from a line source are not infinitely sharp. Their shape is governed by the actual conditions and physical processes occurring in the source. The variation of the radiant intensity with frequency across a line whose center frequency is  $\nu_0$  is described by its *lineshape function*  $g(\nu, \nu_0)$ , where:

$$\int_{-\infty}^{\infty} g(\nu, \nu_0) d\nu = 1 \quad (4.203)$$

The extension of the lower limit of this integral to negative frequencies is done for formal theoretical reasons connected with Fourier theory and need not cause any practical problems, since for a sharp line the major contribution to the integral in Equation (4.203) comes from

frequencies close to the center frequency  $\nu_0$ . There are three main types of lineshape function:<sup>59</sup> Lorentzian, Gaussian, and Voigt.

The *Lorentzian lineshape* function is:

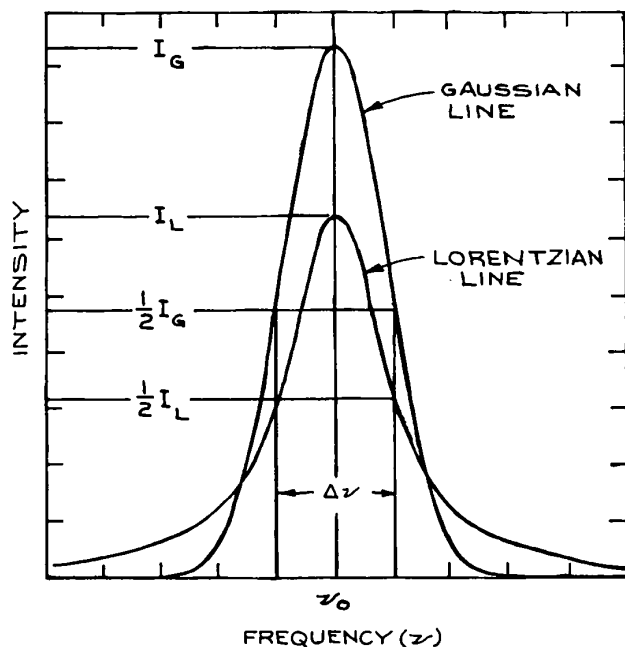
$$g_L(\nu, \nu_0) = \frac{2}{\pi \Delta\nu} \frac{1}{1 + [(\nu - \nu_0)/\Delta\nu]^2} \quad (4.204)$$

where  $\Delta\nu$  is the frequency spacing between the half-intensity points of the line (the full width at half maximum height, or FWHM). Spectral lines at long wavelengths (in the middle and far infrared) and lines emitted by heavy atoms at high pressures and/or low temperatures frequently show this type of lineshape.

The *Gaussian lineshape* function is:

$$g_D(\nu, \nu_0) = \frac{2}{\Delta\nu_D} \left( \frac{\ln 2}{\pi} \right)^{1/2} \exp \left\{ - \left[ 2 \frac{\nu - \nu_0}{\Delta\nu_D} \right]^2 \ln 2 \right\} \quad (4.205)$$

where  $\Delta\nu_D$  is the FWHM. Spectral calibration lamps are available from Avantes, Cathodeon, Gamma Scientific, Heraeus Nobleight, Newport/Oriel, Ocean Optics, Spectral Products, and Resonance. Gaussian lineshapes are usually associated with visible and near-infrared lines emitted by light atoms in discharge-tube sources at moderate pressures. In this case, the broadening comes from the varying Doppler shifts of emitting species, whose velocity distribution in the gas is Maxwellian. Emitting ions in real crystals sometimes have this type of lineshape because of the random variations of ion environment within a real crystal produced by dislocations, impurities, and other lattice defects. A Lorentzian and a Gaussian lineshape are compared in Figure 4.107.



**Figure 4.107** Gaussian and Lorentzian lineshapes of the same FWHM,  $\Delta\nu$ .

Frequently, the broadening processes responsible for Lorentzian and Gaussian broadening are simultaneously operative, in which case the resultant lineshape is a convolution of the two and is called a *Voigt profile*.<sup>15</sup>

The low-pressure mercury lamp is the most commonly used narrow-line source. These lamps actually operate with a mercury–argon or mercury–neon mixture. The principal lines from a mercury–argon lamp are listed in Table 4.7. Numerous other line sources are also available. In particular, hollow-cathode lamps emit the strongest spectral line of any desired element for use in atomic absorption spectrometry. Such lamps are available from Bulbtronics, GBC Scientific Equipment, Hamamatsu, Hellma International, REFLEX Analytical, Semicon Associates, and Vitro Technology, Ltd.

#### 4.5.5 Continuum Sources

A continuum source in conjunction with a monochromator can be used to obtain radiation whose wavelength is tunable throughout the emission range of the source. If the

**Table 4.7** Characteristic lines from a mercury lamp

Wavelength <sup>a</sup> ( $\mu\text{m}$ )
0.253652
0.313156
0.313184
0.365015
0.365483
0.366328
0.404656
0.435835
0.546074
0.576960
0.579066
0.69075
0.70820
0.77292
1.0140
1.1287
3.9425

*Note:* Extensive listings of calibration lines from other sources can be found in C.R. Harrison *M.I.T. Wavelength Tables*, M.I.T. Press, Cambridge, Mass, 1969; and in A. R. Striganov and N. S. Sventitskii, *Tables of Spectral Lines of Neutral and Ionized Atoms*, IFT/Plenum Press, New York, 1968.

<sup>a</sup> *In vacuo.*

wavelength region transmitted by the monochromator is made very small, not very much energy will be available in the wavelength region selected. Even so, continuum sources find extensive use in this way in absorption and fluorescence spectrometers. Certain continuum sources, called blackbody sources, have very well-characterized radiance as a function of wavelength and are used for calibrating both the absolute sensitivity of detectors and the absolute radiance of other sources.

**Blackbody Sources.** All objects are continuously emitting and absorbing radiation. The fraction of incident radiation in a spectral band that is absorbed by an object is called its *absorptivity*. When an object is in thermal equilibrium with its surroundings, it emits and absorbs radiation in any spectral interval at equal rates. An object that absorbs all radiation incident on it is called a *blackbody* – its absorptivity  $\alpha$  is equal to unity.

The ability of a body to radiate energy in a particular spectral band compared to a blackbody is its *emissivity*. A blackbody is also the most efficient of all emitters – its emissivity  $\epsilon$  is also unity. In general, for any object emitting and absorbing radiation at wavelength  $\lambda$ ,  $\epsilon_\lambda = \alpha_\lambda$ . Highly reflecting, opaque objects, such as polished metal surfaces, do not absorb radiation efficiently; nor, when heated, do they emit radiation efficiently.

The simplest model of a blackbody source is a heated hollow object with a small hole in it. Any radiation entering the hole has minimal chance of re-emerging. Consequently, the radiation leaving the hole will be characteristic of the interior temperature of the object. The energy density distribution of this *blackbody radiation* in frequency is:

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1} \quad (4.206)$$

where  $\rho(\nu)d\nu$  is the energy stored ( $\text{J/m}^3$ ) in a small frequency band  $d\nu$  at  $\nu$ . The energy density distribution in wavelength is:

$$\rho(\lambda) = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \quad (4.207)$$

This translates into a spectral emittance (the total power emitted per unit wavelength interval into a solid angle  $2\pi$  by unit area of the blackbody) given by:

$$M_{e\lambda} = \frac{C_1}{\lambda^5 (e^{C_2/\lambda T} - 1)} \quad (4.208)$$

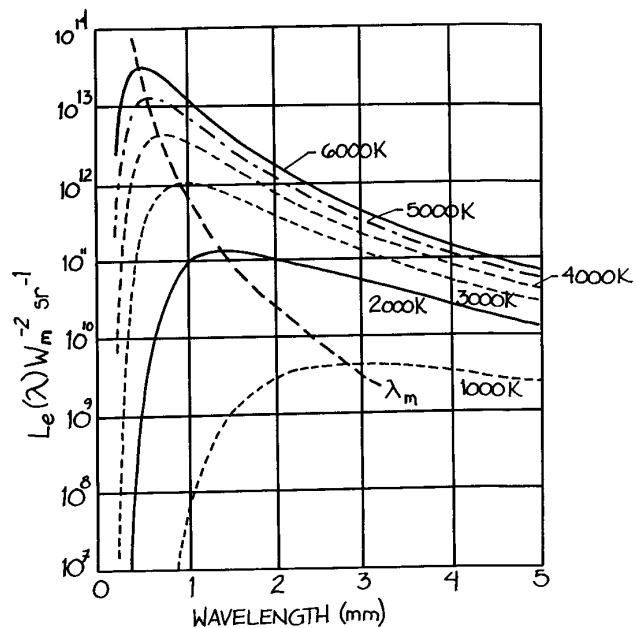
where  $C_1 = 2\pi hc^2$ , called the *first radiation constant*, has the value  $3.7418 \times 10^{-16} \text{ Wm}^2/\text{s}$ , and  $C_2 = ch/k$ , called the *second radiation constant*, has the value  $1.43877 \times 10^{-2} \text{ mK}$ .

A true blackbody is also a diffuse (Lambertian) radiator. Its radiance is independent of the viewing angle. For such a source:

$$M_{e\lambda} = \pi L_{e\lambda} \quad (4.209)$$

The variation of  $L_{e\lambda}$  with wavelength for various values of the temperature is shown in Figure 4.108. The wavelength of maximum emittance,  $\lambda_m$  at temperature  $T$  obeys Wien's displacement law:

$$\lambda_m T = 2.8978 \times 10^{-3} \text{ mK} \quad (4.210)$$



**Figure 4.108** Spectral radiance  $L_{e\lambda}$  of a blackbody source at various temperatures.

The total radiant emittance of a blackbody at temperature  $T$  is:

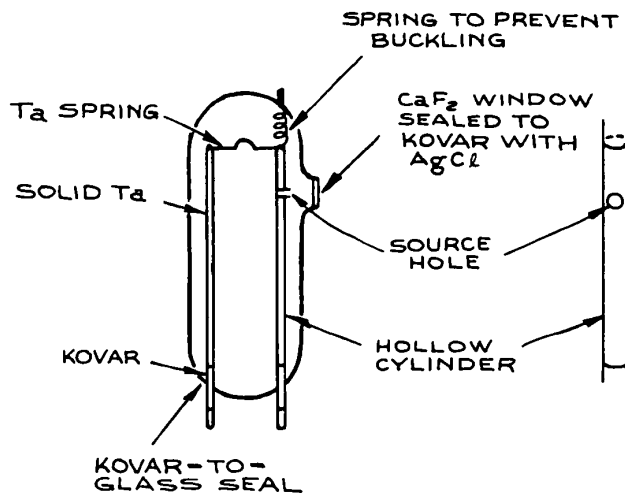
$$M_e = \int_0^{\infty} M_{e\lambda} d\lambda = \frac{2\pi^5 k^4}{15c^2 h^3} T^4 = \sigma T^4 \quad (4.211)$$

This is a statement of the Stefan–Boltzmann law. The coefficient  $\sigma$ , called the *Stefan–Boltzmann constant*, has a value of  $5.6705 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ . The known parameters  $M_{e\lambda}$  and  $M_e$  of a blackbody allow it to be used as an absolute calibration source in radiometry. If a detector responds to photons, the spectral emittance in terms of photons,  $N_\lambda$ , may be useful:

$$N_\lambda = \frac{M_{e\lambda}}{hc/\lambda} \quad (4.212)$$

Curves of  $N_\lambda$  are given by Kruse, McGlauchlin, and McQuistan.<sup>60</sup>

A source whose spectral emittance is identical to that of a blackbody apart from a constant multiplicative factor is called a *graybody*. The constant of proportionality,  $\epsilon$ , is its



**Figure 4.109** Construction details of a simple high-temperature blackbody source. (From P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology: Generation, Transmission, and Detection*, John Wiley & Sons, Inc., New York, 1962; by permission of John Wiley & Sons, Inc.)

emissivity. Several continuum sources, such as tungsten-filament lamps, carbon arcs, and flashlamps, are approximately graybody emitters within certain wavelength regions.

**Practical Blackbody Sources.** The radiant emittance of a blackbody increases at all wavelengths as the temperature of the blackbody is raised, so a practical blackbody should, ideally, be a heated body with a small emitting aperture that is kept as small as possible. Kruse, McGlauchlin, and McQuistan<sup>60</sup> describe such a source, illustrated in Figure 4.109, that can be operated at temperatures as high as 3000 K. A 25- $\mu\text{m}$ -thick tungsten ribbon 2 cm wide is rolled on a 3-mm-diameter copper mandrel and seamed with a series of overlapping spot welds. A hole about 0.75 mm in diameter is made in the foil, and the copper dissolved out with nitric acid under a fume hood. The resulting cylinder is mounted on 1-mm-diameter Kovar or tungsten rod feedthroughs in a glass envelope and heated from a high-current, low-voltage power supply. The glass envelope should be fitted with a window that is transmissive to the wavelength region desired from the source.

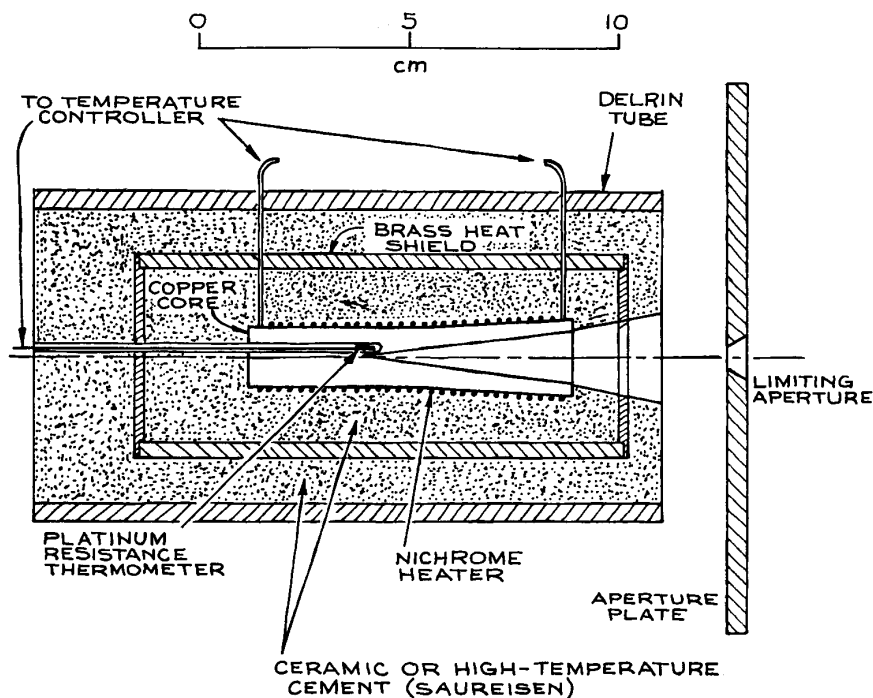
Another design of blackbody source is shown in Figure 4.110. This design is based on a heated copper cylinder, containing a conical cavity of 15° semivertical angle, that is allowed to oxidize during operation (so that it becomes nonreflective and consequently of high emissivity). The cylinder is heated by an insulated heater wire wrapped around its circumference. If nichrome wire is used, the cylinder can be heated to about 1400 K. This assembly is mounted in a ceramic tube (alumina is quite satisfactory) or potted in high-temperature ceramic cement. For high-temperature operation, the whole assembly can be mounted inside a water-cooled block.

A popular blackbody source uses a “Globar,” a rod of bonded silicon carbide. In the high-temperature blackbody source supplied by Newport/Oriel this has been replaced by pyrolytic graphite. For further details of the advantages and disadvantages of this and various other blackbody sources, the reader is referred to Hudson.<sup>61</sup> Blackbody sources are available commercially from The Eppley Laboratory, HGH Optronics, Infrared Industries, Isotech, Micron, Omega, and Newport/Oriel.

**Tungsten-Filament Lamps.** Tungsten-filament lamps are approximately graybodies in the visible with an emittance between 0.45 and 0.5. Such lamps are frequently described in terms of their *color temperature*  $T_c$ , which is the temperature at which a blackbody would have a spectral emittance closest in shape to the lamp’s. The color temperature will depend on the operating conditions of the lamp.

Tungsten-filament lamps can most conveniently be operated in the laboratory with a variable transformer. For best stability and freedom from ripple on their output, however, they should be operated from a stabilized d.c. supply. Typical supply requirements range up to a few hundred volts. Lamps with wattage ratings up to 1 kW are readily available.

Small tungsten-filament lamps that can be used as point sources are available from Newport/Oriel. Very-long-life, constant-efficiency tungsten-halogen lamps are available, in which the lamp envelope usually contains a small amount of iodine. In operation, the iodine vaporizes and recombines with tungsten that has evaporated from the filament and deposited on the inside of the lamp envelope. The tungsten iodide thus formed diffuses to the hot

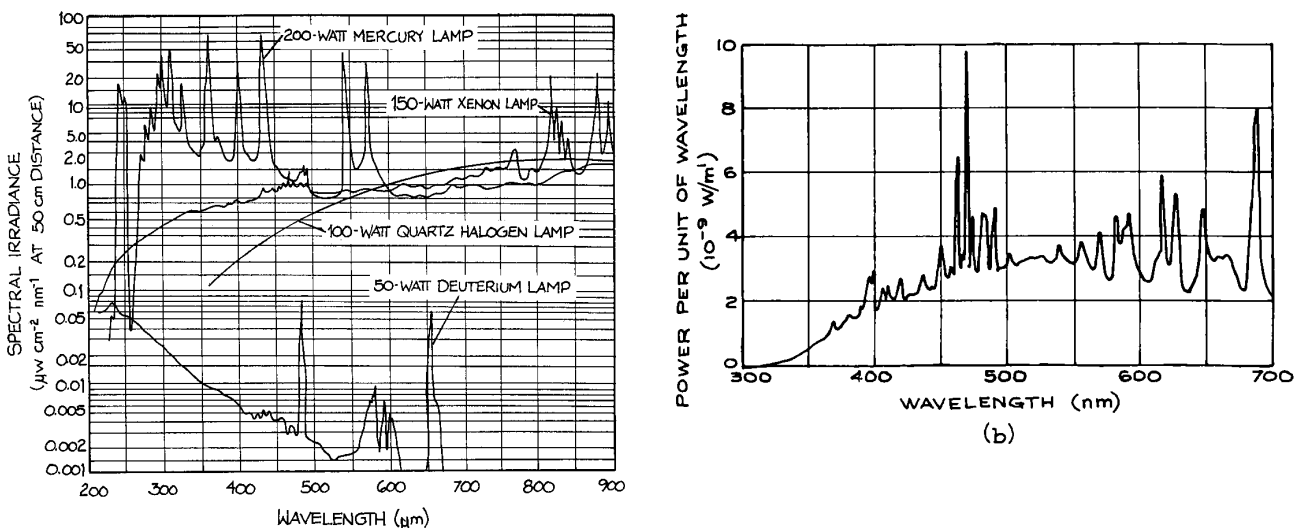


**Figure 4.110** Construction details of an NBS-type blackbody source.

filament, where it decomposes, redepositing tungsten on the filament. The constant replacement of the filament in this way allows it to be operated at very high temperature and radiant emittance. Because the lamp envelope must withstand the chemical action of hot iodine vapor and high temperatures, it is made of quartz. Hence such lamps are frequently called quartz-iodine lamps. Such lamps can be quite compact: a 1 kW lamp will have a filament length of about 1 cm. They are available from Bulb Direct and Ocean Optics. The National Institute of Standards and Technology (NIST) standard of spectral irradiance consists of a quartz-iodine lamp with a coiled-coil tungsten filament operating at about 3000 K and calibrated from 250 nm to 2.6  $\mu\text{m}$  against a blackbody source. Such calibrated lamps are available from EG&G and Newport/Oriel. Because they are intense sources of radiant energy, these lamps can also be used for heating. In particular, they are often placed inside complex vacuum systems to bake out internal components that are well insulated thermally from the chamber walls.

**Continuous Arc Lamps.** High-current electrical discharges in gases, with currents that typically range from 1 to 100 A, can be intense sources of continuum or line emission, and sometimes both at the same time. For substantial continuum emission the most popular such lamps are the high-pressure xenon, high-pressure mercury, and high-pressure mercury-xenon lamps. The arc in these lamps typically ranges up to about 5 cm long and 6.2 mm in diameter (for a 10 kW lamp – 100 V, 100 A input). Because of their small size, arc lamps have much higher spectral radiance (brightness) than quartz-iodine lamps of comparable wattage. In the visible region at 500 nm, a typical xenon arc lamp shows 1.9 times the output of a quartz-iodine lamp; at 350 nm, 14 times; and at 250 nm, 200 times. In addition, because of their small size, high-pressure arc lamps work well in the illumination of monochromator slits in spectroscopic applications. Lower-wattage arc lamps come close to being point sources and are ideal for use in projection systems and for obtaining well-collimated beams.





**Figure 4.111** (a) Spectral irradiance of various moderate-power gas-discharge lamps (courtesy of Oriel Corporation, Stamford, Conn.); (b) spectral energy distribution for a high-power (10 kW) xenon arc lamp with a long discharge column (from A. A. Kruithof, "Modern Light Sources," in *Advanced Optical Techniques*, A. C. S. Van Heel, Ed., North-Holland, Amsterdam, 1967; by permission of North-Holland Publishing Company).

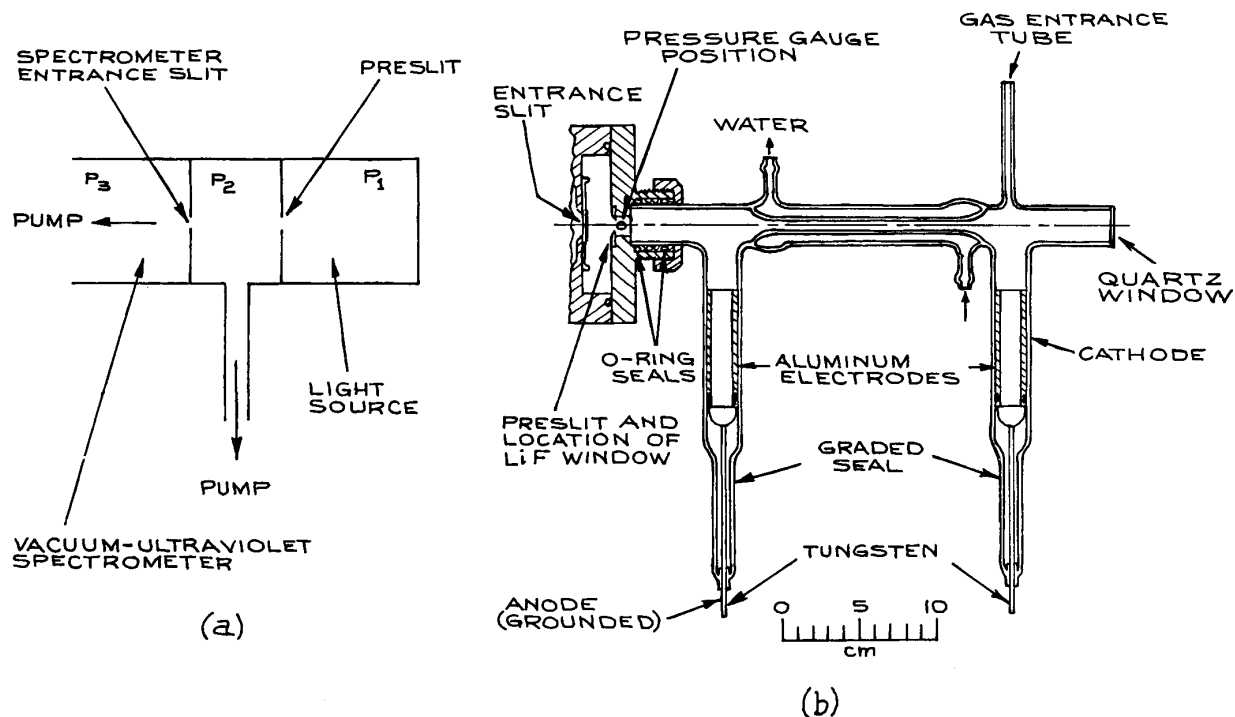
There are two different kinds of high-pressure arc lamps: those where the discharge is confined to a narrow quartz capillary (which must be water-cooled), and those where the discharge is not confined (which usually operate with forced-air cooling). The former are available from Flashlamps (Verre et Quartz), ILC Technology, Xenon Corp., and Ushio and are used for pumping CW solid-state lasers. (Krypton arc lamps are better than xenon arc lamps for pumping  $\text{Nd}^{3+}$  lasers, as their emission is better matched to the absorption spectrum of  $\text{Nd}^{3+}$  ions.)

Because high-pressure arc lamps operate at very high pressures when hot (up to 200 bar), they must be housed in a rugged metal enclosure to contain a possible lamp explosion. The mounting must be such as to allow stress-free expansion during warmup. Generally speaking, commercial lamp assemblies should be used. The power-supply requirements are somewhat unusual. An initial high-voltage pulse is necessary to strike the arc, and then a lower voltage, typically in the range 70–120 V, to establish the arc. When the arc is fully established, the operating voltage will drop to perhaps as low as 10 V. Arcs containing mercury need a further increase in operating voltage as

they warm up and their internal mercury pressure increases. Complete lamp assemblies and power supplies are available from several companies, among them, ILC Technology, Newport/Oriel, ORC Lighting Products, Perkin Elmer, and Spectral Energy.

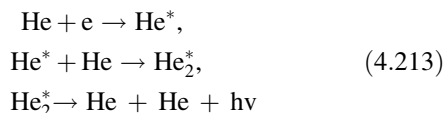
High-pressure arc lamps give substantial continuum emission with superimposed line structure, as can be seen in Figure 4.111. These lamps are not efficient sources of infrared radiation. They give substantial UV emission, however, and care should be taken in their use to avoid eye or skin exposure. Their UV output will also generate ozone, and provision should be made for venting this safely from the lamp housing. They are available from Bulbtronics, Hamamatsu, Laser Drive, Lighting Technologies International, and Newport/Oriel.

Deuterium lamps are efficient sources of ultraviolet emission with very little emission at longer wavelengths, as shown in Figure 4.111(a). They are available from Cathodeon, Hamamatsu, Newport/Oriel, Ocean Optics, and Spectral Energy. Hollow cathode lamps, available from Hamamatsu and Hellma, emit strong spectral lines characteristic of the material of which their cathode is made. They can be used as wavelength calibration sources.



**Figure 4.112** (a) Differential pumping arrangement ( $P_1$ ,  $P_2$ ,  $P_3$  are the light source, differential pumping unit, and spectrometer operating pressures, respectively); (b) light source for the production of the noble-gas continua – in a differentially pumped mode, no LiF window would be used. (From R. E. Huffman, Y. Tanaka, and J. C. Larrabee, "Helium Continuum Light Source for Photoelectric Scanning in the 600–1100 Å Region," *Appl. Opt.*, **2**, 617–623, 1963; by permission of the Optical Society of America.)

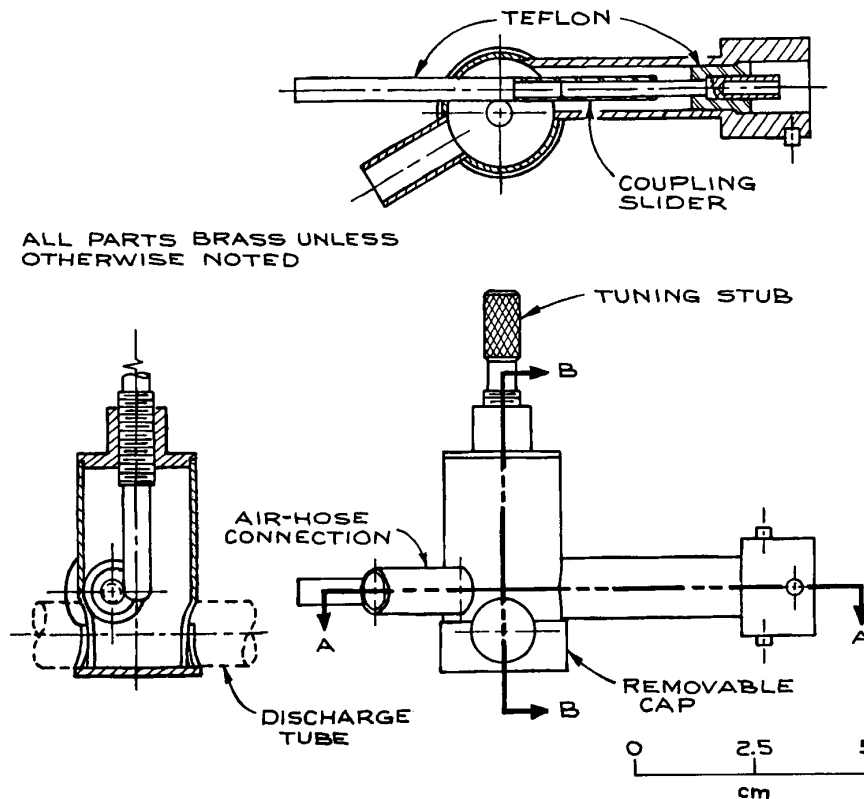
Discharges in high-pressure noble gases can also be used as intense continuum sources of vacuum-ultraviolet radiation. This radiation arises from noble-gas excimer emission, which in the case of helium, for example, arises from a series of processes that can be represented as:



The spectral regions covered by the excimer continua are He, 105–400 nm; Ar, 107–165 nm; Kr, 124–185 nm; and Xe, 147–225 nm. Because there are no transmissive materials available for wavelengths below about 110 nm, sources below this wavelength are used without windows. Radiation leaves the lamp through a small slit, or through a

multicapillary array. The latter is a close-packed array of many capillary tubes, which presents considerable resistance to the passage of gas, but is highly transparent to light. Multicapillary arrays can be obtained from Burle Electro-Optics. To maintain the lamp pressure, gas is continuously admitted. Gas that passes from the lamp into the rest of the experiment (for example a vacuum-ultraviolet monochromator) is continuously pumped away by a high-speed vacuum pump, as shown in Figure 4.112. This technique is called *differential pumping* (see Section 3.6.2). For further details of vacuum-ultraviolet sources and technology, the reader is referred to Samson.<sup>62</sup>

**Microwave Lamps.** Small microwave-driven discharge lamps are very useful where a low-power atomic-emission line source is desired, particularly if the atomic



**Figure 4.113** Microwave cavity for exciting a gas discharge in a cylindrical quartz tube (design used by Opthos). (After F. C. Fehsenfeld, K. M. Evenson, and H. P. Broida, "Microwave Discharge Cavities Operating at 2450 MHz," *Rev. Sci. Instr.*, **36**, 294–298, 1965; by permission of the American Institute of Physics.)

emission is desired from some reactive species such as atomic chlorine or iodine. A small cylindrical quartz cell containing the material to be excited, usually with the addition of a buffer of helium or argon, is excited by a microwave source inside a small tunable microwave cavity, as shown in Figure 4.113. Suitable cavities for this purpose, designed for operation at 2450 MHz, are available from Opthos Instruments; power supplies for these lamps are available from Opthos and Cathodeon.

**Flashlamps.** The highest-brightness incoherent radiation is obtained from short-pulse flashlamps. By discharging a capacitor through a gas-discharge tube, much higher discharge currents and input powers are

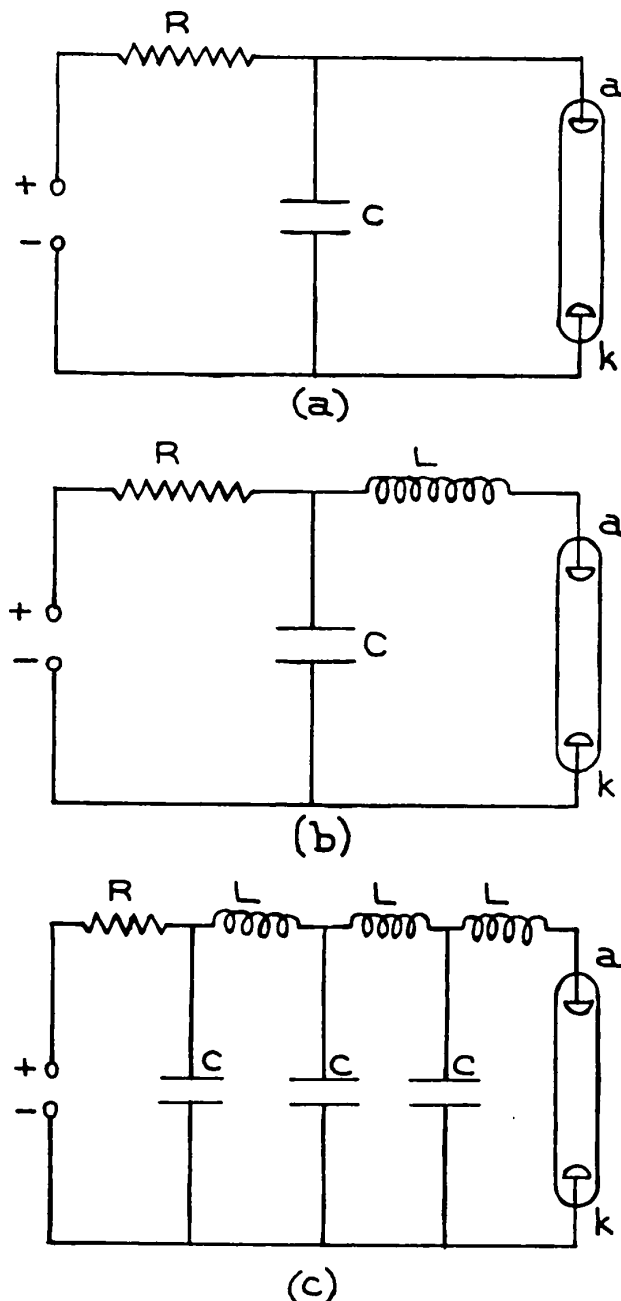
possible than in d.c. operation. As the current density or pressure of a flashlamp is increased, the emission from the lamp shifts from radiation that is characteristic of the fill gas, with lines superimposed on a continuum, to an increasingly close approximation to blackbody emission corresponding to the temperature of the discharge gas. Commercial flashlamps can be roughly divided into two categories. In long-pulse lamps, fairly large capacitors (100–10 000  $\mu\text{F}$ ), charged to moderately high voltages (typically up to about 5 kV), are discharged slowly (on time scales from 100  $\mu\text{s}$  to 1 ms) through high-pressure discharge tubes. In short-pulse, high-peak-power lamps, smaller, low-inductance capacitors (typically 0.1–10  $\mu\text{F}$ ), charged to high voltages (10–80 kV), are discharged

rapidly (on time scales down to 1  $\mu\text{s}$ ) through lower-pressure discharge tubes. Suppliers of linear flashlamps include Continuum, Laser SOS Group, New Source Technology, Perkin-Elmer Optoelectronics, and Xenon Corporation.

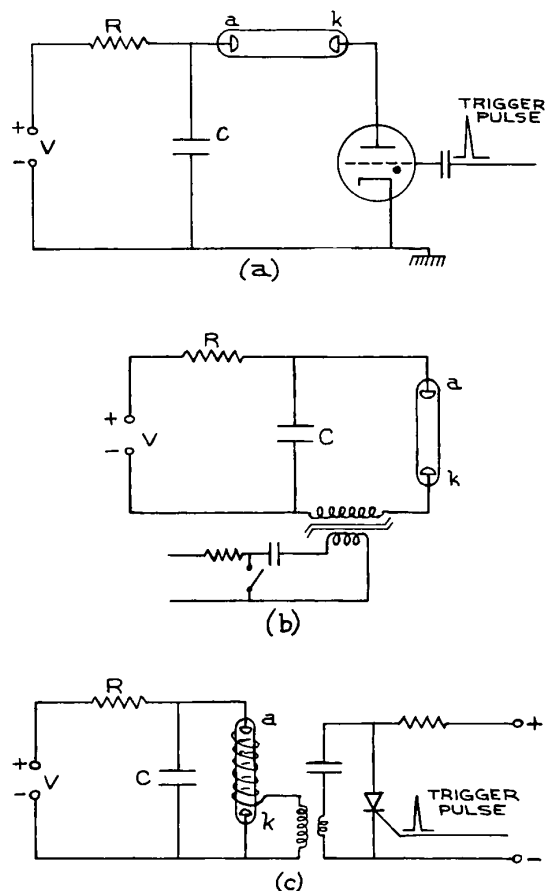
**(i) Long-pulse lamps.** Long-pulse lamps are generally filled with xenon,<sup>63</sup> but krypton lamps<sup>64</sup> (for pumping  $\text{Nd}^{3+}$  lasers) and alkali-metal lamps are also available. Fill pressures are typically on the order of 0.1–1 bar. Although the discharge current in such lamps can run to tens of thousands of  $\text{A}/\text{cm}^2$ , at low repetition frequencies ( $< 0.1$  Hz) ambient cooling is all that is necessary. Three of the most commonly used circuits for operating such lamps are shown in Figure 4.114. In all three of these circuits, the capacitor, or pulse-forming network shown in Figure 4.114 (c), is charged through a resistor  $R$ . The capacitor is discharged through the lamp by triggering the flashlamp with one of the trigger circuits shown in Figure 4.115. The lamp itself behaves both resistively and inductively when it is fired. If the inductance and resistance of the lamp are not sufficiently large, the capacitor may discharge too rapidly, which can lead to damage of both lamp and capacitor. Generally speaking, an additional series inductor will be desirable to control the discharge. The problem of selecting the appropriate inductor for a particular capacitor size and lamp has been dealt with in detail by Markiewicz and Emmett.<sup>65</sup> Slow flashlamps have a nonlinear V-I characteristic, which can be approximated by:

$$V = K_0 \sqrt{|I|} \quad (4.214)$$

where the sign of  $V$  is taken to be the same as the sign of  $I$ . The value of  $K_0$ , measured in  $\Omega\text{A}^{1/2}$ , is a parameter specified by a manufacturer for a given lamp. Given this value and the capacitor size to be used, the calculations of Markiewicz and Emmett allow a suitable value of series inductor to be chosen. Other factors must be taken into account in designing the flashlamp circuit: the maximum power loading, usually specified as the explosion energy of the lamp (which will depend on the discharge-pulse duration), and the maximum repetition frequency of the lamp. Usually, the larger the capacitor-stored energy, the lower the permitted repetition frequency will be. As lamps and discharge energies get smaller, repetition frequencies can be extended, to several kilohertz for the smallest low-energy

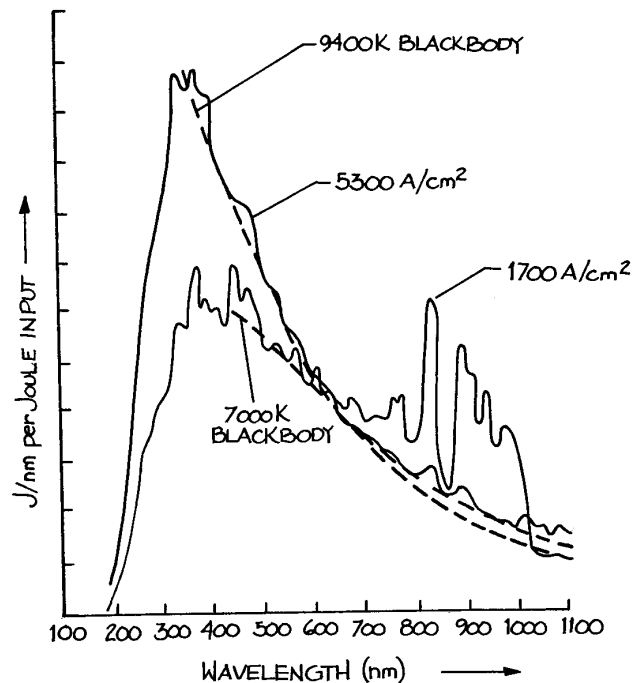


**Figure 4.114** Operating circuits for flashlamps ( $a$  = lamp anode;  $k$  = lamp cathode): (a) RC discharge; (b) RLC critically damped discharge; (c) pulse-forming network.



**Figure 4.115** Flashlamp triggering schemes ( $a$  = lamps anode;  $k$  = lamp cathode): (a) overvoltage triggering ( $V >$  self-flash voltage of flashlamp; switching is accomplished in this case with a thyatron, but a triggered spark gap or ignitron can also be used); (b) series triggering (a saturable transformer is used with a fast-risetime 10–20 kV pulse with sufficient energy to trigger the lamp and saturate the core); (c) external triggering (the flashlamp is ionized by a trigger wire wrapped around the outside of the lamp and connected to the 5–15 kV secondary of a high-voltage pulse transformer).

lamps, such as those used in stroboscopes. The explosion energy of the lamp is the minimum input required to cause lamp failure in one shot. To obtain long flashlamp life, a lamp should be operated only at a fraction of its explosion

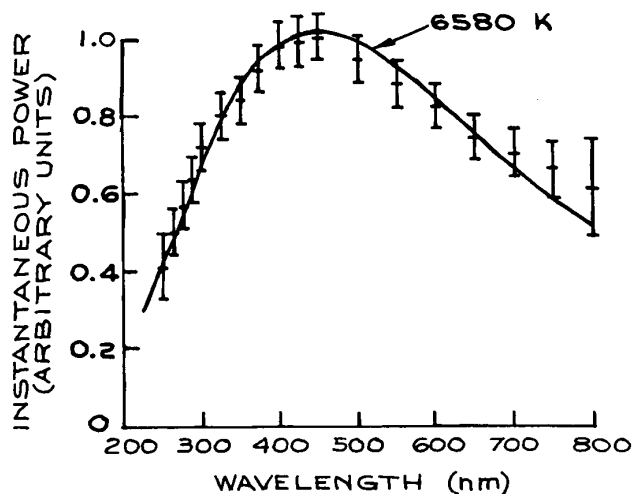


**Figure 4.116** Output spectrum of a Perkin-Elmer FX-47A xenon flashlamp (0.4 atm) at two current densities: 1700A/cm<sup>2</sup> and 5300A/cm<sup>2</sup>.

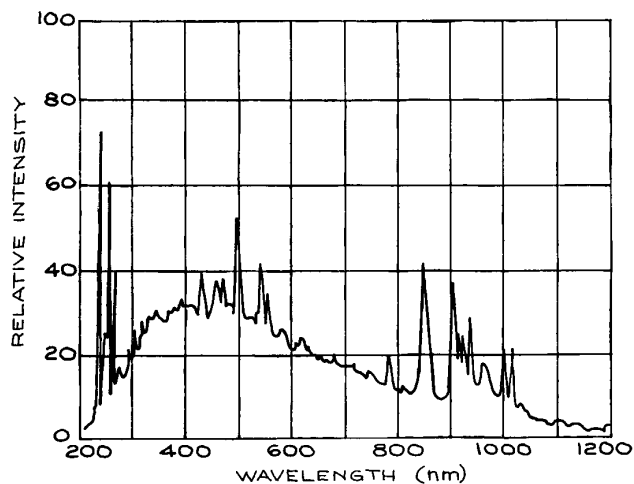
energy. For example, at 50% of the explosion energy the expected lifetime is from 100 to 1000 flashes, while at 20% it is from 10<sup>5</sup> to 10<sup>6</sup> flashes. The lamp should be mounted freely and not clamped rigidly at its ends when it is operated.

The spectral output of a flashlamp varies slightly during the flash. For moderate- and high-power lamps (> 100 J in 1 ms), the spectral output approaches that of a blackbody, as shown in Figures 4.116 and 4.117.<sup>66</sup> Low-power lamps exhibit the spectral feature of their fill gas, as shown in Figure 4.118. Figure 4.116 shows the transition from a spectrum with some line structure to a blackbody continuum as the current density through the lamp is increased.

A word about the triggering schemes shown in Figure 4.115. External triggering, accomplished by switching a high-voltage pulse from a transformer to a trigger wire wrapped around the lamp, is perhaps the simplest scheme; it is used only with long-pulse lamps and does not give quite so good time synchronization as series-spark-gap or



**Figure 4.117** Spectral distribution of power radiated by an EG&G FX42 xenon-filled flashlamp (76 mm long by 7 mm bore) operated in a critically damped mode with 500 J discharged in 1 ms. The spectral distribution was observed 0.7 ms from flash initiation. The line is the blackbody radiation curve of best fit. (After J. G. Edwards, "Some Factors Affecting the Pumping Efficiency of Optically Pumped Lasers," *Appl. Opt.*, **6**, 837–843, 1967; by permission of the Optical Society of America.)

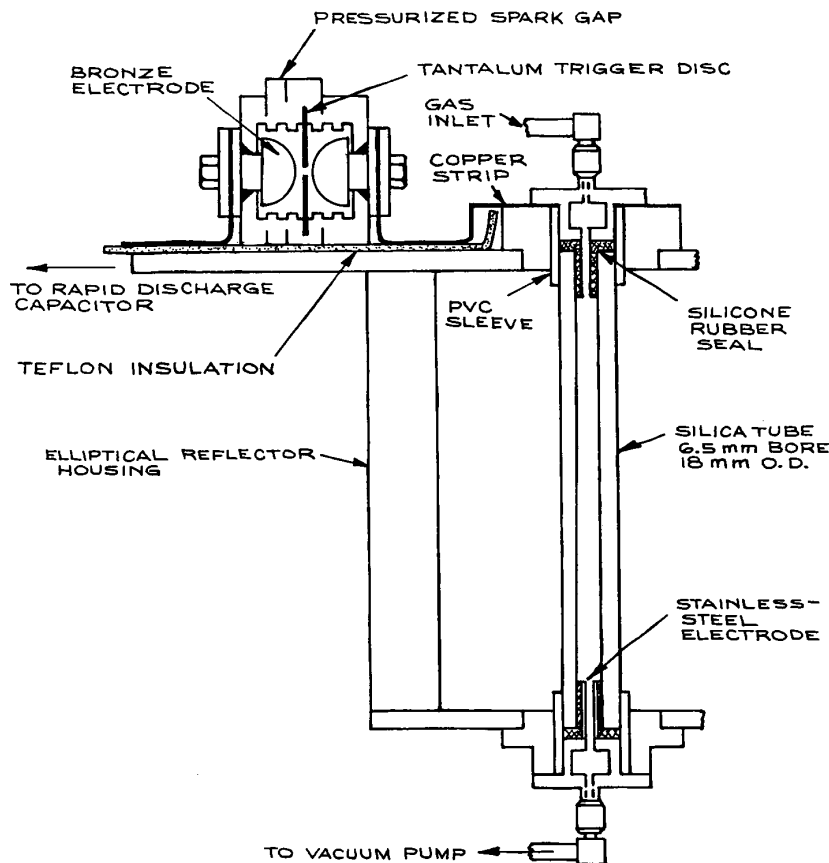


**Figure 4.118** Spectral distribution of intensity from an ILC 4L2 xenon flashlamp (51 mm long by 4 mm bore) operated in a critically damped mode with 10 J discharged in 115  $\mu$ s. (Courtesy of ILC).

thyatron-switched operation. The other methods are more complex, but can be also used with short-pulse, high-power lamps. Series triggering allows the incorporation of triggering and lamp series inductance in a single unit. Trigger transformers of various types are available from EG&G and ILC. EG&G and EEV supply a range of excellent thyratrons.

**(ii) Short-pulse high-power lamps.** Flashlamps that can handle the rapid discharge of hundreds of joules on time-scales down to a few microseconds are available commercially from Flashlamps (Verre et Quartz), ILC Technology, Perkin Elmer, and Xenon Corporation, among others. To achieve such rapid discharges, these lamps are generally operated in a low-inductance discharge circuit. When these lamps are fired at high-peak-power inputs, a severe shock wave is generated in the lamp. To withstand the shock wave, the lamps are designed with special re-entrant electrode seals. If lamps designed for long-pulse operation are discharged too rapidly, their electrodes are quite likely to break off because the electrode seals are not shock-resistant.

The spectral emission of short-pulse, high-energy lamps is generally quite close to that of a high-temperature blackbody, perhaps as hot as 30 000 K. The fill pressure in these lamps is often lower than in long-pulse lamps, typically from 0.1 to a few tens of millibars. The fill gas is usually xenon, but other noble gases such as krypton or argon can be used. If a short-pulse lamp has a narrow discharge-tube cross-section, at high power inputs material can ablate from the wall and a substantial part of the emission from the lamp can come from this material. Such lamps can be made with heavy-walled discharge-tube bodies of silica, glass, or even Plexiglas. These lamps can be made fairly simply. A good design described by Baker and King<sup>67</sup> is illustrated in Figure 4.119. It can be operated in a nonablating or ablating mode (at high or low pressure, respectively). When operated in the ablating regime, such lamps can be filled with air as the nature of the fill gas is unimportant. Figure 4.119 also shows the use of a field-distortion-triggered spark gap for firing the lamp – a spark-gap design that offers quiet, efficient switching of rapid discharge capacitors. Rapid-discharge (low-inductance) capacitors suitable for fast flashlamp and other applications are available from CSI Technologies, Hipotronics, and Maxwell Energy Products.



**Figure 4.119** High-pulse-energy, fast flashlamp design for conventional or ablation-mode operation. (From H. J. Baker and T. A. King, "Optimization of Pulsed UV Radiation from Linear Flashtubes," *J. Phys. E.: Sci. Instr.*, **8**, 219–223, 1975. Copyright 1975 by the Institute of Physics, used with permission).

## 4.6 LASERS

Lasers are now so widely used in physics, chemistry, the life sciences, and engineering that they must be regarded as the experimentalist's most important type of optical source. With few exceptions, anyone who wants a laser for an experiment should buy one. This is certainly true in the case of helium–neon, argon–ion, and helium–cadmium gas lasers and all solid-state crystalline, glass, or semiconductor lasers. True, these lasers can be built in the laboratory, but this is the province of the laser specialist and will generally be found time-consuming and unpro-

ductive for scientists in other disciplines. On the other hand, there are some lasers, such as nitrogen, exciplex, CO<sub>2</sub>, and CO gas lasers, and dye lasers, that it might occasionally pay to build for oneself, even though models are available commercially. A detailed discussion of how to construct all these lasers will not be given here, but, to illustrate how it is accomplished, design features of some specific systems will be discussed.

Table 4.8 lists the primary wavelengths available from the more commonly used lasers presently available commercially. For more detailed information about suppliers

**Table 4.8 Wavelengths available from important lasers**

<i>CW Gas Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (nm)</i>
Ar ion	229
	244 <sup>a</sup>
	348 <sup>a</sup>
	257 <sup>a</sup>
	275 <sup>a</sup>
	333.4
	333.6
	351.1 <sup>b</sup>
	363.8 <sup>b</sup>
	454.5
	457.9
	465.8
	476.5
	488.0 <sup>b</sup>
	496.5
	50.17
	514.5 <sup>b</sup>
528.7	
1092.3	
Kr ion	337.5
	350.7
	356.4
	406.7
	413.1
	461.9
	468.0
	476.2
	482.5
	530.9 <sup>b</sup>
	568.1 <sup>b</sup>
	647.0
	676.4
	752.5
793.1	
799.3	
He-Ne	543
	594.1
	612
	632.8 <sup>c</sup>
	1152
	3392
	5400
11552	
He-Cd	325.0
	441.6
CO	5-6.5 <sup>d</sup>
CO <sub>2</sub>	9-11 <sup>d</sup>
	10.6 <sup>e</sup>



**Table 4.8 (contd.)**

<i>Pulsed Gas Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (nm)</i>
Excimer	
ArF	193.3
KrF	248.4–249.1
XeCl	307–308.43
XeF	348.8–354
Cu vapor	510.54 570 578.2
F <sub>2</sub>	157
N <sub>2</sub>	337.1
<i>Type</i>	<i>Operating Wavelengths (μm)</i>
CO <sub>2</sub>	10.6 9–11
<i>CW Solid-State Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (μm)</i>
Alexandrite	365 0.370 0.730 0.760
Nd:YAG	266.8 <sup>f</sup> 355 <sup>g</sup> 0.532 <sup>h</sup> 1.064 (This is the primary line, Nd:YLF gives 1.047 μm and 1.053 μm, Nd:YVO4 also gives 1064 nm)
Ti:Sapphire	1.318 0.7–1.02 Tunable
<i>Pulsed Solid-State Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (μm)</i>
Alexandrite	0.720–0.780 (tunable, can also provide 2nd, 3rd and 4th harmonic wavelengths)
Er: fiber lasers	1.2–2.0
Er: glass	1.534
Er: YAG	2.940
Er: glass	1.534
Ho: YAG	2.1
Nd: glass	0.266 <sup>f</sup> 0.355 <sup>g</sup> 0.532 <sup>h</sup> 1.06 <sup>i</sup>

**Table 4.8 (contd.)**

<i>Pulsed Solid-State Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (μm)</i>
Ti:sapphire	0.68–1 Tunable
Nd:YAG	0.213 <sup>i</sup> 0.265 <sup>f</sup> 0.355 <sup>g</sup> 0.53 <sup>h</sup> 1.064
Ruby	0.6943
Ytterbium:glass	1.030

<i>Diode Pumped Solid-State Lasers</i>	
<i>Type</i>	<i>Operating Wavelengths (nm)</i>
	266 <sup>f</sup> 355 <sup>g</sup> 460 473 488 515 532 670 1064 1030 1480 1535–1565

*CW Tunable Dye Lasers*

(pumped with ion lasers, frequency-doubled Nd:YAG lasers at 530nm, or frequency tripled Nd:YAG laser at 355 nm)

Tuning range (μm)

0.4–1.1 (The tuning range depends on the dye used. A typical dye provides 50–100 nm of tuning)

*Pulsed Tunable Dye Lasers*

Tuning Range (μm)

0.197–1.036

<sup>a</sup> Produced by second harmonic generation of blue-green argon ion laser lines.

<sup>b</sup> Strongest lines.

<sup>c</sup> Most readily available wavelength.

<sup>d</sup> Molecular lasers that offer discrete tunability over several lines.

<sup>e</sup> Strongest line. Industrial CO<sub>2</sub> lasers generally operate only at 10.6 μm.

<sup>f</sup> Frequency quadrupled from 1.06 μm output.

<sup>g</sup> Frequency tripled from 1.06 μm output.

<sup>h</sup> Frequency doubled from 1.06 μm output.

<sup>i</sup> Phosphate glass gives 1.05 μm. Nd:YLF gives 1.053

<sup>j</sup> Fifth harmonic from 1.06 μm output.

and specifications consult a trade reference, such as the Laser Focus or Photonics Spectra Buyers Guides. Laser suppliers come and go, as does the availability of specific types of laser. Before briefly describing some of these laser systems, some background material will be presented that is pertinent to a discussion of lasers in general. More detailed information regarding the physics of laser operation is given in the books by Davis,<sup>15</sup> Silfvast,<sup>68</sup> Saleh and Teich,<sup>69</sup> Yariv,<sup>14</sup> and Siegman.<sup>70</sup>

The lasers listed in the table fall into two categories: pulsed and CW (continuous wave). Many can be operated in both ways, with the exception of nitrogen, copper vapor, and excimer lasers, which can only operate in a pulsed mode. These three laser types are *self-terminating*: once they lase they destroy their own population inversion, and a period of time must elapse before the upper laser level can be re-excited.

For pulsed lasers, the available energy outputs per pulse can be classified as low (< 10 mJ) medium (10 mJ–1 J), high (1 J–100 J), and very high (> 100 J). Pulse lengths depend on the type of laser and its mode of operation. Almost all pulsed solid-state lasers operate in a *Q-switched* mode,<sup>15</sup> which typically provides a pulse length in the 10 ns range, but varies somewhat from manufacturer to manufacturer. NonQ-switched long-pulse (LP) solid-state lasers have pulse lengths > 0.1 ms. Solid-state and dye lasers are frequently operated in a mode-locked (ML) configuration, which provides pulse lengths in the 0.1–10 ps range. A given laser can often be operated in ML, Q, and LP modes. The energy output per pulse will decrease as LP > Q > ML.

Continuous wave lasers can be classified by output power as low (< 10 mW), medium (10 mW–1 W), high (1–10 W), very high (10–100 W), and industrial (> 100 W). Continuous wave gas lasers, in particular, provide power outputs at many different wavelengths, and the available power varies from line to line. These lasers are generally supplied to operate in a specific wavelength range with discrete tunability within that range. Ultraviolet gas lasers, for example, are equipped with special optics, so do not generally operate in the visible range as well.

The energy outputs available from pulsed lasers, together with their available pulse-repetition rates and pulse lengths, cover a very wide range. The highest available energy outputs are generally available only at the lowest pulse-repetition rates. In the case of CW lasers, available power outputs vary from line to line and

from manufacturer to manufacturer. To the intended purchaser of a laser system, we recommend the following questions:

- (1) What fixed wavelength or wavelengths must the laser supply?
- (2) Is tunability required?
- (3) What frequency and amplitude stability are required?
- (4) What laser linewidth is required?
- (5) What power or pulse energy is required?
- (6) Are high operating reliability and long operating lifetime required?
- (7) For time-resolved experiments, what pulse length is needed?
- (8) Is the spatial profile of the output beam important?

The desirable attributes in each of these areas are unlikely to occur simultaneously. For example, frequency tunability is not readily available throughout the spectrum without resorting to the specialized techniques of nonlinear optics.<sup>14,15,71–75</sup>

In Table 4.8 the various categories of laser: CW gas, pulsed gas, CW solid-state, pulsed solid state, CW dye and pulsed dye frequently show tuning ranges and wavelengths that are obtained from these lasers by the use of nonlinear optics. For example, shorter wavelengths from argon ion, Nd:YAG, Ti:sapphire, and CO<sub>2</sub> can be obtained by frequency doubling (tripling or even quadrupling) by the use of a nonlinear crystal. The nonlinear crystal is generally external to the primary laser, although it may be housed in its own resonant structure. Generating second or third harmonic light from a sufficiently powerful laser is not especially difficult, although this is still a relatively specialized endeavor.

New frequency generation from a primary pump laser can also be obtained by the use of an optical parametric oscillator (OPO). In such a device an appropriate nonlinear crystal is pumped with a frequency  $\nu_p$  and in this process generates two new frequencies  $\nu_s$  and  $\nu_i$ , called the *signal* and *idler* frequencies, respectively. In this process photon energy must be conserved, so:

$$\nu_p = \nu_s + \nu_i \quad (4.215)$$

High-energy and high-power lasers are generally less stable, both in amplitude and in frequency, and will generally be less reliable and need more “hands on” attention than

their low-energy or low-power counterparts. When a decision is made to purchase a laser system, compare the specifications of lasers supplied by different manufacturers; discuss the advantages and disadvantages of different systems with others who have purchased them previously. Reputable laser manufacturers are generally very willing to supply the names of previous purchasers.

To some experimental scientists a laser is merely a rather monochromatic directional lightbulb; for others, detailed knowledge of its operating principles and characteristics is essential. Certain aspects of laser designs and operating characteristics, however, are very general and are worthy of some discussion.

#### 4.6.1 General Principles of Laser Operation

A laser is an optical-frequency oscillator: in common with electronic circuit oscillators, it consists of an amplifier with feedback. The optical-frequency amplifying part of a laser can be a gas, a crystalline or glassy solid, a liquid, or a semiconductor. This medium is maintained in an amplifying state, either continuously or on a pulsed basis, by pumping energy into it appropriately. In a gas laser the input energy comes from electrons (in a gas discharge or an electron beam), or from an optical pump (which may be a lamp or a laser). Solid-state crystalline or glassy lasers receive their pumping energy from continuous or pulsed lamps, and in more and more cases from semiconductor lasers. Liquid lasers can be pumped with a flashlamp or, on a continuous or pulsed basis, by another laser. Semiconductor lasers are  $p$ - $n$  junction devices and are operated by passing pulsed or continuous electrical currents through them.

The amplifying state in a laser medium results if a population inversion can be achieved between two sublevels of the medium. The energy *sublevels* of a system are single states with their own characteristic energy. A set of sublevels having the same energy is called an energy *level*. This can be seen with reference to Figure 4.120, which shows a schematic partial energy-level diagram of a typical laser system. Input energy excites ground-state particles (atoms, molecules, or ions) of the medium into the state (or states) indicated as 3. These particles then transfer themselves or excite other particles preferentially to sublevel 2. In an ideal system, negligible excitation of par-

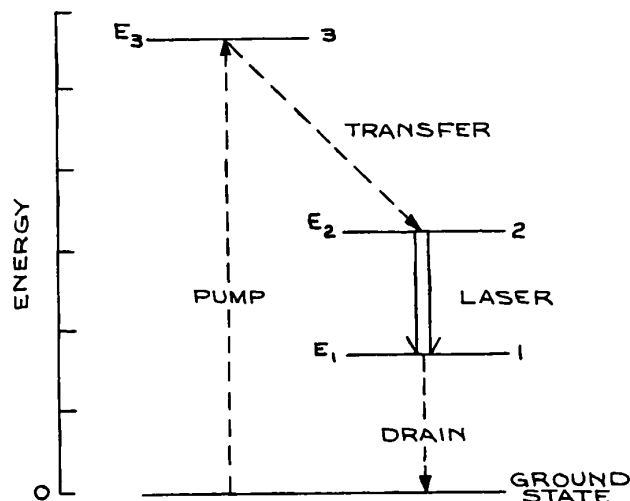
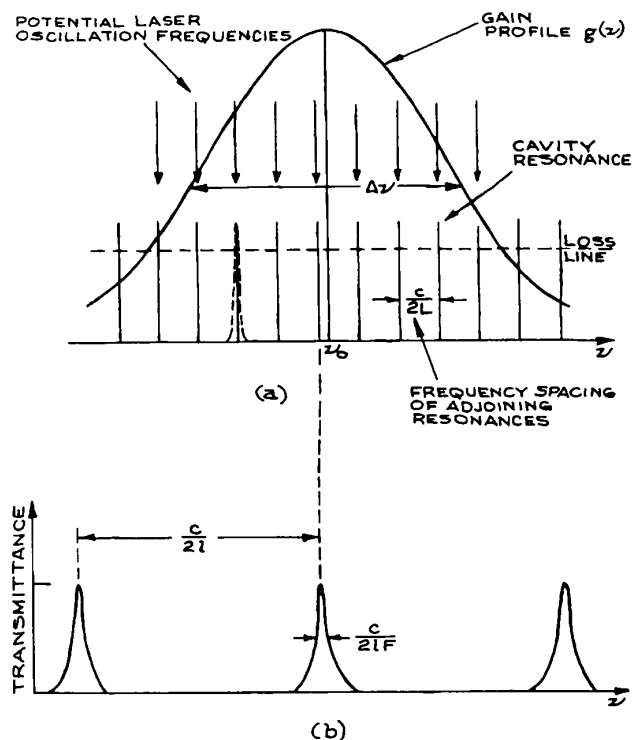


Figure 4.120 Schematic partial energy-level diagram of a laser system.

ticles into sublevel 1 should occur. If the populations of the sublevels 2 and 1 are  $n_2$  and  $n_1$ , respectively, and  $n_2 > n_1$ , this is called a *population inversion*. The medium then becomes capable of amplifying radiation of frequency:

$$\nu = \frac{E_2 - E_1}{h} \quad (4.216)$$

where  $h$  is Planck's constant. In practice, energy levels of real systems are of finite width, so the medium will amplify radiation over a finite bandwidth. The gain of the amplifier varies with frequency in this band and is specified by a gain profile  $g(\nu)$ , as shown in Figure 4.121. Actual laser oscillation occurs when the amplifying medium is placed in an optical resonator, which provides the necessary positive feedback to turn the amplifier into an oscillator. Most optical resonators consist of a pair of concave mirrors, or one concave and one flat mirror, placed at opposite ends of the amplifying medium and aligned parallel. A laser resonator is in essence a Fabry-Perot (see Section 4.7.4), generally of large spacing. The radii and spacing of the mirrors will determine, to a large degree, the type of Gaussian beam that the laser will emit. At least one of the mirrors is made partially transmitting, so that useful output power can be extracted. The choice of optimum mirror reflectance depends on the type of laser and its gain.



**Figure 4.121** (a) Laser gain profile showing position of cavity resonances and potential laser oscillation frequencies near cavity resonances where gain lies above loss ( $L$  is length of laser cavity); (b) transmission characteristic of intracavity etalon, on same frequency scale as (a), for single-mode operation ( $l$  is the etalon thickness and  $F$  its finesse).

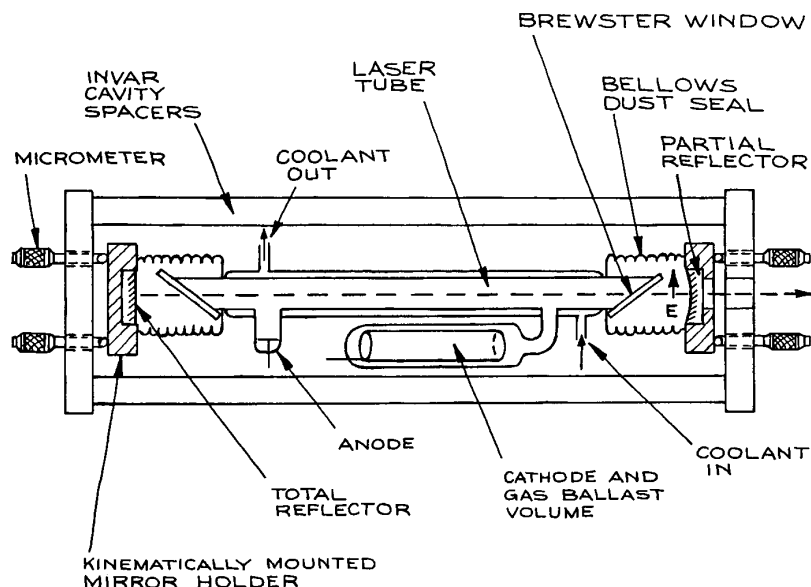
Low-gain lasers such as the helium-neon have high-reflectance mirrors, (98–99%), while higher-gain lasers such as pulsed nitrogen,  $\text{CO}_2$ , or  $\text{Nd}^{3+}$ :YAG can operate with much lower reflectance values. In certain circumstances, when their gain is very high, lasers will emit laser radiation without any deliberately applied feedback. Lasers that operate in this fashion are generally operating in an *amplified spontaneous emission* (ASE) mode. Such lasers are just amplifying their own spontaneous emission very greatly in a single pass.

#### 4.6.2 General Features of Laser Design

Even though the majority of laser users do not build their own, an awareness of some general design features of laser

systems will help the user to understand what can and cannot be done with a laser. It will also assist the user who wishes to make modifications to a commercial laser system to increase its usefulness or convenience in a particular experimental situation. Figure 4.122 shows a schematic diagram of a typical gas laser, which incorporates most of the desirable design features of a precision system. The amplifying medium of a gas laser is generally a high or low-pressure gas excited in a discharge tube. The excitation may be pulsed (usually by capacitor discharge through the tube), d.c., or occasionally a.c. Some gas lasers are also excited by electron beams, by radiofrequency energy, or by optical pumping with either a lamp or another laser. Far infrared lasers, in particular, are frequently pumped with  $\text{CO}_2$  lasers. Depending on the currents that must be passed through the discharge tube of the laser, it can be made of Pyrex, quartz, beryllium oxide, graphite, or segmented metal. These last three have been used in high-current-density discharge tubes for CW ion lasers. Several types of pulsed high-pressure or chemical lasers, such as  $\text{CO}_2$ , HF, DF, and HCl, can be excited in structures built of Plexiglas or Kel-F.

Because most gas lasers are electrically inefficient, a large portion of their discharge power is dissipated as heat. If ambient or forced-air cooling cannot keep the discharge tube cool enough, water cooling is used. This can be done in a closed cycle, using a heat exchanger. Some lasers need the discharge tube to run at temperatures below ambient. In such cases a refrigerated coolant such as ethylene glycol can be used. Many gas lasers use discharge tubes fitted with windows placed at Brewster's angle. This permits linearly polarized laser oscillation to take place in the direction indicated in Figure 4.122; light bouncing back and forth between the laser mirrors passes through the windows without reflection loss. Mirrors fixed directly on the discharge tube can also be used – this is common in commercial He–Ne lasers. In principle, such lasers should be unpolarized, although in practice various slight anisotropies of the structure often lead to at least some polarization of the output beam. The two mirrors that constitute the laser resonator, unless these are fixed directly to the discharge tube, should be mounted in kinematically designed mounts and held at fixed spacing  $\ell$  by a thermally stable resonator structure made of Invar or quartz. In this structure an important parameter is the *optical length*,  $L$  of



**Figure 4.122** Schematic diagram of a typical gas laser showing some of the desirable features of a well-engineered research system. In high-discharge-current systems, some form of internal or external gas-return path from cathode to anode must be incorporated into the structure.

the structure. For a laser in which the whole space between the two resonator mirrors is filled with a medium of refractive index  $n$ , the optical length is  $L = n\ell$ . For a composite structure with regions of different refractive index this result is easily generalized. Because the laser generates one or more output frequencies which are close to integral multiples of  $c/2L$ , any drift in mirror spacing  $L$  causes changes in output frequency  $\nu$ . This frequency change  $\Delta\nu$  satisfies:

$$\frac{\Delta\nu}{\nu} = \frac{\Delta L}{L} \quad (4.217)$$

For a laser 1 m long, even with an Invar-spaced resonator structure, the temperature of the structure would need to be held constant within 10 mK to achieve a frequency stability of only 1 part in  $10^8$ . The discharge tube should be thermally isolated from the resonator structure unless the whole forms part of a temperature-stabilized arrangement.

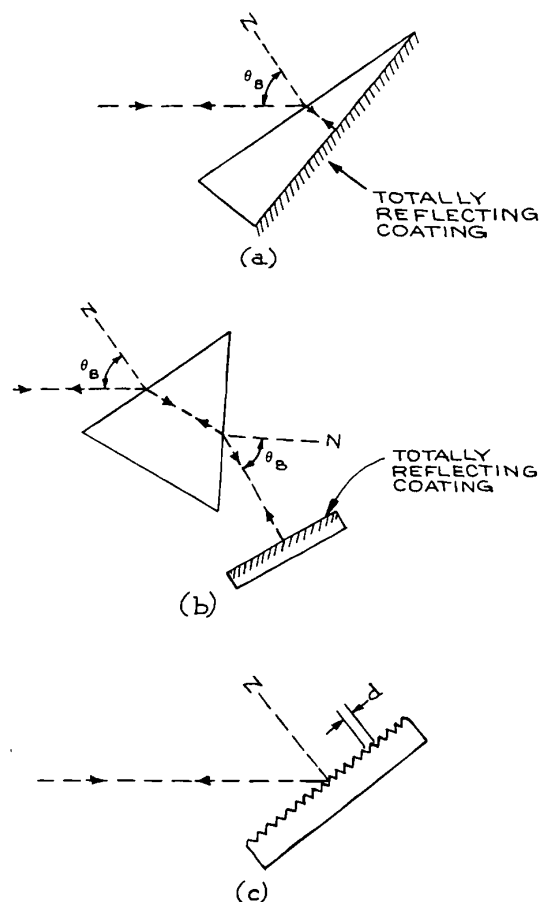
If single-frequency operation is desired, the laser should be made very short – so that  $c/2L$  becomes larger than  $\Delta\nu$  in Figure 4.121 – or operated with an intracavity etalon (see

Sections 4.6.4 and 4.7.4). Visible lasers that have the capability of oscillating at several different wavelengths are generally tuned from line to line by replacing one laser mirror with a Littrow prism whose front face is set at Brewster's angle and whose back face is given a high-reflectance coating, as shown in Figure 4.123(a). Alternatively, as shown in Figure 4.123(b), one can use a separate intracavity prism, designed so the intracavity beam passes through both its faces at Brewster's angle. Infrared gas lasers that can oscillate at several different wavelengths are tuned by replacing one resonator mirror with a gold-coated or solid metal diffraction grating mounted in Littrow, as shown in Figure 4.123(c). The laser oscillation wavelength then satisfies:

$$m\lambda = 2d \sin \theta \quad (4.218)$$

where  $d$  is the spacing of the grating grooves,  $\theta$  is the angle of incidence, and  $m$  is an integer.

A laser will oscillate only if its gain exceeds all the losses in the system, which include mirror transmission losses, absorption and scattering at mirrors and windows,



**Figure 4.123** Methods for wavelength-selective reflection in laser systems: (a) Brewster's-angle Littrow prism; (b) intracavity Brewster's-angle prism; (c) reflective diffraction grating.  $N$  = surface normal.

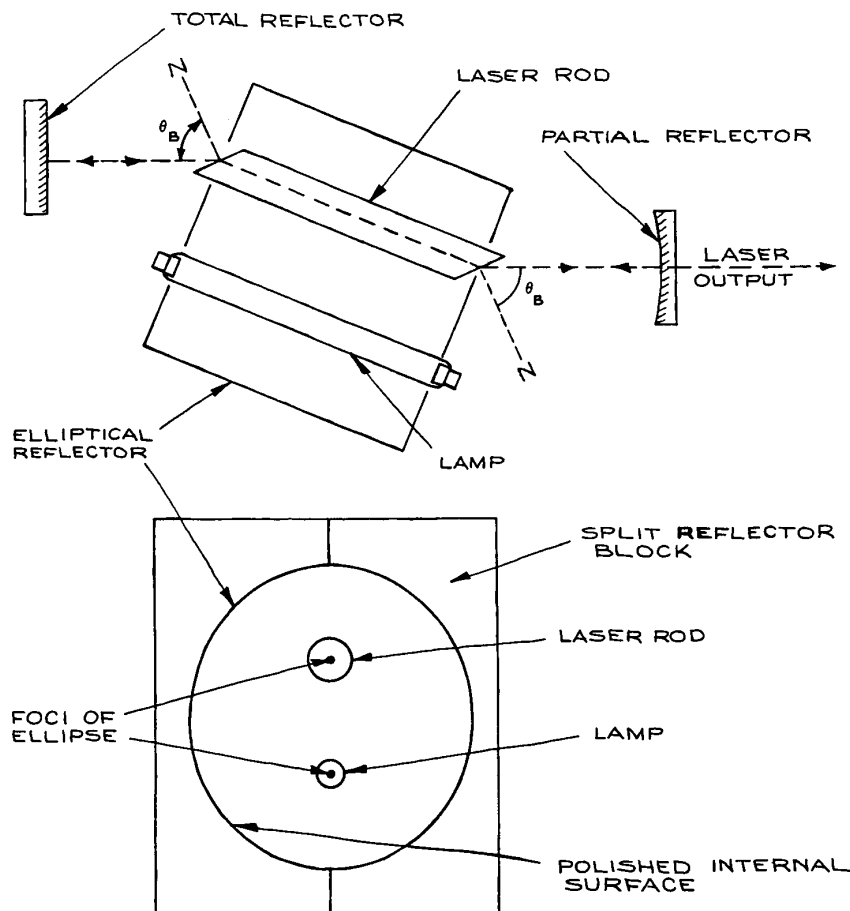
and any inherent losses of the amplifying medium itself – the last generally being significant only in solid-state, liquid, and semiconductor lasers. Consequently, it is particularly important to keep the mirrors and windows of a laser system clean and free of dust. Most commercial systems incorporate a flexible, sealed enclosure between Brewster windows and mirrors to exclude contaminants.

Figure 4.124 shows some design features of a simple lamp-pumped solid-state laser oscillator. This design incorporates a linear pulsed or continuous lamp and a crystalline or glass laser rod with Brewster windows mounted inside a

metal elliptical reflector, which serves to reflect pumping light efficiently into the laser rod. Although such solid-state lasers are rarely used in experiments where extreme frequency stability and narrow linewidth operation are required, it is still good practice to incorporate features such as stable, kinematic resonator design and thermal isolation of the hot lamp from the resonator structure. Elliptical reflector housings for solid-state lasers (and flashlamp-pumped dye lasers) can be made in two halves by horizontally milling a rectangular aluminum block with the axis of rotation of the milling cutter set at an angle of  $\arccos(a/b)$ , where  $a$  and  $b$  are the semiminor and semi-major axes of the ellipse. This process is repeated for a second aluminum block. The two halves are then given a high polish, or plated, and joined together with locating dowel pins. For further details of solid-state laser design and particulars of other arrangements for optical pumping the reader should consult Röss<sup>76</sup> or Koechner.<sup>77</sup> Flashlamp-pumped dye lasers share some design features with solid-state lasers. The main difference is that the former require excitation of short pulse duration ( $< 1 \mu\text{s}$ ) with special flashlamps in low-inductance discharge circuitry.<sup>78–81</sup>

### 4.6.3 Specific Laser Systems

**Gaseous Ion Lasers.** Gaseous ion lasers, of which the argon, krypton, and helium–cadmium are the most important, generate narrow-linewidth, highly coherent visible and ultraviolet radiation with powers in the range from milliwatts to a few tens of watts CW. Pulsed ion lasers operate at high current densities, but have low duty cycles, so that ambient cooling is adequate. These lasers are not in widespread experimental use. Helium–cadmium lasers use low-current-density (a few  $\text{A}/\text{cm}^2$ ) air-cooled discharge structures, which in many respects are similar to those of helium–neon lasers. They are limited to a few tens of milliwatts in conveniently available output power. They do, however, provide UV output at 325 nm. Argon and krypton ion lasers use very high-current-density discharges ( $100$ – $2000 \text{ A}/\text{cm}^2$ ) in special refractory discharge structures usually made of tungsten disks with ceramic spacers, as shown in Figure 4.125. In operation, these disks cool themselves by radiating through an outer fused-silica envelope surrounded by coolant water. Commercially available powers range up to a few tens of watts, but outputs up to 1 kW in



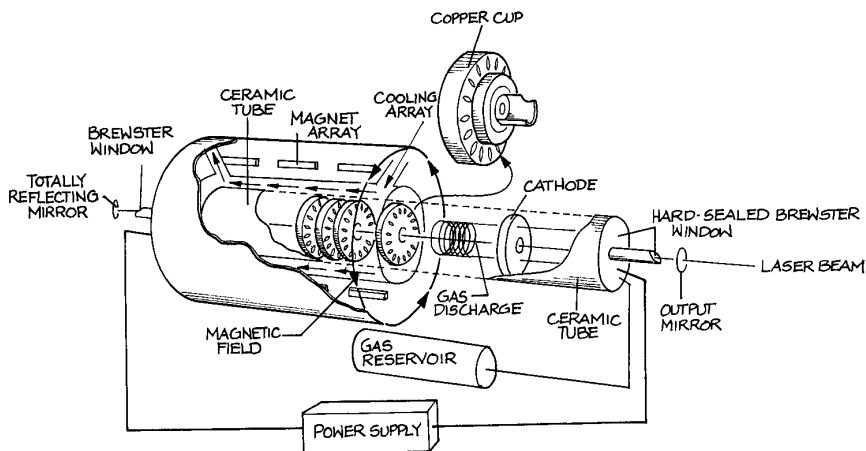
**Figure 4.124** Schematic diagram of a simple solid-state laser system.

the visible have been reported. Low power ( $< 100$  mW) argon ion lasers can be forced-air cooled. Comprehensive information about ion lasers can be found elsewhere.<sup>82,83</sup> Argon and krypton ion lasers are available commercially from Coherent, JSC Plasma, JDSU, Melles Griot, and Spectra Physics (Newport Corporation). Argon and krypton ion lasers have in the past been widely used for pumping CW dye lasers and Ti:sapphire lasers. They are, however, rapidly being replaced in these applications by frequency-doubled Nd lasers, which are more and more likely to be pumped themselves by semiconductor diode lasers. Because of the large currents at which they operate, argon and krypton ion lasers regularly need their plasma

tubes to be replaced (typically every few thousand hours of operation), which is an expensive business. It appears likely that these lasers will disappear from use and be universally replaced by DPSS lasers.

**Helium-Neon Lasers.** Of all types of laser, helium-neon lasers come closest to being the ideal classical monochromatic source. They can have very good amplitude ( $\approx 0.1\%$ ) and frequency stability (1 part in  $10^8$  without servo frequency control). Servo-frequency-controlled versions are already in use as secondary frequency standards, and have excellent temporal and spatial coherence.<sup>84</sup> Typical available power outputs range up to 50 mW.





**Figure 4.125** Construction of a high-power, water-cooled ion laser.

Single-frequency versions with power outputs of 1 mW, which are ideal for interferometry and optical heterodyne experiments, are available from Spectra-Physics and Aero-tech. Although 632.8 nm is the routinely available wavelength from He–Ne lasers, other wavelengths, such as 543 nm or 1.15 or 3.39  $\mu\text{m}$  are also available. Helium–neon lasers of low power are very inexpensive, costing from about \$100, and are ideal for alignment purposes. Because of their superior spectral qualities compared to semiconductor lasers that can operate at the same wavelength, they remain a valuable source in applications such as interferometry and metrology. Other suppliers include Newport/Spectra Physics, Meles Griot, and Micro-Controle.

**Helium–Cadmium Lasers.** These lasers are similar in many ways to helium–neon lasers. They operate at relatively low current densities and use discharge tube structures that do not require water cooling. They are quite important in applications requiring deep blue and ultraviolet light at powers of ten of milliwatts. They operate at 325 nm and 441.6 nm. These wavelengths, especially 325 nm, are useful in applications where a photochemically active source is required. They are widely used in photolithography. Helium–cadmium lasers are available from Kimmon Electric Co. and Melles Griot.

**CO<sub>2</sub> Lasers.** Both pulsed and CW CO<sub>2</sub> lasers are easy to construct in the laboratory. A typical low-pressure CW

version would incorporate most of the features shown in Figure 4.122. A water-cooled Pyrex discharge tube of internal diameter 10–15 mm, 1–2 m long, and operating at a current of about 50 mA, is capable of generating tens of watts output at 10.6  $\mu\text{m}$ . The best operating gas mixture is CO<sub>2</sub>–N<sub>2</sub>–He in the ratio 1:1:8 or 1:2:8, at a total pressure of about 25 mbar in a 10 mm-diameter tube. ZnSe is the best material to use for the Brewster windows, as it is transparent to red light and permits easy mirror alignment. The optimum output mirror reflectance depends on the ratio of the tube length to diameter,  $L/d$ , roughly according to:<sup>85</sup>

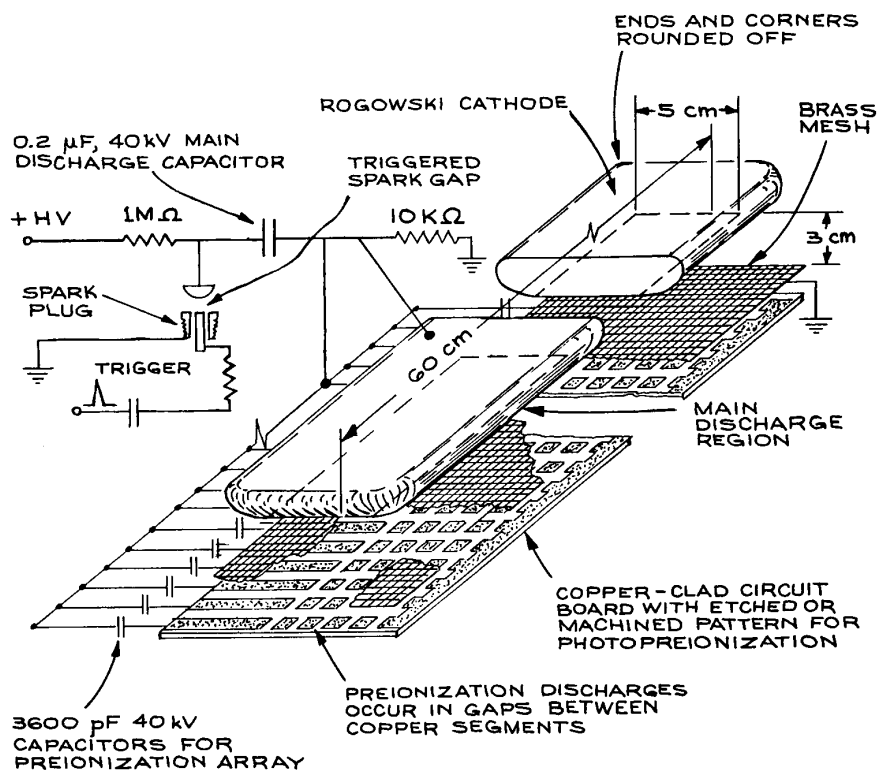
$$R \approx 1 - L/500d \quad (4.219)$$

The output mirror is usually a dielectric-coated germanium, gallium arsenide, or zinc selenide substrate. Such mirrors are available from Janos, Laser Optics, Coherent, Infrared Industries, Laser Research Optics, Unique Optical, and II-VI Infrared, and Rocky Mountain Instruments, among others. The total reflector can be gold-coated in lasers with outputs of a few tens of watts. For tunable operation, gratings are available from American Holographic, Gentec, Jobin-Yvon, Perkin-Elmer, Richardson Grating Laboratory, and Rochester Photonics. When a grating is used to tune a laser with Brewster windows, the grating should be mounted so that the  $E$  vector of the laser beam is orthogonal to the grating grooves. An excellent review of low pressure CW CO<sub>2</sub> gas lasers has been given by Tyte.<sup>85</sup>

Continuous wave CO<sub>2</sub> lasers can also be operated in a waveguide configuration, in which a much higher pressure gas mixture is excited in a small size discharge capillary.<sup>86</sup> These lasers provide substantially greater power output per volume than the low pressure variety. They can be excited by direct current, but r. f. excitation is now popular, and has now solved the unreliability problems that afflicted early commercial versions. Current suppliers of these compact lasers are Access Laser, Coherent, Edinburgh Instruments, Infrared Optical Products, Rofin-Sinar, Synrad, Trumpf, and Universal Laser Systems.

Transversely excited atmospheric-pressure (TEA) CO<sub>2</sub> lasers are widely used when high-energy, pulsed infrared energy near 10 μm is required. These lasers operate in discharge structures where the current flow is transverse to the resonator axis through a high-pressure (≥ 0.5 bar) mixture of CO<sub>2</sub>, N<sub>2</sub>, and He, typically in the ratio 1:1:5.

To achieve a uniform, pulsed glow discharge through a high-pressure gas, special electrode structures and pre-ionization techniques are used to inhibit the formation of localized spark discharges. Many very-high-energy types utilize electron-beam excitation. Discharge-excited TEA lasers are relatively easy to construct in the laboratory. A device with a discharge volume 60 cm long × 5 cm wide with a 3 cm electrode spacing excited with a 0.2 μF capacitor charged to 30 kV, will generate several joules in a pulse about 150 ns long. There are very many different designs for CO<sub>2</sub> TEA lasers; particulars of various types can be found in the *IEEE Journal of Quantum Electronics* and the *Journal of Applied Physics*. A particularly good design that is easy to construct is derived from a vacuum-ultraviolet photopreionized design described by Seguin and Tulip.<sup>87</sup> A diagram of the discharge structure is shown in Figure 4.126, together with its associated capacitor-discharge



**Figure 4.126** Cutaway view of vacuum-ultraviolet photopreionized Rogowski TEA laser structure.

circuitry. The cathode is an aluminum Rogowski  $2\pi/3$  profile,<sup>88</sup> the anode is a brass mesh, beneath which is a section of double-sided copper-clad printed circuit board machined into a matrix of separate copper sections on its top side. This printed circuit board provides a source of very many surface sparks for photopre ionization of the main discharge. The laser is fitted with two separate capacitor banks: each circuit-board spark channel is energized by a 3600 pF Sprague or similar “doorknob” capacitor, while the main discharge is energized with a 0.2  $\mu$ F low-inductance capacitor. Such capacitors are available from CSI Technologies, Hi Voltage Components, and Maxwell Labs. Both capacitor banks are switched with a single spark gap, which incorporates a Champion marine spark plug for triggering. The end of the spark plug, which has an annular gap, is ground down to its ceramic insulator and is triggered by an EG&G Model TM-11A Trigger generator. A good double Rogowski TEA laser design has been given by Sequin, Manes, and Tulip.<sup>89</sup> TEA lasers based on the multiple-pin-discharge design first described by Beaulieu<sup>90</sup> are also easy to construct, but they do not give output energies comparable to those of volume-excited designs. They do, however, lend themselves well to laser oscillation in many different gases – for example, HF formed by discharge excitation of  $H_2-F_2$  mixtures.

**CO Lasers.** Continuous wave CO lasers are essentially similar in construction to  $CO_2$  lasers except that the discharge tube must be maintained at low temperatures – certainly below 0°C. They operate in a complex mixture of He, CO,  $N_2$ ,  $O_2$ , and Xe with (for example) 21 mbar He, 0.67 mbar  $N_2$ , 0.013 mbar  $O_2$ , and 0.4 mbar Xe. They are available from Access Laser and Edinburgh Instruments.

**Exciplex (Excimer) Lasers.** Exciplex lasers operate in high-pressure mixtures such as Xe- $F_2$ , Xe- $Cl_2$ , Kr- $Cl_2$ , Kr- $F_2$ , and Ar- $F_2$ . They are commonly also called “excimer” lasers, although this is not strictly correct scientific terminology except in the case of the  $F_2^*$  laser. Both discharge TEA and E-beam-pumped configurations are used. Under electron bombardment, a series of reactions occur that lead to the formation of an excited complex (exciplex) such as  $XeF^*$ , which is unstable in its ground state and so dissociates immediately on emitting light. Excimer lasers available commercially are sources of intense pulsed ultra-

violet radiation. The radiation from these lasers is not inherently of narrow linewidth, because the laser transition takes place from a bound to a repulsive state of the exciplex, and therefore is not of well-defined energy. Excimer lasers are attractive sources for pumping dye lasers and whenever intense pulsed UV radiation is required in, for example, photolithography or photochemistry.

The beam quality from an excimer laser is not normally very high. These are “multimode” lasers, with often a rectangular output beam profile. These lasers are widely available commercially from companies such as Gam Laser, JP Sercel Associates (JPSA), Lambda Physik (now Coherent), Optec, MPB Communications, and SOPRA.

Excimer lasers share technological design features in common with both  $CO_2$  TEA lasers and nitrogen lasers. They use rapid pulsed electrical discharges through the appropriate high-pressure gas mixture. The discharge design must be of low inductance and the excitation is generally faster than is necessary to operate a  $CO_2$  TEA laser. Because these lasers use toxic gases, such as fluorine and chlorine, appropriate gas-handling safety precautions must be taken in their use. The gases used, which are usually diluted mixtures of a halogen in a noble gas, should be housed in a vertical gas cabinet and exhaust gases from the laser should be sent through a chemical “scrubber” or filter cartridge before venting into the atmosphere.

**Nitrogen ( $N_2$ ) Lasers.** Molecular-nitrogen lasers are sources of pulsed ultraviolet radiation at 337.1 nm. Commercially available low-energy devices usually operate on a flowing nitrogen fill at pressures of several tens of mbar, and generate output powers up to about 1 MW in pulses up to about 10 ns long. These lasers are reliable and easy to operate, and are widely used for pumping low-energy pulsed dye lasers. Nitrogen lasers, by virtue of their internal kinetics, only operate in a pulsed mode. They use very rapid transversely excited discharges, usually between two parallel, rounded electrodes energized with small, low-inductance capacitors charged to about 20 kV. The whole discharge arrangement must be constructed to have very low inductance. Special care must be taken with insulation in these devices; the very rapidly changing electric fields present can punch through an insulating material to ground under circumstances where the same applied d.c. voltage

would be very adequately insulated. There are several good  $N_2$ -laser designs in the literature.<sup>91–96</sup> These designs fall into two general categories: Blumlein-type distributed-capacitance discharge structures, and designs based on discrete capacitors. The latter, although they do not give such high-energy outputs from a given laser as the very rapid-discharge Blumlein-type structures, are much easier to construct in the laboratory and are likely to be much more reliable. The capacitors used in these lasers need to be of low inductance and capable of withstanding the high voltage reversal and rapid discharge to which they will be subjected. Suitable capacitors for this application can be obtained from Murata or Sprague.

The output beam from a transversely excited  $N_2$  laser is by no means Gaussian. It generally takes the form of a rectangular beam, typically  $\simeq 0.5 \text{ cm} \times 3 \text{ cm}$ , with poor spatial coherence. Such a beam can be focused into a line image with a cylindrical lens for pumping a dye cell in a pulsed dye laser. Nitrogen lasers are available commercially from LTB Lasertechnik and Spectr-Physics (now Newport).

**Copper Vapor Lasers.** These are high-pulse-repetition-frequency (prf) pulsed gas lasers that operate using high-temperature gas mixtures of copper vapor in a helium or neon buffer gas. With individual pulse energies up to several millijoules and prf up to 20 kHz, these lasers provide average powers up to several watts. They are used for pumping pulsed dye lasers. The principal operating wavelengths are 510.6 and 578.2 nm. Suppliers include Laser Consultants and Oxford Laser. Pulsed gold vapor lasers are also available commercially, which are similar in most respects. They operate at 627.8 nm.

**Continuous wave Solid-State Lasers.** The most important laser in this category is the Nd:YAG laser, where YAG (yttrium aluminum garnet  $Y_3Al_5O_{12}$  is the host material for the actual lasing species – neodymium ions,  $Nd^{3+}$ ). Other host materials are also used: YLF (yttrium lithium fluoride),  $LiYF_4$ ,  $YVO_4$  (yttrium vanadate), YALO (yttrium aluminum oxide  $YAlO_3$ ) and GSGG (scandium-substituted gadolinium gallium garnet,  $(Gd_3Sc_2Ga_3O_{12})$ ). The principal neodymium output wavelength is 1.06  $\mu\text{m}$ , but other infrared wavelengths, especially 1.32  $\mu\text{m}$ , are available. There are many suppliers.<sup>31</sup> The 1.06  $\mu\text{m}$  can

be efficiently doubled to yield a CW source of green radiation. This source of green radiation has replaced the argon-ion laser in many applications, especially since the neodymium laser itself can be pumped so conveniently with appropriate semiconductor lasers.

Ti:sapphire is an attractive laser material for generating tunable radiation in the 700–1060 nm range. These lasers are generally pumped by argon-ion lasers on frequency-doubled neodymium lasers. Suppliers include Coherent, Inc., Continuum, Photonics Industries International, and Quantronix.

Fiber lasers use a gain medium that is a fiber doped with rare-earth ions such as erbium ( $Er^{3+}$ ), neodymium ( $Nd^{3+}$ ), ytterbium ( $Yb^{3+}$ ), thulium ( $Tm^{3+}$ ), or praseodymium ( $Pr^{3+}$ ). They are pumped by one or more semiconductor laser diodes.<sup>97</sup> A simple schematic of how they are constructed is shown in Figure 4.127. In practice, the pump light may be coupled in through the cladding of the fiber. Very high-power fiber lasers with powers above 100 W are increasingly being used in industrial applications. Suppliers include B&W Tek, Fibercore, Fibertek, IPG Photonics, Micro-Controle, Newport, Nufren, and Toptica. Fiber lasers can be mode-locked to generate very short pulses ( $< 1 \text{ ps}$ ) in the infrared. Pulsed  $Er^{3+}$  fiber lasers are available from Calmar Optcom, PriTel, and IPG Photonics. Pulsed  $Yb^{3+}$  lasers are available from Calmar Optcom, Clark-MXR, IPG Photonics, Laser Photonics, and Quantel. An important application of erbium-doped fiber is in erbium-doped optical fiber amplifiers (EDFAs), widely used in fiber optic communication networks to amplify 1.55  $\mu\text{m}$  semiconductor laser radiation.

*F*-center lasers<sup>98</sup> are doped-crystal lasers that are pumped with argon, krypton, or dye lasers. They provide tunable operation in two regions: between 1.43 and 1.58  $\mu\text{m}$  and between 2.2 and 3.3  $\mu\text{m}$ , but are somewhat inconvenient, because the crystal must be cooled with liquid nitrogen. They no longer appear to be available commercially.

**Pulsed Solid-State Lasers.** Three lasers are most important in this category Nd:YAG, Nd:glass and ruby. Nd:YAG and Nd:glass lasers oscillate at the same principal wavelength, 1.06  $\mu\text{m}$ . Nd:YAG is generally used in high-pulse-repetition-rate systems and/or where good beam quality is desired. High-energy Nd laser systems frequently

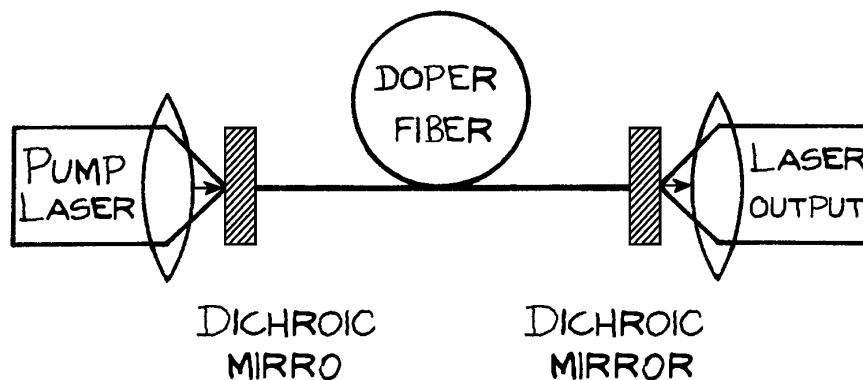


Figure 4.127 Schematic construction of a fiber laser.

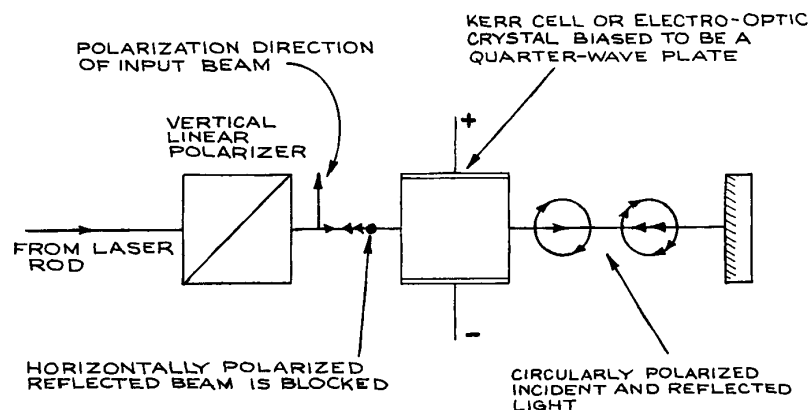


Figure 4.128 Schematic arrangement for  $Q$ -switching. Laser oscillation occurs when the voltage bias on the Kerr cell or electro-optic crystal is removed, thereby allowing the reflected beam to remain linearly polarized in the vertical direction.

incorporate an Nd:YAG oscillator and a series of Nd:glass amplifiers. There are many suppliers.<sup>31</sup> Ruby lasers, despite being the first laser to be demonstrated, are little used in research. They have medical applications because they provide a high-energy pulsed deep-red output (694.3 nm). They are available from Continuum.

All these pulsed solid-state lasers use fairly long flash excitation ( $\approx 1$  ms). Unless the laser pulse output is controlled, they generate a very untidy optical output, consisting of several hundred microseconds of random optical pulses about  $1 \mu\text{s}$  wide, spaced a few microseconds apart. This mode of operation, called *spiking*, is rarely used. Usually the laser is operated in a  $Q$ -switched mode. In this

mode the laser cavity is blocked with an optical shutter such as an electro-optic modulator or Kerr cell, as illustrated in Figure 4.128. An appropriate time after flashlamp ignition, after the population inversion in the laser rod has had time to build to a high value, the shutter is opened. This operation, which changes the  $Q$  of the laser cavity from a low to a high value, causes the emission of a very large short pulse of laser energy. The  $Q$ -switched pulse can contain nearly as much energy as would be emitted in the “spiking” mode and is typically 10–20 ns long. Passive  $Q$ -switching can also be accomplished with a “bleachable” dye solution placed in an optical cell inside the laser cavity. The dye is opaque at low incident light intensities

and inhibits the buildup of laser oscillation. When, however, the laser acquires enough gain to overcome the intracavity dye absorption loss, laser oscillation begins, the dye bleaches, and a  $Q$ -switched pulse results. A bleachable dye cell with a short relaxation time, placed close to one of the laser mirrors, will frequently lead to mode-locked operation<sup>99–101</sup> within the  $Q$ -switched pulse (see Section 4.6.4).

**Semiconductor Lasers.** Five principal types of semiconductor laser are available commercially:

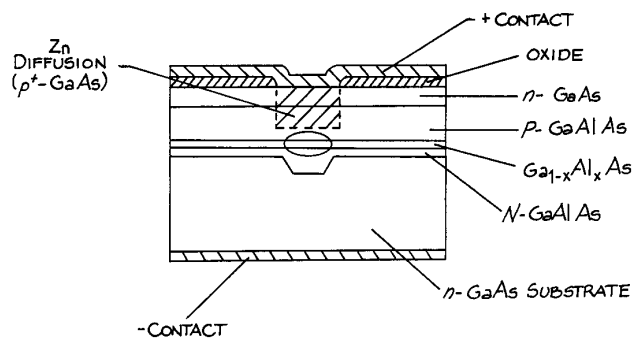
- (1) Wide bandgap semiconductor lasers based on GaN/AlGaN, which, depending on stoichiometry, can emit down to 375 nm. These UV, blue, and green diode lasers have power outputs of several milliwatt and above. The developer of the shortest wavelength laser diodes based on GaN is Nichia, other manufacturers include Coherent, Cree, Koheras, Micro Laser Systems, Sony, and Toptica. Wavelengths above 375 nm are obtained from diodes with elements such as indium, aluminum, gallium, or nitrogen added, which produces laser diodes with wavelengths from 480 to 520 nm. The availability of short wavelength diode and diode pumped solid state lasers has significantly reduced the market and applications for visible and UV gas lasers such as argon ion, helium-cadmium, helium–neon, and krypton ion.
- (2) GaAs or GaAlAs lasers, which operate at fixed wavelengths in the region between 0.75 and 0.9  $\mu\text{m}$ . These lasers are very inexpensive and reliable; they are widely used in communication and information-processing systems, and in compact-disc players. High-power versions are used as the pumping source in DPSS lasers.
- (3) InGaAsP lasers operate principally in the 1.3 and 1.55  $\mu\text{m}$  regions for telecommunications applications. They have been developed to operate at the optimum wavelengths for minimum dispersion and lowest loss in optical-fiber communication systems. Near infrared semiconductor lasers are described in detail in many specialized texts.<sup>102–110</sup>
- (4) Tunable diode lasers, which utilize lead-salt semiconductors such as PbSSe, PbSnSe, and PbSnTe can provide tunable operation over limited regions anywhere between 3 and 30  $\mu\text{m}$ . The tunable lead-salt diode

lasers are now commonly used in high-resolution spectroscopy.<sup>111</sup> They generally emit several modes, which, because of the very short length of the laser cavity, are spaced far enough in wavelength for a single one to be isolated by a monochromator. The output laser linewidths that are obtainable in this way are about  $10^{-4}/\text{cm}^1$  (3 MHz). Scanning is accomplished by changing the temperature of the cryogenically cooled semiconductor or the operating current. Complete tunable-semiconductor-laser spectroscopic systems are available from Quantum Associates or Spectra-Physics (Laser Analytics).

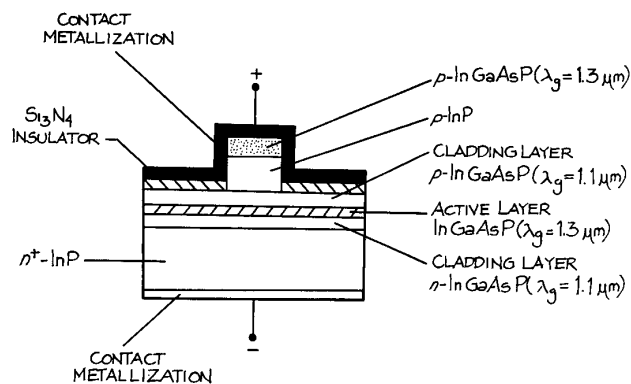
- (5) Quantum cascade lasers<sup>112</sup> can provide tunable operation over specific limited spectral regions within a range from 3 to 211  $\mu\text{m}$ . The wavelength range varies from manufacturer to manufacturer. These lasers use multiple quantum wells, a so-called *superlattice*, to provide laser operation between *sub-bands* in the conduction bands of the structure. These lasers do not use  $p$ - $n$  junctions for their operation. Quantum cascade lasers are available from Alpes, Boston Electronics, Cascade Technologies, and Sacher Lasertechnik.

**Semiconductor Laser Properties.** Only researchers involved in semiconductor laser development will fabricate these lasers. Their laser performance is, however, strongly influenced by their generally small size, and device structure. Figures 4.129–4.132 give examples of some representative structures that are used in current semiconductor diode lasers. The bias laser structure involves a  $p$ - $n$  junction through which a drive current flows. In all these structures the current flow through the device is controlled so that a specific volume of the active region is excited. This region is generally very small. The active length is not generally more than 1–2 mm. The lateral dimensions of the active region are generally in the 5–10  $\mu\text{m}$  range, however, in single quantum well (SQW) and multiple quantum well (MQW) lasers, the dimension of the active region in the direction of current flow can be as small as 10–100 nm.

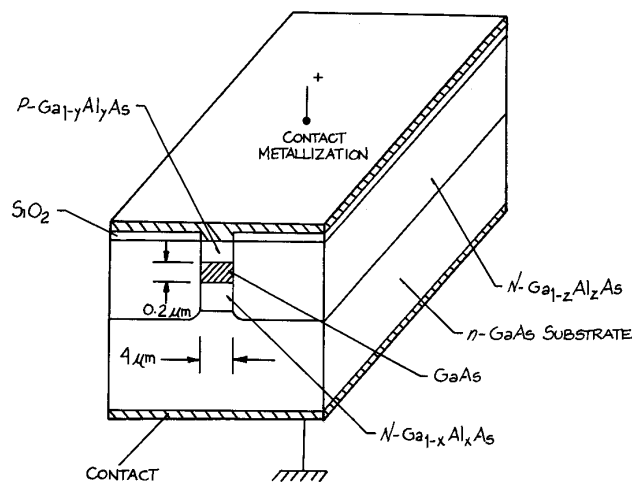
Laser oscillation occurs in a resonant Fabry–Perot structure made up of the two end facets of the laser. The active region acts like a waveguide to confine the laser beam, which ideally emerges from the emitting facet as an elliptical Gaussian beam. This is a Gaussian beam whose spot



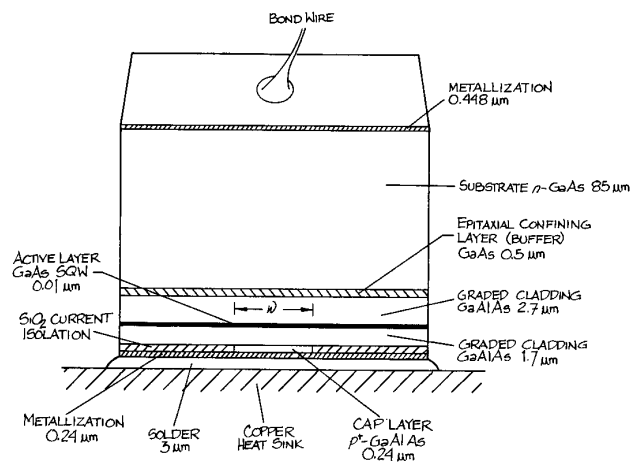
**Figure 4.129** Double heterostructure laser diode design using a stripe excitation region confined by an oxide layer and deep zinc diffusion.



**Figure 4.131** Cladded ridge double heterostructure semiconductor laser design.



**Figure 4.130** Double heterostructure laser of a buried stripe design.

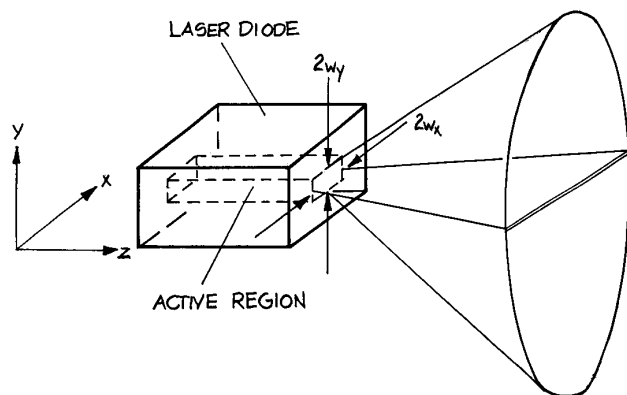


**Figure 4.132** Graded index separate confinement heterostructure (GRINSCH) semiconductor laser design. Typical dimensions of the various layers are indicated. For a narrow stripe laser  $w \sim 5 \mu\text{m}$ , for a broad stripe laser  $w \sim 60 \mu\text{m}$ .

sizes  $w_x$ ,  $w_y$  are different in the two orthogonal ( $x$ ,  $y$ ) directions perpendicular to the  $z$  propagation direction of the laser beam. The smaller spot size is generally in the direction perpendicular to the junction. This asymmetry of the laser beam causes its beam divergence to be larger in the direction perpendicular to the junction than it is in the orthogonal direction, as shown schematically in Figure 4.133. Commercial diode lasers are often supplied with an *anamorphic* beam expander, which circularizes the laser beam and equalizes its orthogonal beam divergence angles, however, in general the beam quality from any

semiconductor laser is inferior to that from most gas lasers. A circular, clean, output beam can be obtained from a fiber-coupled diode laser. Vertical cavity surface emitting lasers (VCSELs) also provide circular output beams.

The spectral purity of Fabry–Perot semiconductor lasers is generally inferior to helium–neon lasers or diode-pumped solid-state lasers. Semiconductor laser linewidths are typically several MHz or more, so their coherence lengths are short. This makes them less suitable in most



**Figure 4.133** Schematic diagram of how a semiconductor laser emits a quasi-elliptical output beam because of asymmetrical beam confinement in two orthogonal directions inside the laser structure.

interferometric applications than most gas lasers or diode-pumped solid-state (DPSS) lasers. Narrower linewidth semiconductor lasers use an external cavity incorporating a grating structure, or have a Bragg grating structure incorporated into the semiconductor material itself: so-called distributed feedback (DFB) or distributed Bragg reflection (DBR) lasers. Suppliers include Agere, AMS Technologies, EXFO, Infineon, Lasermate, Sacher Lasertechnik, and Toptica. These lasers are mostly available at the telecommunications wavelengths of 1300 nm and 1550 nm.

Diode lasers have achieved their greatest commercial importance in telecommunications systems, but the availability of diode lasers at many discrete wavelengths from 375 nm to beyond  $2\ \mu\text{m}$ <sup>31</sup> has increased their use in spectroscopic applications. They can provide some degree of tunability by changing their operating temperature or drive current. Suppliers include Applied Optonics, Coherent, DILAS, Laser Diode, Lasermate Group, Lasertel, Melles Griot, New Focus, Newport, Opton, Photonic Products, Point Source, and QPC Lasers. Manufacturers' data sheets should be consulted for wavelengths, mode quality (single or multimode), beam quality (circular or elliptical), powers, and operational currents.<sup>31</sup> Many of these lasers can be obtained coupled to an optical fiber, which provides a clean, circular beam.

Some of the commonly available wavelengths (in nm) from diode lasers are: 375, 405, 408, 440, 635, 640, 650,

670, 690, 730, 780, 808, 830, 915, 940, 980, 1300, 1370, 1470, 1550, 1720, 1850, 1907, 1950, 2004, and 2350.

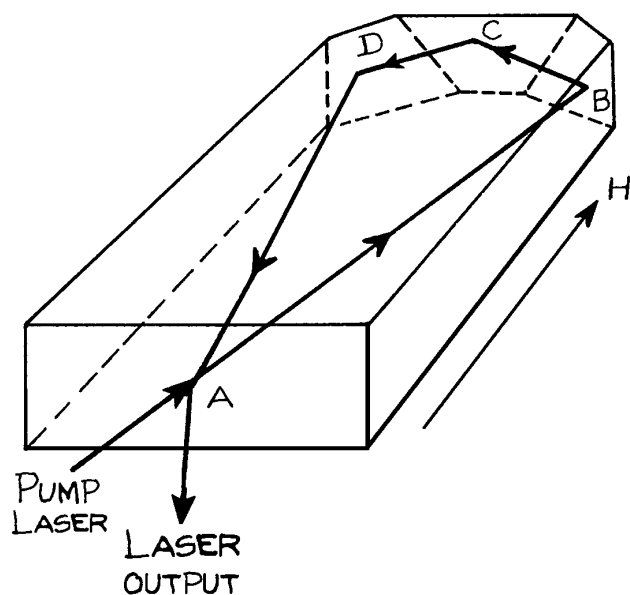
High-power diode lasers are available with powers up to and beyond 1 kW. They are widely used as pump sources in diode-pumped solid-state lasers; however, as far as spectroscopy purity is concerned they are closer in character to a light-emitting diode than they are to high-coherence lasers such as helium–neon, helium–cadmium or diode-pumped Nd:YAG.

**Diode-Pumped CW Solid-State Lasers.** Because of a fortuitous coincidence between the output of GaAlAs semiconductor lasers working near 809 nm and a strong absorption of the neodymium ions ( $\text{Nd}^{3+}$ ) in YAG, YLF, or  $\text{YVO}_4$ , it is very efficient to optically pump  $\text{Nd}^{3+}$  lasers with diode lasers. These lasers have become widely available commercially. They can also be efficiently frequency doubled to 530 nm, so have begun to replace ion lasers in applications where intense green light is required. Green output powers up to 70 W are available.

Diode-pumped lasers in which the semiconductor lasers are arranged to inject their pump radiation through the cylindrical faces of a laser rod are in many ways akin to lamp-pumped lasers of the same kind. Because of the high electrical efficiency of GaAlAs laser diodes (>10%) the overall electrical-to-optical conversion efficiency of these lasers is very high, approaching 10%. If the diode laser pump radiation is injected along the axis and matched to the transverse mode geometry of the solid-state laser then very stable, narrow linewidth laser oscillation can be obtained: Figure 4.134 shows the clever monolithic Nd:YAG laser design of Kane and Byer.<sup>113</sup> In this laser the magneto-optical properties of YAG are used to obtain unidirectional amplification of a traveling wave in the laser cavity. This prevents the spatial hole-burning that can occur in a standing wave homogeneously broadened laser, which can allow multiple longitudinal modes to oscillate.<sup>15</sup>

Frequency doubling of a DPSS laser is generally carried out with an intracavity nonlinear crystal such as BBO (beta- $\text{BaB}_2\text{O}_4$ ), PPLN (periodically poled lithium niobate), or PPKTP [periodically poled potassium titanium oxide phosphate ( $\text{KTiOPO}_4$ )], or a with an external resonant cavity containing a nonlinear crystal. Figure 4.135 shows how intracavity frequency doubling is done with a



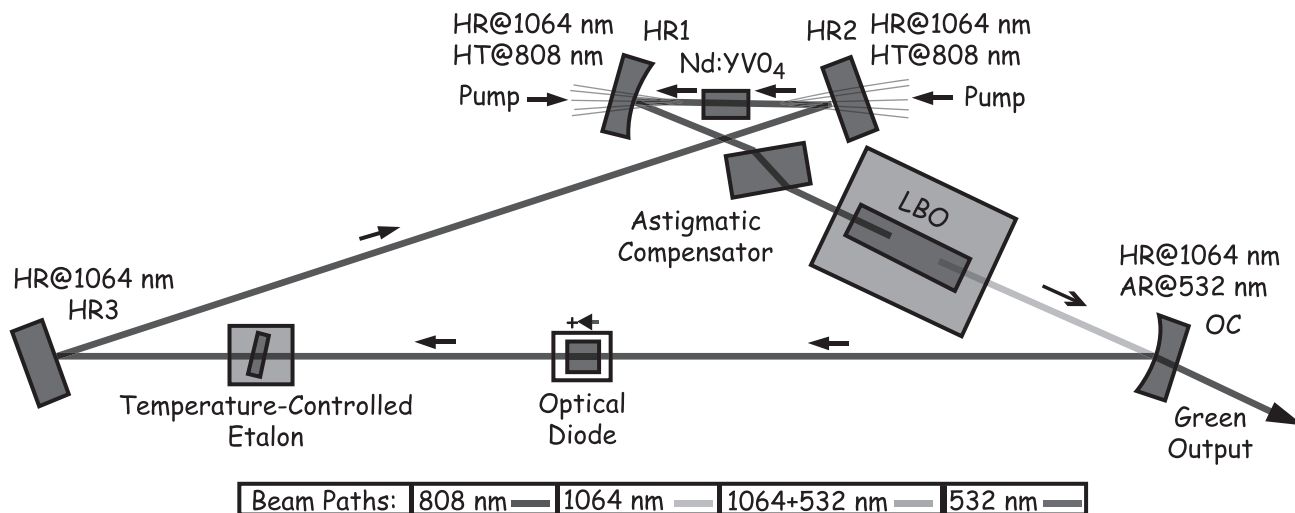


**Figure 4.134** The "MISER" ring laser design of Kane and Byer.

Nd:YVO<sub>4</sub> coherent Verdi laser. This use of a ring laser resonator with unidirectional laser oscillation provides single longitudinal mode operation and a narrow linewidth. Frequency-tripled operation of diode-pumped Nd<sup>3+</sup> laser at 355 nm is available at powers in excess of 3 W, frequency-quadrupled operation at 266 nm provides powers up to 500 mW. There are many suppliers of diode-pumped Nd<sup>3+</sup> lasers, including A-B Lasers, Alphas, Cobolt AB, Coherent, Continuum, Lee Laser, Lightwave Electronics, Melles Griot, Quantronix, Spectra-Physics, TecOptics, and TRUMPF, among others.<sup>31</sup> Diode-pumped solid-state lasers are also available at other wavelengths, notably 2.1 μm in holmium (Ho<sup>3+</sup>), 1.6 μm and 2.8 μm in erbium (Er<sup>3+</sup>) and 2.3 μm in thulium (Tm<sup>3+</sup>).

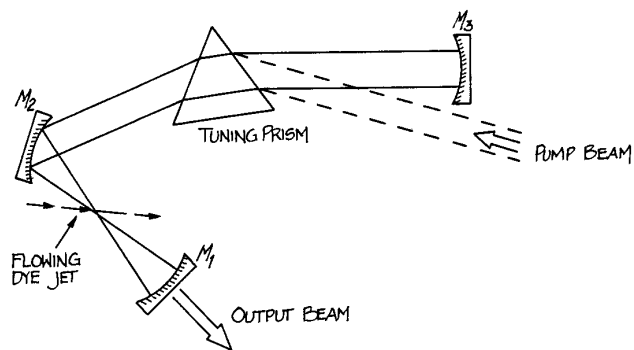
The principal advantages of this type of laser are reliability, narrow linewidth, and single-frequency operation.

**Continuous Wave Dye Lasers.** Continuous Wave dye lasers usually take the form of a planar jet of dye solution continuously sprayed at Brewster's angle from a slit nozzle, collected, and recirculated. The jet stream is in



### Ring Cavity Resonator of Coherent, Inc. Verdi Green DPSS Laser

**Figure 4.135** The layout of the Coherent, Inc. "Verdi" laser where intracavity frequency doubling with a lithium triborate LiB<sub>3</sub>O<sub>5</sub> (LBO) crystal is used to generate 530 nm from 1.06 μm. The initial laser oscillation is at 1.064 μm using a neodymium-doped yttrium vanadate (Nd:YVO<sub>4</sub>) crystal pumped with an 808 nm diode laser. In the diagram HR – high reflectance mirror; OC – output coupling mirror.

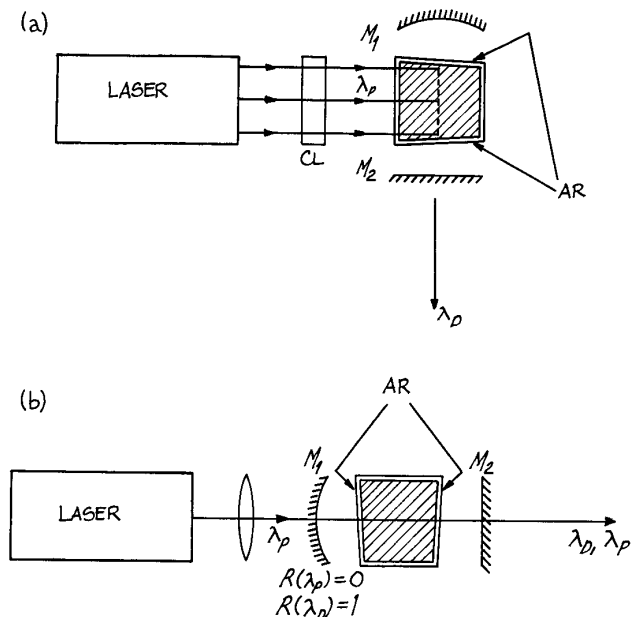


**Figure 4.136** Layout of a CW dye laser using a flowing dye jet.

a spherical mirror laser cavity, where it is pumped by the focused radiation of a CW ion laser or frequency-doubled Nd:YAG laser, as shown in Figure 4.136. These lasers can generally be regarded as wavelength converters – they have essentially the linewidth, frequency, and amplitude stability of the pumping source. For example, narrow-line-width operation requires an etalon-controlled single-frequency ion laser. Continuous Wave dye lasers are quite efficient at converting the wavelength of the pump laser; at pump powers far enough above threshold, the conversion efficiency ranges from 10 to 45%. Laser dyes are available from Exciton Lambda Physik, and Radiant Dyes Laser Accessories, among others. Dye lasers are rapidly being replaced, wherever possible, with tunable solid-state devices. The dyes used photodegrade and liquid spills are common. The laboratories of most dye-laser users are stained with the dyes that they use! For further details of CW-dye-laser operation consult Schafer.<sup>78</sup> Continuous wave dye lasers are available from Big Sky Laser Technologies, Coherent, Quantel and Spectra-Physics.

**Pulsed Dye Lasers.** Continuous wave dye lasers are not as frequently constructed in the laboratory as are pulsed dye lasers. In its simplest form, a pulsed dye laser consists of a small rectangular glass or quartz cell placed on the axis of an optical resonator and excited with a focused line image from a pump laser. Two simple geometrics are shown in Figure 4.137.

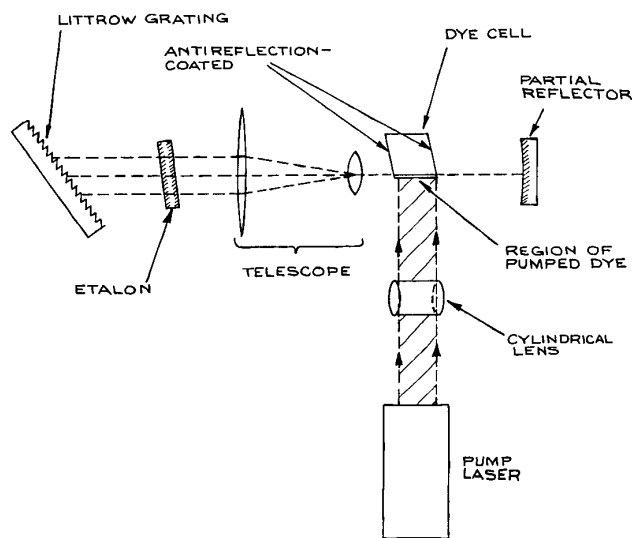
The most commonly used pump lasers are  $N_2$  and frequency-doubled Nd:YAG, although ruby, excimer and



**Figure 4.137** Two simple pulsed dye laser arrangements: (a) transverse pumping; (b) longitudinal pumping.

copper-vapor lasers are also used. In its simplest form, this type of pulsed dye laser will generate a broad-band laser output (5–10 nm). Various additional components are added to the basic design to give tunable, narrow-line-width operation. A very popular design that incorporates all the desirable features of a precision-tunable visible source has been described by Hänsch.<sup>114</sup> The essential components of a Hänsch-type pulsed dye laser are shown in Figure 4.138.

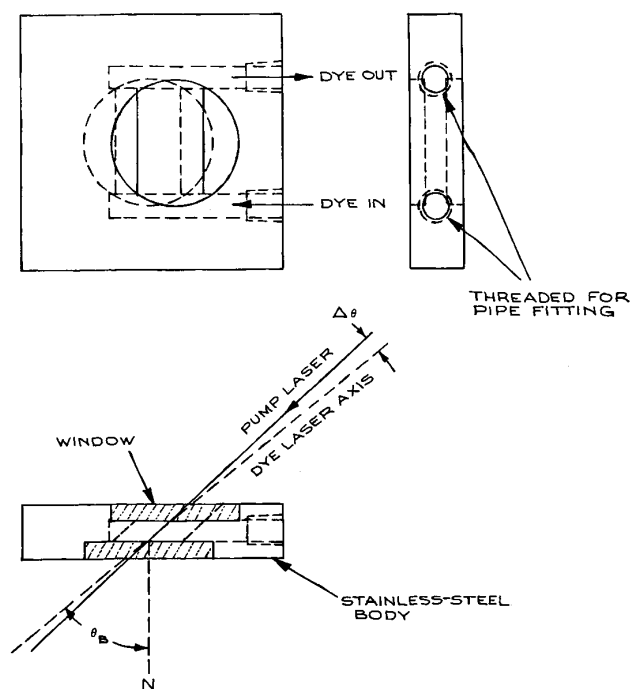
The pump laser is focused, with a cylindrical and/or a spherical lens, to a line image in the front of a dye cell. The line of excited dye solution (for example a  $5 \times 10^{-3}$  molar solution of Rhodamine 6G in ethanol) is the gain region of the laser. The laser beam from the gain region is expanded with a telescope in order to illuminate a large area of a high dispersion, Littrow-mounted echelle grating. Rotation of the grating tunes the center wavelength of the laser. Without the telescope, only a small portion of the grating is illuminated, high resolution is not achieved, and the laser linewidth will not be very narrow. With telescope and grating alone, the laser linewidth will be on the order of 0.005 nm at 500 nm ( $\approx 0.2/\text{cm}$ ). Even narrower linewidth can be achieved by including a tilted Fabry–Perot etalon



**Figure 4.138** Hänsch-type pulsed dye laser.

in the cavity. A suitable etalon will have a finesse of 20 and a free spectral range (see Section 4.7.4) below about 1/cm in this case. Line widths as narrow as  $4 \times 10^{-4}$  nm can be achieved. A few experimental points are worthy of note: the laser is tuned by adjusting grating and/or etalon; alignment of the telescope to give a parallel beam at the grating is quite critical; the dye should be contained in a cell with slightly skewed faces to prevent spurious oscillation; the dye solution can remain static when low-power (10–100 kW) pump lasers are used, magnetically stirred when pump lasers up to about 1 MW are used, or continuously circulated from a reservoir when higher-energy or high-repetition-rate pump lasers are used. Suitable dye cells are available from Anderson Lasers, Esco, Hellma Cells and NSG Precision Cells. Cells can be constructed from stainless steel with Brewster window faces when end-on pumping with a high-energy laser is used (as shown in Figure 4.139). The dye can be magnetically stirred by placing a small Teflon-coated stirring button in the bottom of the dye cell. When continuous circulation of dye is used, suitable pumps are available from Fluid-o-Tech and Micropump Corporation.

Once a narrow-linewidth pulsed-dye-laser oscillator has been constructed, its output power can be boosted by pass-



**Figure 4.139** Flowing dye cell for end-pumped operation.  $N$  = normal to windows;  $\theta_B$  = Brewster's angle, approximately the same for both pump and dye laser wavelengths;  $\Delta\theta$  = a small angle. Dye solution flows laminarily in the channel between the windows.

ing the laser beam through one or more additional dye cells, which can be pumped with the same laser as the oscillator. In a practical oscillator-amplifier configuration, 10% of the pump power will be used to drive the oscillator and 90% to drive the amplifier(s). Further details of Hänsch-type dye-laser construction are available in several publications.<sup>115–119</sup>

Two other approaches to achieving narrow-linewidth laser oscillation are worthy of note. The intracavity-telescope beam expander can be replaced, by a multiple-prism beam expander (see Section 4.3.4), which expands the beam onto the grating in one direction only. Such an arrangement does not need to be focused. An alternative, simple approach described by Littman<sup>118,119</sup> is to eliminate the beam expander and instead use a Littrow-mounted holographic grating in grazing incidence, as shown in

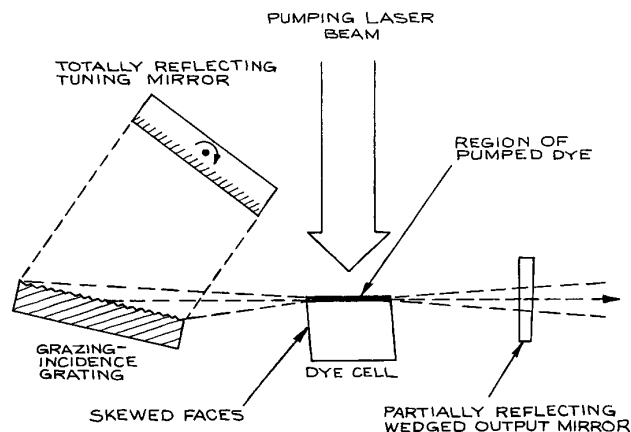


Figure 4.140 Littman-type dye laser.

Figure 4.140. Linewidths below 0.08 cm can be obtained without the additional use of an intracavity etalon.

**Optical Parametric Oscillators.** At one time, *optical parametric oscillators* (OPOs) were not readily available commercially, and were tricky to use. This has changed. Optical parametric oscillators provide tunable radiation typically in the range 210–5000 nm. They operate by using nonlinear optical processes in a crystal.<sup>115,121–125</sup>

The OPO requires a *pump* laser of frequency  $\nu_p$ , which is passed through a nonlinear crystal such as PPLN. Two new frequencies are generated in this process, a *signal* of frequency  $\nu_s$ , and an *idler* of frequency  $\nu_i$ . Energy is conserved in this process so that  $\nu_s = \nu_p - \nu_i$ . Tunability results by changing the angle of the pump beam with respect to the optic axis of the crystal, or by changing the temperature of the crystal. One or more of the three frequencies, pump, signal, or idler, is resonated inside a resonant structure that surrounds the nonlinear crystal. Figure 4.141 shows a schematic of an OPO with a resonant signal beam. An OPO generally generates tunable radiation with relatively broad linewidths ( $>1$  cm). Special techniques are needed to generate narrow linewidths.<sup>125</sup> Optical parametric oscillators are available from Coherent, Continuum, EKSPLA, Linos Photonics, Oportek, Quantel, Radiantis and Spectra Physics. Pulsed OPOs are more readily available than CW OPs and have larger tuning ranges. They can also generate very short (femtosecond) pulses.

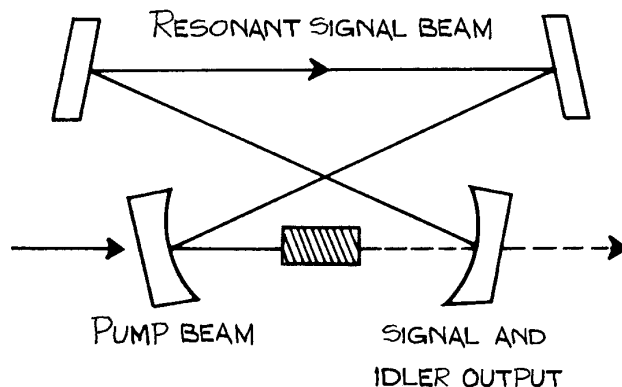


Figure 4.141 Schematic layout of an optical parametric oscillator (OPO).

#### 4.6.4 Laser Radiation

Laser radiation is highly monochromatic in most cases, although flashlamp-pumped dye lasers in a worst case may have linewidths as large as 100 cm (2.5 nm at 500 nm). The spectral characteristics of the radiation vary from laser to laser, but will be specified by the manufacturer. Lasers are usually specified as *single-mode lasers* or *multimode lasers*.

A *single-mode* laser emits a single-frequency output. Continuous wave CO<sub>2</sub> lasers and other middle- and far-infrared lasers usually operate in this way, and helium–neon and argon–ion lasers can also be caused to do so. The single output frequency will itself fluctuate randomly, typically over a bandwidth of 100 kHz–1 MHz, unless special precautions are taken to stabilize the laser – usually by stabilizing the spacing between the two mirrors that constitute its resonant cavity.

*Multimode* lasers emit several modes, called longitudinal modes, spaced in frequency by  $c_0/2L$ , where  $c_0$  is the velocity of light in vacuo and  $L$  is the optical path length between the resonator mirrors. In a medium of refractive index  $n$  and geometric length  $l$ ,  $L = nl$ . There may, in fact, be several superposed combs of equally spaced modes in the output of a multimode laser if it is not operating in a single transverse mode. The transverse modes specify the different spatial distributions of intensity that are possible in the output beam. The most desirable such mode, which is standard in most good commercial lasers, is the

fundamental TEM<sub>00</sub> mode, which has a Gaussian radial intensity distribution (see Section 4.2.4). A laser will generally operate on multiple longitudinal modes if the linewidth  $\Delta\nu$  of the amplifying transition is much greater than  $c_0/2L$ . In this case, the output frequencies will span a frequency range on the order of  $\Delta\nu$ . In short-pulse lasers the radiation oscillating in the laser cavity may not be able to make many passes during the duration of the laser pulse, and (particularly if the linewidth of the amplifying transition is very large) the individual modes may not have time to become well characterized in frequency. This is the situation that prevails in pulsed dye and excimer lasers. Additional frequency-selective components must be included in such lasers, such as etalons and/or diffraction gratings, to obtain narrow output linewidths.

Laser radiation is highly collimated: beam divergence angles as small as 1 mrad are quite common. This makes a laser the ideal way to align an optical system, far superior to traditional methods using point sources and autocollimators. As mentioned previously, lasers are highly coherent. Single-mode lasers have the highest temporal coherence – with coherence lengths that can extend to hundreds of kilometers. Fundamental-transverse-mode lasers have the highest spatial coherence.

In multimode lasers, the many output frequencies lie underneath the gain profile, as shown in Figure 4.121. Laser oscillation can be restricted to a single transverse mode by appropriate choice of mirror radii and spacing. Then, the available output frequencies are uniformly spaced longitudinal modes a frequency  $c_0/2L$  apart, where  $L$  is the optical spacing of the laser mirrors. Laser oscillation can be restricted to a single one of these longitudinal modes by placing an etalon inside the laser cavity. The etalon should be of such a length  $l$  that its free spectral range  $c/2l$  is greater than the width of the gain profile,  $\Delta\nu$ . Its finesse should be high enough so that  $c/2lF \leq c_0/2L$ . The etalon acts as a filter that allows only one longitudinal mode to pass without incurring high loss.

Many lasers can be operated in a *mode-locked* fashion. Then the longitudinal modes of the laser cavity become locked together in phase and the output of the laser becomes a train of uniformly spaced, very narrow pulses. The spacing between these pulses corresponds to the round-trip time in the laser cavity,  $c_0/2L$ . The temporal width of the pulses is inversely proportional to the gain

bandwidth; thus, the broader the gain profile, the narrower the output pulses in mode-locked operation. For example, an argon-ion laser with a gain profile 5 GHz wide can yield mode-locked pulses 200  $\mu\text{s}$  wide. A mode-locked dye laser with a linewidth of 100 cm ( $3 \times 10^{12}$  Hz) can, in principle, give mode-locked pulses as short as about 0.3 ps.

Even shorter pulses than this can be obtained with ring-dye lasers, with mode-locked Ti:sapphire lasers, or with colliding pulse mode-locking techniques,<sup>15</sup> which, with pulse compression techniques, can generate pulses a few tens of femtoseconds ( $10^{-15}$  s) long. Ultrafast laser systems are available from Big Sky Laxer, Clark-MXR, Coherent, Quantronix, and Spectra Physics (now Newport).

Pulse compressors are available from Calmar Optcom, Femtochrome Research, and Swamp Optics.

#### 4.6.5 Coupling Light from a Source to an Aperture

A common problem in optical experiment design involves either the delivery of light from a source to a target, or collection of light from a source and delivery of the light to a collection aperture. The collection aperture might be the active area of a photodetector, the entrance slit of a spectrometer, or the facet of an optical fiber. A problem of specific importance is the focusing of a laser beam to a small spot. Sometimes the solution to one of these problems will involve an imaging system, but it might also involve a nonimaging light collector.

- (1) *Collection of light from a point source and its delivery to a target aperture.* This situation is illustrated in Figure 4.142. The solid angle  $\Delta\Omega$  subtended by a circular aperture  $S = \pi a^2$  at distance  $R$  from the source is:

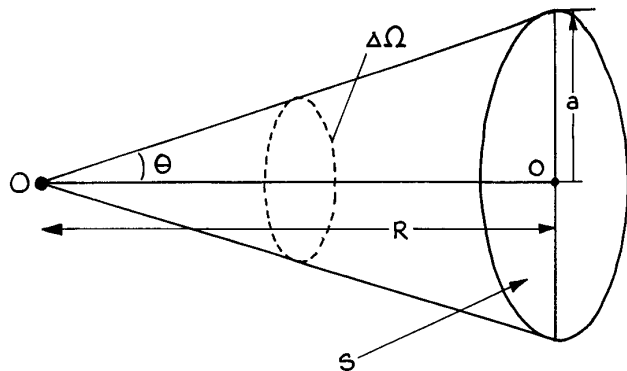
$$\Delta\Omega = 2\pi(1 - \cos\theta) \quad (4.220)$$

which for  $R \gg a$  can be written as:

$$\Delta\Omega = \frac{S}{R^2} \quad (4.221)$$

If the radiant power of the source is  $P$  watts then the power, collected by the aperture is:

$$P' = P \frac{\Delta\Omega}{4\pi} \quad (4.222)$$



**Figure 4.142** Geometry used to determine the collection efficiency of an aperture when it is illuminated by light from a point source.

To increase the light delivered to the aperture  $S$ , a lens system can be placed between  $O$  and  $O'$  as shown in Figure 4.143. To maximize light collection, the lens used should have a small  $f$ /number ( $f/\#$ ). The point source is placed close to the focal point of the lens, so that:

$$\cos \theta \simeq f / \sqrt{a^2 + f^2} \quad (4.223)$$

and

$$P' = \frac{P}{2} \left( 1 - \frac{f}{\sqrt{a^2 + f^2}} \right) \quad (4.224)$$

Since the  $f$ /number is:

$$f/\# = \frac{f}{2a} \quad (4.225)$$

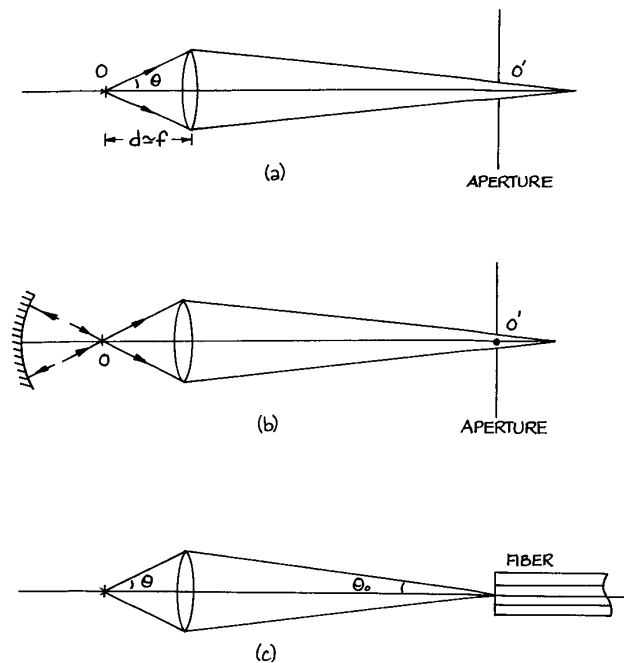
Equation (124) becomes:

$$P' = P \left( \frac{1}{2} - \frac{f/\#}{\sqrt{1 + 4f/\#}} \right) \quad (4.226)$$

With an  $f/1$  lens

$$P' = P \left( \frac{1}{2} - \frac{1}{\sqrt{5}} \right) = 0.053P \quad (4.227)$$

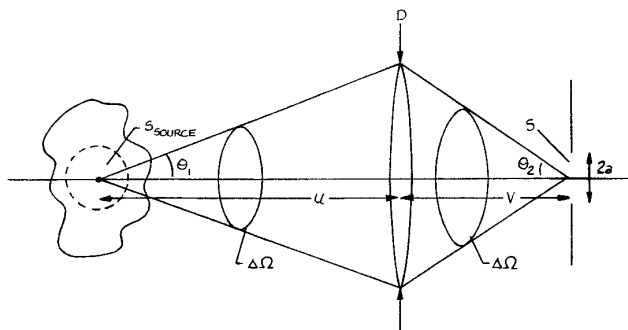
In practice an aspheric lens should be used in this application, especially if the aperture  $S$  is of small size.



**Figure 4.143** Optical arrangements for optimizing coupling of a point source to an aperture: (a) lens coupling; (b) lens and spherical mirror coupling; (c) coupling of a point source to an optical fiber.

If a spherical reflector is placed behind the source, as shown in Figure 4.143(b), so that the source lies at its center of curvature, then the collection efficiency can be doubled.

- (2) *Coupling light from a point source to an optical fiber.* Whether the source is actually a point source or is of small finite size relative to other distances in the geometry of Figure 4.143(c) is not very important. For coupling light to a fiber, the source must be imaged onto the front end of the cleaved fiber, but the light rays must remain within the numerical aperture of the fiber:  $\sin \theta_0 \leq \text{NA}$ . In principle, this can be accomplished by placing the source close to the focal point of the lens so that linear magnification ( $v/u$ ) is large. In practice, the maximum magnification that can be used will be limited by the core diameter of the fiber and the finite size of the source.
- (3) *Collection of light from an extended source and its delivery to a aperture.* This situation is shown in



**Figure 4.144** Geometry of an extended source and collection aperture used to describe the collection efficiency.

Figure 4.144. It is not possible in this case to image the extended source to a point, the best that can be done is to image as large an area of the source as possible onto the aperture  $S$ . If an axial point on the source is imaged to the center of the aperture  $S$ , then the area of the source that can be imaged onto the aperture  $S$  is

$$S_{\text{source}} = \frac{S}{m^2} \quad (4.228)$$

since  $m = (v/u)$  is the linear magnification of the system. A high-quality imaging system would need to be used in this situation, especially if the aperture  $S$  is small, otherwise, aberrations will affect the size of the image.

If the brightness of the extended source in Figure 4.144 is  $B_1(\text{Wm}^{-2} \text{str})$ , then the power collected by the lens is, approximately:

$$P_1 = 2\pi B_1 S_{\text{source}} (1 - \cos \theta_1) \quad (4.229)$$

This light is imaged onto the collection aperture of area  $S$  where it occupies a solid angle  $\Delta\Omega_2$  determined by the angle  $\theta_2$ , where:

$$\Delta\Omega_2 = 2\pi(1 - \cos \theta_2) \quad (4.230)$$

The brightness of the image is:

$$\begin{aligned} B_2 &= \frac{P_1}{2\pi S(1 - \cos \theta_2)} \\ &= \frac{B_1 S_{\text{source}} (1 - \cos \theta_1)}{S (1 - \cos \theta_2)} \end{aligned} \quad (4.231)$$

which gives:

$$B_2 = \frac{B_1 \sin^2(\theta_1/2)}{m^2 \sin^2(\theta_2/2)} \quad (4.232)$$

For a paraxial system  $\sin(\theta_1/2) \simeq (\theta_1/2)$ , so:

$$B_2 = \frac{B_1 (\theta_1^2)}{m^2 (\theta_2^2)} \quad (4.233)$$

The ratio  $\left(\frac{\theta_2}{\theta_1}\right)$  is the angular magnification,  $m'$ , of the system. Therefore, since  $mm' = 1$ :

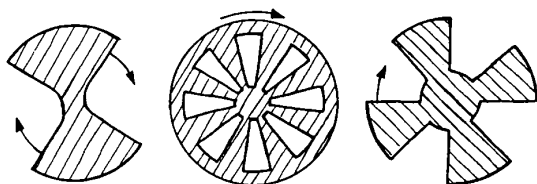
$$B_2 = B_1 \quad (4.234)$$

This is an example of the *brightness theorem*, which states that the brightness of an image can not be greater than the brightness of the object.<sup>k</sup>

#### 4.6.6 Optical Modulators

In many optical experiments, particularly those involving the detection of weak light signals or weak electrical signals generated by some light-stimulated phenomenon, the signal-to-noise ratio can be considerably improved by the use of phase-sensitive detection. The principles underlying this technique are discussed in Section 6.8.2. To modulate the intensity of a weak light signal falling on a detector, the signal must be periodically interrupted. This is most easily done with a mechanical chopper. Weak electrical signals that result from optical stimulation of some phenomenon can also be modulated in this way by chopping the radiation from the stimulating source.

Modulation of narrow beams of light such as laser beams, or of extended sources that can be focused onto a small aperture with a lens, is easily accomplished with a tuning-fork chopper. Such devices are available from Boston Electronics, Electro-Optical Products, and Scitec Instruments. The region chopped can range up to several millimeters wide and a few centimeters long at frequencies from 5 Hz to 3 kHz. For chopping emission from extended sources, or over large apertures, rotating chopping wheels are very convenient. The chopping wheel can be made of any suitable rigid, opaque material and should have radially cut apertures of one of the forms indicated in Figure 4.145. This form of aperture ensures that the mark-to-space



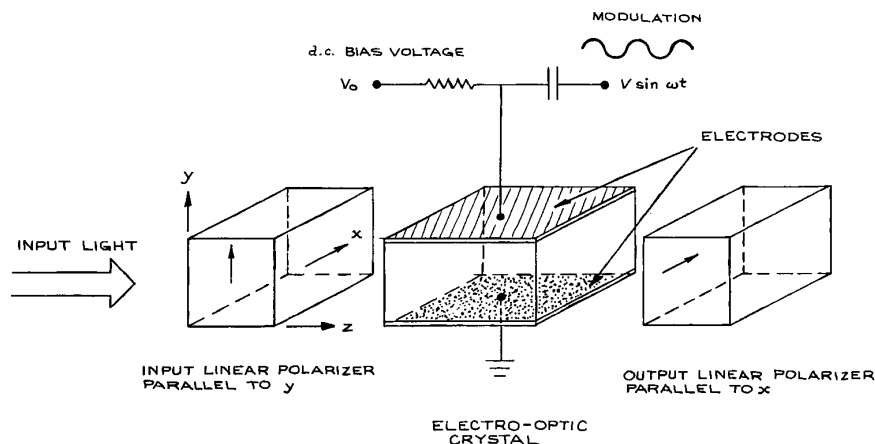
**Figure 4.145** Some chopping-wheel designs.

ratio of the modulated intensity is independent of where the light passes through the wheels. By cutting very many slots in the wheel, very high modulation rates can be achieved – for example, up to 100 kHz with a 20 000 rpm motor and a wheel with 300 slots. The chopping wheel should usually be painted or anodized matt black. For chopping intense laser beams (in excess of perhaps 1 W), it may be better to make the wheel reflective so that unwanted beam energy can be reflected into a beam dump. Reflective glass chopping wheels can be used at low speeds in experiments where a light beam must be periodically routed along two different paths.

To obtain an electrical reference signal that is synchronous with a mechanical chopping wheel, a small portion of the transmitted beam, if the latter is intense enough, can be reflected onto a photodiode or phototransistor. Alternately,

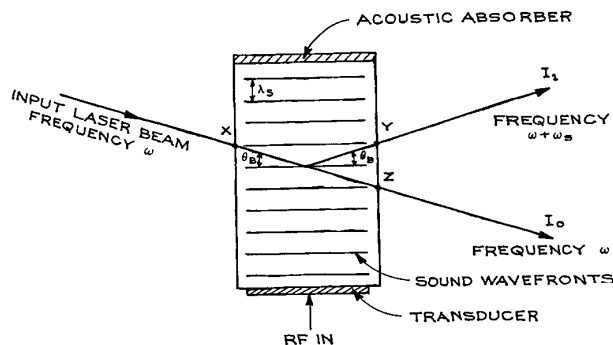
an auxiliary tungsten filament lamp and photodiode can be mounted on opposite sides of the wheel. If these are mounted on a rotatable arm, the phase of the reference signal they provide can be adjusted mechanically. Chopping wheels with built-in speed control and electrical rotation reference signals are available commercially from Boston Electronics, DL Instrument, EG&G Princeton Applied Research, Electro-Optical Products Corp., New Focus, Stanford Research Systems, and Thorlabs.

*Electro-optic modulators (Pockel's cells)* can also be used in special circumstances, particularly for modulating laser beams, or when mechanical vibration is undesirable. The operation of these devices is shown schematically in Figure 4.146. Light passing through the electro-optic crystal is plane-polarized before entry and then has its state of polarization altered by an amount that depends on the voltage applied to the crystal. The effect is to modulate the intensity of the light transmitted through a second linear polarizer placed behind the electro-optic crystal. By choosing the correct orientation of the input polarizer relative to the axes of the electro-optic crystal and omitting the output polarizer, the device becomes an optical phase modulator. For further details of these electro-optic amplitude and phase modulators the reader is referred to the books by Yariv,<sup>14</sup> Kaminow,<sup>125</sup> and Davis.<sup>15</sup> The most



**Figure 4.146** An electro-optic amplitude modulator using a transversely operated electro-optic crystal. The d.c. bias voltage can be used to adjust the operating point of the modulator so that, in the absence of modulation, the output intensity is a maximum, a minimum, or some intermediate value.





**Figure 4.147** Schematic diagram showing the operation of an acousto-optic device in the Bragg regime. For simplicity, the refraction of the laser beam at X, Y, and Z is not shown.

widely used materials for these devices are  $\text{LiNbO}_3$ ,  $\text{TeO}_2$ , and fused quartz. In their operation a radio-frequency (r.f.) sound wave is driven through the material from a piezoelectric transducer, usually fabricated from  $\text{LiNbO}_3$  or  $\text{ZnO}$ . A schematic diagram of how such an acousto-optic modulator works is given in Figure 4.147. Incident light that makes an appropriate angle  $\theta_B$ , with the sound wavefronts will be diffracted if it simultaneously satisfies the condition for constructive interference and reflection from the sound wavefronts. This condition is  $\sin \theta_B = \lambda / \lambda_s$ , where  $\lambda$  is the laser wavelength in the acousto-optic material, and  $\lambda_s$  is the sound wavelength. The device is used as an amplitude modulator by amplitude-modulating the input r.f., which will be set to a specific optimum frequency for the device being used. Increase of drive power increases the diffracted power  $I_1$  and reduces the power of the undeviated beam  $I_0$  and vice versa. The device also functions as a frequency shifter. In Figure 4.147 the laser beam reflects off a moving sound wave and is Doppler shifted so that the beam  $I_1$  is at frequency  $\omega + \omega_s$ , where  $\omega_s$  is the frequency of the sound wave. Acousto-optic devices are available from many suppliers, including Crystal Technology, Electro-Optical Products, IntraAction, Isomet, and NEOS Technologies. Further details about these devices can be found in References 15, 126, and 127.

Very many liquids become optically active on application of an electric field; that is, they rotate the plane of polarization of a linearly polarized beam passing through them. This phenomena is the basis of the *Kerr cell*, which can be used as a modulator, but is more commonly used as

an optical shutter (for example, in laser *Q*-switching applications as discussed in Section 4.6.3). A typical Kerr cell uses nitrobenzene placed between two plane electrodes across which a high voltage is applied. This voltage is typically several kilovolts and is sufficient to rotate the plane of polarization of the incident light passing between the plates by 45 or 90°. Kerr cells are rarely used these days and have been superseded by electro-optic modulators.

Closely related to optical modulators are optical deflectors, which can be used for the spatial scanning or switching of light beams. Two principal types are commonly used: electromechanical devices that utilize small mirrors mounted on galvanometer suspensions (available from Cambridge Technology, GSI Lumonics, and NEOS Technologies) and beam deflectors that utilize the acousto-optic effect. In the latter the deflection angle shown in Figure 4.147 is controlled by changing the r.f. drive frequency. These devices are available from Crystal Technology, IntraAction, Isomet, and NEOS Technologies. Electro-optic scanners are also available, but are less widely used. These devices are generally designed for small deflections. They are available from Conoptics.

Spatial light modulators are multiple element devices, which generally use liquid crystals, work in transmission, and can modify the characteristics of an entire wavefront in a pixellated fashion. They are available from Display Tech and Meadowlark Optics.

### 4.6.7 How to Work Safely with Light Sources

Light sources, whether coherent or incoherent, can present several potential safety hazards in the laboratory. The primary hazard associated with the use of home-built optical sources is usually their power supply. Lasers and flash-lamps in particular, generally operate with potentially lethal high voltages, and the usual safety considerations for constructing and operating such power supplies should be followed in their design:

- (1) Provide a good ground connection to the power supply and light-source housing.
- (2) Screen all areas where high voltages are present.
- (3) Install a clearly visible indicator that shows when the power supply is activated.

- (4) HV power supplies, particularly those that operate with pulsed power, can remain dangerous even after the power is turned off, unless energy-storage capacitors are automatically shunted to ground. Always short the capacitors in such a unit to ground after the power supply is turned off before working on the unit.
- (5) As a rule of thumb, keep a gap of about 25 mm for every 10 kV between high-voltage points and ground.
- (6) To avoid excessive corona at voltages above about 20 kV, make sure that high-voltage components and connections have no sharp edges. Where such points must be exposed, corona can be reduced by installing a spherical metal corona cap on exposed items such as bolts or capacitor terminals. Resistor and diode stacks can be potted in epoxy or silicone rubber to prevent corona.
- (7) Sources that are powered inductively or capacitively with r.f. or microwave power can burn fingers that come too close to the power source, even without making contact.

Commercial optical sources are generally fairly safe electrically and are likely to be equipped with safety features, such as interlocks, which the experimentalist may not bother to incorporate into home-built equipment. *Caveat emptor*. Remember the maxim “it’s the volts that jolts but it’s the mills that kills.”

Other general precautions with regard to electric shock include the following:

- (1) Avoid wearing metallic objects such as watches, watchbands, and rings.
- (2) If any operations must be performed on a line circuit, wear well-insulated shoes and, if possible, use only one hand – “Always keep one hand in your pocket.”
- (3) Keep hands dry; do not handle electrical equipment if you are sweating.
- (4) Learn rescue and resuscitation procedures for victims of electric shock: Turn off the equipment, remove the victim by using insulated material; if the victim is not breathing, start mouth-to-mouth resuscitation; if there is no pulse, begin CPR procedures immediately; summon medical assistance; continue resuscitation procedures until relieved by a physician. If the victim is conscious, but continues to show symptoms of shock, keep the person warm. Always call for profession help. (Dial 911 in the USA.)

Other hazards in the use of light sources include the possibility of eye damage, direct burning (particularly, by exposure to the beam from a high-average-power laser), and the production of toxic fumes. The last is most important in the use of high-power CW arc lamps, which can generate substantial amounts of ozone. The lamp housing should be suitably ventilated and the ozone discharged into a fume hood or into the open air. Some lasers operate using a supply of toxic gas, such as hydrogen fluoride and the halogens. Workers operating such systems must be sufficiently experienced to work safely with these materials. The *Matheson Gas Data Book*, published by Matheson Gas Products, is a comprehensive guide to laboratory gases, detailing the potential hazards and handling methods appropriate to each.

The potential eye hazard presented by most incoherent sources is not great. If a source appears very bright, one should not look at it, just as one should not look directly at the sun. Do not look at sources that emit substantial ultraviolet radiation; at the least, severe eye irritation will result – imagine having your eyes full of sand particles for several days! Long-term exposure to UV radiation should be kept below  $0.5 \mu\text{W}/\text{cm}^2$ . Ordinary eyeglasses will protect the eyes from ultraviolet exposure to some extent, but plastic goggles that wrap around the sides are better. Colored plastic or glass provides better protection than clear material. The manufacturer’s specification of ultraviolet transmission should be checked, since radiation below 320 nm must be excluded from the cornea.

The use of lasers in the laboratory presents an optical hazard of a different order. Because a laser beam is generally highly collimated and at least partially coherent, if the beam from a visible or near-infrared laser enters the pupil of the eye it will be focused to a very small spot on the retina (unless the observer is very short-sighted). If this focused spot happens to be on the optic nerve, total blindness may result: if it falls elsewhere on the retina, an extra blind spot may be generated. Although the constant motion of the human eye tends to prevent the focused spot from remaining at a particular point on the retina for very long, always obey the following universal rule: *Never look directly down any laser beam either directly or by specular reflection*. In practice, laser beams below 1 mW CW are probably not an eye hazard, but they should still be treated with respect. A beam must be below about  $10 \mu\text{W}$  before most workers would regard it as really safe.

The safe exposure level for CW laser radiation depends on the exposure time. For pulsed lasers, however, it is the maximum energy that can enter the eye without causing damage that is the important parameter. In the spectral region between 380 and 1.5  $\mu\text{m}$ , where the interior material of the eye is transparent, the maximum safe dose is on the order of  $10^{-7}$   $\text{J}/\text{cm}^2$  for  $Q$ -switched lasers and about  $10^{-6}$   $\text{J}/\text{cm}^2$  for non- $Q$ -switched lasers. If any potential for eye exposure to a laser beam or its direct or diffuse reflection exists, it is advisable to carry out the experiment in a lighted laboratory: in a darkened laboratory, the fully dark-adapted human eye with a pupil area of about 0.5  $\text{cm}^2$  presents a much larger target for accidental exposure. The Laser Institute of America publishes a guide that shows the maximum permitted exposure in terms of power and direction for a range of laser wavelengths.<sup>129</sup>

Infrared laser beams beyond about 1.5  $\mu\text{m}$  are not a retinal hazard, as they will not penetrate to the retina. Beams between 1.5 and 3  $\mu\text{m}$  penetrate the interior of the eye to some degree. Because its energy is absorbed in a distributed fashion in the interior of the eye, 1.55  $\mu\text{m}$  is one of the safest laser wavelengths. Lasers beyond 3  $\mu\text{m}$  are a burn hazard, and eye exposure must be avoided for this reason.  $\text{CO}_2$  lasers, for example, which operate in the 10  $\mu\text{m}$  region, are less hazardous than equivalent-intensity argon-ion or neodymium lasers. Neodymium lasers represent a particularly severe hazard: they are widely used, emit substantial powers and energies, and operate at the invisible wavelength of 1.06  $\mu\text{m}$ . This wavelength easily penetrates to the retina, and the careless worker can suffer severe eye damage without any warning. The use of safety goggles is strongly recommended when using such lasers. These goggles absorb or reflect 1.06  $\mu\text{m}$ , but allow normal transmission in at least part of the visible spectrum. Laser safety goggles are available commercially from several sources, such as Bollé, Control Optics, Glendale, Kentek, Lase-R Shield, Melles Griot, Newport, NOIR, Thorlabs, and UVEX. The purchaser must usually specify the laser wavelength(s) for which the goggles are to be used.

Unfortunately, one problem with goggles prevents their universal use. Because the goggles prevent the laser wavelength from reaching the eyes, they prevent the alignment of a visible laser beam through an experiment by observation of diffuse reflection of the beam from components in the system. When such a procedure must be carried out, we

recommend caution and the operation of the laser at its lowest practical power level during alignment. Frequently, a weak, subsidiary alignment laser can be used to check the potential path of a high-power beam before this is turned on – a strongly recommended procedure. For infrared laser beams, thermal sensor cards are available that will reveal the location of an infrared laser beam through thermally activated fluorescence or fluorescence quenching. They are available from Applied Scintillation Technology, Laser S.O.S., Macken Instruments, Newport, and Solar TII. Some of these sensor cards or plates, especially for  $\text{CO}_2$  laser beam location, require illumination with a UV source. Where an infrared laser beam strikes the illuminated surface a dark spot appears. Used Polaroid film and thermally sensitive duplicating paper are also useful for infrared-beam tracking.

The beam from an ultraviolet laser can generally be tracked with white paper, which fluoresces where the beam strikes it. A potential burn and fire hazard exists for lasers with power levels above about 1  $\text{W}/\text{cm}^2$ , although the fire hazard will depend on the target. Black paper burns most easily. Very intense beams can be safely dumped onto pieces of firebrick. Pulsed lasers will burn at output energies above about 1  $\text{J}/\text{cm}^2$ .

In the United States, commercial lasers are assigned a rating by the FDA Center for Devices and Radiological Health, which identifies their type, power output range, and potential hazard. It must be stressed, however, that lasers are easy to use safely: accidents of any sort have been rare. Their increasing use in laboratories requires that workers be well-informed of the standard safety practices. For further discussion of this and all aspects of laser safety and related topics, we recommend both the book by Sliney and Wolbarsht<sup>130</sup> and a series of articles on laser safety in the *CRC Handbook of Laser Science and Technology*, Vol. 1.<sup>131</sup>

## 4.7 OPTICAL DISPERSING INSTRUMENTS

Optical dispersing instruments allow the spectral analysis of optical radiation or the extraction of radiation in a narrow spectral band from some broader spectral region. In this general category we include interference filters, prism and grating monochromators, spectrographs and

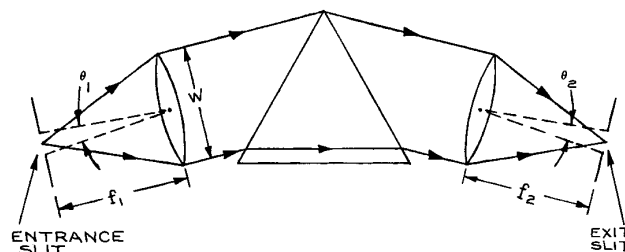
spectrophotometers, and interferometers. Interferometers have additional uses over and above direct spectral analysis – including studies of the phase variation over an optical wavefront, which allow the optical quality of optical components to be measured. Spectrometers, or *monochromators* as they are generally called, are optical filters of tunable center wavelength and bandwidth. The output narrow-band radiation from these devices is generally detected by a photon or thermal detector, which generates an electrical signal output. *Spectrographs*, on the other hand, record the entire spectral content of an extended bandwidth region either photographically, or more commonly these days with a linear detector area. *Spectrophotometers* are complete commercial instruments, which generally incorporate a source or sources, a dispersing system (which may involve interchangeable prisms and/or gratings), and a detector. They are designed for recording ultraviolet, visible, or infrared absorption and emission spectra.

Spectrofluometers are instruments to record fluorescence spectra. In such an application two principal modes of spectral usage exist. The emission fluorescence spectrum is the spectral distribution of emission produced by a particular monochromatic excitation wavelength. The excitation spectrum is the total fluorescence emission recorded as the wavelength of the excitation source is scanned.

Before giving a discussion of important design considerations in the construction of various spectrometers and interferometers, it is worthwhile mentioning two important figures of merit that allow the evaluation and comparison of the performance of different types of optical dispersing instruments. The first is the resolving power, ( $\mathfrak{R} = \lambda/\Delta\lambda$ ), which has already been discussed in connection with diffraction gratings (Section 4.3.5). The second is the luminosity, which is the flux collected by a detector at the output of a spectrometer when the source at the input has a radiance of unity. The interrelation between resolving power and luminosity for spectrometers using prisms, gratings, and Fabry–Perot etalons has been dealt with in detail by Jacquinot.<sup>131</sup>

Prism spectrometers have almost disappeared from use in recent years, however, because of their simple mode of operation they serve as an archetype in discussing spectrometer performance.

Figure 4.148, which shows the essential components of a prism monochromator, will serve to illustrate the points



**Figure 4.148** Basic elements of a prism monochromator.

made here. High resolution is clearly obtained in this arrangement by using narrow entrance and exit slits. The maximum light throughput of the spectrometer results when the respective angular widths  $W_1$ ,  $W_2$  of the input and output slits  $\theta_1$ ,  $\theta_2$  satisfy:

$$W_1 = W_2 \quad (4.235)$$

where

$$W_1 = \frac{\theta_1}{(d\alpha/d\lambda)_\delta}, \quad W_2 = \frac{\theta_2}{(d\delta/d\lambda)_\alpha} \quad (4.236)$$

The angles  $\alpha$ , are the same as those used in Figure 4.47. The input and output dispersions  $(d\alpha/d\lambda)_\delta$  and  $(d\delta/d\lambda)_\alpha$  are only equal in a position of minimum deviation, or with a prism (or grating) used in a Littrow arrangement. If diffraction at the slits is negligible, the intensity distribution at the output slit when a monochromatic input is used is a triangular function, as shown in Figure 4.149, where  $W$  is the spectral width of the slits given by:

$$W = W_1 = W_2 \quad (4.237)$$

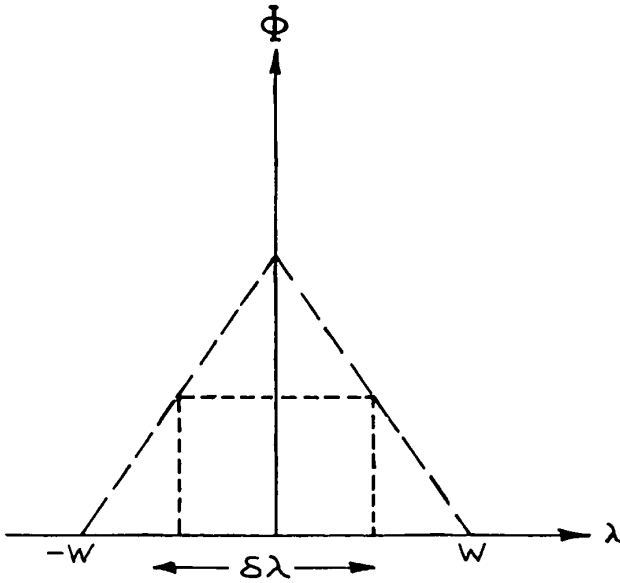
The limit of resolution of the monochromator is:

$$\delta\lambda = W = \frac{\theta_2}{(d\delta/d\lambda)_\alpha} \quad (4.238)$$

The maximum flux passing through the output slit is:

$$\Phi = TE_e(\lambda)Sl\theta_2/f_2 \quad (4.239)$$

where  $S$  is the normal area of the output beam,  $T$  is the transmittance of the prism (or efficiency of the grating in the order used) at the wavelength being considered,  $l$  is the height of the entrance and exit slits (equal), and  $E_e(\lambda)$  is the



**Figure 4.149** Intensity distribution at the output slit of a monochromator whose entrance and exit slits have the same spectral width  $W$ .

radiance of the monochromatic source of wavelength  $\lambda$  illuminating the entrance slit.  $l\theta_2/f_2$  is the solid angle that the exit slit subtends at the output focusing lens. Equation (4.239) can be written as:

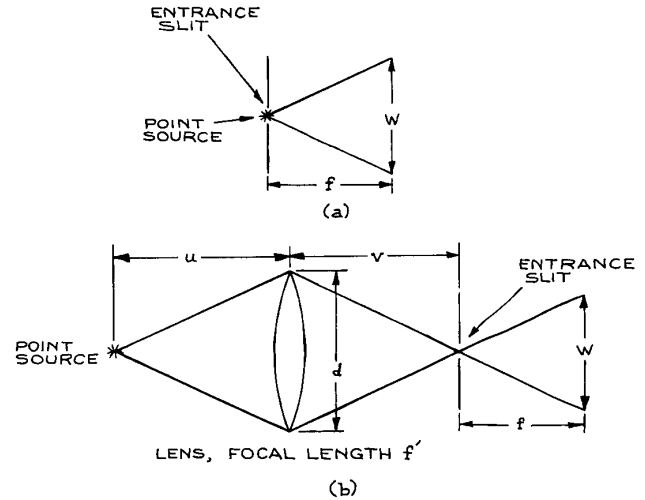
$$\Phi = \frac{TE_c(\lambda)Sl\lambda(d\delta/d\lambda)_x}{f_2\mathcal{R}} \quad (4.240)$$

which clearly shows that the output flux is inversely proportional to the resolving power [for a continuous source at the entrance slit, Equation (4.240) is multiplied by an additional factor  $\lambda/\mathcal{R}$ ]. The efficiency  $E$  of the monochromator is defined by the equation:

$$E = \frac{TSI}{f} \left( \frac{d\delta}{d\lambda} \right)_x \quad (4.241)$$

where we have assumed the usual case in which  $f_1 = f_2 = f$ .

To collect all the light from the collimating optics in the arrangement of Figure 4.148, the prism or grating should have a normal area at least as large as the aperture of the collimating optics. Thus, the height of the prism (or diffraction grating) should be  $h \geq W$ . The ratio  $f/W$ , which is a



**Figure 4.150** Schematic diagrams illustrating geometrical factors involved in optimizing light-collection efficiency by a spectrometer of aperture  $W$ : (a) point source directly at entrance slit; (b) remote point source imaged on entrance slit with a lens.

measure of the light-gathering power of the monochromator, defines its  $f$ /number. The most efficient way to use a spectrometer of any kind is to send the input radiation into the entrance slit within the cone of angles defined by the  $f$ /number. The maximum resolution is obtained if the input  $f$ /number matches the spectrometer  $f$ /number. If radiation enters the spectrometer with an  $f$ /number that is too small, this radiation overfills the dispersing element and some is wasted. This can be illustrated with the aid of Figure 4.150, which shows the use of a point source directly at the entrance slit. In Figure 4.150(b) a point source is imaged on the entrance slit using a lens that matches the input radiation to the  $f$ /number of the spectrometer. In Figure 4.150(a) the fractional useful light collection from the source is:

$$\phi_1 \simeq \frac{\pi W^2}{4f^2} = \frac{\pi}{4F^2} \quad (4.242)$$

where  $F$  is the  $f$ /number of the spectrometer. In contrast, in Figure 4.150(b) it is approximately:

$$\phi_2 \simeq \frac{\pi d^2}{4u^2} \quad (4.243)$$

which can be written as:

$$\phi_2 = \frac{\pi d^2 (v - f)^2}{4f^2 v^2} \quad (4.244)$$

and substituting  $v/d = F$ ,  $f/d = F'$  (the  $f$ number of the lens), we have:

$$\phi_2 = \frac{\pi(F/F' - 1)^2}{4F^2} \quad (F' \leq F) \quad (4.245)$$

Thus, equally efficient light gathering to that which would be obtained with the point source at the entrance slit results if:

$$F' = F/2 \quad (4.246)$$

More efficient light gathering results if  $F'$  is less than this value, but, if the lens and point source are incorrectly positioned, so that  $v/d < F$ , then not all the light collected by the lens reaches the dispersing element of the spectrometer.

It is quite common for spectrometers to be used with a fiber optic light-delivery system. If the NA of the fiber is too large and the fiber optic is placed at the entrance of the spectrometer, then overfilling of the dispersing element will occur. In this case correction optics must be included to match the fiber to the  $f$ number of the spectrometer.

### 4.7.1 Comparison of Prism and Grating Spectrometers

Following Jacquinot,<sup>131</sup> we use Equation (4.240) to compare prism and grating instruments. Assume identical values of  $l$  and  $f_2$ , since, for a given degree of aberration of the system, their values are identical for prisms and gratings. Blazed-grating efficiencies can be high, so  $T$  is assumed to be similar for prism and grating. Thus, the comparison of luminosity under given operating conditions of wavelength and resolution depends solely on the quantity  $S(d\delta/d\lambda)_x$  for a prism and  $S(d\beta/d\lambda)_x$  for a grating.

For the prism:

$$\frac{d\delta}{d\lambda} = \frac{t}{W} \frac{dn}{d\lambda} \quad (4.247)$$

where, for maximum dispersion, the whole prism face is illuminated, so that  $t$  is the width of the prism base.  $S = hW$

and  $th = A$ , the area of the prism base, and so:

$$S \frac{d\delta}{d\lambda} = A \frac{dn}{d\lambda} \quad (4.248)$$

For the grating:

$$\frac{d\beta}{d\lambda} = \frac{m}{d \cos \beta} \quad (4.249)$$

$S = A \cos \beta$ , where  $A$  is the area of the grating. If it is assumed that the grating is used in Littrow, then  $m\lambda = 2d \sin \beta$ , and therefore:

$$S \frac{d\beta}{d\lambda} = \frac{2A \sin \beta}{\lambda} \quad (4.250)$$

The ratio of luminosities for a prism and grating of comparable size is therefore:

$$\rho = \frac{\Phi(\text{prism})}{\Phi(\text{grating})} = \frac{\lambda dn/d\lambda}{2 \sin \beta} \quad (4.251)$$

This ratio can be improved by a factor of two if the prism is also used in Littrow. For a typical value of  $30^\circ$  for  $\beta$ , Equation (4.251) predicts that a grating is always superior to a prism instrument. Except for a few materials in restricted regions of wavelength where  $dn/d\lambda$  becomes large and  $\rho$  may reach 0.2–0.3, a grating is more luminous than a prism by a factor of 10 or more. The only potential advantage of a prism instrument over a grating is the absence of overlapping orders. A grating instrument illuminated simultaneously with 300 and 600 nm, for example, will transmit both at the same angular position; a prism instrument will not. This minor problem is easily solved with an appropriate order-sorting color or interference filter.

Jacquinot<sup>131</sup> has demonstrated that a Fabry–Perot etalon or interferometer (see Section 4.7.4) used at a high or moderate resolving power is superior in luminosity to a grating instrument by a factor that can range from 30 to 400 or more. In general, an etalon cannot be used alone, because of its many potential overlapping orders. It is usual to couple it with a grating or prism monochromator in high-resolution spectroscopic applications, except in those cases where the source is already highly monochromatic.

Table 4.9 gives a comparison of the performance characteristics of prism and grating monochromators and Fabry–Perot interferometers. From the vacuum ultraviolet (prism instruments cannot be used below about 1200 Å) to

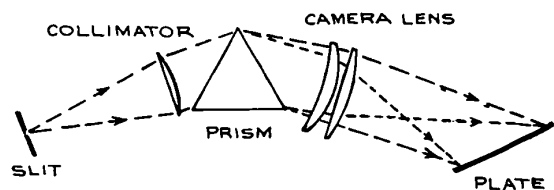
**Table 4.9 Main characteristics of optical dispersing elements**

<i>Element</i>	<i>Advantages</i>	<i>Disadvantages</i>
Multilayer dielectric interference filter	High throughput at center frequency, $\approx 50\%$ . Tunable to some extent by tilting.	Low resolution – minimum bandwidth $\approx 1$ nm. None available at short wave lengths $\leq 200$ nm.
Prism spectrometer	Dispersion and resolution increase near absorption edge of prism (however, transmission falls). No ghosts. Scattered light can be lower than in grating spectrometer. No order sorting necessary.	Cannot be used below 120 nm (with LiF prism). Resolution not as good as grating spectrometer, best $\approx 0.01$ nm nm. Resolution particularly poor in infrared.
Grating spectrometer	Resolution can be high ( $\leq 0.001$ nm). Can be used from vacuum UV to far IR. Dispersion independent of wavelength. Luminosity depends on blaze angle, but generally much higher than for prism spectrometer.	Ghosts can be a problem. Expensive if really low scattered light required – may need double- or triple-grating instrument. Need to remove unwanted orders (not necessarily a serious problem).
Fabry–Perot interferometer	Very high light throughput – much higher than for grating spectrometer. Very high resolution – tens of MHz at optical frequencies ( $\geq 10^{-5}$ nm).	Many orders generally transmitted simultaneously. Cannot be used at short wavelengths ( $\leq 200$ nm). Transmission maximum can be scanned only over narrow range. Most difficult dispersing element to use.

the far infrared, grating instruments are much superior to comparable-size prism instruments in both resolving power and luminosity. In recent years, the advantages of grating instruments over prism instruments has virtually eliminated the latter. This has primarily resulted because of improvements in grating fabrication. Holographic plane gratings have fewer and weaker ghosts than metal gratings, and holographic curved gratings have allowed aberration reduction in spectrometers. Apart from the minor inconvenience of potential overlapping orders, grating instruments are almost always to be preferred over prism instruments. There is one exception worth noting: when a monochromator is being used to observe some weak optical emission in the simultaneous presence of a strong laser beam, unless the optical emission is too close in wave-

length to the laser for prism resolution to be adequate. In this application, a prism avoids the problems of ghost emission or grating scattering. A good prism has very low scattered light. It is common to pre-disperse a light beam with a prism before sending the beam into the entrance slit of a monochromator, in order to avoid both the overlapping-order problem and spurious signals from a very strong light signal at another wavelength. Fabry–Perot interferometers have very high resolution and luminosity, but require careful and regular adjustment to maintain high performance. They are used only where their very high resolution is essential, frequently in conjunction with a grating monochromator for pre-isolation of a narrow spectral region.

A final point to note with regard to the use of grating spectrometers: these instruments are polarization sensitive.



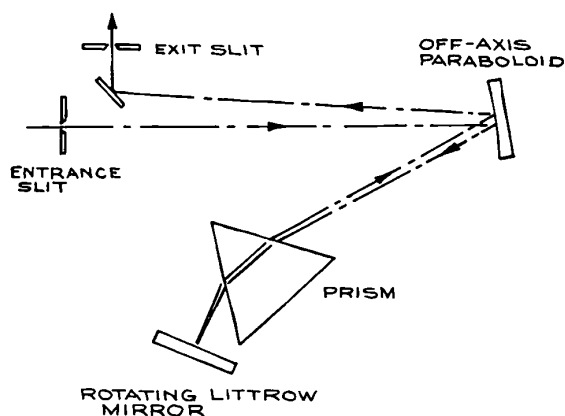
**Figure 4.151** Schematic diagram of a simple prism spectrograph. (From R.J. Meltzer, "Spectrographs and Monochromators," in *Applied Optics and Optical Engineering*, Vol. 5, R. Kingslake (Ed.), Academic Press, New York, 1969; by permission of Academic Press.)

If unpolarized light strikes a grating, the diffracted light will be partially polarized because the grating efficiency is wavelength-, and  $S$  and  $P$  polarization-dependent.

#### 4.7.2 Design of Spectrometers and Spectrographs

The construction of monochromators and spectrographs is a specialized undertaking. The availability of good commercial instruments generally makes it unnecessary for anyone but the enthusiast to contemplate their construction in the laboratory. On the other hand, the principles that underlie good design in a dispersing instrument are not complex. Some essential features of such instruments can be illustrated with reference to particular designs that are used both in laboratory and commercial instruments. Attention will be restricted to monochromators; spectrographs are similar, apart from having an array detector or photographic plate instead of an exit slit, as shown in Figure 4.151. Spectrographs are also simpler in construction because wavelength scanning is not required and all optical components of the system remain fixed. The discussion here should also allow intelligent evaluation and comparison of commercial instruments. For further details of the many different instrument designs that have been used for the construction of monochromators, spectrographs and spectrophotometers, References<sup>133,134</sup>, *The Photonics Design and Applications Handbook*<sup>135</sup>, and the catalogs of instrument manufacturers should be consulted.

Figure 4.152 shows a Littrow prism monochromator. The light from the input slit is collimated with an off-axis

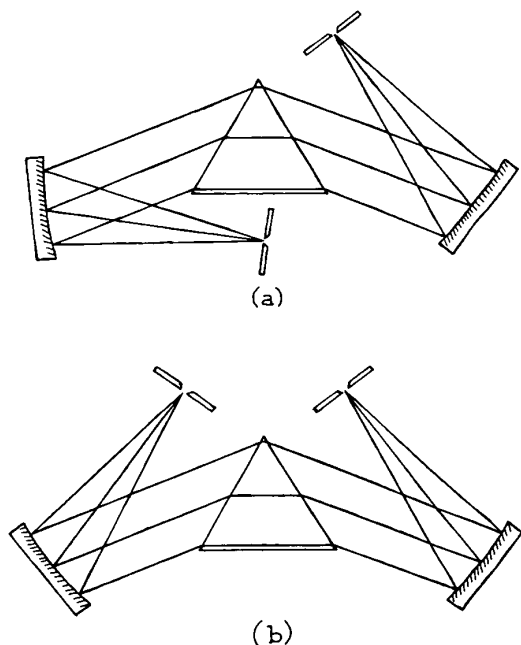


**Figure 4.152** Littrow monochromator. (From R.J. Meltzer, "Spectrographs and Monochromators," in *Applied Optics and Optical Engineering*", Vol. 5, R. Kingslake (Ed.), Academic Press, New York, 1969; by permission of Academic Press.)

paraboloidal mirror. The output wavelength is changed by rotating the Littrow mirror. If the output wavelength must change uniformly with rotation of a drive shaft, then the rotation of the mirror requires a cam. Figure 4.153 shows a Czerny–Turner prism configuration. The collimating mirrors are spherical. In the geometry of Figure 4.153(a), the coma of one mirror compensates that of the other, while in Figure 4.153(b) these comas are additive. The wavelength is changed by rotating the prism.

In any of these spectroscopic instruments, the optical arrangement should be enclosed in a light-tight enclosure painted on its interior with black nonreflective paint. Table 4.10 lists the various "black" coatings that are available for this purpose and in any application where broad reflection over a range of angles from a surface must be minimized. This helps to minimize scattered light. It is essential that light passing through the entrance slit is not able to reach the exit slit without going through the illuminating optics and prism. To this end, it is common practice to incorporate internal blackpainted baffles in the monochromator enclosure. The most likely source of scattered light is overillumination of the input collimator: input radiation should not enter with an  $f$ -number smaller than that of the monochromator. To check for stray light, visual observation through the exit slit in a darkened room when the





**Figure 4.153** Czerny-Turner monochromator arrangements; (a) coma of one mirror compensates that of the other; (b) comas are additive.

entrance slit is illuminated with an intense source will reveal where additional internal baffles would be helpful. Stray light is more likely to be a problem in a Littrow arrangement than in a straight-through type, such as the Czerny–Turner.

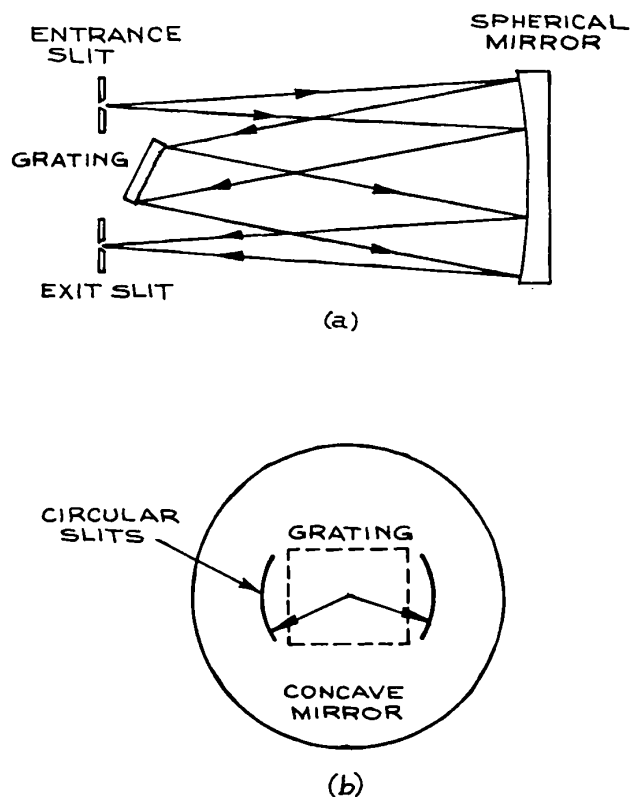
In the visible and near infrared, lens collimating optics can be used, as, for example, in the design shown in Figure 4.151. Good lenses, such as camera lenses, should be used. Mirror collimating optics are useful over a much broader wavelength region. For ultraviolet use they should be aluminum overcoated with  $\text{MgF}_2$ , and they are frequently of this type, even in instruments designed for longer wavelengths.

Slit design is important in achieving high resolution. Entrance and exit slits of equal width are generally used. If the source is small, having a height much less than the height of the prism or grating, parallel-sided slits are adequate. Adjustable versions are available from Melles Griot, Newport, and OptoSigma. Fixed slits can be easily

**Table 4.10** Black materials and coatings

<i>Material</i>	<i>Absorptivity</i>
Anodized black	0.88
Carbon lampblack	0.84
Chemglaze black paint Z306	0.91
Delrin black plastic	0.87
Electro optical industries	0.965
Mid-temperature black coating (up to 200 °C)	
Electro-optical industries	0.93
High-temperature black coating (up to 1400 °C)	
Martin black velvet paint	0.94
3M black velvet paint	0.91
Parsons black paint	0.91
Polyethylene black plastic	0.92
Anodized aluminum	0.84

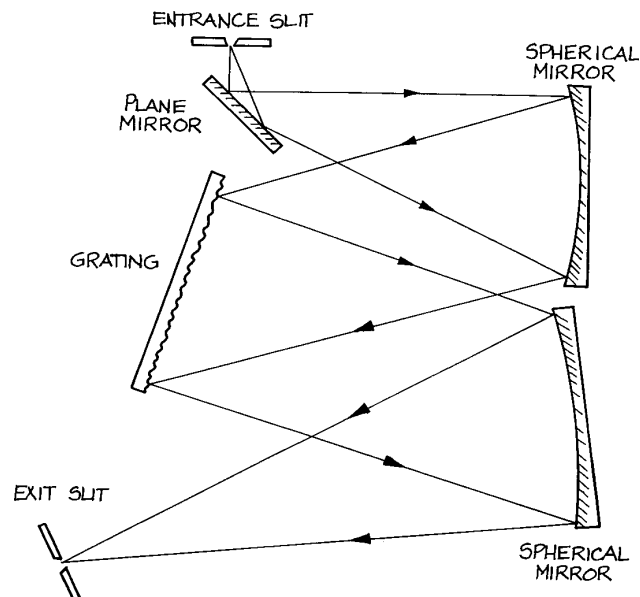
made using razor blades or purchased from Melles Griot or National Aperture. If the source is high compared to the height of the dispersing element, a straight entrance slit produces a curved image at the exit slit. In the case of a prism, this happens because light passing through the prism at an angle inclined to the meridian plane suffers slightly more deviation than light in the meridian plane. Light striking a grating in a plane that is not perpendicular to the grating grooves sees an apparently smaller groove spacing, and is consequently deviated more than light in the plane perpendicular to the grating. To maintain high resolution in the presence of image curvature, curved entrance and exit slits should be used, each with a radius equal to the distance of the slit from the axis of the system. This slit arrangement is shown in Figure 4.154, which also shows a very good, and widely used, grating monochromator design due to Fastie<sup>135</sup> (based on a design originally described by Ebert<sup>136</sup> for use with a prism). It is also quite common for the single spherical mirror in Figure 4.154 to be replaced by a pair, in which case the design becomes the Czerny–Turner one. Such an arrangement is shown in Figure 4.155. In either case the wavelength is tuned by rotating the grating. The wavelength variation at the exit slit is much more nearly linear with grating rotation than with prism rotation. If linearity is required, a cam, sine-drive, or other such arrangement is used for rotating the grating or



**Figure 4.154** (a) Ebert-Fastie monochromator; (b) curved entrance- and exit-slit configuration for use with the above.

prism.<sup>134</sup> In a Czerny-Turner grating instrument the entrance and exit slits are generally arranged asymmetrically with respect to the grating, as this allows coma to be corrected in the working wavelength range. Czerny-Turner monochromators are available from Edmund Optics, Genesis Laboratory Systems (who now manufacture the former Jarrel-Ash line of spectrometers), Horiba Yobin Yvon, McPherson, Minuteman Laboratories, Ocean Optics, Princeton Instruments/Acton, and Thorlabs.

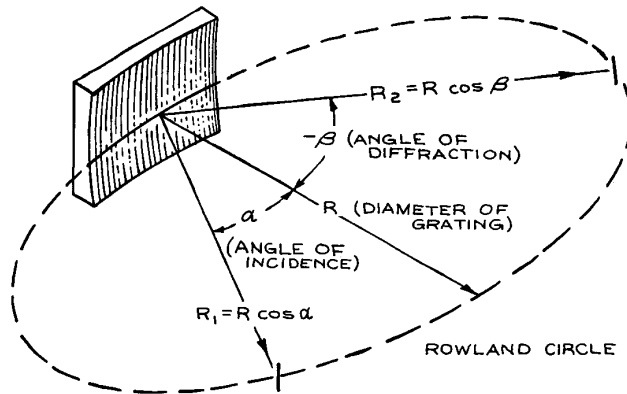
Grating monochromators for use in the vacuum ultraviolet generally have a concave grating. Such a grating, which has equally spaced grooves on the chord of a spherical mirror, combines both dispersion and focusing in one element. For maximum resolution, both the object and image formed by a concave grating must be on a circle called the *Rowland circle*, whose diameter is the principal



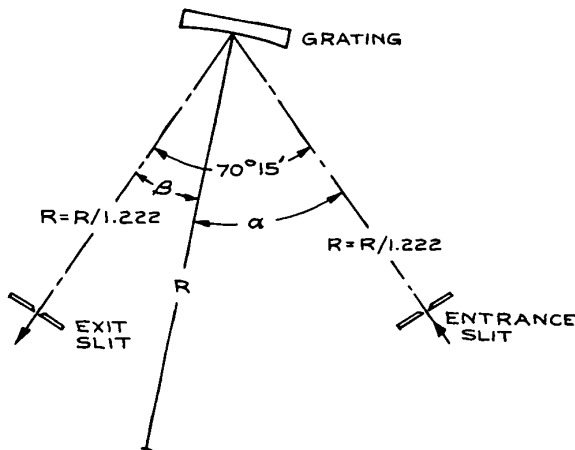
**Figure 4.155** Arrangement of a Czerny-Turner grating monochromator.

radius of the spherical surface on which the grating is ruled, as shown in Figure 4.156. The use of a concave-grating monochromator (or spectrograph) entails mounting grating and slits (or grating, slits, and photographic plate) on a Rowland circle. Adjustment of the instrument will involve moving the grating, slits, or plate on the circle. Various mounting configurations, such as the Abney, Eagle, Paschen, Rowland, and Wadsworth, and grazing-incidence mounting have been used to accomplish this.

A convenient concave-grating design widely used in vacuum-ultraviolet monochromators is the Seya-Namioka,<sup>137,138</sup> shown in Figure 4.157. In this arrangement, by judicious choice of slit distances from the grating and slit angular separation, the only adjustment needed for wavelength adjustment is rotation of the grating. In comparison with an ideal Rowland-circle mounting, only a small resolution loss results over a wide wavelength range. Vacuum-ultraviolet instruments possess the added complication that all interior adjustments of the instrument must be transmitted through the walls of the vacuum chamber that houses the optical arrangement. Various kinds of rotary-shaft and bellows seals and magnetically coupled drives exist for this purpose.<sup>140</sup> See also Section 3.5.4.



**Figure 4.156** Rowland circle of a concave diffraction grating. For maximum resolution, both object and image must be on the Rowland circle.  $R$  = object (entrance-slit) distance:  $R_2$  = image (exit-slit) distance. (From R. J. Meltzer, "Spectrographs and Monochromators," in *Applied Optics and Optical Engineering*, Vol. 5, R. Kingslake (Ed.), Academic Press, New York, 1969; by permission of Academic Press.)



**Figure 4.157** Layout of the Seya-Namioka monochromator.

Multiple monochromators use multiple dispersing elements in series to obtain greater dispersion and/or lower scattered light. For example, in a double-grating monochromator, two diffraction gratings move simultaneously; the desired diffracted wavelength from the first grating is diffracted once again from the second grating before being

focused on the exit slit. The considerable reduction in scattered light that can be obtained with multiple monochromators makes them particularly useful for analyzing weak emissions close in wavelength to a strong emission, as in Raman spectroscopy. Instruments with as many as four gratings in series are available commercially. Grating monochromators, spectrographs, and spectrophotometers are available from many manufacturers,<sup>31</sup> including American Holographic, Coherent Scientific, CVI Spectral Products, Horiba Jobin-Yvon (formerly Spex or Instruments SA), Minuteman Laboratories, Newport/Oriel, Olis, Princeton Instruments/Acton, and Thermo Scientific (which incorporates the old companies Jarrell-Ash, Aminco-Bowman, Unicam, and Varian Associates). Fiber-coupled spectrometers are manufactured by Ocean Optics, Photon Control, and Stellar Net. These instruments either consist of a complete spectrometer on a PC card or are compact units connected to a computer through a USB port. McPherson, Minuteman Laboratories, and Princeton Instruments/Acton are noted for their vacuum instruments, Coherent Scientific, Horiba Jobin-Yvon, and Princeton Instruments/Acton for their multiple grating spectrometers, and Genesis Laboratory Systems for their high-resolution spectrometers and spectrographs.

Imaging Spectrographs [sometimes called optical multi-channel analysers (OMAs)] image an entire spectrum onto either a linear photodiode area (PDA) or onto a linear CCD. These instruments often have the capability for recording complete spectra from several sources at once. Manufacturers include Horiba Jobin-Yvon, Instrument Systems GmbH, Olis, Nomadics, Princeton Instruments/Acton, and Roper Scientific. Note that CCD arrays are approximately 100 times more sensitive than PDA, so are best used for the lowest light-level applications. Change-coupled devices must be used with appropriate shutters or light modulation otherwise image "smear" will occur. This happens if new light falls on the array while the previous deposited charge distribution is being read out.

Wavemeters are optical devices that are designed to provide direct electronic readout of the wavelength of monochromatic radiation that is directed into them. They are most frequently used in conjunction with a tunable laser system. These convenient instruments incorporate a Fabry-Perot or Fizeau interferometer.<sup>148</sup> Manufacturers include Bristol Instruments, Elliott Scientific, Holo-Spectra, MK Photonics, and Specialise Instruments Marketing.

### 4.7.3 Calibration of Spectrometers and Spectrographs

To calibrate the wavelength scale of a spectrometer or spectrograph, the entrance slit is illuminated with a reference source that has characteristic spectral features of accurately known wavelength. It is quite common to mount a small reference lamp – mercury, neon, and hollow-cathode iron lamps are the most popular – near the entrance slit and use a small removable mirror or beamsplitter to superimpose the spectrum of the reference on the unknown spectrum under study. To calibrate a grating instrument through its various orders, a helium–neon laser is ideal. If the slits are set to their narrowest position and a low-power ( $< 1$  mW) He–Ne laser is directed at the entrance slit, the various grating settings that transmit a signal (which can be safely observed by eye at the exit slit) correspond to the wavelengths 0, 632.8 nm,  $632.8 \times 2$  nm,  $632.8 \times 3$  nm, and so on. By interpolation between the observed values, a reliable calibration curve can be obtained even for instruments operating far into the infrared.

The efficiency of the instrument at a given wavelength can be determined with any suitable detector. A laser or other narrow-band source at the desired wavelength illuminates the entrance slit, or a duplicate of it, and the transmitted power is measured with the detector placed directly behind the slit. A similar measurement is then made at the exit slit of the whole instrument. The relative spectral efficiency can be determined by illuminating the entrance slit with a blackbody source. A lens should not be used for focusing light on the entrance slit, as its collection and focusing efficiency will vary with wavelength. The signal at the output slit is then measured as a function of wavelength with a spectrally flat detector, such as a pyroelectric detector, thermopile, or Golay cell (see Section 7.10), and normalized with respect to the blackbody distribution.

### 4.7.4 Fabry–Perot Interferometers and Etalons

Fabry–Perot interferometers and etalons are optical filters that operate by multiple-beam interference of light reflected and transmitted by a pair of parallel flat or coaxial spherical reflecting interfaces. In its simplest form, an *etalon* consists of a flat, parallel-sided slab of transparent

material, which may or may not have semireflecting coatings on each face. An etalon may also consist of a pair of parallel, air-spaced flat mirrors of fixed spacing. If the reflective surfaces have adjustable spacing, or if their effective optical spacing can be adjusted by changing the gas pressure between the surfaces, the device is called a *Fabry–Perot interferometer*. These devices can be analyzed by the impedance concepts discussed in Section 4.2.6, but it is instructive to discuss their operation from the standpoint of multiple-beam interference.

Figure 4.158 shows the successive reflected and transmitted field amplitudes of a plane electromagnetic wave striking a plane-parallel Fabry–Perot etalon or interferometer at angle of incidence  $\theta'$ . Although the reflection coefficients at the two reflecting interfaces may be different, only the case where they are equal will be analyzed. Almost all practical devices are built this way. The optical-path difference between successive transmitted waves is  $2nl \cos \theta$ , where  $l$  is the interface spacing and  $n$  is the refractive index of the medium between these interfaces (air, glass, quartz, or sapphire, for example). If refraction occurs at the reflecting interfaces,  $\theta$  and  $\theta'$  will be related by Snell's law. The phase difference between successive transmitted waves is:

$$\delta = 2kl \cos \theta + 2\epsilon \quad (4.252)$$

where:

$$k = \frac{2\pi}{\lambda} = \frac{2\pi n}{\lambda_0} = \frac{2\pi n\nu}{c_0} \quad (4.253)$$

and  $\epsilon$  is the phase change (if any) on reflection.

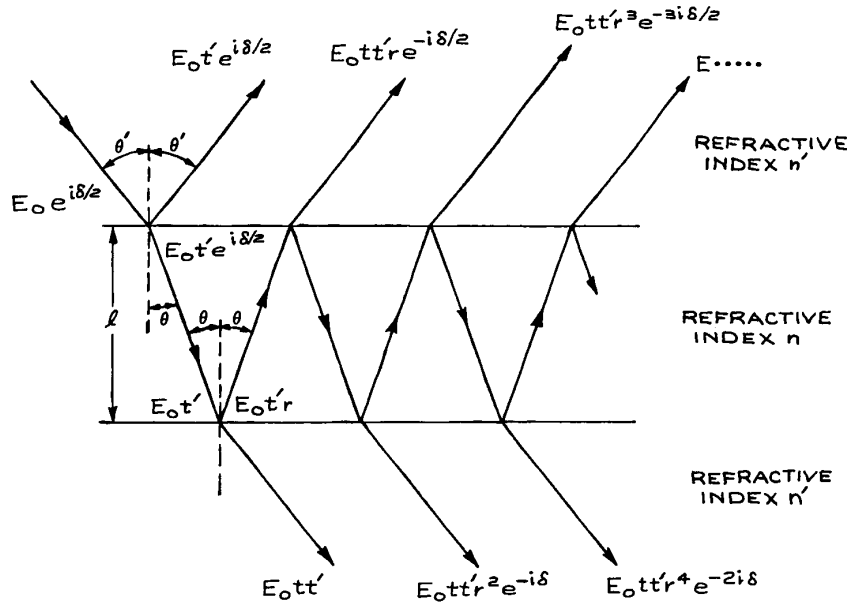
The total resultant transmitted complex amplitude is:

$$\begin{aligned} E_T &= E_0 t t' + E_0 t t' r^2 e^{-i\delta} + E_0 t t' r^4 e^{-2i\delta} + \dots \\ &= \frac{E_0 t t'}{1 - r^2 e^{-i\delta}} \end{aligned} \quad (4.254)$$

Here  $r$  and  $t$  are the reflection and transmission coefficients of the waves passing from the medium of refractive index  $n$  to that of refractive index  $n'$  at each interface,  $r'$  and  $t'$  are the corresponding coefficients for passage from  $n'$  to  $n$ , and  $E_0$  is the amplitude of the incident electric field.

The total transmitted intensity is:

$$I_T = \frac{|E_T|^2}{2Z} = \frac{I_0 |t t'|^2}{|1 - r^2 e^{-i\delta}|^2}. \quad (4.255)$$



**Figure 4.158** Paths of transmitted and reflected rays in a Fabry-Perot etalon. The complex amplitude of the fields associated with the rays at various points is shown.

Here  $|tt'| = T$  and  $|r|^2 = |r'|^2 = R$ , where  $T$  and  $R$  are the transmittance and reflectance of each interface. If there is no energy lost in the reflection process:

$$T = 1 - R \quad (4.256)$$

If the etalon had slightly absorbing reflective layers, this would no longer be true; if each layer absorbed a fraction  $A$  of the incident energy, then:

$$T = 1 - R - A \quad (4.257)$$

If  $A = 0$ , Equation (4.255) gives:

$$\frac{I_T}{I_0} = \frac{1}{1 + \frac{4R}{(1-R)^2} \sin^2 \frac{\delta}{2}} \quad (4.258)$$

This variation of transmitted intensity with  $\delta$  is called an Airy function and is illustrated in Figure 4.159 for several values of  $R$ . The variation of reflected intensity with  $\delta$  is just the inverse of Figure 4.159, since:

$$\frac{I_R}{I_0} = 1 - \frac{I_T}{I_0} \quad (4.259)$$

If  $A \neq 0$ , Equation (4.258) becomes:

$$\frac{I_T}{I_0} = \frac{\left(\frac{T}{1-R}\right)^2}{1 + \frac{4R}{(1-R)^2} \sin^2 \frac{\delta}{2}} \quad (4.260)$$

In either case, maximum transmitted intensity results when:

$$\delta = 2m\pi \quad (4.261)$$

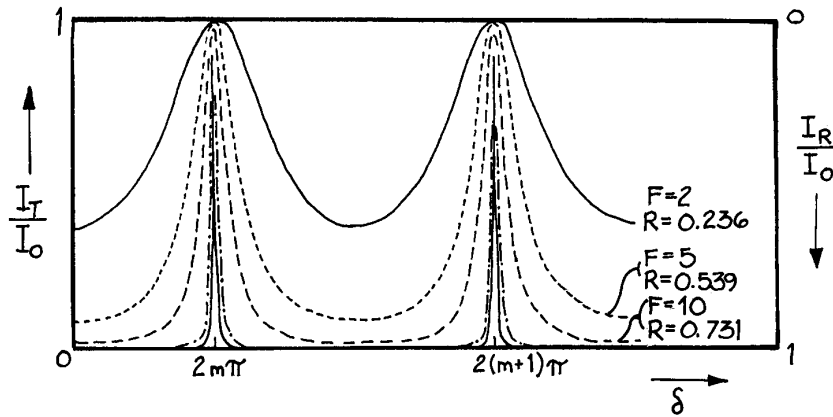
where  $m$  is an integer. If the phase change on reflection is neglected, this reduces to:

$$2l \cos \theta = m\lambda \quad (4.262)$$

where  $\lambda$  is the wavelength of the light in the medium between the two reflecting surfaces (plates). In normal incidence, transmission maxima result when:

$$l = m\lambda/2 \quad (4.263)$$

Even if the phase change on reflection is included, the adjustment in spacing necessary to go from one transmission maximum to the next is one half wavelength. Thus,



**Figure 4.159** Transmission and reflection characteristics of a Fabry-Perot etalon.

for example, the transmitted intensity as a function of plate separation when a Fabry-Perot interferometer is illuminated normally with plane monochromatic light is also as shown in Figure 4.159. In an ideal Fabry-Perot device, the overall transmittance at a transmission-intensity maximum is unity, whereas in a practical device, which will invariably have some losses, the maximum overall transmittance is reduced by a factor  $[T/(1 - R)]^2$ . As is clear from Figure 4.159, the transmission peaks of the device become very sharp as  $R$  approaches unity.

**Free Spectral Range, Finesse, and Resolving Power.** If a particular transmission maximum for fixed  $l$  occurs for normally incident light of frequency  $\nu_0$ , then:

$$\nu_0 = \frac{mc}{2l} \quad (4.264)$$

where  $c$  is the velocity of light in the material between the plates. Of course, the device will also show transmission maxima at all frequencies  $\nu_0 \pm pc/2l$  as well, where  $p$  is an integer. The frequency between successive transmission maxima is called the *free spectral range*:

$$\Delta\nu = \frac{c}{2l} \quad (4.265)$$

When  $R$  is close to unity, all phase angles  $\delta$  within a transmission maximum differ from the value  $2m\pi$  by only a small angle. Thus, we can write:

$$\delta = \frac{4\pi\nu l}{c} = \frac{4\pi\nu_0 l}{c} + \frac{4\pi(\nu - \nu_0)l}{c} \quad (4.266)$$

which can be written in the form:

$$\delta = 2m\pi + \frac{2\pi(\nu - \nu_0)}{\Delta\nu} \quad (4.267)$$

Equation (4.258) becomes:

$$\frac{I_T}{I_0} = \frac{1}{1 + \frac{4R\pi^2}{(1-R)^2} \left(\frac{\nu - \nu_0}{\Delta\nu}\right)^2} \quad (4.268)$$

Writing  $\pi\sqrt{R}/(1-R) = F$ , where  $F$  is called the *finesse*, the shape of a narrow transmission maximum can be written as:

$$\frac{I_T}{I_0} = \frac{1}{1 + [2(\nu - \nu_0)/\Delta\nu_{1/2}]^2} \quad (4.269)$$

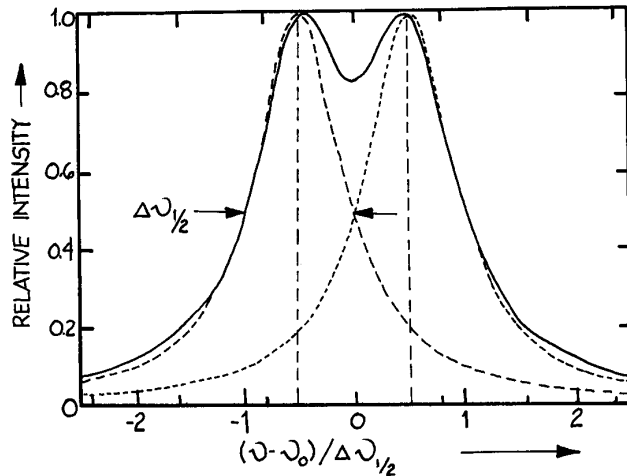
Here  $\Delta\nu_{1/2}$ , the full width at half maximum transmission of the transmission peak, is given by:

$$\Delta\nu_{1/2} = \Delta\nu/F \quad (4.270)$$

The *resolving power* of a Fabry-Perot device is a measure of its ability to distinguish between two closely spaced monochromatic signals. A good criterion for determining this is the *Rayleigh criterion*, which recognizes two closely

spaced lines as distinguishable if their half-intensity points on opposite sides of the two line shapes are coincident, as shown in Figure 4.160. Thus, for a Fabry–Perot device the resolving power is:

$$\mathfrak{R} = \frac{\nu}{\Delta\nu_{1/2}} = \frac{F\nu}{\Delta\nu} \quad (4.271)$$



**Figure 4.160** Two monochromatic spectral lines just resolved according to the Rayleigh criterion.

It would appear that very high resolving powers can be obtained by the use of high-reflectance plates (and consequent high finesse values, as illustrated in Table 4.11) and/or large plate spacings (and consequent small free spectral ranges). Additional practical considerations limit how far these approaches to high resolution can actually be used.

### Practical Operating Configurations and Performance Limitations of Fabry–Perot Systems.

The use of very high reflectance values to obtain improved resolving power is limited by the ability of the optical polisher to achieve ultraflat optical surfaces. Very good-quality plates for use in the visible can be obtained with flatnesses of  $\lambda/200$  over regions a few centimeters in diameter. Some claims of flatness in excess of  $\lambda/200$  are also seen. These should be viewed with skepticism – such a degree of flatness is extremely difficult to verify. If the plates have surface roughness  $\Delta s$ , the average error in plate spacing is  $\sqrt{2}\Delta s$ . This gives rise to a spread in the frequency of transmission maxima equal to:

$$\Delta\nu_s \approx \frac{mc\Delta s}{\sqrt{2}l^2} \approx \sqrt{2}\nu_0 \frac{\Delta s}{l} \quad (4.272)$$

Thus the flatness-limited finesse is:

$$F_s = \frac{\Delta\nu}{\Delta\nu_s} = \frac{1}{\sqrt{2}} \left( \frac{\Delta\nu}{\nu_0} \right) \frac{l}{\Delta s} = \frac{\lambda}{2\sqrt{2}\Delta s} \quad (4.273)$$

**Table 4.11 Properties of Fabry–Perot etalons**

Plate Separation $l$ (cm)	Reflectance $R$ (%)	Free Spectral Range $\Delta\nu = c/2Z$ Hz	Finesse $F = \pi\sqrt{\frac{R}{1-R}}$	Resolving Power $R = F\nu/\Delta\nu$
0.1	80	$1.5 \times 10^{11}$	14	$5.6 \times 10^4$
0.1	90	$1.5 \times 10^{11}$	30	$1.2 \times 10^5$
0.1	95	$1.5 \times 10^{11}$	61	$2.44 \times 10^5$
0.1	99	$1.5 \times 10^{11}$	313	$1.25 \times 10^6$
1.0	80	$1.5 \times 10^{10}$	14	$5.6 \times 10^5$
1.0	90	$1.5 \times 10^{10}$	30	$1.2 \times 10^6$
1.0	95	$1.5 \times 10^{10}$	61	$2.44 \times 10^6$
1.0	99	$1.5 \times 10^{10}$	313	$1.25 \times 10^7$
10.0	80	$1.5 \times 10^9$	14	$5.6 \times 10^6$
10.0	90	$1.5 \times 10^9$	30	$1.2 \times 10^7$
10.0	95	$1.5 \times 10^9$	61	$2.44 \times 10^7$
10.0	99	$1.5 \times 10^9$	313	$1.25 \times 10^8$

Note: Properties shown are at 500 nm (600 THz).

For example, two plates with surface roughness of  $\lambda/200$  would have a flatness-limited finesse of about 71. The use of super mirrors has allowed finesse values to 30 000 to be achieved. In practice, Fabry–Perot plates may not be randomly rough, and their stated flatness figure may represent the average deviation from parallelism of the two plates. In this case, the parallelism-limited finesse is:

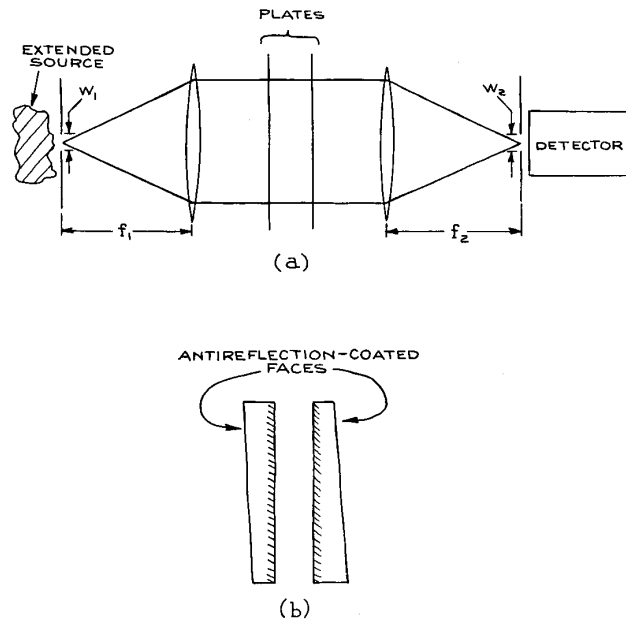
$$F_p = \frac{\lambda}{2\Delta s} \quad (4.274)$$

where  $\Delta s$  is the deviation from parallelism over the used aperture of the system. Thus, for example, with a 1 cm aperture, a 1/100 arc-second deviation from parallelism would imply  $\Delta s = 4.84$  nm and a parallelism-limited finesse of only 56 at 546 nm. Depending on the definition of what is meant by surface roughness and deviation from parallelism, Equations (4.273) and (4.274) may take slightly different forms.<sup>139</sup>

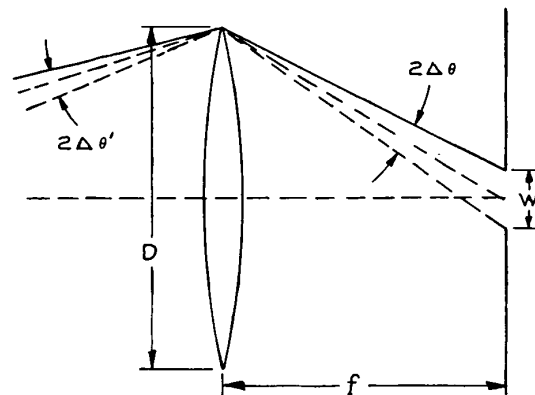
Fabry–Perot plates must be held very nearly parallel to achieve high resolution. Also, their separation should not be allowed to drift. A random variation of  $\Delta s$  will limit the finesse, as predicted by Equation (4.273). A fractional length change of  $10^{-6}$  in a 1 cm spacing Fabry–Perot device will limit the finesse to only about 27. Thus it must be isolated from vibrational interference to achieve high finesse. Further, if the frequencies of the transmission maxima are not to drift, the change in plate spacing due to thermal expansion must be minimized by constructing the device of low-expansion materials such as Invar,<sup>141</sup> Super-Invar,<sup>142</sup> fused silica,<sup>31</sup> or special ceramics.<sup>144</sup>

The resolving power of a Fabry–Perot system is also limited by the range of angles that can be transmitted through the instrument. One popular configuration is shown in Figure 4.161. A small source, or light from an extended source that passes through a small aperture, is collimated by a lens, passes through the interferometer, and is focused onto a second small aperture in front of a detector. The maximum angular spread of rays passing through the system will be governed by the aperture sizes  $W_1$  and  $W_2$  and the two focal lengths  $f_1$  and  $f_2$ . In the paraxial-ray approximation, the angular width  $2\Delta\theta$  of the paraxial-ray bundle that comes from, or is delivered to, an aperture of diameter  $W$  is equal to the angular width  $2\Delta\theta'$  that traverses the system, as illustrated in Figure 4.162. In this case:

$$\Delta\theta \approx W/2f \quad (4.275)$$



**Figure 4.161** (a) Operating scheme for a Fabry–Perot interferometer, using collimated light; (b) detail of typical plate configuration (the wedge angle of the reflecting plates is exaggerated).



**Figure 4.162** Angular factors involved in the collection of light transmitted through a Fabry–Perot interferometer by a circular aperture. In the paraxial-ray approximation,  $\Delta\theta' = \Delta\theta$ .



In Figure 4.161 the resolving power will be limited by  $W_1/f_1$  or  $W_2/f_2$ , whichever is the larger. The transmitted frequency spread associated with an angular spread  $\Delta\theta$  around the normal direction is, from Equations (4.262) and (4.264):

$$\Delta\nu_{1/2} = \nu_0(\Delta\theta)^2/2 \quad (4.276)$$

The corresponding aperture-limited finesse is:

$$F_A = \frac{\lambda}{l(\Delta\theta)^2} = \frac{4\lambda f^2}{lW^2} \quad (4.277)$$

For example, using a 1 mm diameter aperture, a 10 cm focal-length lens, and a 1-cm air-spaced Fabry–Perot device at 546.1 nm, the aperture-limited finesse is 218. The operating conditions of a Fabry–Perot system should be arranged so that  $F_A$  is at least three times larger than the desired operating finesse, which usually is ultimately limited by the flatness-limited finesse. The apertures in Figure 4.161 should be circular and accurately coaxial with both lenses, or the aperture finesse will be further reduced. If the limiting aperture (either a lens or the Fabry–Perot device itself) is  $D$ , there is also a diffraction limit to the finesse, which on axis is:

$$F_D = \frac{2D^2}{\lambda nl}. \quad (4.278)$$

At the  $p$ th fringe off axis:

$$F_D = \frac{D^2}{2p\lambda nl} \quad (4.279)$$

where  $n$  is the refractive index of the material between the plates. The overall finesse of a Fabry–Perot system,  $F_t$ , is related to the individual contributions to the finesse,  $F_i$ , by

$$F_t^{-2} = \sum_i F_i^{-2} \quad (4.280)$$

In the operating configuration of Figure 4.161, if the plate spacing is fixed, then with a broad-band source, all frequencies that satisfy Equation (4.262) will be transmitted. These frequencies become more closely spaced as the plate separation  $l$  is increased. There is a concomitant increase in resolution, but this may not prove useful if any of the spectral features under study are broader than the free spectral range. In this case, simultaneous transmission of two broad-

ened spectral features always occurs. For example, for two lines of wavelength  $\lambda_1$  and  $\lambda_2$  and of FWHM  $\Delta\lambda$ , it is usually possible to find two integers  $m_1$  and  $m_2$  such that:

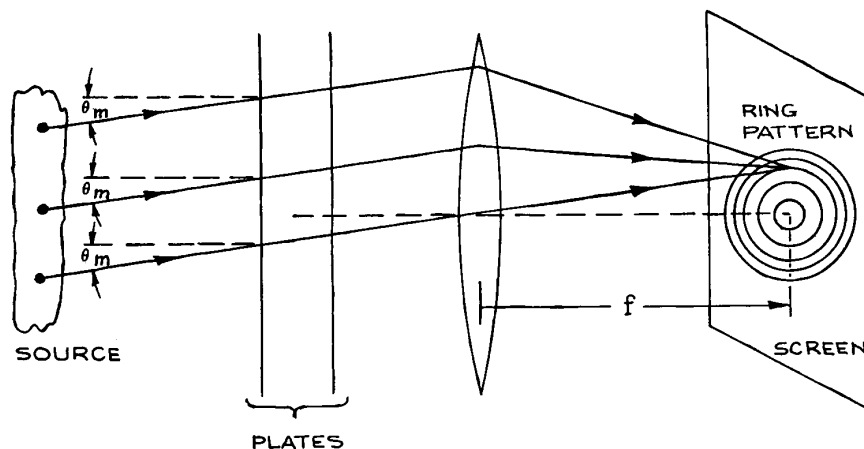
$$\begin{aligned} m_1\lambda'_1 &= 2l, & |\lambda'_1 - \lambda_1| &\leq \Delta\lambda \\ m_2\lambda'_2 &= 2l, & |\lambda'_2 - \lambda_2| &\leq \Delta\lambda \end{aligned} \quad (4.281)$$

Thus, for isolated, high-resolution studies of the spectral feature at  $\lambda_1$ , say, all potential interference features such as  $\lambda_2$  must be filtered out. This is done with a color or interference filter, or a prism or grating monochromator, before the signal is sent to the interferometer. The choice of pre-filtering element will depend on the relative wavelength spacing of  $\lambda_1$  and  $\lambda_2$ . The optical throughput of filters can be very high, 50% or more, but will not allow prefiltering of lines closer than a few nanometers. When very high throughput coupled with high resolution is essential, one or more additional Fabry–Perot devices of appropriate free spectral range can be used. For example, an etalon of spacing 0.01 mm has a free spectral range of about 15 nm. If it has a finesse of 50, its useful transmission bandwidth is 0.3 nm. Such an etalon could be coupled with etalons of spacing 0.1 mm and 1 mm with similar finesse to isolate a band only 0.003 nm wide from an original relatively broad frequency band transmitted through a filter. The final transmitted wavelength of such a combination of etalons can be tuned by tilting.

The wavelength of maximum transmission of a single etalon varies as:

$$\lambda = \lambda_0 \cos \theta \quad (4.282)$$

where  $\lambda_0$  is the wavelength for maximum transmission in normal incidence. In the interferometer arrangement shown in Figure 4.161(a), wavelength scanning can be accomplished in two ways: by pressure scanning, or by adjusting the axial position of one of the plates with a piezoelectric transducer.<sup>145</sup> To accomplish pressure scanning, the interferometer is mounted in a vacuum- or pressure-tight housing into which gas can be admitted at a slow steady rate. The resultant change of refractive index of the gas between the plates, which is approximately linear with pressure, causes a continuous, unidirectional scanning of the center transmitted frequency. The interferometer should be mounted freely within the housing so that the pressure differential between the



**Figure 4.163** Arrangement for observing a ring interference pattern with a Fabry-Perot etalon.

interior and exterior of the housing causes no deformation of the interferometer. Pressure scanning is not convenient where rapid or bidirectional scanning is required. Because no geometrical dimensions of the interferometer change during the scan, no change in finesse should occur during scanning.

Piezoelectric scanning is very convenient, but the transducer material must be specially selected to give uniform translation of the moving plate without any tilting – which would cause finesse reduction during scanning. If separate piezoelectric transducers<sup>145</sup> are mounted on the kinematic-mounted alignment drives of the plate holders, it is possible to trim the finesse electronically after good mechanical alignment has been achieved by micrometer adjustment. The alignment of the plates during scanning can be maintained with a servo system.<sup>146,147</sup>

A Fabry-Perot etalon can be used in the configuration shown in Figure 4.163 to give a high-resolution display of a narrow-band spectral feature that cannot be displayed with a scanning interferometer because (for example) the optical signal is a short pulse. In the arrangement shown in Figure 4.163, all rays from a monochromatic source that are incident at angle  $\theta_m$  on the plates will be transmitted and brought to a focus in a bright ring on a screen if:

$$\cos \theta_m = \frac{m\lambda}{2l} \quad (4.283)$$

In the paraxial approximation,  $\theta_m$  is a small angle and the radius  $\rho_m$  of a ring on the screen is:

$$\rho_m = f\theta_m \quad (4.284)$$

which can be written as:

$$\rho_m = f\sqrt{2 - \frac{m\lambda}{l}} \quad (4.285)$$

The radius of the smallest ring corresponds to the largest integer  $m$ , for which  $m\lambda/2l \leq 1$ . Successive rings going out from the center of the pattern correspond to integers  $m - 1$ ,  $m - 2$ , and so on.

If there is a bright spot at the center of the ring pattern, the radius of the  $p$ th ring from the center is:

$$\rho_p = f\sqrt{\frac{p\lambda}{l}} \quad (4.286)$$

This ring will be diffuse because of the finite finesse of the etalon, even if the source is highly monochromatic. The FWHM at half maximum intensity of the  $p$ th ring in this case is:

$$\Delta\rho_p = \frac{\rho_p}{2pF} \quad (4.287)$$

which can be verified by expanding Equation (4.260) in the angle  $\theta$ . This ring width is the same as would be produced by a source of FWHM  $\Delta\lambda = \lambda^2/2lF$ .

Equation (4.287) predicts that the total flux through each ring of the pattern will be identical, since the area of each ring is:

$$\Delta A = 2\pi\rho_p\Delta\rho_p = \pi f^2/2lF \quad (4.288)$$

The ratio of ring width to ring spacing for the  $p$ th ring in the pattern is:

$$\frac{\Delta\rho_p}{\rho_{p+1} - \rho_p} = \left[ 2Fp \left( 1 - \sqrt{1 + \frac{1}{p}} \right) \right]^{-1} \quad (4.289)$$

which is close to  $1/F$  except for the first few rings. Thus observation of the ring pattern from a quasimonochromatic source will yield its approximate bandwidth  $\Delta\lambda$  as:

$$\Delta\lambda \simeq \frac{\lambda}{2l} \frac{\Delta\rho}{\rho_{p+1} - \rho_p} \quad (4.290)$$

A convenient way of making a qualitative visual observation of this kind is to use an etalon in conjunction with a small telescope, as shown in Figure 4.164.

**Luminosity, Throughput, Contrast Ratio, and Étendue.** The luminosity of a Fabry–Perot system is, as already mentioned, much higher than that of any prism or grating monochromator. In the arrangement of Figure 4.161, the luminosity is:

$$\Phi = \left( \frac{T}{1-R} \right)^2 E_e(\lambda) S \frac{\pi W_2^2}{4f_2} \quad (4.291)$$

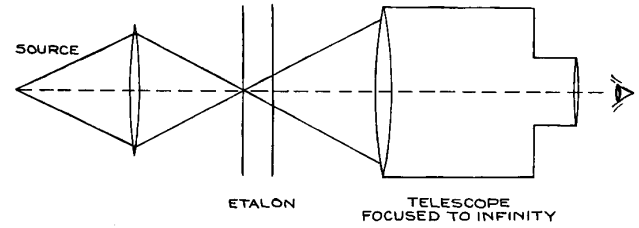
where  $\pi W_2^2/4f_2$  is the solid angle subtended by the output circular aperture at the output focusing lens, and  $S$  is the illuminated area of the plates. If the system has an aperture-limited finesse, then, from Equation (4.277) the luminosity becomes:

$$\Phi = \left( \frac{T}{1-R} \right)^2 E_e(\lambda) \frac{\pi\lambda S}{\rho F_A} \quad (4.292)$$

which can be written in terms of the resolving power  $\mathfrak{R}$  as:

$$\Phi = \left( \frac{T}{1-R} \right)^2 E_e(\lambda) \frac{2\pi S}{\mathfrak{R}} \quad (4.293)$$

The factor  $[T/(1-R)]^2$  is called the throughput of the system. To compare the luminosity predicted by Equation



**Figure 4.164** Arrangement for visual observation of Fabry–Perot fringes.

(4.293) with that of a grating, we shall assume that the throughput of the Fabry–Perot system is equal to the corresponding efficiency factor  $T$  of the grating in Equation (4.240). In this case, for identical resolving powers  $\mathfrak{R}$ , from Equations (4.293) and (4.240), we have:

$$\frac{\Phi(\text{F.P.})}{\Phi(\text{grating})} = \frac{\pi S}{(A \sin \beta)(l/f_2)} \quad (4.294)$$

where  $l/f_2$  is the angular slit height at the output-focusing element of the grating mono-chromator. For a blaze angle of  $30^\circ$  and equal Fabry–Perot aperture and grating area:

$$\frac{\Phi(\text{F.P.})}{\Phi(\text{grating})} = \frac{\pi}{l/f_2} \quad (4.295)$$

For a 0.3 m focal length grating instrument and a slit height of 3 cm (which is a typical maximum value for a small instrument of this size), the luminosity ratio is 71.4. This probably represents an optimistic comparison as far as the grating instrument is concerned, since high-resolution grating monochromators typically have focal lengths in excess of 0.75 m.

The *contrast ratio* of a Fabry–Perot system is defined as:

$$C = \frac{(I_T/I_0)_{\max}}{(I_T/I_0)_{\min}} \quad (4.296)$$

which from Equation (4.258) clearly has the value:

$$C = \left( \frac{1+R}{1-R} \right)^2 = 1 + \frac{4F^2}{\pi^2} \quad (4.297)$$

The *étendue*  $U$  of a Fabry–Perot system is a measure of its light-gathering power for a given frequency bandwidth  $\Delta\nu_{1/2}$ . It is defined by:

$$U = \Omega S \quad (4.298)$$

where  $A$  is the area of the plates and  $\Omega$  is the solid angle within which incident rays can travel and be transmitted with a specified frequency bandwidth. From Equation (4.276), since  $\Omega = \pi\Delta\theta^2$ :

$$\Delta\nu_{1/2} = \frac{\nu_0\Omega}{2\pi} \quad (4.299)$$

Therefore:

$$U = \Omega S = \frac{2\pi\Delta\nu_{1/2} \pi D^2}{\nu_0} \frac{1}{4} \quad (4.300)$$

which can be written in the form:

$$U = \Omega S = \frac{\pi^2 D^2 \lambda}{4lF_1} \quad (4.301)$$

### Spherical Mirror Fabry–Perot Interferometers.

Fabry–Perot interferometers with spherical concave mirrors are generally used in a confocal arrangement, as shown in Figure 4.165. In this configuration the various characteristics of the interferometer can be summarized as follows:

$$\left(\frac{I_T}{I_0}\right)_{\max} = \frac{T^2}{2(1-R)^2} \quad (4.302)$$

$$\Delta\nu(\text{FSR}) = c/4l \quad (4.303)$$

$$\text{reflectivity-limited finesse} = \frac{\pi\sqrt{R}}{(1-R)} \quad (4.304)$$

The approximate optimum radius of the apertures used in an arrangement similar to Figure 4.161 is  $1.15(\lambda r_3/ff)^{1/4}$ , where  $r$  is the radius of the spherical mirrors. Fabry–Perot interferometers with spherical mirrors can achieve very high finesse values.

The characteristics of all the possible combinations of two spherical mirrors that can be used to form a Fabry–Perot interferometer have been very extensively studied in connection with the use of such mirror arrangements to form the resonant cavities of laser oscillators.<sup>14–16</sup> In that application, these spherical-mirror resonators support Gaussian beam modes, and it is with such Gaussian beams that highest performance can be achieved with a spherical-mirror Fabry–Perot interferometer. Thus, a

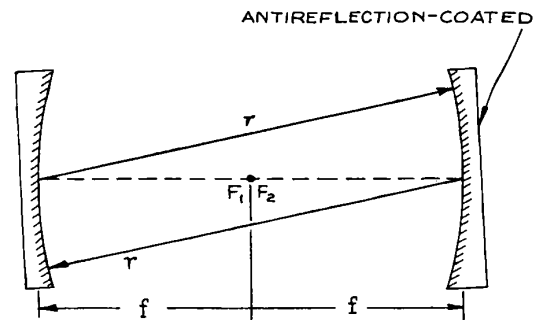
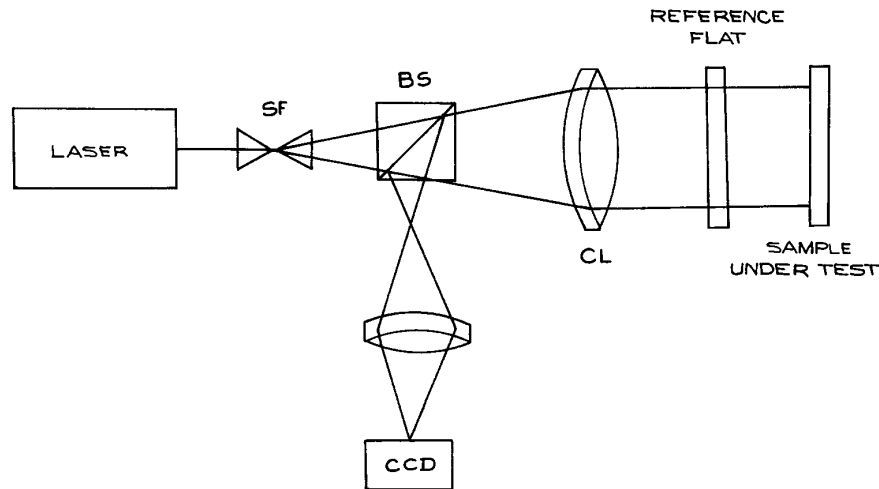


Figure 4.165 Confocal Fabry–Perot interferometer.

confocal Fabry–Perot interferometer is ideal for examining the spectral content of a Gaussian laser beam. The beam to be studied should be focused into the interferometer with a lens that matches the phase-front curvature and spot size of the laser beam to the radius of the Fabry–Perot mirror and the spot size of the resonant modes. (See Section 4.2.7.)

### 4.7.5 Design Considerations for Fabry–Perot Systems

Fabry–Perot interferometers and etalons are extremely sensitive to angular misalignment and fluctuations in plate spacing. Consequently, their design and construction are more demanding than for other laboratory optical instruments. The plates must be extremely flat, as already mentioned, and must be held without any distortion in high-precision alignment holders. The plate spacing must be held constant to about 1 nm in a typical high-finesse device for use in the visible, and the angular orientation must be maintained to better than 0.01 second of arc. Small-plate-spacing etalons are generally made commercially by optically contacting two plates to a precision spacer, or spacers, made of quartz. Such etalons are available from CVI Laser, II-VI Infrared, Michigan Aerospace, OFR, Research Electro-Optics, Rocky Mountain Instrument Co., TecOptics, Scientific Solutions, and SLS Optics, among others. Etalons of thickness in excess of 1 or 2 mm for use in the visible can be made from solid plane-parallel pieces of quartz or sapphire with dielectric coatings on their faces, or from materials, such as germanium or zinc



**Figure 4.166** Fizeau interferometer. SF – spatial filter, BS – beam splitter cube, CL – collimating lens.

selenide, for infrared use. Fabry–Perot interferometer plates are almost always slightly wedge-shaped and anti-reflection-coated on their outside faces to eliminate unwanted reflections, as shown in Figure 4.161(b). To prevent distortion of the plates they should be kinematically mounted in their holders. A good way to do this is to use three quartz or sapphire balls cemented to the circumferences of the plates and held in ball, slot, and plane locations with small retaining clips. For further details see Section 1.7.1 and Figures 1.52 and 1.53.

For the construction of Fabry–Perot interferometers, low-expansion materials should be used. The plate holders can be held at a fixed spacing by having them spring-loaded against three quartz or ceramic spacer rods or by the use of an Invar or Super-Invar spacer structure.

After a Fabry–Perot interferometer is aligned mechanically, it will creep slowly out of alignment, because of the easing of micrometer threads, for example. The alignment can be returned by having a piezoelectric element<sup>145</sup> incorporated in each plate-orientation adjustment screw.

Fabry–Perot interferometers are made commercially by Bernhard Halle Nachfl. GmbH, Bristol Instruments (who still provide service and parts for Burleigh interferometers), Hovemere, SSI, and Thorlabs.

The Fizeau interferometer is closely related to the Fabry–Perot, except that one of the mirrors is made totally reflecting and a beam splitter is used to extract the interfering waves. They measure the phase difference between a reference surface and the object under test, as shown schematically in Figure 4.166. These interferometers are commonly used to test the flatness of optical surfaces. They are available commercially from 4D Technology, Davidson Optronics, Logitech, Moeller-Wedel Optical GmbH, Sigma Koki Co, Silo, and Zygo.

#### 4.7.6 Double-Beam Interferometers

In a double-beam interferometer, waves from a source are divided into two parts at a beamsplitter and then recombined after traveling along different optical paths. The most practically useful of these interferometers are the Michelson and the Mach–Zehnder.

**The Michelson Interferometer.** The operation of a Michelson interferometer can be illustrated with reference to Figure 4.167. Light from an extended source is divided at a beamsplitter and sent to two plane mirrors  $M_1$  and  $M_2$ . If observations of a broad-band

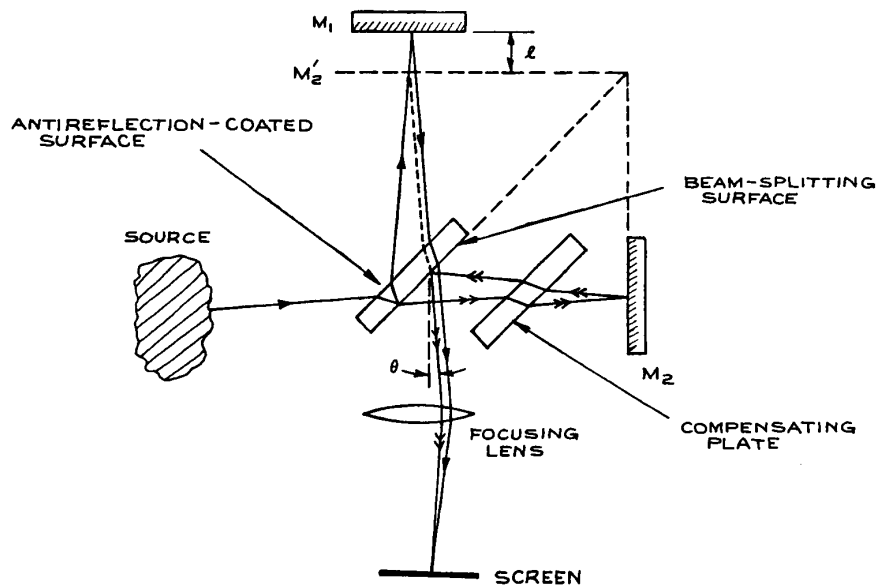


Figure 4.167 Michelson interferometer.

(temporally incoherent) source are being made, a compensating plate of the same material and thickness as the beamsplitter is included in arm 2 of the interferometer. This plate compensates for the two additional traverses of the beamsplitter substrate made by the beam in arm 1. Because of the dispersion of the beamsplitter material, it produces a phase difference that is wavelength-dependent, and without the compensating plate, interference fringes would not be observable with broad-band illumination (such as white light).

A maximum of illumination results at angle  $\theta$  in Figure 4.167 if the phase difference between the two beams coming from  $M_1$  and  $M_2$  is an integral multiple of  $2\pi$ . If  $M_2'$  represents the location of the reflection of  $M_2$  in the beamsplitter, the condition for a maximum at angle  $\theta$  is:

$$2l \cos \theta = m\lambda. \quad (4.305)$$

The positioning of  $M_1$  at  $M_2$  corresponds to  $l = 0, m = 0$ . If the output radiation is focused with a lens, a ring pattern is produced. A clear ring pattern is only visible if:

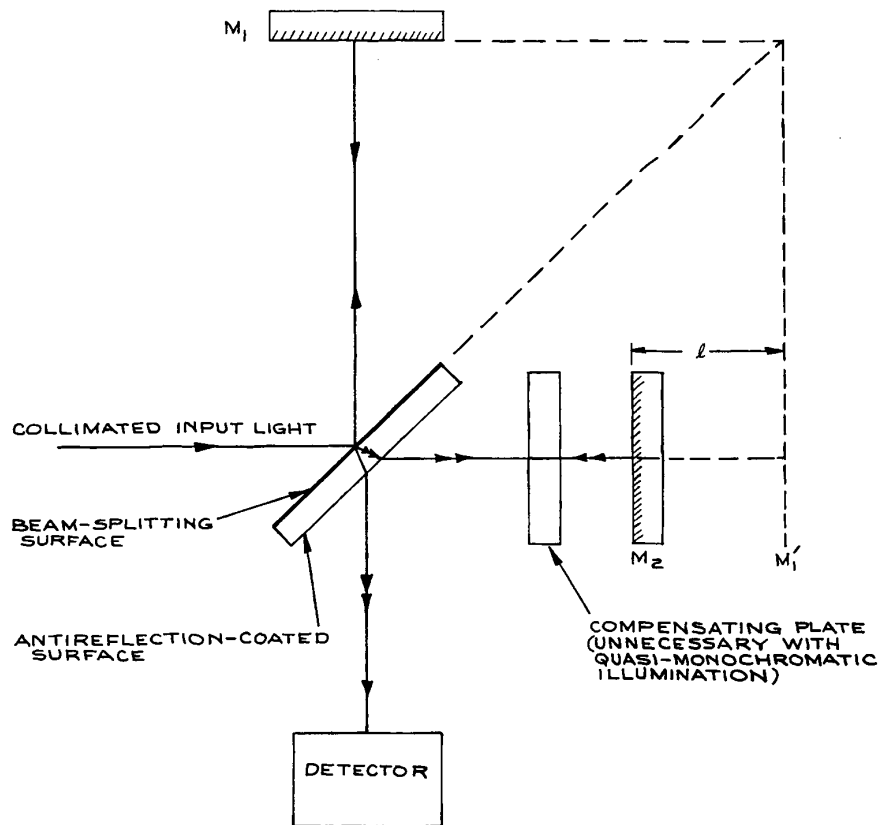
$$2l \cos \theta \lesssim l_c \quad (4.306)$$

where  $l_c$  is the coherence length of the radiation from the source. The clearness of the rings is usually described in terms of their *visibility*  $V$ , defined by the relation:

$$V_{\max} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \quad (4.307)$$

To obtain a ring-shaped interference pattern,  $M_1$  and  $M_2$  must be very nearly parallel; otherwise a pattern of curved dark and light bands will result. To study the fringes, mirror  $M_1$  must be scanned in a direction perpendicular to its surface. This can be done with a precision micrometer, or a piezoelectric transducer, if only small motion is desired. Because interference fringes are only observed with a broad-band source illuminating a Michelson (or other double-beam) interferometer when the two interferometer arms are of almost equal length, this mode of operation is extraordinarily sensitive. Such “white-light interferometry” forms the basis of a number of sensor systems for phenomena such as vibration, temperature, and pressure.<sup>149–151</sup>

In practice, Michelson interferometers are rarely used in the configuration shown in Figure 4.167, but are used instead with collimated illumination (a laser beam, or



**Figure 4.168** Simple Michelson interferometer for use with collimated input radiation.

parallel light obtained by placing a point source at the focal point of a converging lens) as shown in Figure 4.168. In this case, the output illumination is (almost) perfectly uniform, being a maximum if:

$$2l = m\lambda \quad (4.308)$$

where  $2l$  is the path difference between the two arms.

If monochromatic illumination of angular frequency  $\omega$  is used, we can write the fields of the waves from arms 1 and 2 as:

$$\begin{aligned} V_1 &= V_0 e^{i(\omega t - \phi_1)} \\ V_2 &= V_0 e^{i(\omega t - \phi_2)} \end{aligned} \quad (4.309)$$

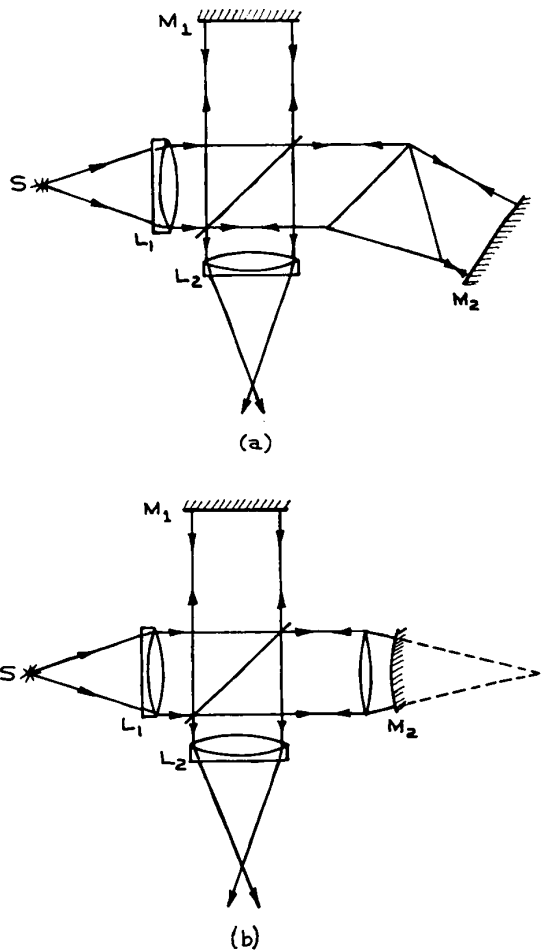
where it is assumed that the beamsplitter divides the incident beam into two equal parts.  $\phi_1$  and  $\phi_2$  are the optical phases of these two beams, which are dependent on the optical path lengths traveled in arms 1 and 2.

The output intensity is:

$$I \propto (V_1 + V_2) * (V_1 + V_2) \quad (4.310)$$

which gives

$$\begin{aligned} I &= I_0 [1 + \cos(\phi_2 - \phi_1)] \\ &= I_0 \left[ 1 + \cos\left(\frac{4\pi vl}{c}\right) \right] \end{aligned} \quad (4.311)$$



**Figure 4.169** Twyman–Green interferometer: (a) arrangement for testing a prism; (b) arrangement for testing a lens.  $S$  is a monochromatic source,  $L_1$  is a collimating lens, and  $L_2$  is a focusing lens. (Adapted with permission from M. Born and E. Wolf, *Principles of Optics*, 3rd edn., Pergamon Press, Oxford, copyright © 1965.)

in which  $I_0$  is the intensity of the light from one arm alone. Suppose an optical component such as a prism or lens is placed in one arm of the interferometer, as shown in Figure 4.169, with the plane mirror replaced by an appropriate radius spherical mirror in the case of the lens, and quasi-monochromatic illumination is used. Then the output fringe pattern will reveal inhomogeneities and defects in

the inserted component. A Michelson interferometer used in this way for testing optical components is called a *Twyman–Green interferometer*.<sup>11</sup>

Equation (4.302) illustrates why Michelson (and all other double-beam) interferometers are not directly suitable for spectroscopy. The distance the movable mirror must be adjusted to go from one maximum to the next is:

$$\Delta l = \lambda/2 \quad (4.312)$$

The distance required to go from a maximum- to a half-maximum-intensity point is:

$$\Delta l = \lambda/8 \quad (4.313)$$

Thus the effective finesse is only 4: sharp spectral peaks with monochromatic illumination are not obtained, as they are in multiple-beam interferometers.

**Fourier Transform Spectroscopy.** In an arrangement such as Figure 4.168, if the movable mirror is scanned at a steady velocity  $v$ , the resultant intensity as a function of position can be used to obtain considerable information about the spectrum of the source. If the source has intensity  $I(\nu)$  at  $\nu$ , the contribution to the intensity from the small band  $\nu d\nu$  at  $\nu$  is:

$$I_\nu(l) = I(\nu) \left( 1 + \cos \frac{4\pi\nu l}{c} \right) d\nu \quad (4.314)$$

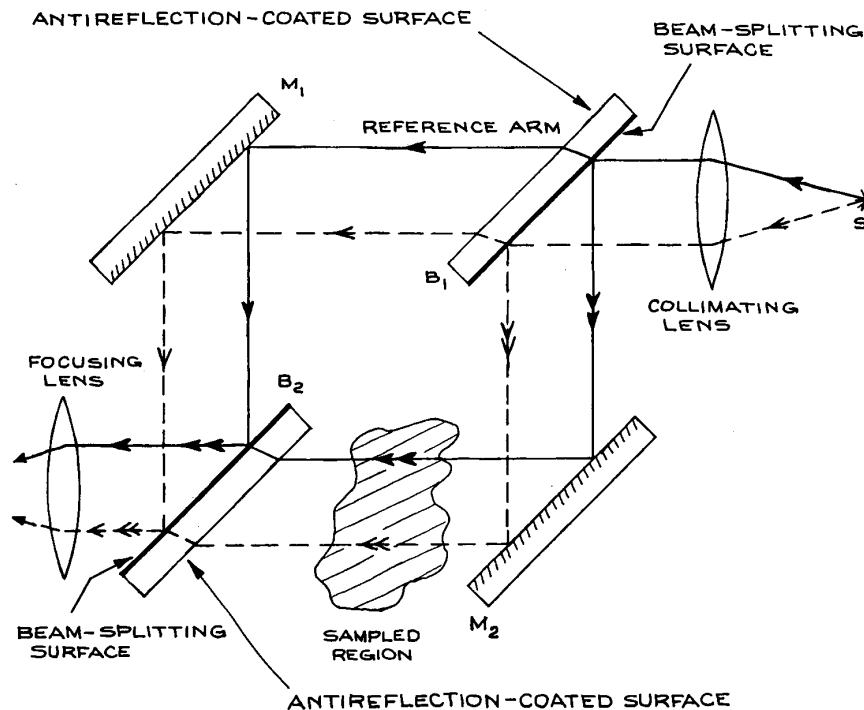
where  $l = \nu t$ . The total observed output from the full spectrum of the source is:

$$\begin{aligned} I(l) &= \int_{\nu=0}^{\infty} I(\nu) \left( 1 + \cos \frac{4\pi\nu l}{c} \right) d\nu \\ &= \frac{I_0}{2} + \int_{\nu=0}^{\infty} I(\nu) \cos \frac{4\pi\nu l}{c} d\nu \end{aligned} \quad (4.315)$$

The second term on the second line is essentially the cosine transform of  $I(\nu)$ . Thus  $I(\nu)$  can be found as the cosine transform of  $I(l) - (I_0/2)$ :

$$I(\nu) = \frac{2}{c} \int_{l=0}^{\infty} \left( I(l) - \frac{I_0}{2} \right) \cos \frac{4\pi l \nu}{c} dl \quad (4.316)$$





**Figure 4.170** Mach-Zehnder interferometer.  $M_1$ ,  $M_2$  are totally reflecting mirrors;  $B_1$ ,  $B_2$  are beamsplitters (usually equally reflective and transmissive). The distribution of output illumination reflects the refractive-index distribution in the sampled region.

This method for obtaining the spectrum of a source as the Fourier transform of an intensity recorded as a function of scanning mirror position is called *Fourier transform spectroscopy*. Such Fourier transform spectrometers have high luminosity and collect spectral information very efficiently: the whole emission spectrum is in reality sampled all at once, rather than one feature at a time, as in a conventional dispersive grating spectrometer. This is often called the Fellgett or Jacquinot advantage.

Commercial Fourier transform spectrometers have become widely available and easy to use. Suppliers include ABB Bomen, Bruker Optics, Melles Griot, MIDAC, Newport/Oriel, Perkin-Elmer, ScienceTech, Shimadzu, Thermo-Fischer Scientific (formerly Nicolet), and Varian (previously Bio-Rad and Digilab). These instruments have replaced grating instruments in many analytical applica-

tions, especially for infrared absorption spectroscopy. They must be “zeroed” by running a spectrum of the source itself before any absorbing material is included in the sample path, and the absorption of the sample is obtained by ratioing with respect to the empty sample path. They can also be checked by measuring the absorption spectrum of a known sample, such as a thin polystyrene film. Commercial instruments often have interchangeable sources and detectors to cover em spectral regions. For example they might use a hot ceramic (globar type) source for infrared work, and germanium, InGaAs, HgCdTe, or triglycine sulphate (TGS) detectors, depending on the spectral region of interest.

**The Mach-Zehnder Interferometer.** The Mach-Zehnder interferometer, shown in Figure 4.170, is widely

used for studying refractive-index distributions in gases, liquids, and solids. It is particularly widely used for studying the density variations in compressible gas flows,<sup>152</sup> the thermally induced density (and consequent refractive-index) changes behind shock fronts, and thermally induced density changes produced by laser beams propagating through transparent media.<sup>153,154</sup> If the light entering the interferometer is perfectly collimated, uniform output illumination results, which is a maximum if the path difference between the two arms is an integral number of wavelengths. If the image of  $M_1$  in beamsplitter  $B_2$  is not parallel to  $M_2$ , the output fringe pattern is essentially the interference pattern observed from a wedge – a series of parallel bright and dark bands. If the medium in arm 2 is perturbed in some way, so that a spatial variation in refractive index results, the output fringe pattern changes. Analysis of the new fringe pattern reveals the spatial distribution of refractive-index variation along the beam path in arm 2. Various design variations on the Mach–Zehnder interferometer exist, notably the Jamin and the Sirks–Pringsheim.<sup>11,154</sup>

**Design Considerations for Double-Beam Interferometers.** Michelson and Mach–Zehnder interferometers are much easier to build in the laboratory than multiple-beam interferometers, such as the Fabry–Perot. Because the beams only reflect off the interferometer mirrors once, mirrors of flatness  $\lambda/10$  or  $\lambda/20$  are quite adequate for building a good instrument. The usual precautions should be taken in construction: stable rigid mounts for the mirrors and beamsplitters and, if long-term thermal stability is required, low-expansion materials. Piezoelectric control of mirror alignment is not generally necessary, as these instruments are less sensitive to small angular misalignments than Fabry–Perot interferometers. To achieve freedom from alignment disturbances along one or two orthogonal axes, the plane mirrors of a Michelson interferometer can be replaced with right-angle prisms or corner-cube retroreflectors, respectively. Some care should be taken to avoid source-polarization effects in these instruments; the beamsplitters are generally designed to reflect and transmit equal amounts only for a specific input polarization. If polarizing materials are placed in either arm of the interferometer, no fringes will be seen if, by any chance, the beams from the two arms emerge orthogonally polarized.

## Endnotes

- a 1 GHz (GigaHertz) =  $10^9$  Hz, 1 THz (TeraHertz) =  $10^{12}$ . The infrared range can also be somewhat arbitrarily divided up into the submillimeter region ( $\lambda = 0.1$ – $1$  mm), far-infrared ( $\lambda = 20$ – $100$   $\mu\text{m}$ ), middle infrared ( $\lambda = 3$ – $20$   $\mu\text{m}$ ), and the near infrared ( $\lambda = 0.7$ – $3$   $\mu\text{m}$ ).
- b At the boundary separating one or more anisotropic media, Snell's Law must be used with some care as, in anisotropic media, the ray direction is not in general parallel to the wave vector (or wave normal), which is also used to characterize the direction of the ray.<sup>15</sup>
- c Used by physicians for examining the retina of a patient's eye.<sup>17</sup>
- d A pinhole camera is free of all aberrations in a geometrical optics description.
- e A simple lens with just two spherical surfaces.
- f Often called a TM wave.
- g Often called a TE wave
- i The word rugate means corrugated, having alternating ridges and depressions
- h The old pressure unit of pounds per square inch (psi) = 6895 Pa
- j The phosphorous is often omitted, and the precise stoichiometry of (In, Ga) and (As, P) determines the laser wavelength
- k In practice, in any real system, transmission losses, and aberrations will cause  $B_2 < B_1$

## Cited References

1. *Proceedings of the Symposium on Quasi-Optics*, New York, June 8–10, 1964, J. Fox (Ed.), Polytechnic Press of the Polytechnic Institute of Brooklyn, New York, 1964.
2. *Proceedings of the Symposium on Submillimeter Waves*, New York, March 31–April 2, 1970, J. Fox (Ed.), Polytechnic Press of the Polytechnic Institute of Brooklyn, New York, 1964.
3. Amir Mortazwi, Tatsuo Itoh, and James Harvey, *Active Antennas and Quasi-Optical Arrays*, Wiley-IEEE Press, New York, 1998.
4. J. Lesurf, *Millimeter-Wave Optics, Devices and Systems*, Lavoisier, Chachan, 1990.

5. K. M. Baird, D. S. Smith, and B. G. Whitford, Confirmation of the currently accepted value 299792458 metres per second for the speed of light, *Opt. Commun.*, **31**, 367–368, 1979.
6. E. R. Cohen and B. N. Taylor, *Committee on Data for Science and Technology, CODATA Bulletin*, No. **63**, 1986.
7. G. W. C. Kaye and T. H. Laby, *Tables of Physical and Chemical Constants*, 16th edn., Longman, London: 1995. Now available free online at <http://www.kayelaby.npl.co.uk/>
8. C. D. Coleman, W. R. Bozman, and W. F. Meggers, Table of Wavenumbers, *US Nat'l. Bur. Std. Monograph 3*, Vols. 1–2, 1960.
9. M. W. Urban, *Attenuated Total Reflectance Spectroscopy of Polymers: Theory and Practice*, American Chemical Society, Washington, DC, 1996.
10. R. Wiesendanger, *Scanning Probe Microscopy and Spectroscopy*, Cambridge, University Press, Cambridge 1994.
11. M. Born and E. Wolf, *Principles of Optics*, 7th ed., Cambridge University Press, Cambridge, 1999.
12. E. E. Wahlstrom, *Optical Crystallography*, 5th edn., John Wiley & Sons, Inc., New York, 1979.
13. W. A. Shurcliff, *Polarized Light: Production and Use*, Harvard University Press, Cambridge, MA 1962.
14. A. Yariv, *Optical Electronics in Modern Communications*, 5th edn., Oxford University Press, New York, 1997.
15. C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
16. H. Kogelnik and T. Li, *Laser beams and resonators*, *Proc. IEEE*, **54**, 1312–1329, 1966.
17. R. Kingslake, *Optical System Design*, Academic Press, Orlando, FL, 1983.
18. Code V. Available from Optical Research Associates.
19. Zemax. Available from Zemax Development Corporation.
20. Oslo developed by Sinclair Optics, Inc. Distributed by Lambda Research Corporation.
21. Solstis Odyssey. Available from Optis.
22. Optikwerk. Available from Optikwerk, Inc.
23. R. W. Ditchburn, *Light*, 3rd edn., Academic Press, New York, 1976.
24. L. Levi, *Applied Optics*, John Wiley & Sons, Inc., New York, Vol. 1, 1968.
25. W. J. Smith, *Modern Optical Engineering*, 3rd edn., SPIE Press, Bellingham, WA, 2000.
26. W. J. Smith, *Modern Lens Design: A Resource Manual*, McGraw-Hill, New York, 1992.
27. R. R. Shannon, *The Art and Science of Optical Design*, Cambridge University Press, Cambridge 1997.
28. M. Laikin, *Lens Design*, 3rd edn., Marcel Dekker, New York, 2001.
29. S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, 3rd edn., John Wiley & Sons, Inc., New York, 1994.
30. Diffraction-limited spherical lenses are available from Melles Griot, Optics for Research, Newport, Special Optics, and J. L. Wood Optical Systems, among others (see following reference).
31. *Laser Focus Buyers Guide*, published annually by Pennwell Publishing Co., 1421 South Sheridan, Tulsa, OK 74112 01460, lists a large number of suppliers of a wide range of optical components and systems as does the *The Photonics Buyers' Guide*, published annually by Photonics Spectra, Laurin Publishing Co., Berkshire Common, P.O. Box 4949, Pittsfield, MA 01202-4949. Additional valuable listings of this sort are: *Lasers and Optronics Buying Guide*, published annually by Cahners, 301 Gibraltar Drive, Box 650, Morris Plains, NJ 07950-0650; and Lightwave, 98 Spit Brook Road, Suite 100, Nashua, NH 03062–5737.
32. The “float” method for manufacturing plate glass (developed by Pilkington Glass) involves the drawing of the molten glass from a furnace, where the molten glass floats on the surface of liquid tin. The naturally flat surface of the liquid metal ensures the production of much larger sheets of better-flatness glass than was possible by earlier techniques.
33. Alexander Hornberg (Ed.), *Handbook of Machine Vision*, John Wiley & Sons, Inc., New York, 2006.
34. G. Rempe, R. J. Thompson, H. J. Kimble, and R. Lalezari “Measurement of ultralow losses in an optical interferometer”, *Opt. Lett.*, **17**, 363–365, 1992.
35. C. J. Hood, H. J. Kimble, and J. Ye, Characterization of high-finesse mirrors: loss, phase shifts, and mode structure in an optical cavity, *Phys. Rev. A*, **64**, 033804, 2001.
36. H. R. Bilger, P. V. Wells, and G. E. Stedman, Origins of fundamental limits for reflection losses at multilayer dielectric mirrors, *Appl. Opt.*, **33**, 7390–7396, 1994.
37. H.-J. Cho, M.-J. Shin, and J.-C. Lee, Effects of substrate and deposition method onto the mirror scattering, *Appl. Opt.*, **45**, 1440–1446, 2006.
38. A. B. Meinel, Astronomical telescopes, in *Applied Optics and Optical Engineering*, Vol. 5, R. Kingslake (Ed.), Academic Press, New York, 1969.

39. W. Brouwer and A. Walther, Design of optical instruments, in *Advanced Optical Techniques*, A. C. S. Van Heel, (Ed.), North-Bouand, Amsterdam, 1967.
40. Celestron Telescopes are available from Celestron International, 2835 Columbia Street, P.O. Box 3578, Torrance, CA 90503; (310) 328-9560.
41. Meade Telescopes are available from Meade Instruments Corporation, 6001 Oak Canyon, Irvine, California 92618-5200, (949) 451-1450.
42. W. T. Wellford and R. Winston, *High Collection Nonimaging Optics*, Academic Press, San Diego, CA, 1989.
43. Roland Winston, Juan C. Miñano, and Pablo G. Benitez, *Nonimaging Optics*, Elsevier, Amsterdam 2005.
44. A. Girard and P. Jacquinot, Principles of instrumental methods in spectroscopy, in *Advanced Optical Techniques*, A. C. S. Van Heel (Ed.), North-Bouand, Amsterdam 1967.
45. J. D. Strong, *Procedures in Experimental Physics*, Prentice-Hall, Englewood Cliffs, NJ 1938.
46. David R. Lide (Ed.), *Handbook of Chemistry and Physics*, 81st. edn., CRC Press, Boca Raton, FL, 2000.
47. D. R. Goff, *Fiber Optic Reference Guide*, 3rd edn., Focal Press (Elsevier), Amsterdam, 2002.
48. U. Hochuli and P. Haldemann, Indium sealing techniques, *Rev. Sci. Instr.*, **43**, 1088-1089, 1972.
49. Epoxy-removing solvents are available from Electron Microscopy Services, Oakite, or Hydroclean.
50. T. C. Poulter, A glass window mounting for withstanding pressure of 30 000 atmospheres, *Phys. Rev.*, **35**, 297, 1930.
51. W. Paul, W. W. Meis, and J. M. Besson, Windows for optical measurements at high pressures and long infrared wavelengths, *Rev. Sci. Instr.*, **39**, 928-930, 1968.
52. *High Pressure Technology*, I. L. Spain and J. Paawe (Eds.), Dekker, New York, 1967.
53. *Compilation of ASTM Standard Definitions*, 3rd edn., American Society for Testing and Materials, Philadelphia, 1976.
54. Glass, quartz, and sapphire vacuum window assemblies are available from Adolf Meller, Ceramaseal, Varian, and Vacuum Generators.
55. D. H. Martin (Ed.), *Spectroscopic Techniques for Far Infa-Red, Submillimetre and Millimetre Waves*, NorthHolland, Amsterdam, 1967.
56. K. D. Moller and W. G. Rothschild, *Far-Infrared Spectroscopy*, Wiley-Interscience, New York, 1971.
57. A. Hadni, *Essentials of Modern Physics Applied to the Study of the Infrared*, Pergamon Press, Oxford, 1967.
58. G. A. Fry, The eye and vision, in *Applied Optics and Optical Engineering*, Vol. 2, R. Kingslake (Ed.), Academic Press, 1965.
59. C. C. Davis and R. A. McFarlane, Lineshape effects in atomic absorption spectroscopy, *J. Quant. Spect. Rad. Trans.*, **18**, 151-170, 1977.
60. P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology: Generation, Transmission and Detection*, John Wiley & Sons, Inc., New York, 1962.
61. R. D. Hudson, Jr., *Infrared System Engineering*, Wiley-Interscience, New York, 1969.
62. J. A. R. Samson, *Techniques of Vacuum Ultraviolet Spectroscopy*, John Wiley & Sons, Inc., New York, 1967.
63. Available from EG&G Optoelectronics, ILC, Verre & Quartz, and Xenon Corporation.
64. Available from EG&G Optoelectronics, Verre & Quartz, Resonance, and Xenon Corporation.
65. J. P. Markiewicz and J. L. Emmett, Design of flashlamp driving circuits, *IEEE J. Quant. Electron.*, **QE-2**, 707-711, 1966.
66. J. G. Edwards, Some factors affecting the pumping efficiency of optically pumped lasers, *Appl. Opt.*, **6**, 837-843, 1967.
67. H. J. Baker and T. A. King, Optimization of pulsed UV radiation from linear flashtubes, *J. Phys. E.: Sci. Instr.*, **8**, 219-223, 1975.
68. William T. Silfvast, *Laser Fundamentals*, 2nd edn., Cambridge University Press, Cambridge, 2004.
69. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, John Wiley & Sons, Inc., New York, 1991.
70. A. E. Siegman, *Lasers*, University Science Books, Mill Valley, CA, 1986.
71. H. Rabin and CL. Tang (Eds.), *Quantum Electronics*, Vols. 1A and 2A, *Nonlinear Optics*, Academic Press, New York, 1975.
72. *Nonlinear Optics*, Proceedings of the Sixteenth Scottish Universities Summer School in Physics, 1975, P. G. Harper and B. S. Wherrett (Eds.), Academic Press, London, 1977.
73. N. Bloembergen, *Nonlinear Optics*, Benjamin, New York, 1965.
74. F. Zernike and J. E. Midwinter, *Applied Nonlinear Optics*, John Wiley & Sons, Inc., New York, 1973.
75. Robert W. Boyd, *Nonlinear Optics*, Academic Press, Boston, 1992.

76. D. Röss, *Lasers, Light Amplifiers and Oscillators*, Academic Press, New York, 1969.
77. W. Koechner, *Solid State Laser Engineering*, 5th revised and updated edn., Springer-Verlag, Berlin, 1999.
78. *Topics in Applied Phys.*, F. P. Schäfer (Ed.), Vol. 1, *Dye Lasers*, 3rd revised and enlarged edn., Springer-Verlag, Berlin, 1990.
79. H. W. Furumoto and H. L. Cecon, Optical pumps for organic dye lasers, *Appl. Opt.*, **8**, 1613–1623, 1969.
80. J. F. Holzrichter and A. L. Schawlow, Design and analysis of flashlamp systems for pumping organic dye lasers, *Ann. N.Y. Acad. Sci.*, **168**, 703–714, 1970.
81. T. B. Lucatoro, T. J. McIlrath, S. Mayo, and H. W. Furumoto, High-stability coaxial flashlamp-pumped dye laser, *Appl. Opt.*, **19**, 3178–3180, 1980.
82. C. C. Davis and T. A. King, Gaseous ion lasers, in *Advances in Quantum Electronics*, Vol. 3, D. W. Goodwin (Ed.), Academic Press, London, 1975, pp. 169–454.
83. W. B. Bridges, Ion lasers, in *Handbook of Laser Science and Technology*, Vol. 1: *Lasers in All Media*, M. J. Weber (Ed.), CRC Press, Boca Raton, FL, 1982.
84. C. C. Davis, Neutral gas lasers, in *Handbook of Lasers Science and Technology*, Vol. 1, *Lasers in All Media*, M. J. Weber (Ed.), CRC Press, Boca Raton, FL, 1982.
85. D. C. Tyte, Carbon dioxide lasers, in *Advances in Quantum Electronics*, Vol. 1, D. W. Goodwin (Ed.), Academic Press, London, 1970.
86. J. J. Degnan, The waveguide laser: a review, *Appl. Phys.*, **11**, 1–33, 1976.
87. H. Seguin and J. Tulip, Photoinitiated and photosustained laser, *Appl. Phys. Lett.*, **21**, 414–415, 1972.
88. J. D. Cobine, *Gaseous Conductors*, Dover, New York, 1958.
89. H. J. Seguin, K. Manes, and J. Tulip, Simple inexpensive laboratory-quality Rogowski TEA laser, *Rev. Sci. Instr.*, **43**, 1134–1139, 1972.
90. A. J. Beaulieu, Transversely excited atmospheric pressure CO<sub>2</sub> laser, *Appl. Phys. Lett.*, **16**, 504–505, 1970.
91. D. Basting, F. P. Schäfer, and B. Steyer, A simple, high power nitrogen laser, *Opto-electron.*, **4**, 43–49, 1972.
92. P. Schenck and H. Metcalf, Low cost nitrogen laser design for dye laser pumping, *Appl. Opt.*, **12**, 183–186, 1973.
93. C. P. Wang, Simple fast-discharge device for high-power pulsed lasers, *Rev. Sci. Instr.*, **47**, 92–95, 1976.
94. A. J. Schwab and F. W. Bouinger, Compact high-power N<sub>2</sub> laser: circuit theory and design, *IEEE J. Quant. Electron.*, **QE-12**, 183–188, 1976.
95. C. L. Sam, Small-size discrete-capacitor N<sub>2</sub> laser, *Appl. Phys. Lett.*, **29**, 505–506, 1976.
96. M. Feldman, P. Lebow, F. Raab, and H. Metcalf, Improvements to a home-built nitrogen laser, *Appl. Opt.*, **17**, 774–777, 1978.
97. Michel J. F. Dignonnet (Ed.), *Rare Earth Doped Fiber Lasers And Amplifiers*, 2nd Revised edn., Taylor & Francis Ltd., Abingdon, 2001.
98. L. F. Mollenauer, Dyelike Lasers for the 0.9–2  $\mu\text{m}$  region using F<sub>2</sub><sup>+</sup> centers in alkali halides, *Opt. Lett.*, **1**, 164, 1977 (see also *Opt. Lett.* **3**, 48–50, 1978; **4**, 247–299, 1979; **5**, 188–190, 1980).
99. R. C. Greenhow and A. J. Schmidt, Picosecond light pulses, in *Advances in Quantum Electronics*, Vol. 2, D. W. Goodwin (Ed.), Academic Press, London, 1973.
100. S. L. Shapiro (Ed.), *Ultrashort Light Pulses, Picosecond Techniques and Applications*, Topics in Applied Physics, Vol. 18, Springer, Berlin, 1977.
101. *Picosecond Optoelectronic Devices*, C. H. Lee (Ed.), Academic Press, Orlando, FL, 1984.
102. P. Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd edn., Prentice-Hall, Upper Saddle River, NJ, 1997.
103. J. Gowar, *Optical Communication Systems*, 2nd edn., Prentice-Hall, Englewood Cliffs, NJ, 1993.
104. D. Wood, *Optoelectronic Semiconductor Devices*, Prentice-Hall, New York, 1994.
105. G. P. Agrawal and N. K. Dutta, *Long-Wavelength Semiconductor Lasers*, Van Nostrand Reinhold, New York, 1986.
106. G. H. B. Thompson, *Physics of Semiconductor Laser Devices*, John Wiley & Sons, Ltd, Chichester, 1980.
107. S. L. Chuang, *Physics of Optoelectronic Devices*, John Wiley & Sons, Inc., New York, 1995.
108. J. Singh, *Semiconductor Optoelectronics*, McGraw-Hill, New York, 1995.
109. *Semiconductor Devices for Optical Communication*, Vol. 39, H. Kressel Ed., Topics in Applied Physics, 2nd updated edn., Springer-Verlag, Berlin, 1982.
110. *Semiconductor Lasers*, E. Kapon (Ed.), Vols. 1 and 2, Academic Press, San Diego, CA 1999.
111. E. D. Hinkley, K. W. Nill, and F. A. Blum, Infrared spectroscopy with tunable lasers, in *Topics in Applied Physics*, Vol. 2, H. Walther (Ed.), Springer, Berlin, 1976, pp. 125–196.
112. J. Faist, F. Capasso, D. L. Sivco, et al., Quantum cascade laser. *Science*, **264** (5158), 553–556, 1994.
113. T. J. Kane and R. L. Byer, Monolithic, unidirectional single-mode Nd:YAG ring laser, *Optics Lett.*, **10**, 65–67, 1985.

114. T. W. Hänsch, Repetitively pulsed tunable dye laser for high resolution spectroscopy, *Appl. Opt.*, **11**, 895–898, 1972.
115. R. Wallenstein and T. W. Hänsch, Linear pressure tuning of a multielement dye laser spectrometer, *Appl. Opt.*, **13**, 1625–1628, 1974.
116. R. Wallenstein and T. W. Hänsch, Powerful dye laser oscillator-amplifier system for high-resolution spectroscopy, *Opt. Commun.*, **14**, 353–357, 1975.
117. G. L. Eesley and M. D. Levenson, Dye-laser cavity employing a reflective beam expander, *IEEE J. Quant. Electron.* **QE-12**, 440–442, 1976.
118. M. G. Littman and H. J. Metcalf, Spectrally narrow pulsed dye laser without beam expander, *Appl. Opt.*, **17**, 2224–2227, 1978.
119. M. Littman and J. Montgomery, Grazing-incidence designs improve pulsed dye lasers, *Laser Focus/Electro-optics*, **24**, 70–86, 1988.
120. M. B. Radunsky Improving high power OPO performance, *Laser Focus World*, p. 107, October 1995.
121. B. J. Orr et al., Spectroscopic applications of pulsed tunable parametric oscillators, in *Tunable Laser Applications*, F. J. Duarte (Ed.), Marcel Dekker, New York, 1995.
122. R. G. Batchko, D. R. Weise, T. Plettner, et al., Continuous-wave 532-nm-pumped singly resonant optical parametric oscillator based on periodically poled lithium niobate, *Opt. Lett.* **23**(3), 168, 1998.
123. M. E. Klein, D.-H. Lee, J.-P. Meyn, K.-J. Boller, and R. Wallenstein, Singly resonant continuous-wave optical parametric oscillator pumped by a diode laser, *Opt. Lett.*, **24**, 1142–1144, 1999.
124. F. Müller, G. von Basum, A. Popp, et al., Long-term frequency stability and linewidth properties of continuous-wave pump-resonant optical parametric oscillators, *Appl. Phys.* **80**, 307–313, 2004.
125. I. P. Kaminow, *An Introduction to Electro-optic Devices*, Academic Press, New York, 1974.
126. A. Korpel, *Acousto-Optics*, Marcel Dekker, New York, 1988, see also Acousto-Optics – a review of fundamentals, *Proc. IEEE.*, **69**, 48–53, 1981.
127. I. C. Chang, Acousto-optic devices and applications, *IEEE Trans. Sonics and Ultrasonics*, **SU-23**, 2–22, 1976.
128. American National Standard for Safe Use of Lasers, ANSI Z136-2-2000, Published by the Laser Institute of America, Suite 128, 13501 Ingenuity Drive, Orlando, FL 32826.
129. D. Sliney and M. Wolbarsht, *Safety with Lasers and Other Optical Sources: A Comprehensive Handbook*, Plenum, New York, 1980.
130. M. J. Weber (Ed.), *CRC Handbook of Laser Science and Technology*, Vol. 1, Lasers and Masers, CRC Press, Boca Raton, FL, 1982.
131. P. Jacquinot, The luminosity of spectrometers with prisms, gratings, or Fabry–Perot etalons, *J. Opt. Soc. Am.*, **44**, 761–765, 1954.
132. Applied Optics and Optical Engineering, Vol. 5: Optical Instruments Part II, R. Kingslake (Ed.), Academic Press, New York, 1969.
133. J. F. James and R. S. Sternberg, *The Design of Optical Spectrometers*, Chapman and Hall, London, 1969.
134. *The Photonics Design and Applications Handbook*, published annually by Laurin Publishing Co. Inc., Pittsfield, MA.
135. W. G. Fastie, A small plane grating monochromator, *J. Opt. Soc. Am.*, **42**, 641–647, 1952.
136. H. Ebert, Zwei Formen von Spectrographen, *Annalen der Physik und Chemie*, **38**, 489–493, 1889.
137. M. Seya, A new mounting of concave grating suitable for a spectrometer, *Sci. Light (Tokyo)*, **2**, 8–17, 1952.
138. T. Namioka, Constitution of a Grating Spectrometer, *Sci. Light (Tokyo)*, **3**, 15–24, 1952.
139. P. Atherton, N. Reay, J. Ring, and T. Hicks, Tunable Fabry–Perot filters, *Opt. Eng.*, **20**, 806, 1981.
140. Available from numerous suppliers of vacuum equipment – for example, Ceram Tec (Cerameseal Division), Edwards, Ferrofluidics Corp., Perkin Elmer, Vacuum Generators (VG), Varian, and Veeco.
141. Available from Carpenter Technology, P.O. Box 14662, Reading, PA 19612–4662; Tel: (800) 338–4592 and (610) 208–2000, FAX: (610) 208–2361.
142. Available from Burleigh.
143. Available from Heraeus-Amersil, Dynasil, Esco, and Quartz Scientific, among others.
144. Available from Corning Glass Works, Optical Products Department.
145. Piezoelectric transducers are available from American Piezo Ceraamics, Burleigh, EDO, Polytec, Queensgate Instruments, and Xinetics, among others.
146. C. F. Bruce, On automatic parallelism control in a scanning Fabry–Perot interferometer, *Appl. Opt.*, **5**, 1447–1452, 1966.
147. Commercial Fabry–Perot Interferometers that can incorporate this feature are available from Burleigh.

148. P. Hariharan, *Optical Interferometry*, 2nd edn., Academic Press, 2003.
149. B. T. Meggitt and K. T. V. Grattan, *Optical Fiber Sensor Technology*, Springer, New York, 1999.
150. James C. Wyant, White light interferometry, paper presented at AeroSense, Orlando, Florida, 1–5 April 2002. (Available online at [http://www.optics.arizona.edu/jcwyant/pdf/Meeting\\_papers/WhiteLightInterferometry.pdf](http://www.optics.arizona.edu/jcwyant/pdf/Meeting_papers/WhiteLightInterferometry.pdf))
151. R. Ladenburg and D. Bershader, Interferometry, in *High Speed Aerodynamics and Jet Propulsion*, Vol. 9, *Physical Measurements in Gas Dynamics and Combustion*, R. Ladenburg (Ed.), Princeton University Press, Princeton, NJ, 1954.
152. P. R. Longaker and M. M. Litvak, Perturbation of the refractive index of absorbing media by a pulsed laser beam, *J. Appl. Phys.*, **40**, 4033–4041, 1969.
153. D. C. Smith, Thermal defocusing of CO<sub>2</sub> laser radiation in gases, *IEEE J. Quant. Electron.*, **QE-5**, 600–607, 1969.
154. H. Kuhn, New techniques in optical interferometry, *Rep. Prog. Phys.*, **14**, 64–94, 1951.

## General References

### Comprehensive (General) Optics Texts

- M. Born and E. Wolf, *Principles of Optics*, 7th edn., Cambridge University Press, Cambridge, 1999.
- E. Hecht, *Optics*, 3rd edn., Addison-Wesley, Reading, MA, 1998.
- R. W. Ditchburn, *Light*, 3rd edn., Academic Press, New York, 1976.
- F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th edn., McGraw-Hill, New York, 1976.
- M. V. Klein and T. E. Furtak, *Optics*, 2nd edn., John Wiley & Sons, Inc., New York, 1986.
- R. S. Longhurst, *Geometrical and Physical Optics*, 3rd edn., Longman, London, 1973.
- F. G. Smith and J. H. Thomson, *Optics*, 2nd edn., John Wiley & Sons, Inc., Chichester, 1988.
- J. Strong, *Concepts of Classical Optics*, Freeman, San Francisco, CA, 1958.

### Applied Optics

- W. J. Smith, *Modern Optical Engineering*, 3rd edn., SPIE Press, Bellingham, WA, 2000.

- R. Kingslake (Ed.), *Applied Optics and Optical Engineering* Academic, New York, Vol. 1, 1965; Vol. 2, 1965; Vol. 3, 1965; Vol. 4, 1967; Vol. 5, 1969.
- L. Levi, *Applied Optics*, John Wiley & Sons, Inc., New York, Vol. 1, 1968; Vol. 2, 1980.

### Lens Design

- W. J. Smith, *Modern Lens Design: A Resource Manual*, McGraw-Hill, New York, 1992.
- R. R. Shannon, *The Art and Science of Optical Design*, Cambridge University Press, Cambridge, 1997.
- M. Laikin, *Lens Design*, 3rd edn., Marcel Dekker, New York, 2001.

### Electro-Optic Devices

- M. A. Karim, *Electro-Optical Devices and Systems*, PWS-Kent Publishing, Boston, MA, 1990.
- I. P. Kaminow, *An Introduction to Electro-Optics*, Academic Press, New York, 1974.
- A. Yariv, *Optical Electronics in Modern Communications*, 5th edn., Oxford University Press, New York, 1997.

### Far Infrared Techniques

- M. F. Kimmitt, *Far-Infrared Techniques*, Routledge, 1970.
- A. Hadni, *Essentials of Modern Physics Applied to the Study of the Infrared*, Pergamon Press, Oxford, 1967.
- K. D. Möller and W. G. Rothschild, *Far-Infrared Spectroscopy*, Wiley-Interscience, New York, 1971.
- L. C. Robinson, *Physical Principles of Far-Infrared Radiation, Methods in Experimental Physics*, Vol. 10, L. Marton (Ed.), Academic Press, New York, 1973.

### Fiber Optics

- J. Hecht, *Understanding Fiber Optics*, 3rd edn., Prentice-Hall, Upper Saddle River, NJ, 1999.
- A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998. A. Ghatak, A. Sharma, and R. Tewari, *Understanding Fiber Optics on a PC*, Viva Books Private Ltd., New Delhi, 1994. P.K. Cheo, *Fiber Optics and Optoelectronics*, 2nd edn., Prentice-Hall, Englewood Cliffs, NJ, 1990. L. Kazovsky, S. Benedetto, and A. Willner, *Optical Fiber Communication Systems*, Artech House, Boston, MA, 1996.

- J. Gowar, *Optical Communication Systems*, Prentice-Hall, Englewood Cliffs, NJ, 2nd edn. 1993.
- Introduction to Integrated Optics*, M. K. Barnoski, (Ed.), Plenum, New York, 1974.
- D. Marcuse, *Principles of Optical Fiber Measurements*, Academic Press, San Diego, CA., 1981.
- T. Okoshi, *Optical Fibers*, Academic Press, New York, 1982.
- D. Gloge (Ed.), *Optical Fiber Technology*, IEEE, New York, 1976.
- J. C. Palais, *Fiber Optic Communications*, 5th edn., Prentice-Hall, Upper Saddle River, NJ, 2004.
- A. D. Snyder and J. D. Love, *Optical Waveguide Theory*, Chapman and Hall, London and New York, 1983.

## Filters

- David R. Lide (Ed.), *Handbook of Chemistry and Physics*, 87th edn., Taylor and Francis, Boca Raton, Florida, 2000. Also available on line at <http://www.hbcnpnetbase.com/help/default.asp>
- Handbook of Lasers*, R. J. Pressley (Ed.), CRC Press, Cleveland, OH, 1971.
- L. Levi, *Applied Optics*, Vol. 2, John Wiley & Sons, Inc., New York, 1980.
- H. A. Macleod, *Thin-Film Optical Filters*, American Elsevier, New York, 1969.
- E. J. Bowen, *Chemical Aspects of Light*, 2nd revised edn., The Clarendon Press, Oxford, 1946.

## Fourier Transform Spectroscopy

- P. Griffiths and J. deHaseth, *Fourier Transform Infrared Spectroscopy*, John Wiley & Sons, Inc., New York, 1986.
- B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*, CRC Press Inc., Boca Raton, IL, 1996.

## Incoherent Light Sources

- Advanced Optical Techniques*, A. C. S. Van Heel (Ed.), North-Bouand, Amsterdam, 1967.
- R. Kingslake (Ed.), *Applied Optics and Optical Engineering*, Vol. 1, Academic Press, New York, 1965.
- Handbook of Lasers*, R. J. Pressley (Ed.), CRC Press, West Palm Beach, FL, 1971.
- L. Levi, *Applied Optics*, Vol. 1, John Wiley & Sons, Inc., New York, 1968.

## Photometry

- C. DeCusatis, *Handbook of Applied Photometry*, AIP Press, New York, (1997).
- Lighting Handbook: Reference and Application*, 8th edn., M. Rea (Ed.), Illuminating Engineering Society of North America, New York, 1993.

## Infrared Technology

- William L. Wolfe and George J. Zissis (Eds.), *The Infrared Handbook*, Office of Naval Research, Washington, DC, 1985.
- A. R. Jha, *Infrared Technology*, John Wiley & Sons, Inc., New York, 2000.
- A. Hadni, *Essentials of Modern Physics Applied to the Study of the Infrared*, Pergamon Press, Oxford, 1967.
- R. D. Hudson, *Infrared System Engineering*, Wiley-Interscience, New York, 1969.
- P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, 1962.

## Interferometers and Interferometry

- P. Hariharan, *Optical Interferometry*, Academic Press, Orlando, FL, 1985.
- G. Hernandez, *Fabry-Perot Interferometers*, Cambridge University Press, Cambridge, 1986.
- R. Jones and C. Wykes, *Holographic and Speckle Interferometry*, 2nd edn., Cambridge University Press, Cambridge, 1989.
- J. C. Dainty (Ed.), *Laser Speckle and Related Phenomena, Topics in Applied Physics*, Vol. 9, Springer-Verlag, Berlin; New York, 1975.
- M. Francon, *Laser Speckle and Applications in Optics*, Academic Press, New York, 1979.
- M. Born and E. Wolf, *Principles of Optics*, 7th edn., Cambridge University Press, Cambridge, 1999.
- M. Francon, *Optical Interferometry*, Academic Press, New York, 1966.
- W. H. Steel, *Interferometry*, 2nd edn., Cambridge University Press, Cambridge, 1983.
- S. Tolansky, *An Introduction to Interferometry*, Longman, London, 1955.
- J. M. Vaughan, *The Fabry-Perot Interferometer: History, Theory, Practice and Applications*, Adam Hilger, Bristol, UK, 1989.



## Fourier Optics and Holography

- Joseph Goodman, *Introduction to Fourier Optics*, 2nd edn., McGraw-Hill, New York, 1996.
- P. Hariharan, *Optical Holography*, Cambridge University Press, Cambridge, 1984.

## Lasers

- A. Yariv, *Optical Electronics in Modern Communications*, 5th edn., Oxford University Press, New York, 1997.
- C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
- Clifford R. Pollack, *Fundamentals of Optoelectronics*, Irwin, Chicago, 1995.
- William T. Silfvast, *Laser Fundamentals*, 2nd edn., Cambridge University Press, 2004.
- B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 2nd edn., John Wiley & Sons, Inc., New York, 2007.
- M. Weber (Ed.), *Handbook of Laser Science and Technology*, Vol. 1, *Lasers and Masers*; Vol. II, *Gas Lasers*, CRC Press, Boca Raton, FL, 1982.
- A. Maitland and M. H. Dunn, *Laser Physics*, North Holland, Amsterdam, 1969.
- D. C. O'Shea, W. R. Callen, and W. T. Rhodes, *Introduction to Lasers and Their Applications*, Addison-Wesley, Reading, MA, 1977.
- J. T. Verdeyen, *Laser Electronics*, 3rd edn., Prentice-Hall, Englewood Cliffs, NJ, 1995.
- A. Yariv, *Quantum Electronics*, 3rd edn., John Wiley & Sons, Inc., New York, 1989.
- A. E. Siegman, *Lasers*, University Science Books, Mill Valley, CA, 1986.

## Nonlinear Optics

- Robert W. Boyd, *Nonlinear Optics*, Academic Press, Boston, MA, 1992.
- N. Bloembergen, *Nonlinear Optics*, Benjamin, New York, 1965.
- P. G. Harper and B. S. Wherrett (Eds.), *Nonlinear Optics*, Academic Press, New York, 1977.
- C. L. Tang and H. Rabin, (Eds.), *Quantum Electronics*, Vols. 1A and 2A *Nonlinear Optics*, Academic Press, New York, 1975.
- F. Zernike and J. E. Midwinter, *Applied Non-Linear Optics*, John Wiley & Sons, Inc., New York, 1973.
- Y. R. Shen, *The Principles of Nonlinear Optics*, John Wiley & Sons, Inc., New York, 1984.

## Optical Component and Instrument Design

- A. C. S. Van Heel (Ed.), *Advanced Optical Techniques*, North Holland, Amsterdam, 1967.
- R. Kingslake (Ed.), *Applied Optics and Optical Engineering*, Academic Press, New York, Vol. 3, 1965; Vol. 4, 1967; Vol. 5, 1969.

## Optical Detectors

- A. Rogalski, *Infrared Detectors*, Gordon and Breach, Amsterdam, 2000.
- J. Gowar, *Optical Communication Systems*, 2nd edn., Prentice-Hall, Englewood Cliffs, NJ, 1993.
- E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, John Wiley & Sons, Inc., New York, 1996.
- E. L. Dereniak and Devon G. Crowe, *Optical Radiation Detectors*, John Wiley & Sons, Inc., New York, 1984.
- J. Graeme, *Photodiode Amplifiers*, McGraw-Hill, New York, 1996.
- J. B. Dance, *Photoelectronic Devices*, Iliffe, London, 1969.
- P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, 1962.
- L. Levi, *Applied Optics*, Vol. 2, John Wiley & Sons, Inc., New York, 1980.
- R. S. Keyes (Ed.), *Optical and Infrared Detectors*, Topics in Applied Physics, Vol. 19, Springer, Berlin, 1977.

## Optical Materials

- R. Kingslake (Ed.), *Applied Optics and Optical Engineering*, Vol. 1, Academic Press, New York, 1965.
- M. Weber (Ed.), *Handbook of Laser Science and Technology*, Vols. III–V, *Optical Materials*, CRC Press, Boca Raton, FL, 1986–87.
- P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, 1962.
- A. J. Moses, *Optical Material Properties*, IFI/Plenum, New York, 1971.

## Optical Safety

- M. J. Weber (Ed.), *Handbook of Laser Science and Technology*, Vol. 1, *Lasers and Masers*, CRC Press, Boca Raton, FL, 1982.

D. Sliney and M. Wolbarsht, *Safety with Lasers and Other Optical Sources: A Comprehensive Handbook*, Plenum, New York, 1980.

## Polarized Light and Crystal Optics

- R. M. A. Azzam and N. M. Bashara, *Ellipsometry and Polarized Light*, North-Bouand, Amsterdam, 1977.  
 Dennis Goldstein, *Polarized Light*, CRC Press, Boca Raton, FL, 2003.  
 W. A. Shurcliff, *Polarized Light*, Harvard University Press, Cambridge, MA, 1962.  
 E. Wahlstrom, *Optical Crystallography*, 5th edn., John Wiley & Sons, Inc., New York, 1979.  
 A. Yariv and P. Yeh, *Optical Waves in Crystals*, John Wiley & Sons, Inc., New York, 1983.

## Spectrometers

- N. B. Colthup, L. H. Daly, and S. E. Wiberley, *Introduction to Infrared and Raman Spectroscopy*, 2nd edn., Academic Press, New York, 1975.  
 Francis M. Mirabella Jr., (Ed.), *Internal Reflection Spectroscopy. Theory and Applications*, CRC Press, Boca Raton, 1993.  
 J. F. James and R. S. Sternberg, *The Design of Optical Spectrometers*, Chapman & Hall, London, 1969.  
 H. S. Strobel and W. R. Heineman, *Chemical Instrumentation*, 3rd edn., John Wiley & Sons, Inc., New York, 1989.

## Spectroscopy

- M. Pinta (Ed.), *Atomic Absorption Spectrometry*, John Wiley & Sons, Inc., New York, 1975.  
 J. R. Edisbury, *Practical Hints on Absorption Spectrometry*, Hilger and Watts, London, 1966.  
 R. J. Reynolds and K. Aldous, *Atomic Absorption Spectroscopy*, Barnes and Noble, New York, 1970.  
 A. Lee Smith, *Applied Infrared Spectroscopy: Fundamentals, Techniques and Analytical Problem-Solving*, Vol. 54 (Chemical Analysis, Vol. 21), John Wiley & Sons, Inc., New York, 1979.  
 R. A. Sawyer, *Experimental Spectroscopy*, Prentice-Hall, Englewood Cliffs, NJ, 1951.  
 D. Williams (Ed.), *Spectroscopy, Methods of Experimental Physics*, Vol. 13, Parts A and B, Academic Press, New York, 1968.

S. Walker and H. Straw, *Spectroscopy*, Vol. I, *Microwave and Radio Frequency Spectroscopy*; Vol. II, *Ultraviolet, Visible, Infrared and Raman Spectroscopy*, Macmillan, New York, 1962.

## Submillimeter Wave Techniques

- Kenneth J. Button, *Infrared and Millimeter Waves: Submillimeter Techniques*, Academic Press, New York, 1980.  
 G. W. Chantry, *Submillimeter Spectroscopy*, Academic Press, New York, 1971.  
 E. Kollberg (Ed.), *Instrumentation for submillimeter spectroscopy*, SPIE Proceedings, Vol. 598, SPIE, Bellingham, WA, 1986.  
 D. H. Martin (Ed.), *Spectroscopic Techniques*, North-Holland, Amsterdam, 1967.

## Tables of Physical and Chemical Constants

G. W. C. Kaye and T. H. Laby, *Tables of Physical and Chemical Constants*, 16th edn., Longman, London, 1995. Now available free online at <http://www.kayelaby.npl.co.uk/>

## Tables of Spectral and Laser Lines

- M. J. Weber, *Handbook of Lasers*, CRC Press, Boca Raton, FL, 2001.  
 M. J. Weber, *Handbook of Laser Wavelengths*, CRC Press, Boca Raton, FL, 1999.  
 G. R. Harrison, *MIT Wavelength Tables*, MIT Press, Cambridge, MA, 1969.  
 A. R. Striganov and N. S. Sventitskii, *Tables of Spectral Lines of Neutral and Ionized Atoms*, IFI/Plenum, New York, 1968.  
 A. N. Zaidel', V. K. Prokof'ev, S. M. Raiskii, V. A. Slavnyi, and E. Ya. Shreider, *Tables of Spectral Lines*, IFI/Plenum, New York, 1970.

## Ultraviolet and Vacuum-Ultraviolet Technology

A. E. S. Green (Ed.), *The Middle Ultraviolet: Its Science and Technology*, John Wiley & Sons, Inc., New York, 1966.

J. A. R. Samson, *Techniques of Vacuum Ultra-violet Spectroscopy*, John Wiley & Sons, Inc., New York, 1967.

## Suppliers of Optical Windows

Several of the suppliers listed will supply lenses, prisms, and other components fabricated from these materials:

AMTIR (GeAsSe Glass): Harrick Scientific, Janos, REFLEX Analytical

Arsenic trisulphide: Infrared Optical Products, REFLEX Analytical, Spectrum Thin Films Corp.

Arsenic triselenide: Infrared Optical Products, REFLEX Analytical, Spectrum Thin Films Corp.

Barium fluoride: Crystran, Del Mar Photonics, Harrick Scientific, Infrared Optical Products, ISP Optics, Janos, Koch Crystal Finishing, Molecular Technology.

Cadmium sulfide: Cleveland Crystals, Molecular Technology.

Cadmium selenide: Cleveland Crystals, Molecular Technology.

Cadmium telluride (Irtran 6): Cleveland Crystals, ISP Optics, Janos, Laser Research Optics, Molecular Technology.

Calcium carbonate (calcite): Crystran, Karl Lambrecht Corp (KLC), Photox, Thin Film Lab.

Calcium fluoride (Irtran 3): Argus International, Coherent, Inc., Crystran, Edmund Optics, EKSPLA, Gooch and Housego, Hellma International, Infrared Optical Products, ISP Optics, Janos, KLC, Koch Crystal Finishing, Lambda Research Optics, Meller Optics, Molecular Technology, Newport, Optimax, Opto-Sigma, Photox, REFLEX Analytical, Rocky Mountain Instruments, Spectrum Thin Films, Thorlabs.

Cesium bromide: Argus International, Crystran, Harrick Scientific, Janos.

Cesium iodide: Crystran, Koch Crystal Finishing, Molecular Technology, REFLEX Analytical.

Diamond: Argus International, Gist Optics, Coherent Photonics Group, II-VI Infrared, ISP Optics, Laser Power Optics, REFLEX Analytical, Optics for Research, Newport/Oriel.

Gallium arsenide: Argus International, Crystran, Infrared Optical Products, II-VI Infrared, ISP Optics, Lambda Research Optics, Laser Power Optics, Laser Research Optics, Meller Optics, REFLEX Analytical, Rocky Mountain Instruments, Sterling Precision Optics.

Germanium: Argus International, Coherent, Inc., Crystran, Edmund Optics, Gooch and Housego, II-VI Infrared, ISP Optics, Janos, Laser Power Optics, Laser Research Optics, Meller Optics, Photox, REFLEX Analytical, Rocky Mountain

Instruments, Spectrogon, Spectrum Thin Films Corporation, Sterling Precision Optics, Unicore.

Glasses: Coherent, Ealing, Edmund Optics, Newport, Opto-Sigma, Rocky Mountain Instruments, Rolyn, Schott, Sterling Precision Optics, Thorlabs.

Lithium fluoride: Argus International, Crystran, Coherent, Inc., EKSPLA, Hellma International, Infrared Optical Products, ISP Optics, Lambda Research Optics, Macrooptica, Molecular Technology, OPCO, Photox, REFLEX Analytical, Rocky Mountain Instrument Co., Sterling Precision Optics.

Magnesium fluoride (Irtran 1): Argus International, Coherent, Inc., Crystran, Edmund Optics, EKSPLA, Gooch and Housego, Hellma International, Infrared Optical Products, ISP Optics, KLC, Lambda Research Optics, Macrooptica, Meller Optics, Molecular Technology, Newport, Optimax, Photox, REFLEX Analytical, Rocky Mountain Instruments, Sterling Precision Optics.

Magnesium oxide (Irtran 5): Crystran, Harrick Scientific.

Potassium bromide: Argus International, Crystran, EKSPLA, Hilger Crystals, Infrared Optical products, ISP Optics, Janos, Koch Crystal Finishing, Lambda Research Optics, Macrooptica, Molecular Technology, OPCO, Photox, REFLEX Analytical Corp., Spectrum Thin Films Corp., Thorlabs.

Potassium chloride: Crystran, ISP Optics, Janos, Koch Crystal Finishing, Macrooptica, Molecular Technology, Optovac, Photox.

Potassium iodide: Crystran.

Quartz (crystalline): Crystran, Esco, Infrared Optical Products, ISP Optics, Janos, Meller Optics, Molecular Technology, Newport, Opto-Sigma, Photox, REFLEX Analytical, Sterling Precision Optics, Continental Optical, Optics for Research, Newport/Oriel, Adolf Meller.

Sapphire: Crystran, Infrared Optical Products, ISP Optics, Laser Power Optics, Meller Optics, Newport, Opto-Sigma, REFLEX Analytical, Rocky Mountain Instruments, Rolyn, Sterling Precision Optics.

Silica (fused): Esco, ISP Optics, Janos, Macrooptica, Meller Optics, Newport, Opto-Sigma, Photox, REFLEX Analytical, Rolyn, Schott, Sterling Precision Optics, Thorlabs, Continental Optical, Optical Coating Lab (OCLI).

Silicon: Crystran, Infrared Optical Products, ISP Optics, Janos, Laser Power Optics, Macrooptica, Meller Optics, Molecular Technology, Photox, REFLEX Analytical, Rocky Mountain Instruments, Sterling Precision Optics.

Silver bromide: Crystran, Harrick Scientific, ISP Optics, REFLEX Analytical.

Silver chloride: Crystran, Harrick Scientific, ISP Optics, REFLEX Analytical.

Sodium chloride: Crystran, ISP Optics, Koch Crystal Finishing,

Macrooptica, Molecular Technology, Photox.

Sodium fluoride: Crystran.

Strontium fluoride: Crystran, Harrick Scientific.

Strontium titanate: Commercial Crystal Labs, Harrick Scientific,

Hibshman-Pacific Optical, Photox.

Tellurium: Molecular Technology.

Thallium bromide: Crystran, Korth Kristalle GmbH.

Thallium bromiodide (KRS-5): Argus International, Crystran,

Infrared Optical Products, ISP Optics, Janos, Koch Crystal

Finishing, Macrooptica, Molecular Technology, Perkin-Elmer.

Thallium chlorobromide (KRS-6): Crystran, Macrooptica,

Molecular Technology.

Titanium dioxide (rutile): Commercial Crystal Labs, Crystran, Harrick Scientific, ISP Optics, Molecular Technology, Thin Film Lab.

Zinc selenide (Irtran 4): Argus International, Coherent, Inc., Crystran, CVI Laser, Edmund Optics, EKSPLA, Gooch and Housego, Harrick Scientific, Hellma International, II-VI Infrared, Infrared Optical Products, ISP Optics, Janos, Laser Power Optics, Laser Research Optics, Meller Optics, Molecular Technology, REFLEX Analytical, Rocky Mountain Instruments, Sterling Precision Optics.

Zinc sulphide (Irtran 2): Crystran, Gooch and Housego, Infrared Optical Products, Laser Power Optics, Laser Research Optics, Meller, OPCO, Optimax, REFLEX Analytical, Rocky Mountain Instruments, Sterling Precision Optics,

Zirconium dioxide: Insaco.

## CHARGED-PARTICLE OPTICS

The science of charged-particle optics owes a great deal to the principles of design formulated by C. E. Kuyatt (1930–1998). The authors are especially indebted to Dr. Kuyatt for his advice and for permission to present parts of the contents of his “Electron Optics Lectures” in the beginning of this chapter.

Fifty years ago devices employing charged-particle beams were confined to the purview of a small group of physicists studying elementary processes. Today chemists, biologists, and engineers employ beams of ions or electrons to probe various materials and to investigate discrete processes. Physicists are constructing beam machines to control the momentum of interacting particles with energies from a few tenths of an electron volt to trillions of electron volts. Chemists routinely use mass spectrometers as analytical tools and various electron spectrometers to probe molecular structures. The electron microscope is one of the primary tools of the modern biologist. Furthermore charged-particle beam technology has spread to industry, where electron-beam machines are used for cleaning surfaces and welding, and ion-beam devices are used in the preparation of semiconductors.

The properties of charged-particle beams are analogous in many respects to those of photon beams: hence the appellation *charged-particle optics*. In the following sections the laws of geometrical optics will be covered insofar as they apply to charged-particle beams. The consequences of the coulombic interaction of charged particles will be considered. In addition we shall discuss the design of electron and ion sources, as well as the design of electrodes that constitute optical elements for manipulating beams of

charged particles. We shall consider primarily electrostatic focusing by elements of cylindrical symmetry and restrict discussion to particles of sufficiently low kinetic energies that relativistic effects can be ignored.

A number of excellent books and review articles are available to the reader who wishes to pursue the topic of electron and ion optics in greater detail.<sup>1–6</sup>

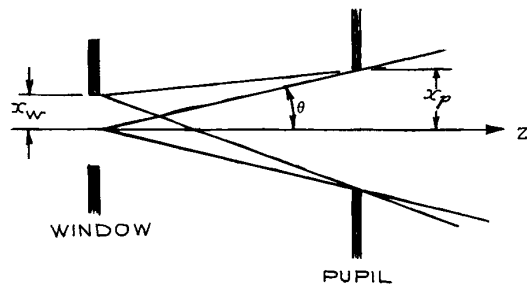
### 5.1 BASIC CONCEPTS OF CHARGED-PARTICLE OPTICS

#### 5.1.1 Brightness

The *brightness*,  $\beta$ , of a point on a luminous object is determined by the differential of current  $dI$  that passes through an increment of area  $dA$  about the point and shines into a solid angle  $d\Omega$ :

$$\beta = \frac{dI}{dA d\Omega} (\mu\text{A}/\text{cm}^2/\text{sr}^1). \quad (5.1)$$

In electron or ion optics, a luminous object is usually defined by an aperture, called a *window*, which is uniformly illuminated from one side by a stream of charged particles. A second aperture called a *pupil* limits the angular spread of particles emanating from the window. This situation, in the  $xz$  plane, is illustrated in Figure 5.1. It will be assumed that the system is cylindrically symmetric about the  $z$ -axis and the distance between the window and pupil is sufficiently great that the angular spread of rays emanating from each point on the object is the same.



**Figure 5.1** A pupil establishes the half angle  $\theta$  of the bundle of rays emanating from each point on an object defined by a window.

Then the integrated brightness of the object outlined by the window is

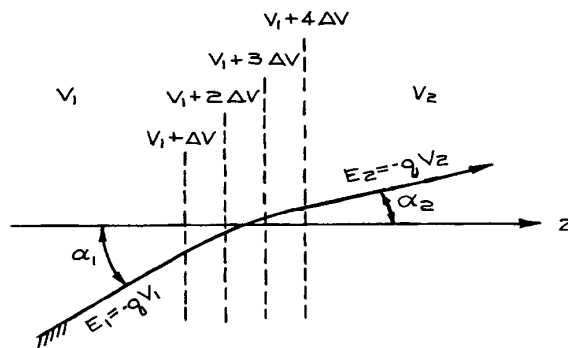
$$\beta = \frac{I}{\pi^2 x_w^2 \theta^2} \quad (5.2)$$

where  $x_w$  is the radius of the window and  $\theta$  is the half angle defined by the pupil relative to a point at the window.

### 5.1.2 Snell's Law

When a beam of charged particles enters an electric field, the particles will be accelerated or decelerated; the trajectory will depend on the angle of incidence upon the equipotential surfaces of the field. This effect is analogous to the situation in optics when a light ray passes through a medium in which there is a change of refractive index. Figure 5.2 illustrates the behavior of a charged-particle beam as it passes from a region of uniform potential  $V_1$  to a region of uniform potential  $V_2$ . The initial and final kinetic energies of a particle of charge  $q$  that originates at *ground potential* with no kinetic energy are  $E_1 = -qV_1$  and  $E_2 = -qV_2$ , respectively (i.e., for  $V$  in volts and  $E$  in eV, the charge  $q$  on an electron is  $-1$ , and the charge of a singly charged positive ion is  $+1$ ).  $\alpha_1$  and  $\alpha_2$  are the angles of incidence and refraction with respect to the normals to the equipotential surfaces that separate the field-free regions.

The quantity in charged-particle optics that corresponds to the index of refraction is the particle velocity, propor-



**Figure 5.2** A charged-particle trajectory exhibits "refraction" at a potential gradient. (The energies,  $E_i$  specified in the figure assume the particle originated at ground potential with no kinetic energy.)

tional to the square root of the particle energy. Thus, the charged-particle analog of Snell's law is

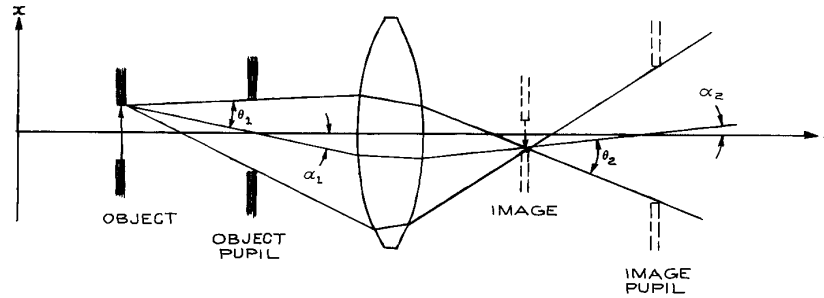
$$\sqrt{E_1} \sin \alpha_1 = \sqrt{E_2} \sin \alpha_2. \quad (5.3)$$

Clearly, this property can be exploited in charged-particle optics, as in light optics, to make lenses by shaping the equipotential surfaces. This is equivalent to varying the refractive index and shape of lenses for light.

### 5.1.3 The Helmholtz-Lagrange Law

Consider the situation illustrated in Figure 5.3, in which an object defined by a window at  $z_1$  is imaged at  $z_2$ . The rays emanating from a point  $(x_1, z_1)$  at the edge of the object are limited by a pupil to fall within a cone defined by a half angle  $\theta_1$ ; this is the *pencil angle*. The angle of incidence  $\alpha_1$  of the central ray from  $(x_1, z_1)$  on the plane of the pupil is referred to as the *beam angle*. The rays emanating from each point on the image appear to be limited by an aperture that corresponds to the image of the object pupil. An image pencil angle  $\theta_2$  and beam angle  $\alpha_2$  from a point  $(x_2, z_2)$  at the edge of the image are defined with respect to this image pupil.

For most practical electron-optical systems, the pencil and beam angles are small and the approximation  $\sin \theta = \theta$  is valid. The treatment of optics based on this approximation is called *Gaussian* or *paraxial* optics.



**Figure 5.3** Relation of image pencil angle  $\theta_2$  and beam angle  $\alpha_2$  to object pencil angle  $\theta_1$  and beam angle  $\alpha_1$ .

The image pencil angle is determined by the object pencil angle. This relation is given to first order by the Helmholtz–Lagrange law:

$$x_1 \theta_1 \sqrt{E_1} = x_2 \theta_2 \sqrt{E_2} \quad (5.4)$$

$$\sqrt{\frac{E_1}{E_2}} = Mm; \quad M = \frac{x_2}{x_1}, \quad m = \frac{\theta_2}{\theta_1}$$

where  $M$  and  $m$  are the linear and angular magnifications, respectively.

The current  $I_1$  through the object window, that falls within the object pupil is the same as the current  $I_2$  through the image window and pupil. Setting  $I_1 = I_2 = I$ , it follows that:

$$\frac{I}{E_1 \theta_1^2 x_1^2} = \frac{I}{E_2 \theta_2^2 x_2^2} \quad (5.5)$$

Furthermore, if  $\beta_1$  and  $\beta_2$  are the brightnesses of the object and image, respectively, then this equality can be written:

$$\frac{\beta_1}{E_1} = \frac{\beta_2}{E_2} \quad (5.6)$$

demonstrating that *the ratio of brightness to energy is conserved from object to image.*

### 5.1.4 Vignetting

The foregoing discussion assumes that the object radiates uniformly; to the extent that the pencil angles are the same for all points on the object, it follows that the image is uniformly illuminated. However, if, as illustrated in Figure 5.4, a second pupil is added to the system, the illumination

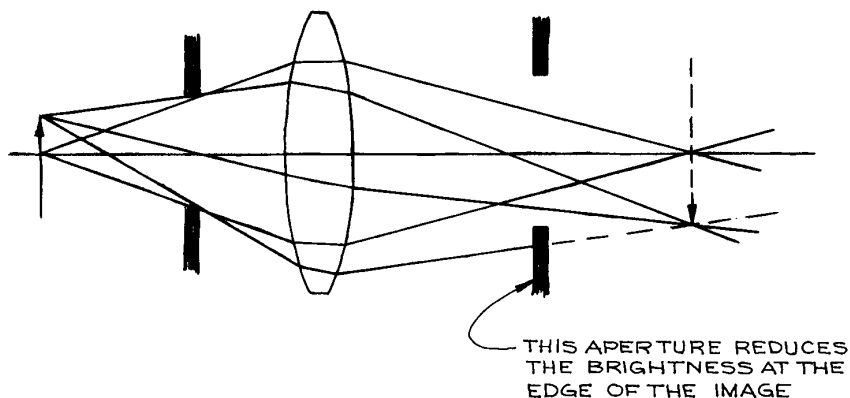
will vary across the image. Such a situation is called *vignetting*. The one case in which an aperture, in addition to the object window and pupil, does not produce vignetting is when an aperture is placed at the location of the image of the original window or pupil. Such an aperture is sometimes employed to skim off stray current resulting from scattering from slit edges and from aberrations. Also, an aperture is placed at the exit window of an energy or momentum analyzer to define its resolution. Except for “spatter apertures” or resolving apertures, an electro-optical system should have only two apertures.

## 5.2 ELECTROSTATIC LENSES

It is a simple matter to produce axially symmetric electrodes that, when electrically biased, will produce equipotential surfaces with shapes similar to those of optical lenses. A charged particle passing across these surfaces will be accelerated or decelerated, and its path will be curved so as to produce a focusing effect. The chief difference between such a charged-particle lens and an optical lens is that the quantity analogous to the refractive index, namely the particle velocity, varies continuously across an electrostatic lens, whereas a discontinuous change of refractive index occurs at the surfaces of an optical lens. Charged-particle lenses are “thick” lenses, meaning that their axial dimensions are comparable to their focal lengths.

### 5.2.1 Geometrical Optics of Thick Lenses

In the case of a thick lens, it is not correct to measure focal distances from a plane perpendicular to the axis through



**Figure 5.4** Vignetting.

the center of the lens. The focal points of a thick lens are located by *focal lengths*,  $f_1$  and  $f_2$ , measured from *principal planes*,  $H_1$  and  $H_2$ , respectively. As shown in Figure 5.5, the principal planes of a charged-particle lens are always crossed, and they are both located on the low-voltage side of the midplane,  $M$ , of the physical lens.  $F_1$  and  $F_2$  locate the focal points with respect to the central plane and hence the distances from the central plane to the principal planes are  $F_1 - f_1$  and  $F_2 - f_2$ , respectively. The object and image distances with respect to the principal planes are  $p$  and  $q$ , and, with respect to the central plane, are  $P$  and  $Q$ , respectively. All distances are positive in the direction indicated by the arrows in Figure 5.5.

As in light optics, it is possible to graphically construct the image produced by a lens if the cardinal points of the lens are known. This procedure is illustrated in Figure 5.6. It is only necessary to trace two principal rays. From a point on the object draw a ray through the first focal point and thence to the first principal plane. From the point of intersection with this plane draw a ray parallel to the axis. Draw a second ray through the object point and parallel to the axis. From the point of intersection of this ray with the second principal plane, draw a ray through the second focal point. The intersection of the first and second principal rays gives the location of the image point corresponding to the object point.

A number of important relationships can be derived geometrically from Figure 5.6. The linear and angular magnifications,  $M$  and  $m$ , respectively, are given by:

$$M = \frac{f_2 - q}{f_2} = \frac{f_1}{f_1 - p} \quad (5.7)$$

and

$$m = \frac{f_1 - p}{f_2} = \frac{f_1}{f_2 - q} \quad (5.8)$$

where negative magnification implies an inverted image.

The object and image distances from the central plane are

$$P = F_1 - \frac{f_1}{M} \quad (5.9)$$

and

$$Q = F_2 - Mf_2 \quad (5.10)$$

The object and image distances from the principal planes are related by *Newton's law*:

$$(p - f_1)(q - f_2) = (P - F_1)(Q - F_2) = f_1 f_2 \quad (5.11)$$

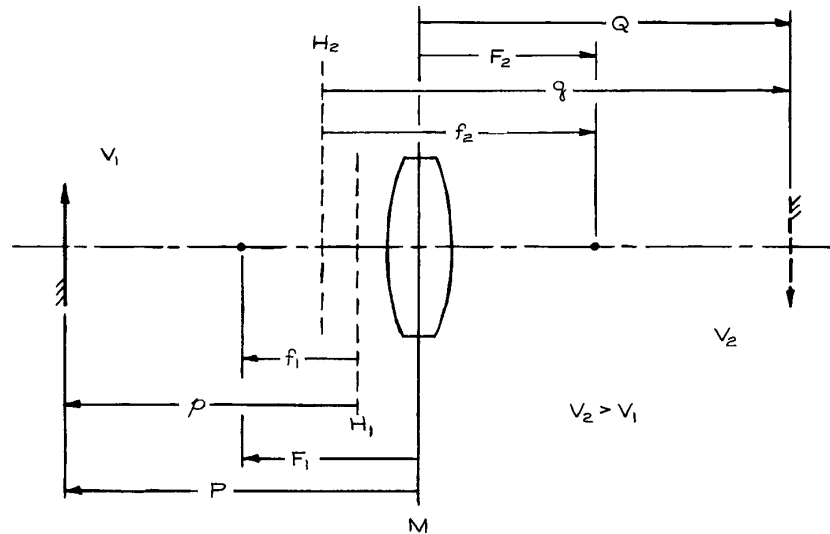
and *Newton's formula*:

$$\frac{f_1}{p} + \frac{f_2}{q} = 1 \quad (5.12)$$

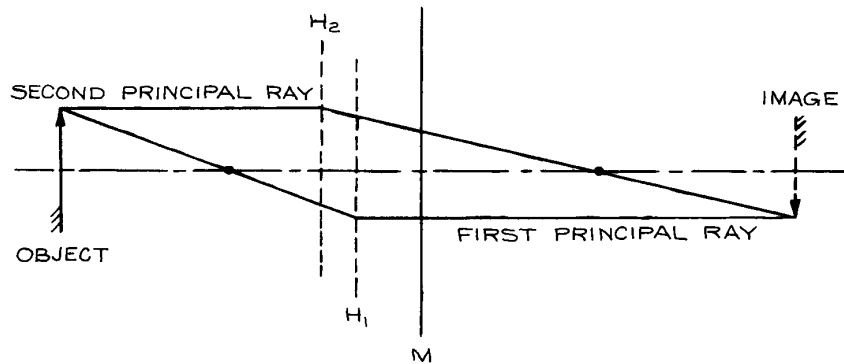
For lenses that are not very strong, the principal planes are close to the central plane. Then:

$$\begin{aligned} p &\rightarrow P \\ q &\rightarrow Q \end{aligned} \quad (5.13)$$





**Figure 5.5** Lens parameters.  $H_1$  and  $H_2$  are the principal planes.  $M$  is the midplane of the physical lens.



**Figure 5.6** Graphical construction of an image.

and

$$f_1 \approx f_2 \approx f$$

$$M \approx -0.8 \frac{Q}{P} \quad (5.15)$$

to give an approximate form of Newton's formula that is quite useful in the initial stages of design work:

$$\frac{1}{P} + \frac{1}{Q} = \frac{1}{f}. \quad (5.14)$$

Furthermore, Spangenberg<sup>1</sup> has observed that for the lenses discussed in succeeding sections:

## 5.2.2 Cylinder Lenses

The most widely used lens for focusing charged particles with energies of a few eV to several keV is that produced by two cylindrical coaxial electrodes biased at voltages corresponding to the desired initial and final particle energies. Figure 5.7 illustrates the equipotential surfaces and

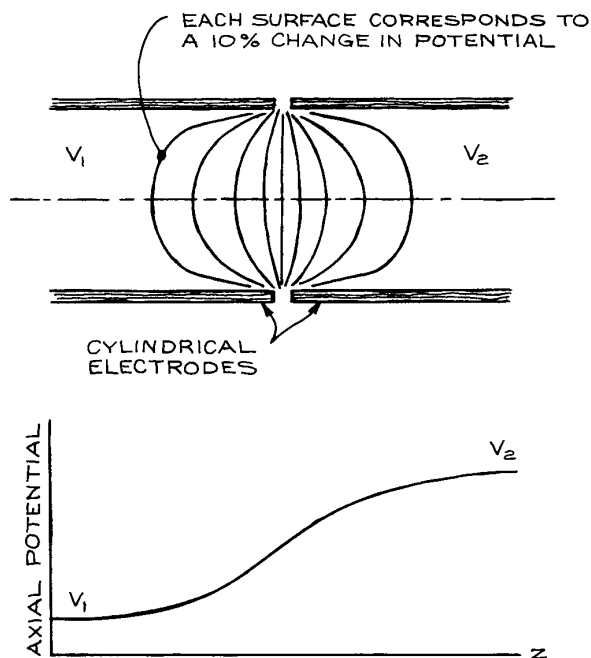


Figure 5.7 Potential distribution in a two-cylinder lens.

the axial electrical potential associated with the field produced by two coaxial cylinders biased at voltage  $V_1$  and  $V_2$ , respectively.

The focal properties of a *two-cylinder lens* depend upon the diameters of the cylinders, the spacing between them, and the ratio  $E_2/E_1$  of the final to initial kinetic energies of the transmitted particles. The lens properties scale with the diameter, so all dimensions are taken in units of the diameter  $D$  of the larger cylinder. If the gap between cylinders is large, the focal properties become sensitive to the cylinder wall thickness and external fields are liable to penetrate to the lens. Most lenses are constructed with a gap  $g = 0.1D$ . The particle energies depend upon the electrical potentials (bias voltages) applied to the elements. It is assumed that the bias voltage supplies are referenced to the particle source, so that  $V_2/V_1 = E_2/E_1$ .

Focal properties are usually presented as graphs (Figure 5.8) or tables of  $F_1, f_1, F_2$ , and  $f_2$  as a function of  $V_2/V_1$ . For simple lenses it is also convenient to plot  $P$  versus  $Q$  for different  $V_2/V_1$  ratios, as in Figure 5.9. The chief sources of information on the focal properties of electrostatic lenses

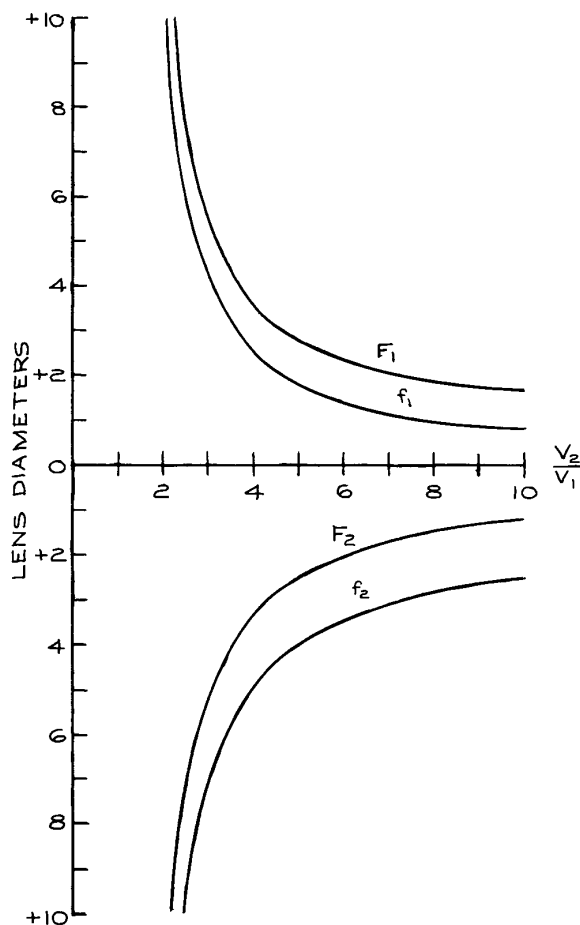
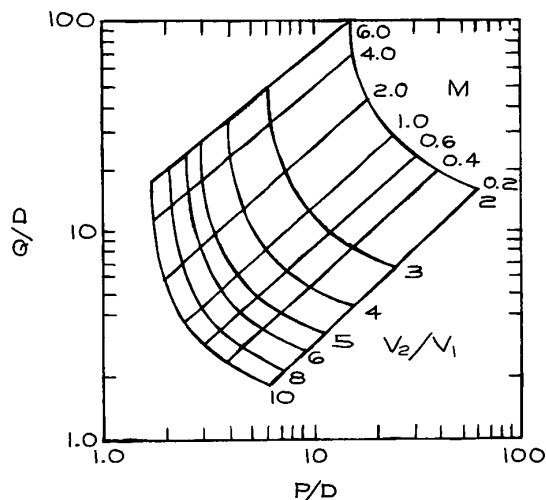


Figure 5.8 Focal properties of a two-cylinder lens with coaxial cylinders of the same diameter  $D$  and a gap between of  $0.1D$ . (From E. Hartung and F. H. Read, *Electrostatic Lenses*, Elsevier, New York, 1976; by permission of Elsevier Publishing Company.)

are the book *Electrostatic Lenses* by Harting and Read<sup>7</sup> and various journal articles by Read and his coworkers.<sup>8</sup> These data are determined from numerical solutions of the equation of motion of an electron in a lens field and are thought to be accurate to 1–3%. Unfortunately, these sources contain no information applicable to systems with virtual objects or images. For two-cylinder lenses, Natali, DiChio, Uva, and Kuyatt<sup>9</sup> have computed  $P$ - $Q$  curves that extend to negative values of the object and image distances.

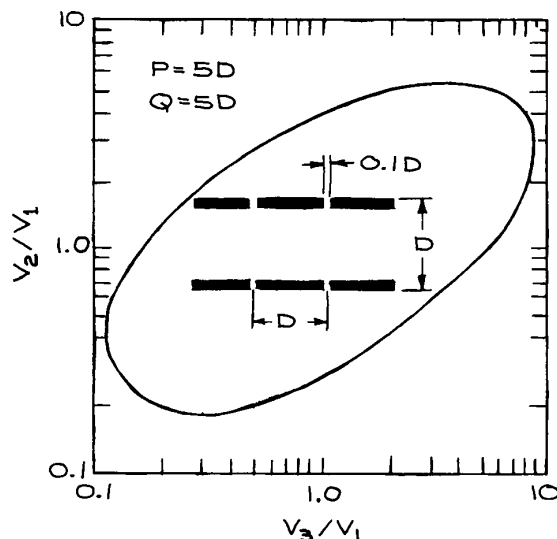


**Figure 5.9**  $P$ - $Q$  curves for a two-cylinder lens with equal diameter cylinders and a gap of  $0.1D$ . (From E. Hartung and F. H. Read, *Electrostatic Lenses*, Elsevier, New York, 1976; by permission of Elsevier Publishing Company.)

To obtain desired focal properties with a given acceleration ratio, two or more two-cylinder lenses can be used in series. The lens field extends about one diameter  $D$  on either side of the midplane gap. When two gaps are sufficiently close that their lens fields overlap, it is convenient to treat the combination as a single lens. Such a *three-cylinder lens* has the advantage that it can produce an image of a fixed object at a fixed image plane for a range of final-to-initial energy ratios. Alternatively, it can produce a variable image location with a fixed energy ratio.

A three-cylinder lens for which the initial and final energies differ is called an *asymmetric lens*. It can be used with an electron or ion source to produce a variable-energy beam. Such lenses are also used to focus an image of the exit slit of a monochromator on a target while allowing for a variation of the particle acceleration between monochromator and target.

The focal properties of three-cylinder lenses are not conveniently presented graphically, since each focal length is a function of two independent variables. These variables are usually taken to be  $V_2/V_1$  and  $V_3/V_1$ , where  $V_1$ ,  $V_2$ , and  $V_3$  are the voltages on the first, second, and third lens elements. Harting and Read<sup>7</sup> and Heddle<sup>10</sup> present data



**Figure 5.10** Typical "zoom-lens" curve. (From A. Adams and F. H. Read, "Electrostatic Cylinder Lenses III," *J. Phys.*, **E5**, 156; copyright 1972 by the Institute of Physics.)

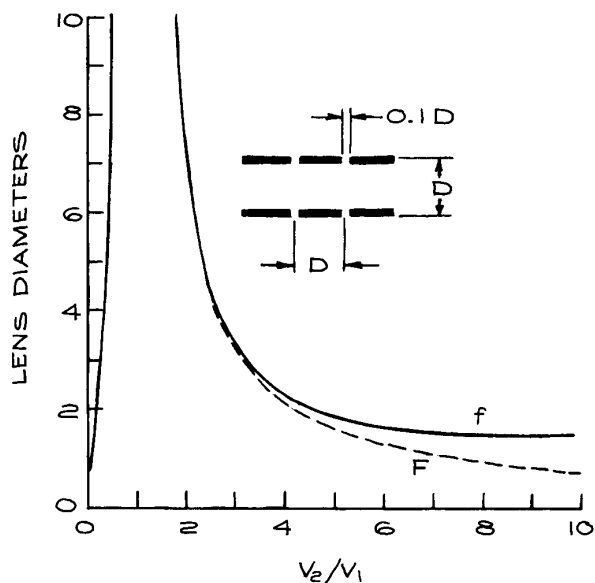
on three-element lenses for  $V_3/V_1$  values between 1.0 and 30.0 in the form of tables of lens parameters ( $f_1, f_2, F_1, F_2$ ) as a function of  $V_2/V_1$ . They also plot "zoom-lens curves" as graphs of  $V_2/V_1$  versus  $V_3/V_1$  for selected values of  $P$  and  $Q$ . An example is given in Figure 5.10.

Read has defined a length  $f$  that approximately satisfies the relation

$$\frac{1}{f} = \frac{1}{P} + \frac{1}{Q} \quad (5.16)$$

for three-element lenses. For use in initial design work,  $V_2/V_1$  is graphed as a function of  $f/D$  for various values of  $V_3/V_1$ .<sup>3</sup>

A three-element lens for which  $V_3/V_1 = 1$  is called an *einzel lens* or *unipotential lens*. Such a lens produces focusing without an overall change in the energy of the transmitted particle. The focal properties are symmetrical:  $f_1 = f_2 = f$  and  $F_1 = F_2 = F$ . Figure 5.11 illustrates the focal properties of a typical einzel lens. For any desired object and image distance there are two focusing conditions: a decelerating mode with  $V_2/V_1 < 1$  and an accelerating mode with  $V_2/V_1 > 1$ . The accelerating mode is



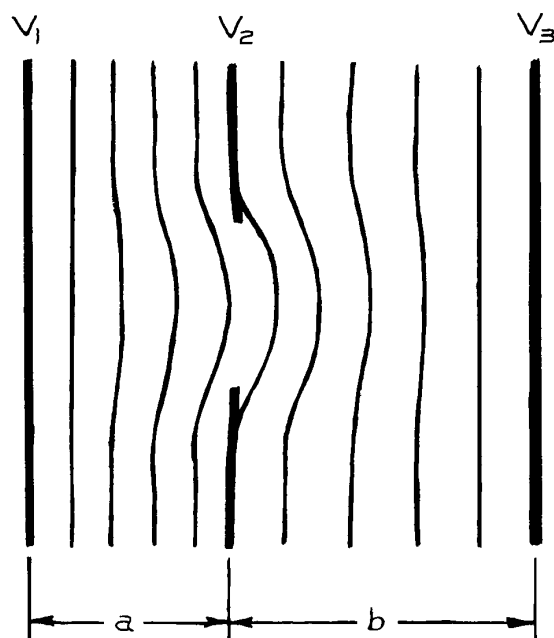
**Figure 5.11** Focal properties of an einzel lens with the intermediate cylinder of length  $A = D$  and lens gaps  $g = 0.1D$ . (From E. Hartung and F. H. Read, *Electrostatic Lenses*, Elsevier, New York, 1976; by permission of Elsevier Publishing Company.)

preferred, because deceleration in the middle element causes expansion of the transmitted beam, resulting in aberrations and unwanted interactions with the lens surfaces.

### 5.2.3 Aperture Lenses

An aperture in a plane electrode separating two uniform-field regions will have a focusing effect if the field on one side is different from that on the other. As illustrated in Figure 5.12, this lens results from a bulge in the equipotential surfaces caused by field penetration through the aperture. A lens of this type is called a *Calbick lens*. Single-aperture focusing is important in electron gun and ion-source design because these devices usually have an aperture at their output.

The Calbick lens is the analog of a thin lens. The electric fields are established by the potentials  $V_1$ ,  $V_2$ , and  $V_3$  applied to the three planar elements, as shown in Figure 5.12. In this case it is assumed that electrons or ions



**Figure 5.12** Protrusion of equipotential surfaces through an aperture separating regions of different field strength.

originate at potential  $V_1$  with essentially no kinetic energy. For a circular aperture:

$$f_1 = f_2 = F_1 = F_2 = \frac{4V_2}{E_B - E_A} \quad (5.17)$$

where the electrical field intensities in the uniform-field regions are:

$$E_A = \frac{V_2 - V_1}{a}, \quad E_B = \frac{V_3 - V_2}{b} \quad (5.18)$$

The ratio of aperture potential to radius must be larger than  $E_A$  or  $E_B$ . For a long slot in the  $xy$  plane:

$$F_x = \frac{2V_2}{E_B - E_A} \quad (5.19)$$

and

$$f_y \approx \infty \quad (5.20)$$

when the short dimension of the slot is in the  $x$ -direction. The focal properties of a slot are important in the design of electron guns that employ a ribbon filament.<sup>11</sup>

Lenses can also be constructed using two or more apertures. Read has computed the properties of two-aperture and three-aperture lenses.<sup>7,8</sup>

In practice, concentric cylindrical mounts support circular-aperture electrodes of an aperture lens. The focusing effects of these cylindrical elements can be ignored if the cylinder diameter is at least three times greater than the aperture diameter.

### 5.2.4 Matrix Methods

A useful description of the charged-particle trajectories through a focusing system can be formulated using the matrix methods introduced in Section 4.2.2. A particle trajectory through a system that is cylindrically symmetrical about the  $z$ -axis is determined by  $(r, dr/dz = r', z)$ , where  $r$  is the radial distance of the trajectory from the axis at  $z$ . A trajectory through a plane perpendicular to the central axis at  $z$  is represented by a vector:

$$\begin{pmatrix} r \\ r' \end{pmatrix} \quad (5.21)$$

The effect of a displacement in a field-free region from  $z_1$  to  $z_2$  is given by:

$$\begin{pmatrix} r_2 \\ r_2' \end{pmatrix} = \mathbf{M}(z_1 \rightarrow z_2) \begin{pmatrix} r_1 \\ r_1' \end{pmatrix} \quad (5.22)$$

where the *ray transfer matrix* is:

$$\mathbf{M}(z_1 \rightarrow z_2) = \begin{pmatrix} 1 & \Delta z \\ 0 & 1 \end{pmatrix}, \quad \Delta z = z_2 - z_1 \quad (5.23)$$

A lens can be represented as if its effect were confined to the region between the principal planes. The matrix operating between the principal planes is:

$$\mathbf{M}(H_1 \rightarrow H_2) = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_2} & \frac{f_1}{f_2} \end{pmatrix} \quad (5.24)$$

Consider, for example, a ray originating at a point on an object at  $z_1$  and passing through a point on an image at  $z_2$ . The trajectory at  $z_2$  can be determined from the trajectory at  $z_1$  by:

$$\begin{pmatrix} r_2 \\ r_2' \end{pmatrix} = \mathbf{M}(H_2 \rightarrow z_2) \mathbf{M}(H_1 \rightarrow H_2) \mathbf{M}(z_1 \rightarrow H_1) \begin{pmatrix} r_1 \\ r_1' \end{pmatrix} \quad (5.25)$$

In terms of the parameters defined in Figure 5.5, the displacement between the object and the first principal plane is  $P - F_1 + f_1$ ; thus:

$$\mathbf{M}(z_1 \rightarrow H_1) = \begin{pmatrix} 1 & P - F_1 + f_1 \\ 0 & 1 \end{pmatrix} \quad (5.26)$$

and similarly:

$$\mathbf{M}(H_2 \rightarrow z_2) = \begin{pmatrix} 1 & Q - F_2 + f_2 \\ 0 & 1 \end{pmatrix} \quad (5.27)$$

Substitution yields:

$$\begin{pmatrix} r_2 \\ r_2' \end{pmatrix} = \begin{pmatrix} \frac{F_2 - Q}{f_2} & \frac{f_1 f_2 - (Q - F_2)(P - F_1)}{f_2} \\ -\frac{1}{f_2} & \frac{F_1 - P}{f_2} \end{pmatrix} \begin{pmatrix} r_1 \\ r_1' \end{pmatrix} \quad (5.28)$$

This transformation matrix is generally applicable to the problem of tracing a trajectory through the field of a lens. For the two-cylinder lens, DiChio, *et al.* have derived simple analytical expressions for the elements of the matrix.<sup>12</sup> When  $r_1$  and  $r_2$  refer to points in the object and image planes, respectively, the diagonal elements of the transformation matrix are the linear and angular magnification. Furthermore, the focusing condition given by Newton's formula requires the numerator of the upper right element to be zero. Thus, the transformation matrix between object and image is:

$$\mathbf{M}(\text{object} \rightarrow \text{image}) = \begin{pmatrix} M & 0 \\ -\frac{1}{f_2} & m \end{pmatrix} \quad (5.29)$$

The matrix formulation is a desirable alternative to graphical ray tracing for a system of lenses that are so close to one another that an image does not appear between each lens and the next. Consider, for example, the system of two lenses in Figure 5.13. In this case the transformation matrix is:

$$\mathbf{M}(z_1 \rightarrow z_2) = \mathbf{M}(H_2' \rightarrow z_2) \mathbf{M}(H_1' \rightarrow H_2') \times \mathbf{M}(H_2 \rightarrow H_1') \mathbf{M}(H_1 \rightarrow H_2) \mathbf{M}(z_1 \rightarrow H_1) \quad (5.30)$$

In general, the image distance  $L_3$  will be unknown initially. It can be determined, however, from the focusing condition

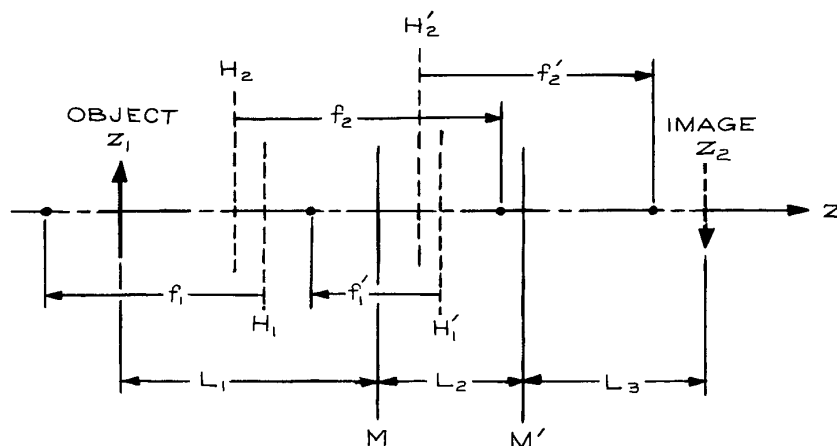


Figure 5.13 A compound lens system.

that requires the upper right element of  $\mathbf{M}(z_1 \rightarrow z_2)$  to be zero.

It is occasionally necessary to determine a particle trajectory in a region of uniform field. The transformation matrix from  $z_1$  to  $z_2$  in a uniform field in the  $z$ -direction is:

$$\mathbf{M}(\text{uniform field}) = \begin{pmatrix} 1 & \frac{2\Delta z V_1}{V_2 - V_1} \left( \sqrt{\frac{V_2}{V_1}} - 1 \right) \\ 0 & \sqrt{\frac{V_1}{V_2}} \end{pmatrix} \quad (5.31)$$

where  $V_1$  and  $V_2$  are the potentials of the planar equipotential surfaces perpendicular to the  $z$ -axis at  $z_1$  and  $z_2$ .

## 5.2.5 Aberrations

There are three types of aberrations that must be considered in designing charged-particle optics:

- (1) *Geometrical aberrations* due to deviations from the paraxial assumption that the angle  $\theta$  between a ray and the central axis is sufficiently small that  $\theta \approx \sin \theta \approx \tan \theta$ .
- (2) *Chromatic aberrations* caused by variations in the kinetic energies of transmitted particles.
- (3) *Space charge* due to repulsive interactions between the charged particles.

The *geometrical aberration* associated with a lens is defined in terms of the radius  $\Delta r_2$  of a spot in the image

plane formed by rays within the pencil angle  $\theta_1$  that emanate from an object point on the axis, as shown in Figure 5.14. The important geometrical errors can be expressed in terms of the coefficients of a power series in  $\theta_1$ :

$$\Delta r_2 = b\theta_1^2 + c\theta_1^3 \dots \quad (5.32)$$

An imaging system is said to be *second-order-focusing* if  $b = 0$ , and *third-order-focusing* if  $c = 0$ . An axially symmetric system is always at least second-order-focusing, since only terms of odd order appear in the expansion of  $\Delta r_2$ . The coefficient  $c$  is referred to as the *third-order-aberration coefficient*. Higher-order coefficients can usually be ignored. For an object point on a lens axis, the error represented by  $c$  is spherical aberration. Harting and Read<sup>7</sup> have computed *spherical-aberration coefficients*  $C_s$  defined by:

$$c = MC_s \quad (5.33)$$

for all of the lenses for which they give focal properties. For object points off axis, other types of aberration must be considered (coma, field curvature, distortion, and astigmatism). Most off-axis aberrations can be related to the magnitude of  $C_s$ . When the object is small compared to the lens diameter, the spherical-aberration coefficient gives a reasonable measure of the total geometrical aberration.

As can be seen from Figure 5.14, the pencil of rays emanating from an object point achieves a minimum

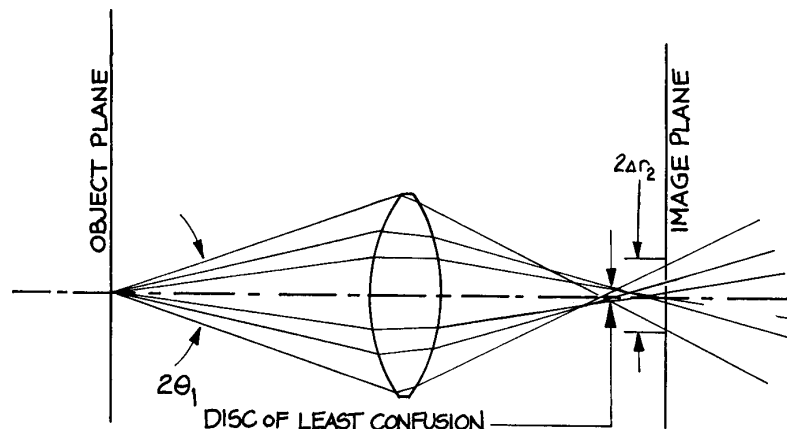


Figure 5.14 Geometrical aberration of the image of a point.

diameter at a point slightly in front of the image plane. This minimum diameter is the *circle of least confusion*. To achieve the sharpest focus it is often possible to vary the lens voltages so as to place the circle of least confusion on the desired image plane. The radius of this circle is  $\frac{1}{4}\Delta r_2$ ; thus  $C_s$  can, in practice, be reduced by a factor of four.

The diameter of the bundle of rays emanating from an object is usually determined by a pupil that limits the portion of the lens that is illuminated by rays from the object. The extent of illumination is specified by a *filling factor*, which is the ratio of the diameter of the bundle of rays near the lens gap to the lens diameter. Because the spherical aberration depends upon the third power of the angle between the limiting ray and the axis, it is obvious that aberration increases rapidly with increasing filling factor. In practice, a lens system should not be designed with a filling factor in excess of about 50%.

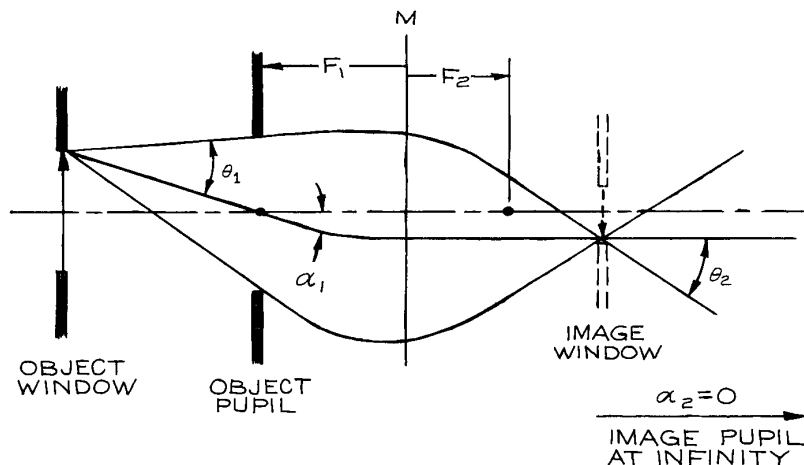
For an object of finite extent, the worst aberration occurs for off-axis image points. The magnitude of the aberration depends upon the maximum value of the angle of a ray from the edge of the object relative to the central axis of the lens system. From Figure 5.3 it can be seen that this angle is the sum of the object pencil angle and beam angle. In many lens systems, the image produced by one lens serves as an object for succeeding lenses. According to the Helmholtz–Lagrange law, the pencil angle associated with this intermediate image is determined by the pencil angle associated with the initial object. There are, how-

ever, no such physical restrictions on the beam angle from the intermediate image. In fact, as shown in Figure 5.15, if an object pupil is placed at the first focal point of a lens, then the corresponding image pupil is at infinity and the image beam angle is zero. Obviously such an arrangement reduces aberrations produced by lenses that treat this image as an object.

*Chromatic aberration* depends upon the relative spread in particle energies,  $\Delta E/E$ , passing through a lens. Variations in particle energies can occur because of conditions in the particle source or within the lens. An energy spread of 0.2 to 0.4 eV is characteristic of electron sources. Some ion sources yield ions in an energy bandwidth as great as 30 eV. Energy variations may also be caused by fluctuations in the voltage from the power supplies that establish both source and lens-element potentials.

For a given lens, the extent of chromatic aberration can best be determined from a calculation of the image distances for particles with energies at the extremes of the anticipated energy distribution. Consider, for example, the case of a source at ground potential that produces particles with energies between 0 and  $\Delta E$ . For a lens of voltage ratio  $V_2/V_1$ , a distance  $P$  from the source, determine the corresponding image position,  $Q = f(P, V_2/V_1)$ . Determine also:

$$Q + \Delta Q = f\left(P, \frac{V_2 - \Delta E/q}{V_1 - \Delta E/q}\right) \quad (5.34)$$



**Figure 5.15** An object pupil at the first focal point moves the image pupil to infinity and reduces the image beam angle  $\alpha_2$  to zero.

Then a point on axis at the source yields an image at  $Q$  of radius:

$$\Delta r_2 = \theta_2 \Delta Q \quad (5.35)$$

or, by the Helmholtz–Lagrange law:

$$\Delta r_2 = \frac{\theta_1 \Delta Q}{M} \sqrt{\frac{V_2}{V_1}} \quad (5.36)$$

where  $\theta_1$  and  $\theta_2$  are the pencil angles at the source and the image, respectively.

In electron-beam devices, the main cause of chromatic aberration is the spread of electron kinetic energies emitted from a thermionic cathode as well as the potential drop across the emitting portion of the cathode. The energy of electrons at the maximum of the current distribution from a cathode is:

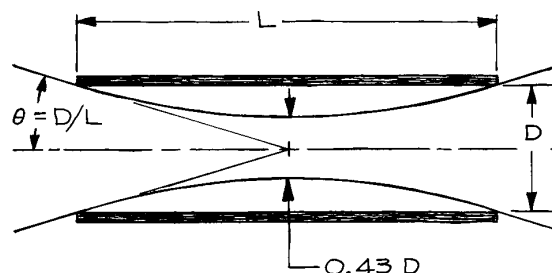
$$E_{\max} = 8.6 \times 10^{-5} T \text{ eV} \quad (5.37)$$

The width of the distribution  $\Delta E \approx E_{\max}$ .

For a bare tungsten filament, a temperature of about 3000 K is required to produce significant emission. At this temperature  $\Delta E \approx 0.25$  eV. An oxide-coated cathode (available from Electron Technology) can be operated at temperatures as low as 1200 K, where  $\Delta E \approx 0.1$  eV. The potential drop across the emitting portion of the cathode may add several tenths of an eV to the energy spread.

The *space-charge* effect arises from the mutual repulsion of particles of like charge. The effect increases with current density. As a result, even a focused beam will give a diffuse image. Furthermore, the beam will diverge away from the image more rapidly than predicted by geometrical optics.

Space charge places a limit on the current in a charged-particle beam. As an example of practical value, consider the problem of transmitting maximum current through a cylindrical element of length  $L$  and diameter  $D$ . As shown in Figure 5.16, the maximum current is achieved by focusing the beam on a point at the center of the tube with a pencil angle  $\theta = D/L$ . It can be shown that



**Figure 5.16** A beam focused to give maximum current through a tube.



the space-charged-limited current of electrons of energy  $E = -qV$  is

$$I_{\max}(\text{electrons}) = 38.5V^{3/2} \left(\frac{D}{L}\right)^2 \mu\text{A} \quad (5.38)$$

and the maximum ion current is:

$$I_{\max}(\text{ions}) = 0.90 \left(\frac{M}{n}\right)^{1/2} V^{3/2} \left(\frac{D}{L}\right)^2 \mu\text{A} \quad (5.39)$$

where  $V$  is in volts,  $M$  is in amu, and  $n$  is the charge state.<sup>2</sup> The minimum beam diameter at the space-charge limit is  $0.43D$ .

Space charge also limits the current available from a surface that emits electrons or ions. The typical source geometry is the plane diode consisting of a planar anode at potential  $V$  a distance  $d$  from the cathode. Charged particles produced by thermionic emission from the cathode are accelerated toward the anode. The maximum electron current density at the anode is:

$$J_{\max}(\text{electrons}) = 2.34 \frac{V^{3/2}}{d^2} \mu\text{A}/\text{cm}^2 \quad (5.40)$$

while for an ion-emitting cathode:

$$J_{\max}(\text{ions}) = 0.054 \left(\frac{n}{M}\right)^{1/2} \frac{|V|^{3/2}}{d^2} \mu\text{A}/\text{cm}^2 \quad (5.41)$$

with  $V$  in volts,  $d$  in cm,  $M$  in amu, and  $n$  the charge state.

## 5.2.6 Lens Design Example

Consider the problem of producing the image of the anode aperture of an electron source on the entrance plane of an

energy analyzer, as illustrated in Figure 5.17. The anode potential is  $V_1 = 100$  V, and the analyzer entrance-plane potential is  $V_2 = 10$  V. The anode aperture radius is  $r_1 = 1.0$  mm, and the distance from anode to analyzer is  $l = 10$  cm. It is desired that the image of the anode aperture serve as the entrance aperture for the analyzer with radius  $r_2 = 0.5$  mm. So as not to overfill the analyzer, one must also have the image pencil angle  $\theta_2 = 0.14$  and the image beam angle  $\alpha_2 = 0$ .

The lens to be designed is a decelerating lens with  $V_2/V_1 = 0.1$  and magnification  $M = 0.5$ . Customarily only the focal properties of accelerating lenses are tabulated. The focal properties for the desired lens are just the reverse of those of the accelerating lens, with  $V_2/V_1 = 10$ . From Harting and Read,<sup>7</sup> the focal properties of the  $V_2/V_1 = 10$  lens (identified by a prime) are:

$$\begin{aligned} f_1' &= 0.80D & F_1' &= 1.62D \\ f_2' &= 2.54D & F_2' &= 1.19D \end{aligned} \quad (5.42)$$

The corresponding parameters for the  $V_2/V_1 = 0.1$  lens are:

$$\begin{aligned} f_1 &= 2.54D & F_1 &= 1.19D \\ f_2 &= 0.80D & F_2 &= 1.62D \end{aligned} \quad (5.43)$$

From the  $P$ - $Q$  curve of Figure 5.9 for  $V_2/V_1 = 10$  and  $M' = 1/M = 2$ , we estimate:

$$\begin{aligned} P &= Q' = 6.20D \\ Q &= P' = 2.05D \end{aligned} \quad (5.44)$$

Taking  $P = 6.20D$ , Newton's law gives:

$$Q = \frac{f_1 f_2}{P - F_1} + F_2 = 2.03D \quad (5.45)$$

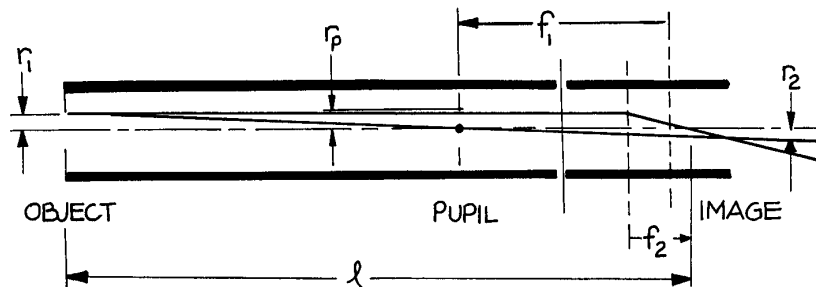


Figure 5.17 A simple lens design.

The overall length of the lens is:

$$l = P + Q = 6.20D + 2.03D = 10 \text{ cm} \quad (5.46)$$

and thus:

$$D = 1.22 \text{ cm}. \quad (5.47)$$

In order for the image beam angle to be zero, the object pupil should be placed at the first focal point of the lens so that the image pupil is at infinity. The object pupil must define a pencil angle  $\theta_1$  that is consistent with the required image pencil angle. By the Helmholtz–Lagrange law:

$$\theta_1 = M\theta_2 \sqrt{\frac{V_2}{V_1}} = 0.022 \quad (5.48)$$

This angle is determined by the radius  $r_p$  of an aperture at the first focal point:

$$r_p = \theta_1(P - F_1) = 0.13 \text{ cm} \quad (5.49)$$

To estimate the geometrical aberration of this lens, we calculate the spherical aberration for the lens used in reverse. This is necessary because spherical-aberration coefficients are available only for lenses with  $V_2/V_1 > 1$ . For  $V_2/V_1 = 10$ , Harting and Read give:

$$\frac{M' C_s}{D} = 20 \quad (5.50)$$

Thus for the lens used in reverse, the image of a point is a spot of diameter

$$\Delta r' = M' C_s \theta_2^3 = 0.067 \text{ cm} \quad (5.51)$$

For the lens used as designed, the size of the image of an object point is:

$$\Delta r = M \Delta r' = 0.033 \text{ cm} \quad (5.52)$$

Thus spherical aberration will enlarge the image by 67% ( $\Delta r/r_2 = 0.033/0.05$ ); however, if the lens voltage is adjusted slightly to bring the circle of least confusion onto the image plane, the spherical aberration can be reduced to 17%.

For a typical electron source,  $E = 0.3 \text{ eV}$ . This energy spread will result in chromatic aberration at the image. To estimate this effect, calculate the image position  $Q + \Delta Q$  for

electrons at the extreme of the energy distribution where the deceleration ratio of the lens is:

$$(V_2 - \Delta E/q)/(V_1 - \Delta E/q) = 10.3/100.3 = 0.103 \quad (5.53)$$

As before, the focal properties can be determined from those of the corresponding accelerating lens. For the lens with  $V_2/V_1 = 0.103$ ,

$$\begin{aligned} f_1 &= 2.58D & F_1 &= 1.22D \\ f_2 &= 0.82D & F_2 &= 1.65D \end{aligned} \quad (5.54)$$

The image position by Newton's law in this case is:

$$Q + \Delta Q = \frac{f_1 f_2}{P - F_1} + F_2 = 2.07D \quad (5.55)$$

so:

$$\Delta Q = 0.04D = 0.05 \text{ cm} \quad (5.56)$$

This displacement of the image plane will cause the image of a point object to appear as a disc of radius:

$$\Delta r = \Delta Q \theta_2 = 0.007 \text{ cm} \quad (5.57)$$

in the original image plane. Chromatic aberration enlarges the image by about 14% ( $\Delta r/r_2 = 0.007/0.05$ ).

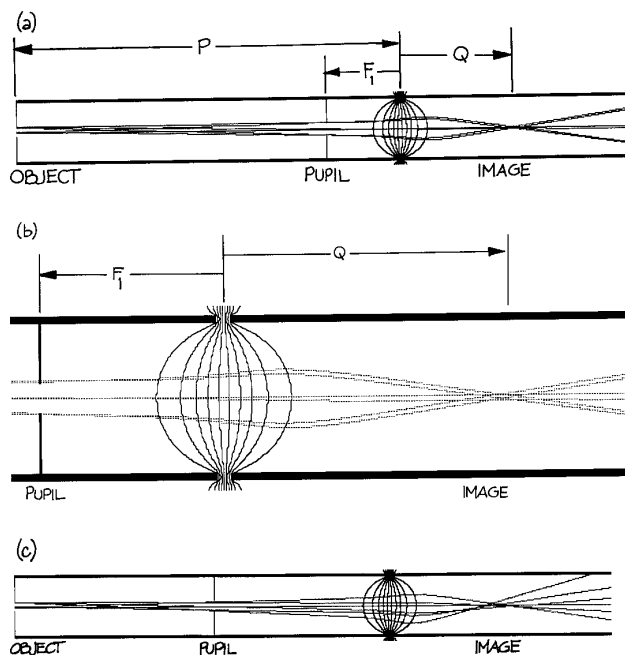
The overall aberration of the lens as designed is about 30%. The filling factor is:

$$\frac{2(\theta_1 + \alpha_1)P}{D} = 54\% \quad (5.58)$$

From the aberration calculations it can be seen that the aberrations would become serious if the filling factor were to exceed this value.

## 5.2.7 Computer Simulations

There are a variety of programs available for the computation of charged-particle trajectories in the field of an array of electrodes. These employ Green's function, finite-element, and surface-charge methods to compute the electrical field produced by the array. All provide for the possibility of an applied magnetic field and some solve for the field produced by an array of magnets. A particular virtue in computer-simulation programs is that



**Figure 5.18** (a) MacSimion simulation of the cylindrical lens shown in Figure 5.17 and discussed in Section 5.2.6. The voltage ratio  $V_2/V_1 = 0.10$ ; the object is defined by an aperture window and the object distance from the lens gap is  $P = 6.20D$  ( $D$  is the diameter); a pupil is placed at the first focal point of the lens. The electric field at the lens gap is shown by equipotential surfaces computed by the program. The trajectories of 100 eV electrons originating at the center and edge of the window have been traced by the computer program. (b) A blowup showing the image formed. The image distance from the lens gap is  $Q = 1.93D$ . (c) The same lens with the pupil moved closer to the object and away from the first focal point of the lens. The result is a significant increase in the beam angle at the image. (MacSimion, is a simulation program for MacIntosh developed by Don McGilvery and coworkers, Department of Chemistry, Monash University, Clayton, Victoria, AUSTRALIA 3168)

they can be used to predict the properties of compound lenses that are more complex than the two- and three-element lenses that are described above, and for which there are tabulated values of focal lengths and focusing properties. Among the most useful and widely used simulation programs are Simion (Scientific Instrument Services, Inc., [www.simion.com](http://www.simion.com)) and CPO (CPO, Ltd., [www.electronoptics.com](http://www.electronoptics.com)).

It must be emphasized that computer simulation programs cannot substitute for an understanding of the basic principles of electron optics described in this chapter. It is impossible to obtain the desired performance in the design of an electron optical system by arbitrarily choosing an array of electrodes and “fiddling” to obtain the desired beam properties; the positions of lenses, windows, and, especially, pupils should be determined from considerations of basic principles before turning to the computer to optimize a design or test its performance.

Figure 5.18(a) shows a computer simulation of the simple cylindrical lens discussed in Section 5.2.6 and illustrated in Figure 5.17. The computed electric field at the lens gap is shown as equipotential surfaces. Electron trajectories are traced from points at the center and the upper edge of the object window. One of the three trajectories from each point passes through the center of the pupil; the other two are the extreme rays that can pass through the pupil. Figure 5.18(b), a blowup of a portion of 5.18(a), shows the location of the image formed. It is worth noting that the size of the image is half that of the object, as predicted in Section 5.2.6. The image distance,  $Q = 1.93D$ , is somewhat short of that predicted. The bundle of rays from both the center and the edge of the object emerge from the image with their central axes parallel to the axis of the lens; this implies that the beam angle  $\alpha_2 = 0$  (see Figure 5.15). Figure 5.18(c) shows a lens system identical to that in 5.18(a), except the pupil has been moved closer to the object and away from the first focal point of the lens. The result is a significant increase in the image beam angle, as well as noticeable geometric aberration.

### 5.3 CHARGED-PARTICLE SOURCES

An electron or ion gun consists of a source of charged particles, such as a hot metal filament or a plasma, and an electrode structure that gathers particles from the source and accelerates them in a particular direction to form a beam. There is a wide range of practical electron and ion guns. Some simple devices and their principles of operation will be described in this section. In addition to simple devices that can be conveniently constructed for laboratory use, there are a number of very sophisticated guns that have been

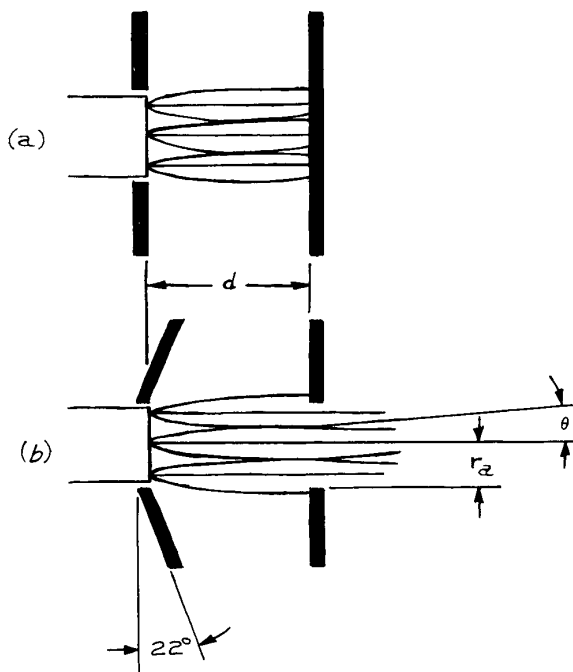
developed commercially. These include electron guns for cathode-ray tubes and electron-beam welders, and ion guns for sputter-cleaning apparatus and for ion-implantation doping of semiconductors. Manufacturers of electron guns and ion sources suitable for use in the laboratory include Kimble Physics (electron and ion guns), Spectra-Mat (alkali-metal ion sources), and Southwest Vacuum Devices and Cliftronic (both manufacturing CRT guns useful as electron sources and both owned by Video Display Corporation).

### 5.3.1 Electron Guns

Electrons are produced by thermionic emission, field emission, photoelectric emission, and electron-impact ionization. Thermionic sources are the most common.<sup>13</sup> These sources typically consist of a wire filament of some refractory metal, such as tungsten or tantalum, which is heated by an electrical current passing through it. In some cases thermionic sources consist of a metal cup or button that is indirectly heated by an electrical heater or by electron bombardment of the back surface. The electrons from a hot filament possess energies of a few tenths of an electron volt. Mutual repulsion will cause the electrons to diffuse away from the source if they are not immediately accelerated into a beam. Many accelerator structures are used. Most use either diode (two-electrode) or triode (three-electrode) geometries, although some TV tubes employ a pentode geometry. The Pierce diode geometry is most common in laboratory devices.

Figure 5.19(a) illustrates a simple diode electron gun consisting of a plane emissive surface and a parallel plane anode. Electrons leave the cathode with a nominal energy  $E_k$ . The anode is biased at a positive potential  $V_a$  relative to the cathode, so that electrons from a spot on the cathode will appear at a spot on the anode with energies of approximately  $E_a = -qV_a$ . To admit the accelerated electrons to the system beyond, a hole is made in the anode. If the cathode and anode were infinite in extent, the space-charge-limited current density given in Section 5.2.5 could be achieved at the anode, and the electron beam emerging from the anode hole would be characterized by a beam angle:

$$\alpha = \frac{r_a}{3d} \quad (5.59)$$



**Figure 5.19** (a) The plane diode; (b) the Pierce geometry that simulates the field of the plane diode.

where  $r_a$  is the radius of the anode hole and  $d$  the cathode-to-anode spacing. Since the most divergent electron arriving at the anode would be one emitted parallel to the cathode with energy  $E_k$ , it can be seen that the pencil angle characterizing the beam from the anode aperture would be

$$\theta = \sqrt{\frac{E_k}{E_a}} \quad (5.60)$$

For a cathode of finite extent, the space-charge interaction causes the beam to spread laterally within the gap between cathode and anode. Pierce<sup>2</sup> has shown that the electric field in an infinite space-charge-limited diode can be reproduced in the region of a finite cathode by means of the conical cathode structure illustrated in Figure 5.19(b). The maximum current of electrons is:

$$I_{\max} = \pi r_a^2 J_{\max} = 7.35 \left( \frac{r_a}{d} \right)^2 V_a^{3/2} \mu\text{A} \quad (5.61)$$

The corresponding ratio of brightness to energy (which is conserved) is:

$$\begin{aligned} \frac{\beta}{E_a} &= \frac{I_{\max}}{E_a \pi^2 r_a^2 \theta^2} \\ &= 0.74 \frac{E_a^{3/2}}{E_k d^2} \mu\text{A}/\text{cm}^2/\text{sr}^1/\text{eV}^1 \end{aligned} \quad (5.62)$$

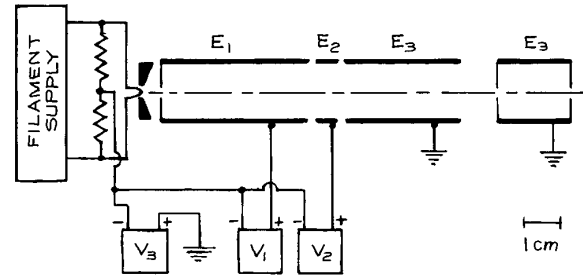
The design of multielectrode guns is a complicated process involving time-consuming experimentation. A number of special guns have been developed for such purposes as producing low-energy electron beams or very narrow beams. Klemperer and Barnett describe many of these,<sup>3</sup> and most can be constructed easily if they are really needed. In general, however, it is best to use the simplest device that will fulfill one's requirements.

The guns developed for TV tubes are the product of much industrial development. These guns will produce a beam of electrons of several hundred eV to several tens of thousands eV that is sharply focused to a spot at a distance of 10 to 20 cm. Cathode-ray-tube guns can be obtained quite inexpensively from commercial suppliers such as Southwest Vacuum Devices and Cliftronic (both owned by Video Display Corporation). Specialized guns of similar design are available from Kimble Physics and other manufacturers of electron spectrometers.

### 5.3.2 Electron-Gun Design Example

Consider the problem of designing an electron gun consisting of a Pierce diode and a lens system that can inject a beam of 20–200 eV electrons into a gas cell, as illustrated in Figure 5.20. It is desired that the beam current approach the space-charge limit at 20 eV. The overall length of the gun is to be about 10 cm, and the cell is to be located 4 cm in front of the gun. The cell is 2 cm long, with an input aperture of radius  $r_2 = 0.1$  cm and an exit aperture of radius 0.15 cm. Treat the exit aperture as though its radius  $r_4$  were 0.1 cm, so as to make the beam tight enough to minimize backscattering of electrons from the edges of this aperture.

Note that the electron gun system in Figure 5.20 is wired in such a way that the energies of the electrons through the system,  $E_1$ ,  $E_2$ , and  $E_3$ , in electronvolts, are numerically equal to the power supply voltages  $V_1$ ,  $V_2$ , and  $V_3$ . This is, in general, a great convenience, since electron energies can



**Figure 5.20** An electron gun system to inject a variable-energy beam into a gas cell.

be read directly from the power supplies. In the following discussion,  $E_i$  and  $V_i$  can be used interchangeably.

The maximum current through the cell can be taken as the space-charge-limited current through a cylinder with a diameter  $D$  equal to that of the gas-cell apertures and length  $L$  equal to the length of the gas cell. From Section 5.2.5, the maximum current of electrons with energy  $E_3 = 20$  eV is, in this case

$$\begin{aligned} I_{\max} &= 38.5 V_3^{3/2} \left( \frac{D}{L} \right)^2 \\ &= 34.4 \mu\text{A} \end{aligned} \quad (5.63)$$

To approach this limit, focus an image of the anode aperture on the center of the cell. Make the image radius  $r_3 = 0.43r_2$ , the minimum size of the space-charge-limited beam (note Section 5.2.5 and Figure 5.16).<sup>14</sup> The image pencil angle will be  $\theta_3 = D/L = 0.1$  and the ratio of brightness to energy at this image is:

$$\begin{aligned} \frac{\beta_3}{E_3} &= \frac{I_{\max}/\pi^2 r_3^2 \theta_3^2}{E_3} \\ &= 9500 \mu\text{A}/\text{cm}^2/\text{sr}/\text{eV} \end{aligned} \quad (5.64)$$

At the anode of the Pierce diode (Section 5.3.1), the ratio of brightness to energy is given by:

$$\frac{\beta_1}{E_1} = 0.74 \frac{E_1^{3/2}}{E_k^2} \quad (5.65)$$

where  $E_k$  is the mean kinetic energy of electrons emitted by the cathode, and  $E_1$  is the energy of electrons at the anode that is held at the potential  $V_1$  relative to the cathode.

Now the ratio of brightness to energy is a conserved quantity, so:

$$\frac{\beta_1}{E_1} = \frac{\beta_3}{E_3} \quad (5.66)$$

Substituting from the previous two equations yields:

$$E_1 = \left[ \left( \frac{\beta_3}{E_3} \right) \frac{E_k d^2}{0.74} \right]^{2/3} \quad (5.67)$$

For  $E_k = 0.25$  eV (Section 5.2.5), a cathode-to-anode spacing  $d = 0.5$  cm, and a final electron energy of 20 eV, the kinetic energy of electrons at the anode,  $E_l = 86$  eV, and the potential to be applied to the anode is:

$$V_1(E_3 = 20 \text{ eV}) = 86 \text{ V} \quad (5.68)$$

A similar calculation yields:

$$V_1(E_3 = 200 \text{ eV}) = 185 \text{ V} \quad (5.69)$$

The current from the diode depends upon the radius of the anode aperture  $r_1$  and the current density  $J$  at the anode:

$$I = J\pi r_1^2 \quad (5.70)$$

With the lens system tuned to inject 20 eV electrons into the target cell ( $E_3 = 20$  eV), the maximum current density at the anode: is

$$\begin{aligned} J_{\max} &= \frac{2.34}{d^2} V_1^{3/2}; \quad V_1 = V_1(E_3 = 20 \text{ eV}) \\ &= 7500 \mu\text{A}/\text{cm}^2 \end{aligned} \quad (5.71)$$

To achieve maximum current in the cell at 20 eV, we choose the anode aperture to be:

$$r_1 = \left( \frac{I_{\max}}{\pi J_{\max}} \right)^{1/2} = 0.04 \text{ cm} \quad (5.72)$$

A similar calculation for  $E_3 = 200$  eV would suggest using a larger anode aperture; a larger aperture, however, would produce an excess of current and thus many stray electrons when the system is tuned to produce a low-energy beam.

Note that when the system is designed to approach the space-charge limit at 20 eV, the pencil angle at the anode is consistent with the pencil angle defined by the

cell apertures. The pencil angle characteristic of the Pierce diode at the low-energy limit is:

$$\theta_1 = \sqrt{\frac{E_k}{E_1}} = \sqrt{\frac{0.25}{86}} = 0.05 \quad (5.73)$$

and the anode pencil angle that gives the space-charge-limited pencil angle in the cell can be determined from the Helmholtz–Lagrange law:

$$\theta_1 = \sqrt{\frac{E_k}{E_1}} \left( \frac{r_3}{r_1} \right) \theta_3 = 0.05 \quad (5.74)$$

For final energies greater than 20 eV, these pencil angles will be less than the angle defined by the cell apertures, and scattering from the edges of the apertures will be minimized.

It remains to design a lens system that will image the anode aperture on the center of the gas cell. A variable-ratio lens is required, and in this case a three-cylinder asymmetric lens seems appropriate. The voltage ratios at the extremes of the desired operating conditions are:

$$\frac{V_3}{V_1}(E_3 = 20 \text{ eV}) = 0.23 \quad (5.75)$$

and

$$\frac{V_3}{V_1}(E_3 = 200 \text{ eV}) = 1.1 \quad (5.76)$$

The desired magnification is

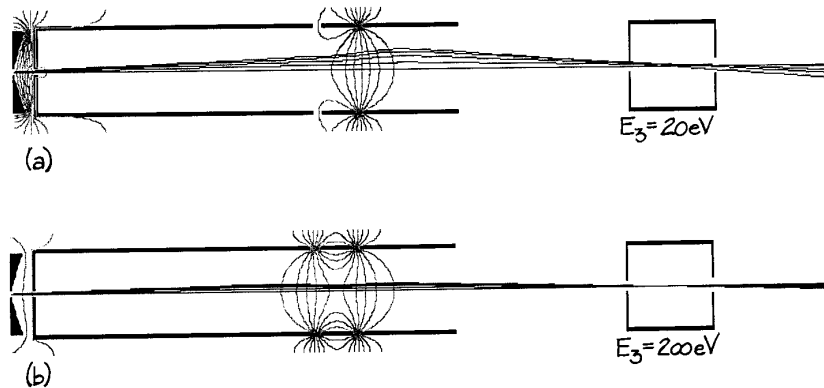
$$M = \frac{r_3}{r_1} \approx 1 \quad (5.77)$$

The lens diameter must be chosen to give an acceptable filling factor. As a starting point, recall that:

$$M \approx 0.8 \frac{Q}{P} \quad (5.78)$$

The lens is to focus the anode aperture on the center of the gas cell, so  $P + Q = 10 \text{ cm} + 4 \text{ cm} + 1 \text{ cm} = 15 \text{ cm}$ . The distance from the anode to the middle of the lens will be:

$$\begin{aligned} P &= \left( \frac{P}{P+Q} \right) \times 15 \text{ cm} \\ &= \left( \frac{P}{P+1.25MP} \right) \times 15 \text{ cm} \\ &= 6.7 \text{ cm} \end{aligned} \quad (5.79)$$



**Figure 5.21** MacSimion simulation of the electric field and electron trajectories in the electron gun system shown schematically in Figure 5.20. The cathode-to-anode distance in the Pierce diode is  $d = 0.5$  cm and the radius of the aperture in the anode is  $r_1 = 0.04$  cm. The radii of the apertures in the target cell are  $r_3 = 0.1$  cm and  $r_4 = 0.15$  cm. The diameter of the lens  $D = 2$  cm, the length of the central element  $A = 0.5D$ , and the gap between lens elements is  $g = 0.1D$ . (a) Trajectories of electrons that pass through the cell with energy  $E_3 = 20$  eV;  $V_1 = 86$  V,  $V_2 = 105$  V, and  $V_3 = 20$  V. (b) Trajectories of electrons that pass through the cell with energy  $E_3 = 200$  eV;  $V_1 = 185$  V,  $V_2 = 960$  V, and  $V_3 = 200$  V.

For a final energy of 20 eV, the maximum diameter of the beam through the lens system will be approximately:

$$2\theta_1 P = 0.67 \text{ cm}; \quad \theta_1(E_3 = 20 \text{ eV}) \quad (5.80)$$

Choose  $D = 2.0$  cm to achieve a filling factor less than 30%. Then:

$$\begin{aligned} P &= 3.35D \\ Q &= 4.15D \end{aligned} \quad (5.81)$$

and the nominal focal length of the desired lens (see Section 5.2.2) is

$$f = \left( \frac{1}{P} + \frac{1}{Q} \right)^{-1} = 1.85D \quad (5.82)$$

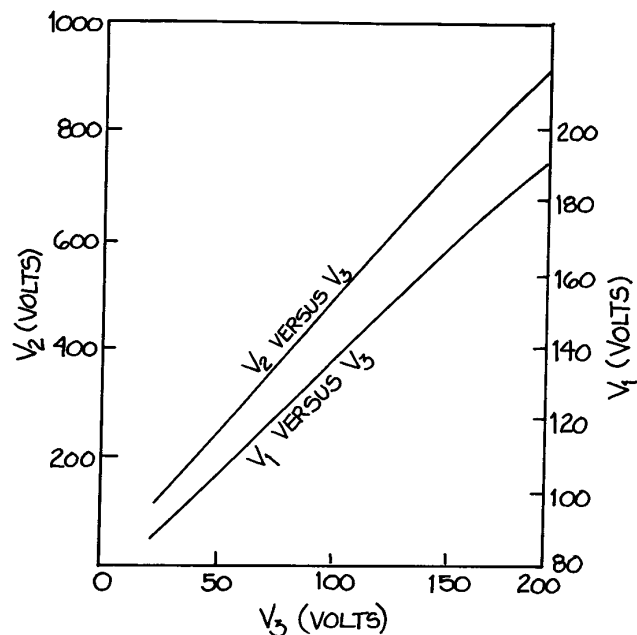
The calculations of Adams and Read<sup>8</sup> or Harting and Read<sup>7</sup> can then be used to determine the lens voltages that correspond to this nominal focal length. Alternatively, the lens system can be modeled by computer simulation and the potential to be applied to the central element of the three-cylinder lens can be determined empirically. Figure 5.21 shows the results of simulations for the projection of 20 and 200 eV electrons into the target cell. Figure 5.22

gives the central lens element voltage  $V_2$  obtained from the simulation.

### 5.3.3 Ion Sources

Many parameters must be considered in choosing or designing an ion source. Foremost is the desired type of ionic species and charge state. The physical and chemical properties of the corresponding parent material determine to a large extent the means of ion production. Other important parameters are the desired current, brightness, and energy distribution. In some cases it is also necessary to consider the efficiency of utilization of the parent material. In view of the number of variables involved, it is not surprising that many different types of ion sources have been developed to meet both scientific and industrial demands.<sup>15</sup>

Ion sources can be classified according to the ion production mechanism employed. The two most common means are surface ionization and electron impact. The electron-impact sources include both electron-bombardment sources and plasma sources. More exotic ion sources employ ion impact, charge exchange, field ionization, and

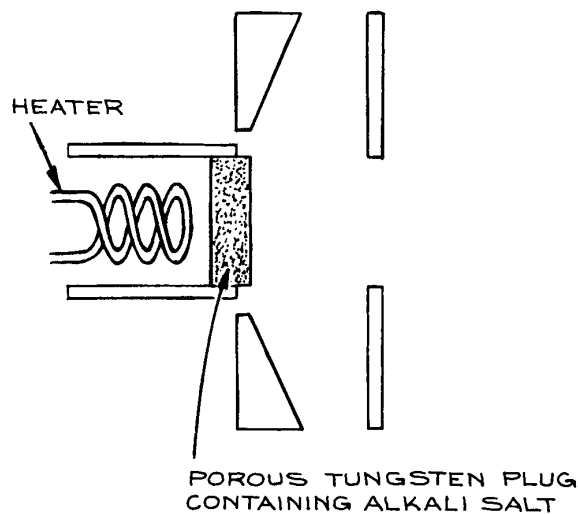


**Figure 5.22** Lens voltages for the lens system of Figure 5.20.

photoionization. Three typical sources, illustrating different ion production mechanisms, will be discussed below.

In practice, the simplest means of producing ions is by thermal excitation of neutral species. This process, known as *surface ionization*, occurs efficiently when an atom or molecule is brought in contact with a heated surface whose work function exceeds the ionization potential of the atom or molecule. This requirement is fulfilled for alkali atoms (Li, Na, K, Rb, Cs) on surfaces of tungsten, iridium, or platinum. In order for ionization to compete with vaporization, it is necessary that the surface be sufficiently hot that the substrate surface is only partially covered with neutrals, so that the work function of the surface is characteristic of the substrate rather than the material to be ionized.

In a surface-ionization ion gun, the source of neutral atoms can either be remote from, or integral with, the ionizing surface. In the remote type, the alkali metal is contained in an oven. A stream of metal vapor from the oven is directed toward the ionizer, which is a heated metal "cathode" located within an electrode structure similar to that of an electron gun. The "anode" is negatively biased to accelerate positive ions into a beam. The most convenient

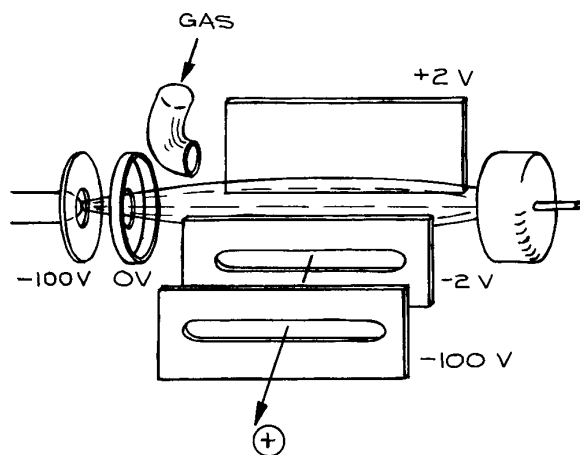


**Figure 5.23** An alkali-ion source consisting of an indirectly heated emitter in a Pierce diode.

alkali-ion emitter consists of a porous tungsten disc that has been infused with an alkali-containing mineral and mounted on an electrical heater. When heated, alkali atoms are generated; the atoms diffuse to the surface of the disc, and are ionized as they leave the surface. These emitters with integral heaters are commercially available (Spectra-Mat). As shown in Figure 5.23, a source is constructed by inserting the emitter into the cathode of a Pierce diode or into some other extractor electrode structure.

The majority of laboratory ion sources employ *electron-impact* ionization. The simplest configuration is illustrated in Figure 5.24. In this source, an electron beam from a simple diode gun is injected into an ionization chamber containing an appropriate parent gas. A transverse electric field of a few V/cm causes ions produced by electron impact to drift across the chamber and out through a slit in the side of the chamber. An extractor electrode system accelerates the ions into a beam. The ionic species produced will depend upon the parent gas, the gas pressure, and the electron energy. Singly charged ions are produced with maximum efficiency at electron energies of about 70 eV. Higher electron energies favor multiply charged ions. Lower electron energies yield only singly charged ions and reduce the extent of fragmentation in the event that the parent gas is a molecular species. Gas pressures are



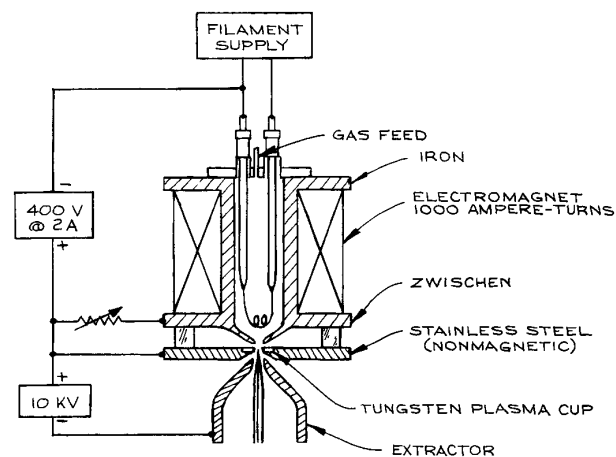


**Figure 5.24** An electron-impact ion source.

typically 1 to 10 mtorr. Higher pressures place an unreasonable gas load on the vacuum system and impede the flow of ions from the ionization chamber.

The efficiency of the electron-impact ionization source can be improved by imposing a magnetic field of a few hundred gauss coaxial with the electron beam. The field confines the electron beam to a spiral around the beam axis, thus increasing the path length and maximizing the probability of collision with the gas molecules. Electron-bombardment sources of this type yield currents of a few tens of microamperes. The chief advantages are simplicity of construction and an energy spread in the product ions of only a few eV.

Relatively high ion densities can be achieved in an electrical discharge through a gas. The ion density in a discharge can be further increased if the discharge is confined and compressed by a magnetic field. The duoplasmatron source illustrated in Figure 5.25 is the prototypical discharge-type ion source.<sup>16</sup> Applying a potential of 300 to 500 V between the heated cathode and the anode produces a discharge through a gas at about 100 mtorr. After the arc is struck, the discharge is maintained by passing a current of 0.5 to 2 A through the ionized gas. In the duoplasmatron, the intermediate element, known as the *zwischen*, is one pole of a magnet. The axial magnetic field at the tip of the *zwischen* confines the discharge to a dense plasma bubble at the anode. Plasma leaks through a hole in the anode to fill a cup on the front face of the anode. An



**Figure 5.25** A duoplasmatron ion source.

extractor electrode, biased at about 10 kV relative to the anode, withdraws ions from the surface of the plasma. Ion currents of 1 to 10 mA are easily obtained. In operation, several hundred watts of electrical power are dissipated in the arc and the electromagnet. In its original form, this source was liquid-cooled. However, the author has found air-cooling to be acceptable for slightly smaller versions. The duoplasmatron is the brightest of ion sources. Copious quantities of singly charged atomic and diatomic ions can be obtained for these species for which an appropriate gaseous parent can be found. It is also possible to obtain a few microamperes of doubly charged ions, and in addition, negative ions can be obtained by reversing the extraction potential.<sup>17</sup> The chief disadvantages of plasma sources are their complicated construction and the fact that stable operation is only possible for a few days before cleaning and filament replacement are necessary.

## 5.4 ENERGY ANALYZERS

There are three basic means of measuring the energy of charged particles in a beam. These involve measuring the time of flight over a known distance, the retarding potential required to stop the particles, or the extent of deflection in an electric or magnetic field.

Because of the great velocities involved, the flight time of a charged particle over any reasonable distance is very

short. Determination of the kinetic energy of a particle from the time required to cover a distance of a few centimeters typically requires electronics with a response time of a few nanoseconds. *Time-of-flight analyzers* are used for energy analysis of electrons with energies less than 10 eV and ions below 1 keV.

Placing a grid or aperture in front of a particle collector and varying the potential on this element will perform an energy analysis of the particles in a beam.<sup>18</sup> This is a *retarding potential analyzer*. The current at the collector is the integrated current of particles whose energy exceeds the potential established by the grid. As the grid potential is reduced from that at which all current is cut off, the collector current increases. To obtain the energy distribution, the integrated current as a function of retarding potential must be differentiated. One drawback to this method is that only the component of velocity normal to the retarding grid is selected. There are also a number of practical difficulties. The ratio of the initial energy to the energy at the potential barrier varies rapidly for particles near the threshold for penetrating the retarding barrier. This gives rise to rapidly varying focusing effects near the threshold, so particles approaching slightly off axis are often deflected away from the collector. The result is that the transmission of the analyzer is unpredictable near the threshold. Another problem with retarding-potential analyzers is that low-energy particles near the retarding grid are seriously affected by space charge and by stray electric and magnetic fields. Retarding potential analyzers are very easily constructed and very compact, but the vagaries of their performance suggests that their use for high-resolution energy analysis be avoided if possible.

The energies of particles in a beam can be determined by passing the beam through an electric or magnetic field, so that the deflection of the particle paths is a function of the particle energy per unit charge or momentum per unit charge. In these *dispersive-type analyzers*, the shape of the deflection field must be carefully controlled. Since it is generally easier to produce a shaped electric field than a shaped magnetic field, dispersive analyzers for particle energies up to several keV are usually of the electrostatic variety. For very high-energy particles, magnetic analyzers are preferred, since production of the required electric fields would demand inconveniently large electrical potentials.

The energy *passband*,  $E$ , of an analyzer may be defined as the full width at half maximum (FWHM) of the peak that

appears in a plot of transmitted current versus energy. Ideally the transmission function is triangular. For a real analyzer it resembles a Gaussian. We shall take  $\Delta E$  to be half the full width of the transmission function; such an approximation overestimates the passband of a real analyzer.<sup>19</sup> To first order, entrance and exit slits or apertures, usually of equal width  $w$ , establish the passband. The transmission function also depends upon the maximum angular extent to which particles can deviate from the central path leading from the entrance to the exit slit. This angular deviation is defined by angles  $\Delta\alpha$  in the plane of deflection and  $\Delta\beta$  in the perpendicular plane. If  $E$  is the central energy of particles transmitted through an analyzer, then the *resolution* is:

$$\frac{\Delta E}{E} = aw + b(\Delta\alpha)^2 + c(\Delta\beta)^2 \quad (5.83)$$

where  $a$ ,  $b$ , and  $c$  are constants characteristic of the particular analyzer.

### 5.4.1 Parallel-Plate Analyzers

The simplest electrostatic analyzer employs a uniform field created by placing a potential difference across a pair of plane parallel plates, as shown in Figure 5.26. With the entrance and exit slits in one of the plates, first-order focusing in the deflection plane is obtained when the angle of incidence of entering particles is about  $\alpha = 45^\circ$ .

The deflection potential  $V_d$  (V), in relation to the incident energy  $E$ (eV), the plate spacing  $d$ , and the slit separation  $L$ , is given by:

$$V_d = (E/q) \frac{2d}{L} \quad (5.84)$$

with the proviso that  $d > L/2$ , to prevent particles from striking the back plate. The resolution is:

$$\frac{\Delta E}{E} = \frac{w}{L} + (\Delta\alpha)^2 + \frac{1}{2}(\Delta\beta)^2 \quad (5.85)$$

A point at the entrance aperture is focused into a line of length  $2\sqrt{2}L\Delta\beta$ , and the length of the exit slit is correspondingly greater than that of the entrance slit.

When the slits are placed in a field-free region,<sup>20</sup> as in Figure 5.26(b), optimum performance is obtained with the angle of incidence  $\alpha = 30^\circ$ . The distance from the slits to

the entrance plate of the analyzer  $d_1 \approx 0.1d_2$ , where  $d_2$  is the plate spacing. The deflection potential is:

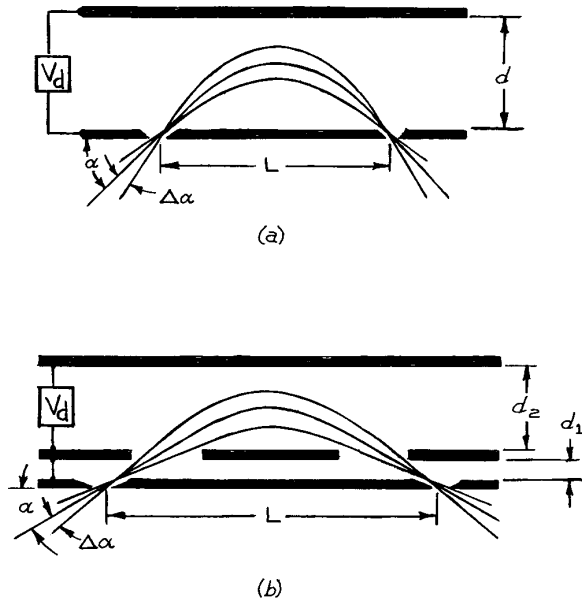
$$V_d = 2.6(E/q) \left( \frac{d_2}{L} \right) \quad (5.86)$$

This arrangement gives second-order focusing in the plane of deflection. The resolution is:

$$\frac{\Delta E}{E} = 1.5 \frac{w}{L} + 4.6(\Delta\alpha)^2 + 0.75(\Delta\beta)^2 \quad (5.87)$$

and the image of a point on the entrance slit is a line of length  $4L\beta$  at the exit slit.

Although the parallel plate is an attractive design because of its simple geometry, there are several problems. The entrance apertures or slits in the front plate are at the boundary of a strong field and act as lenses to produce unwanted aberrations. This problem can be alleviated for the design with the energy-resolving slits in a field-free region by placing a fine wire mesh over these entrance apertures to mend the field. A large electrical potential must be applied to the back plate, creating a strong electrostatic field outside



**Figure 5.26** Parallel-plate analyzers with (a) the energy-resolving slits in one of the plates and (b) the slits in a field-free region.

the analyzer. The apparatus in which the analyzer is installed must often be shielded from this field. In addition, because the gap between the plates is large, the fringing field at the edges of the plates can penetrate into the deflection region. This problem can be solved by extending the edges of the plate well beyond the deflection region, or by placing compensating electrodes at the edges of the gap, as in Figure 5.27.

## 5.4.2 Cylindrical Analyzers

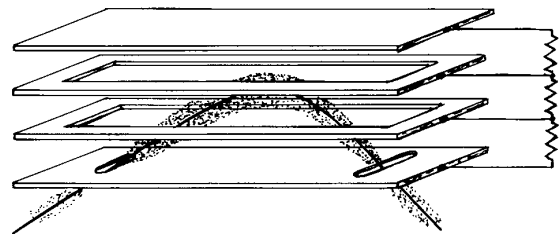
Two electrostatic analyzers employ cylindrical electrodes. These are the radial cylindrical analyzer and the cylindrical mirror analyzer.

The *radial cylindrical analyzer* is shown in Figure 5.28. The radial electric field is produced by an electrical potential placed across concentric cylindrical electrodes. Particles are injected midway between the electrodes in a direction approximately tangent to the circular arc of radius  $R_0$ . First-order focusing is obtained if the cylindrical electrodes subtend an angle of  $\pi/\sqrt{2} = 127^\circ$ . Assuming a charged particle that originates at ground potential with essentially no kinetic energy, and a mean pass energy  $E = -qV$  (that is, the energy of the particle that travels along the central path of radius  $R_0$ ), the potentials to be applied to the outer and inner cylindrical elements are:

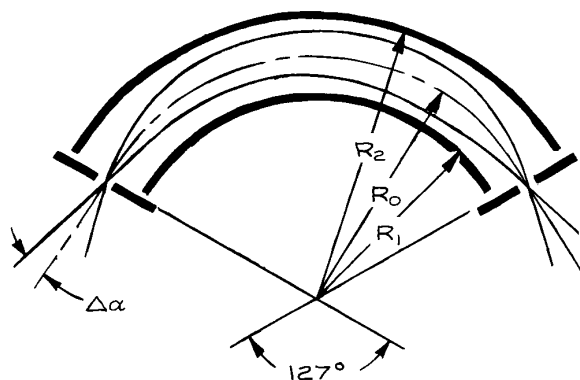
$$V_{\text{outer}} = V \left( 1 - 2 \ln \frac{R_2}{R_0} \right) \quad (5.88)$$

and

$$V_{\text{inner}} = V \left( 1 - 2 \ln \frac{R_1}{R_0} \right) \quad (5.89)$$



**Figure 5.27** Guard electrodes at the edges of a parallel-plate analyzer to offset the distortion of the field caused by fringing.



**Figure 5.28** The radial cylindrical analyzer.

where  $R_2$  and  $R_1$  are the radii of the outer and inner cylinders, respectively. The potentials on the electrodes are such that a charged particle that enters at the midradius  $R_0$  with an energy equal to the pass energy  $E$  will *lose* an amount of energy equal to  $2E \ln R_2/R_0$  were it to travel to the outer electrode, and would *gain* an amount of energy equal to  $2E \ln R_0/R_1$  were it to travel to the inner electrode.

The resolution of this analyzer is:

$$\frac{\Delta E}{E} = \frac{w}{R_0} + \frac{2}{3}(\Delta\alpha)^2 + \frac{1}{2}(\Delta\beta)^2 \quad (5.90)$$

It is good practice to limit the angle of divergence in the plane of deflection so that:

$$\Delta\alpha < \frac{2\sqrt{2}R_2 - R_1}{\pi R_0} \quad (5.91)$$

in order to keep the filling factor below 50%. The radial-field analyzer gives focusing only in the plane of deflection. A point at the entrance slit is imaged as a line of length  $\sqrt{2}\pi R_0 \Delta\beta$  at the exit slit.

The *cylindrical-mirror analyzer* is similar to the parallel-plate analyzer except the deflection plates are coaxial cylinders. In fact, the parallel-plate analyzer can be considered a special case of the cylindrical mirror. In the usual geometry, shown in Figure 5.29, the source is located on the axis, and particles emitted into a cone defined by polar angle  $\alpha$  pass through an annular slot in the inner cylinder. Particles of energy  $E$  are deflected so that they pass through an exit slot and are focused to an image on the axis. The cylindrical mirror is double-focusing (focusing occurs in both the

deflection plane and the perpendicular plane) so that the image of a point at the source appears as a point at the detector. An obvious advantage of this analyzer is that particles at any azimuthal angle can be collected.

For optimum performance the entry angle  $\alpha = 42.3^\circ$ , in which case the distance from source to detector is  $L = 6.12R_1$ .<sup>21</sup> The inner cylindrical plate is at the same potential as the source, and the potential on the outer cylinder, relative to the inner cylinder, is:

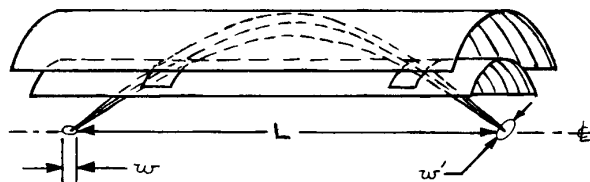
$$V_{\text{outer}} = 0.763(E/q) \ln \frac{R_2}{R_1} \quad (5.92)$$

It is wise to choose  $R_2 > 2.5R_1$  so that particles are not scattered from the outer electrode. The resolution of the axial-focusing cylindrical mirror analyzer is approximately:

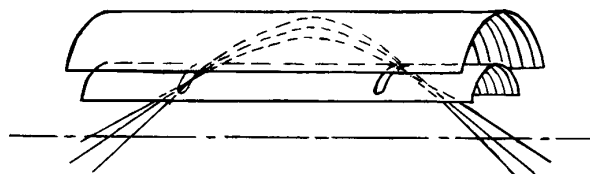
$$\frac{\Delta E}{E} = 1.09 \left( \frac{w}{L} \right) \quad (5.93)$$

for a source of axial extent  $w$  and an energy-resolving aperture of diameter  $w' = w \sin \alpha$  perpendicular to the axis (as shown in Figure 5.29).

If the source is not small and well defined, the entry and exit slots in the inner cylinder can be used to define the resolution of the cylindrical-mirror analyzer, as in Figure 5.30; this arrangement is only first-order-focusing.



**Figure 5.29** Axial-focusing cylindrical-mirror analyzer.



**Figure 5.30** Cylindrical-mirror analyzer with energy-resolving slits in the inner electrode.

### 5.4.3 Spherical Analyzers

For many applications, the most desirable analyzer geometry employs an inverse-square-law field created by placing a potential across a pair of concentric spherical electrodes. Focusing in both the deflection plane and the perpendicular plane can be obtained using any sector portion of a sphere. As shown in Figure 5.31, the object and image lie on lines that are perpendicular to the entrance and exit planes, respectively, and tangent to the circle described by the midradius  $R_0$ . Furthermore, by *Barber's rule*, the object, the center of curvature of the spheres, and the image lie on a common line. The  $180^\circ$  spherical sector (Figure 5.32) is often used because of the compact geometry that results from folding the beam back onto a line parallel to its original path. As with the cylindrical mirror, the *spherical analyzer* can be used to collect all particles emitted from a point source at or near a particular polar angle.

Assuming a charged particle that originates at ground potential with essentially no kinetic energy, and a mean pass energy  $E = -qV$  (that is, the energy of the particle that travels along the central path of radius  $R_0$ ), the

potentials to be applied to the outer and inner spherical elements are:

$$V_{\text{outer}} = V \left( 2 \frac{R_0}{R_2} - 1 \right) \quad (5.94)$$

and:

$$V_{\text{inner}} = V \left( 2 \frac{R_0}{R_1} - 1 \right) \quad (5.95)$$

where  $R_2$  and  $R_1$  are the radii of the outer and inner spheres, respectively; in general, the potentials on the electrodes are such that a charged particle that enters at the midradius  $R_0$  with an energy equal to the pass energy  $E$  will *lose* an amount of energy equal to  $2E(1 - R_0/R_2)$  were it to travel to the outer electrode, and would *gain* an amount of energy equal to  $2E(R_0/R_1 - 1)$  were it to travel to the inner electrode.

The resolution of the  $180^\circ$  spherical sector (Figure 5.32) is:

$$\frac{\Delta E}{E} = \frac{w}{2R_0} + \frac{1}{2}(\Delta\alpha)^2 \quad (5.96)$$

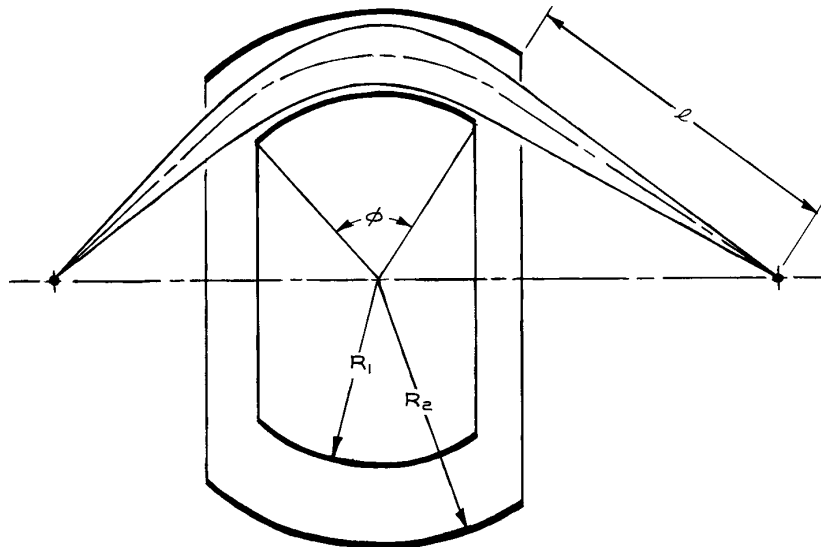
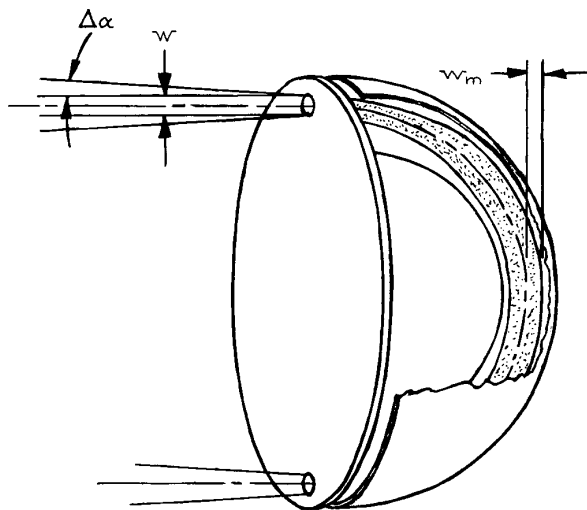


Figure 5.31 Focusing of a spherical analyzer.



**Figure 5.32** The hemispherical analyzer:  $w$  is the width of the entrance and exit apertures,  $\Delta\alpha$  the maximum angular deviation of an incident trajectory, and  $w_m$  the maximum deviation of a trajectory from the central path through the analyzer.

and the maximum deviation,  $w_m$ , of a trajectory from the central path within the analyzer is given by:

$$\frac{w_m}{R_0} = \frac{\Delta E}{E} + \Delta\alpha + \frac{1}{2\Delta\alpha} \left( \frac{w}{2R_0} + \frac{\Delta E}{E} \right)^2 \quad (5.97)$$

For the truncated spherical sector shown in Figure 5.31, the resolution is approximately:

$$\frac{\Delta E}{E} = \frac{w}{R_0(1 - \cos \phi) + \sin \phi} \quad (5.98)$$

where  $\phi$  is the angle subtended by the analyzer sector and  $l$  is the distance from the exit boundary of the analyzer to the exit aperture. Although it would appear that the resolution could be increased arbitrarily by increasing  $l$ , the aberrations increase rapidly as the system becomes asymmetric. It is best to employ the symmetric geometry with  $\phi$  in the range 60–180°.

In comparison with the parallel-plate or cylindrical-mirror analyzers, the spherical analyzer has the advantage of requiring relatively low electrical potentials on the electrodes. Because the electrodes are closely spaced in

the spherical analyzer, fringing fields are less of a problem and more easily controlled. The chief drawback of this type of analyzer is the difficulty of fabrication and mounting.

#### 5.4.4 Preretardation

In all electrostatic-deflection analyzers, the *passband*  $\Delta E$  is a linear function of the transmitted energy. Thus, the absolute energy resolution of these devices can be improved by retarding the incident particles prior to their entering the analyzer. In principle, the passband of an analyzer can be reduced arbitrarily by *preretardation*; in practice, there are limitations on this technique. A decelerating lens as in the example given in Section 5.2.6 can slow charged particles incident on the entrance aperture of an analyzer. It would be very difficult, however, to design a lens system to be used with analyzers that have an annular entrance slit. Space charge and stray electric and magnetic fields must be considered. Because the flight path through an analyzer is long, it is usually not possible to reduce the energy of transmitted particles below about 2 eV before these effects result in severe aberrations. Finally, it is important to recall that if all particles of a particular energy in a beam are to be transmitted through an analyzer, the ratio of brightness to energy must be conserved (Section 5.1.3). As the energy of particles incident on an analyzer is decreased, the current is ultimately reduced because the pencil angle of the beam exceeds the acceptance angle of the analyzer.

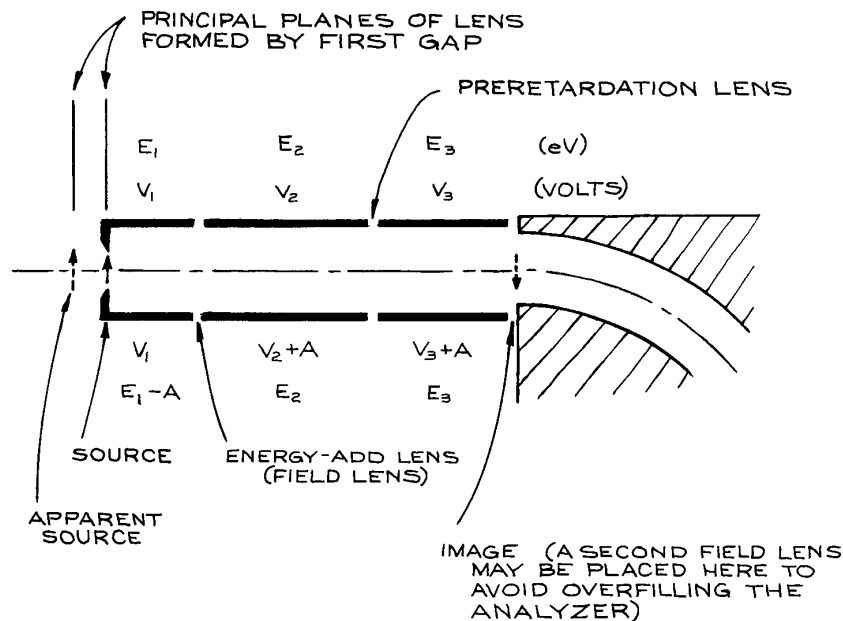
When using a decelerating lens with an analyzer, it is good practice to place an aperture on the high-energy side of the lens and design the lens so that the image of this aperture appears at the entrance plane of the analyzer. The need for a real entrance aperture is thereby eliminated, and there are no metal surfaces in the vicinity of the low-energy beam. Since space charge will cause the low-energy beam to expand at the entrance plane of an analyzer, current would be lost if a real entrance aperture were used. As demonstrated by Kuyatt and Simpson, space-charge expansion at a virtual aperture is compensated by the spreading of the beam, with the result that the beam appears to originate at an aperture of about the same size as that in the absence of space charge<sup>22</sup> (see Figure 5.16).

### 5.4.5 The Energy-Add Lens

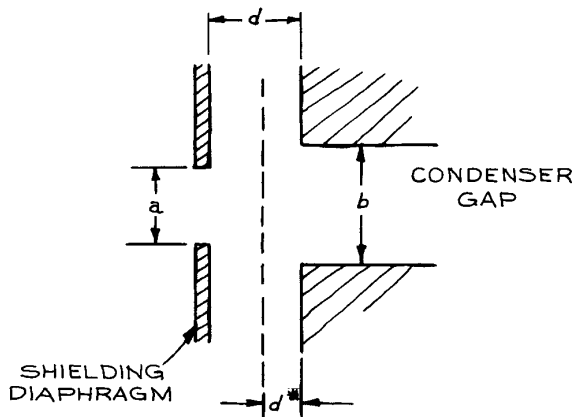
An electrostatic energy analyzer can be used as a monochromator to select the particles in a beam whose energies fall within a narrow range, or as an analyzer to determine the energy distribution of particles originating from a process under study. In the latter application the passband of the analyzer must be scanned over the energy range of interest. Scanning can be accomplished by varying the potential difference between the analyzer electrodes; however, neither the transmission nor the energy resolution will be constant. A more satisfactory method for determining the energy distribution of particles in a beam is to fix the analyzer potential at the voltage that allows transmission of particles of the highest desired energy and *preaccelerate* the incident particles by a variable amount. The energy that must be added to the particles so that they pass through the analyzer is then a measure of the difference between their initial energy and the energy necessary to transmit them with no preaccelera-

tion. In a scattering experiment, for example, the analyzer might be set to transmit elastically-scattered particles, and the distribution of energy lost by inelastically scattered particles would be scanned by monitoring the transmitted current as a function of energy added.

An *energy-add lens* must add back energy without disturbing the optics of the analyzer. Kuyatt<sup>23</sup> showed that the electron-optical analog of the *field lens* is suited to this application. This lens is positioned so that its first principal plane coincides with the particle source to be examined. The source is then imaged onto the second principal plane with unit magnification. For a reasonably strong lens ( $V_2/V_1 > 3$ ), the position of the principal planes is nearly independent of the voltage ratio and the separation of the planes is small. Thus particles from the source can be accelerated by a variable amount without changing the apparent position of the source. An energy-add lens used in conjunction with a fixed-ratio decelerating lens for preretardation at the input of an electron energy analyzer is schematically illustrated



**Figure 5.33** An energy-add lens and fixed-ratio retarding lens at the input of an analyzer. The energies and corresponding lens voltages for electrons transmitted with no added energy are given above the lens. Energies and voltages for transmission of electrons that have lost energy  $A$  are shown below. The "source" is defined by an aperture that is finally imaged at the entrance plane of the analyzer.



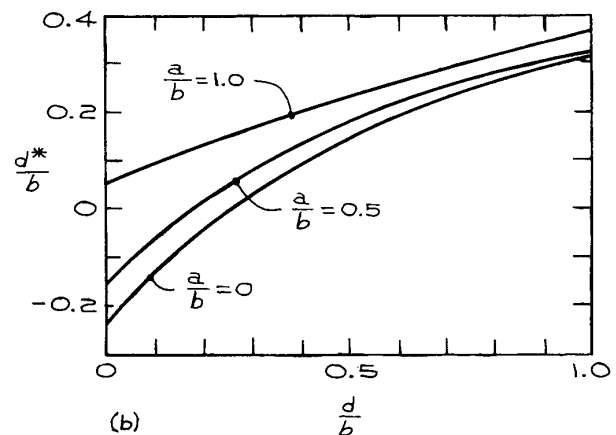
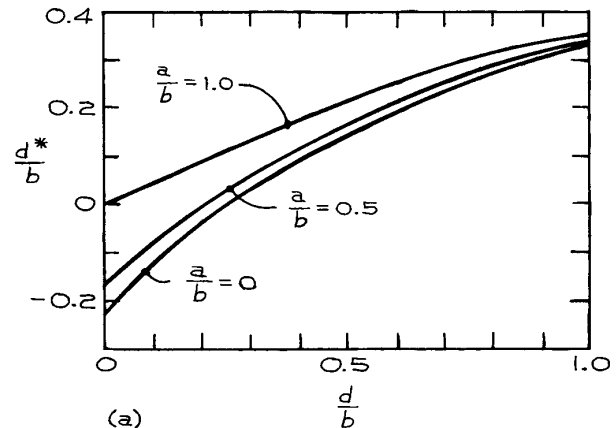
**Figure 5.34** A shielding diaphragm to correct for focusing at the fringing field at the edge of an electrostatic condenser. The field appears as if it were terminated abruptly at a distance  $d^*$  from the gap.

in Figure 5.33. For transmission of electrons with energy  $A$  (eV) less than the maximum energy to be transmitted, the potential of the second element of the energy-add lens, and every electrode thereafter is increased by  $A$  (V). For a positive ion analyzer, the potential would be decreased by  $A/n$  (V), where  $n$  is the charge state of the ions.

A problem with the energy-add lens is that the object pupil moves as the lens voltages change. Thus, while the apparent source remains stationary, the beam angle may vary over a considerable range as the lens is tuned. As a result, the lenses following the energy-add lens may be overfilled. This can be avoided by placing a second field lens downstream from the first at an intermediate image of the apparent source. The second field lens is then tuned to give a zero beam angle. An einzel lens (Section 5.2.2) can be used so that there is no net change of energy as the beam angle is manipulated.

### 5.4.6 Fringing-Field Correction

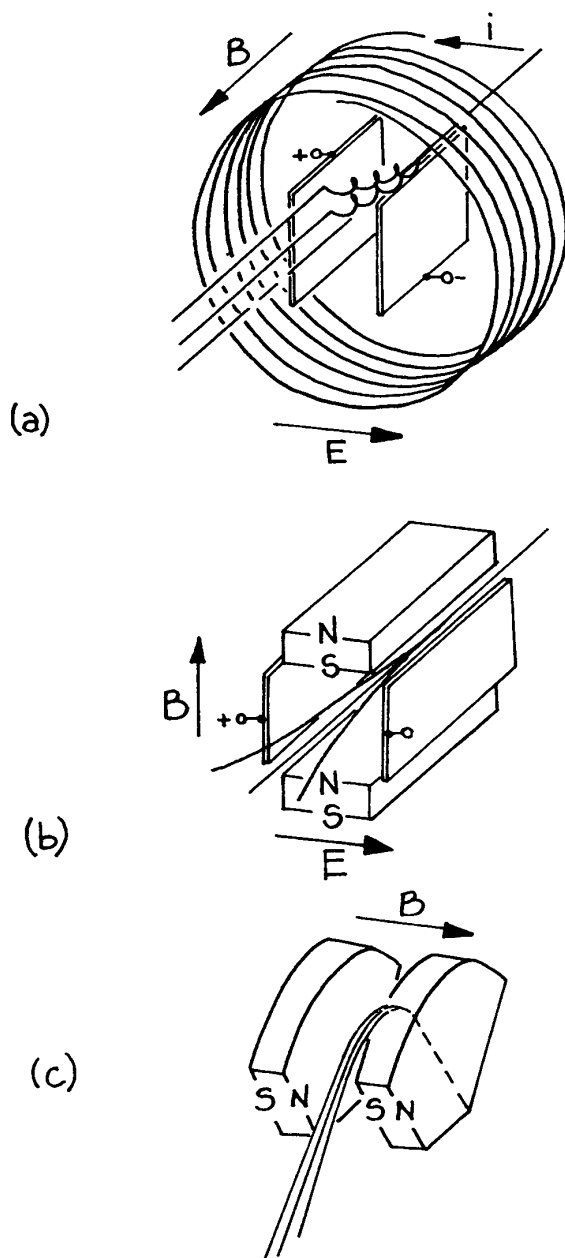
For the radial-field cylindrical analyzer or the spherical analyzer, particles must pass through the fringing field at the edge of the electrode gap as they enter the deflection region. The electric field at the edge bulges out of the gap, and this curvature produces a focusing effect that causes undesirable aberrations. Herzog<sup>24</sup> and Wollnik and Ewald<sup>25</sup> have shown that a diaphragm at the entrance of the condenser gap can



**Figure 5.35** The distance of the apparent field boundary from the edge of the condenser electrodes as a function of the dimensions given in Figure 5.34 for (a) a thick diaphragm and (b) a thin diaphragm. (From H. Wollnik and H. Ewald, "The Influence of Magnetic and Electric Fringing Fields on the Trajectories of Charged Particles," *Nucl. Instr. Meth.*, **36**, 93 (1965); by permission of North-Holland Publishing Company.)

largely eliminate this problem. As illustrated in Figure 5.34, the diaphragm, when properly located, produces a field that has the same effect as a field abruptly terminated at a distance  $d^*$  in front of the condenser gap. The appropriate location of the diaphragm as a function of the dimensions given in Figure 5.34 can be determined for either a thick or thin diaphragm from the graphs in Figure 5.35.





**Figure 5.36** Charged-particle energy analyzers with magnetic fields: (a) the trochoidal analyzer with an electromagnet; (b) the Wien filter; (c) the sector magnet analyzer. Trajectories for electrons of different energies are shown. Magnet polarities are for electrons.

### 5.4.7 Magnetic Energy Analyzers

Magnetic fields are employed in several charged-particle-energy analyzers and filters. The *trochoidal analyzer* [Figure 5.36(a)] has proven quite useful for the dispersion of very-low-energy (0 to 10 eV) electrons and ions. This device employs a magnetic field aligned with the direction of the charged-particle beam, and an electric field perpendicular to this direction.<sup>26</sup> The trajectory of a particle injected into this analyzer describes a spiral and the guiding center of the spiral drifts in the remaining perpendicular direction. The drift rate depends upon the particle energy; a beam entering the device is dispersed in energy at the exit. The projection of the trajectory on a plane perpendicular to the electric field direction is a trochoid, hence the name. The device requires a very uniform magnetic field of about 100 gauss. The field is usually produced by a relatively large Helmholtz pair (see Section 5.6.5) mounted outside the vacuum system containing the analyzer. The result is that the entire experiment is immersed in the field. This may be a drawback in some cases, but, on the other hand, for low-energy electrons or ions, the field helps contain particles that otherwise would be lost from coulombic repulsion (i.e., space charge).

The Wien filter (Figure [5.36(b)]) similarly employs crossed electric and magnetic fields, however, the fields are perpendicular to one another and both are perpendicular to the injected beam direction.<sup>27</sup> The Coulomb force induced by the electric field  $E$  deflects charged particles in one direction and the Lorentz force associated with the magnetic field  $B$  tends to deflect them in the opposite direction. The forces balance for one velocity,  $v = |E|/|B|$ , and particles of the corresponding energy are transmitted straight through to the exit aperture.

A magnetic field alone, perpendicular to the direction of a charged-particle beam will provide energy dispersion, provided that the particles are all of the same mass and charge (Figure [5.36(c)]). Sector magnets such as those used in mass spectrometers are used in electron spectrometers for very-high-energy electrons and ions, the advantage over electrostatic deflectors being that large electrical potentials are not required. Another advantage is that the deflection is in the direction parallel to the magnet pole face, making it possible to view the entire dispersed spectrum at one time. By contrast, energy dispersion in an

electrostatic device results in a significant portion of the dispersed particles' striking one or the other electrode.

## 5.5 MASS ANALYZERS

Mass analysis is more complicated than energy analysis of an isotopically pure beam, since energy and momentum are independent variables in a mixed beam. A detailed discussion of the many types of mass analyzers that have been developed is beyond the scope of this book. The principles of operation of a few representative analyzers will be described.

### 5.5.1 Magnetic Sector Mass Analyzers

The deflection of ions in a perpendicular magnetic field is proportional to the particle momentum per unit charge. If all particles entering a magnetic field have the same energy per unit charge, then the field will separate particles according to their masses. In the usual configuration, a magnetic field is produced between the two parallel, sector-shaped polefaces of an electromagnet, as illustrated in Figure 5.37. The focusing properties of this field<sup>28</sup> are the same as those of the spherical electrostatic analyzer, and the locations of the entrance and exit slits are given by Barber's rule (Section 5.4.3). The radius of curvature of ions of energy  $E$  in a magnetic field of intensity  $B$  is:

$$R = \frac{144}{Bn} \sqrt{mE} \text{ cm} \quad (5.99)$$

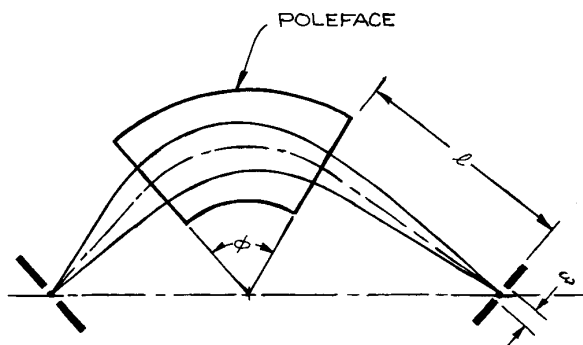


Figure 5.37 Focusing of the magnetic sector.

with  $B$  in gauss,  $E$  in eV,  $m$  in amu, and  $n$  the charge state. In terms of the parameters specified in Figure 5.37, the resolution is:

$$\frac{\Delta E}{E} = \frac{w}{R_0(1 - \cos \phi) + l \sin \phi} \quad (5.100)$$

where  $R_0$  is the radius of curvature of the central trajectory. To first order, there is no focusing in the plane perpendicular to the plane of deflection. When the curvature of the fringing field is taken into consideration, however, some focusing in the perpendicular plane results. Incident ion trajectories should be normal to the entrance plane to avoid focusing effects.

### 5.5.2 Wien Filter

The Wien filter, employing mutually perpendicular electric and magnetic fields normal to the ion trajectory [Figure 5.36(b)] can serve to disperse charged particles according to either mass or energy.<sup>27</sup> Ions of a unique velocity  $v$  will experience equal and opposite forces through interaction with the two fields (the Coulomb force and the Lorentz force) and be transmitted in a straight line. If all of the ions injected into the fields of the Wien filter have the same kinetic energy, ions of a unique mass are transmitted straight through; if the injected ions have all been accelerated through the same potential (relative to their source), ions of a unique mass-to-charge ratio are transmitted. In most Wien filters a pair of permanent magnets supplies the magnetic field, and the electric field is produced by placing an electrical potential  $V_d$  between a pair of parallel electrodes separated by a distance  $d$ . The velocity of ions transmitted straight through the Wien filter is

$$v = 10^8 \frac{V_d}{dB} \text{ cm/s}^1 \quad (5.101)$$

with  $V_d$  in volts,  $d$  in cm, and  $B$  in gauss. Varying  $V_d$  scans the mass spectrum of ions. If slits of width  $w$  are placed at the entrance and exit of a Wien filter of length  $L$ , the mass resolution for ions of energy  $E$  is:

$$\frac{\Delta m}{m} = \frac{2dEw}{qV_d L^2} \quad (5.102)$$

assuming the incident ion trajectories are normal to the entrance plane. Boersch, Geiger, and Stickel have made

a detailed analysis of the focusing properties of the Wien filter.<sup>29</sup>

### 5.5.3 Dynamic Mass Spectrometers

*Dynamic mass spectrometers* use time-varying electric or magnetic fields or timing circuits to disperse ions according to their masses.<sup>30</sup> The quadrupole mass analyzer and the linear time-of-flight mass spectrometer are the two most successful designs of this type.

The *quadrupole mass analyzer* illustrated in Figure 5.38 employs a time-varying electric quadrupole field. For a particular field intensity and frequency, only ions of a unique mass-to-charge ratio will follow a stable path and pass through the field. As shown in the figure, the quadrupole field is approximated by a square array of four cylindrical electrodes parallel to the axis along which ions are injected. The potential applied to the vertical pair of electrodes is:

$$V_v = U + V \cos 2\pi ft \quad (5.103)$$

while for the horizontal pair

$$V_h = -U + V \cos(2\pi ft + \pi) \quad (5.104)$$

where  $f$  refers to an r.f. frequency. For optimum performance the ratio of the d.c. field to the r.f. field is adjusted so

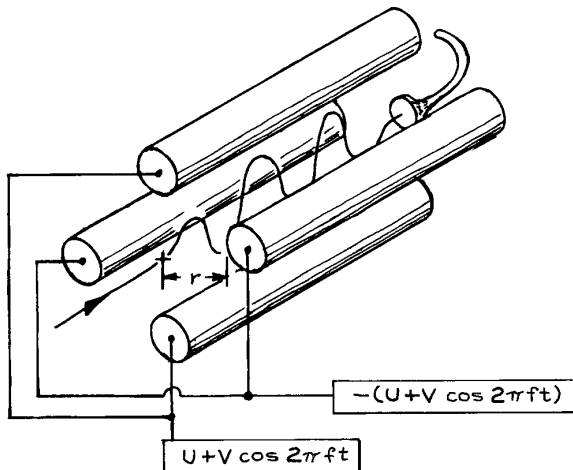


Figure 5.38 The quadrupole mass filter.

that  $U/V \approx 0.17$ . For singly charged ions, the mass of the ions that are transmitted is:

$$m = 0.14 \frac{V}{r^2 f^2} \text{ amu} \quad (5.105)$$

with  $V$  in volts,  $r$  in cm, and  $f$  in MHz. Mass scanning is usually accomplished by varying the r.f. amplitude  $V$ , while fixing the ratio of d.c.-to-r.f. voltage ( $U/V$ ) and the frequency ( $f$ ). Quadrupole analyzers are compact and offer the advantages of high transmission, fast scanning, and insensitivity to the initial ion energy. These instruments are available commercially as residual-gas analyzers for high-vacuum systems and are suitable for use in many scattering experiments.

The *time-of-flight mass spectrometer* depends upon the fact that the velocity of ions of the same kinetic energy is a function of mass. In this spectrometer, ions that have been accelerated to an energy of about 1 keV are admitted to a long drift tube in short bursts by a negative electrical pulse applied to a grid at the entrance of the drift region. The lightest ions travel most rapidly and are the first to arrive at the detector located at the end of the drift tube. Heavier ions arrive in the order of their masses. The obvious advantage of this design is that the entire mass spectrum is scanned in a few microseconds. Furthermore, the spectrum can be scanned thousands of times each second, making this system ideal for studying rapidly varying processes. Its chief drawback is its size, since the length of the flight tube required to achieve good resolution may be greater than a meter.

## 5.6 ELECTRON- AND ION-BEAM DEVICES: CONSTRUCTION

To realize the design of a charged-particle optical system, it is necessary to create an environment where a stream of particles can travel without loss of momentum and to make electrodes and pole pieces that faithfully produce the electrical and magnetic fields necessary to deflect the particles.

### 5.6.1 Vacuum Requirements

Charged particles lose energy through interactions with gas molecules; thus ions or electrons can only be transported in a vacuum. The required pressure depends upon the distance they must travel. As noted in Section 3.1.2, the mean free path at a pressure of 1 mtorr is a few centimeters,

while at  $10^{-5}$  torr this increases to a few meters. It follows that high vacuum is required for the operation of a charged-particle optical system.

Conducting surfaces must be kept clean, since contamination with an insulating material will result in the buildup of surface charges that cause an unpredictable deflection of charged particles passing nearby. Clean electrode surfaces are particularly important for particles of energy less than 100 eV and when high spatial resolution is required. All electrodes must be clean and free of hydrocarbon contamination before installation in the vacuum system (Section 3.6.3). An oil-free vacuum system with turbomolecular pumps, ion pumps, or sorption pumps (Section 3.4.3) is most desirable, although a system evacuated with a properly trapped oil diffusion pump (Sections 3.5.5 and 3.4.2) is adequate in many cases.

Polyphenyl ether diffusion-pump fluids such as Convalex-10 (Consolidated Vacuum) or Neovac Sy (Varian) have been found to be the least offensive diffusion-pump oils for electron-optical systems. In low-energy-beam devices, a daily bake to 200–300 °C will keep electrode surfaces clean and ensure stable operation. Hydrocarbons on aperture surfaces exposed to charged particles create a particularly obnoxious problem. Bombardment of the adsorbed hydrocarbons produces an insulating carbon polymer that adheres tenaciously to the underlying metal. This carbon material can only be removed by high-temperature baking (400 °C), by vigorous application of an abrasive, or by etching with a strong solution of NaOH.

When baking is impractical, electrode surface quality and stability are often improved by a coating of carbon black. A surface can be blacked by brushing with a sooty acetylene flame; however, the coating produced in this manner does not adhere well. A superior coating can be produced by spraying or brushing the surface with a thin water or ethanol slurry of colloidal graphite (Aquadag, made by Acheson Colloids Co.). The tenacity of this coating is improved by preheating the metal surface to about 100 °C before the slurry is applied.

## 5.6.2 Materials

The materials used to construct lens elements must be such that the equipotential surfaces near the electrodes faithfully follow the contours of the electrode surfaces. These surfaces

must be clean, and must withstand periodic cleansing by baking, bead blasting, electropolishing, or harsh chemical action.

The *refractory metals* such as tungsten, tantalum, and particularly molybdenum are probably the best electrode materials. These metals have a low, uniform, surface potential, they do not oxidize at ordinary temperatures, and they are bakeable. Refractory metals are hard and rather brittle. Only the wrought material can be machined or formed easily. Stock produced by sintering is very difficult to machine. Electrodes spun from sheet molybdenum are available in cylindrical and spherical shapes (Bomco, Inc.).

Oxygen-free high-conductivity (OFHC) *copper* (Section 1.2.5) is often used for electrode fabrication, although Kuyatt used commercial, half-hard copper in some applications. Exposed to air, copper forms a surface oxide, but this oxide is conducting. Copper is bakeable and reasonably machinable.

*Stainless steels* (Section 1.3.1) contain domains of different composition, some of which may be ferromagnetic. The magnetic properties of stainless steels vary considerably, even between samples of the same net composition; these materials are unsuitable for use with low-energy charged particles unless they are carefully annealed to remove residual magnetism (Section 1.3.1). Stainless steels have the advantage of being bakeable, strong, and easy to machine.

*Aluminum* (Section 1.3.4) is unsuitable, because its surface is rapidly attacked by oxygen in the air to form a hard, insulating, layer of oxide. On the other hand, aluminum has a low density, and most aluminum alloys are easy to machine or form by bending, spinning or rolling. When these properties are important, the surface can be plated with copper or gold to overcome the problem of oxidation.

Electrodes and lenses must be mounted on nonconducting materials. Glass or ceramic (Section 1.3.7) or, in some cases, plastic (Section 1.3.6) can be used for this purpose. Of the ceramic materials, *alumina* is the best insulator and the strongest. Precision-ground rods and balls are available commercially. Alumina circuit board substrate in the form of thin plates laser-cut in complex shapes can be obtained quite inexpensively. There are also machinable ceramics from which complicated shapes can be formed (MACOR from Corning). As noted in Chapter 3, it is important to be aware that the resistivity of a ceramic is strongly dependent upon temperature. The resistivity of a ceramic

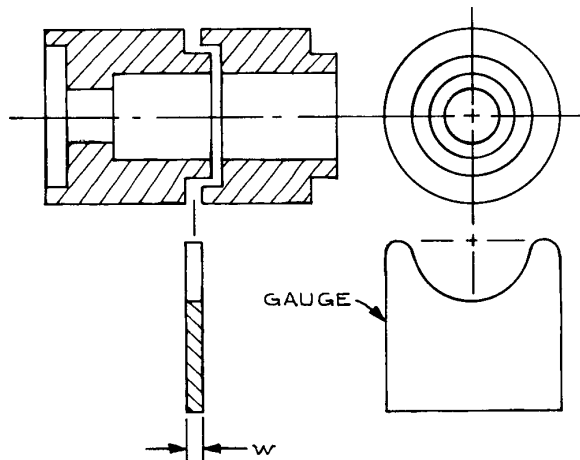
material falls by roughly a factor of 10 for every  $100^{\circ}\text{C}$  increase in temperature.<sup>31</sup> For example, the resistivity of alumina is greater than  $10^{17}$  ohm cm at  $100^{\circ}\text{C}$  and falls below  $10^7$  ohm cm at  $1000^{\circ}\text{C}$ .

*Plastics* are machinable and inexpensive. Unfortunately, they are not bakeable; a further disadvantage is that they contain hydrocarbons that can contaminate the vacuum environment. *Polyimide plastic sheet* (Kapton) is a useful insulator.

*Mica* in sheets is an excellent insulator (“stove mica” from Spruce Pine Mica Co.).

### 5.6.3 Lens and Lens-Mount Design

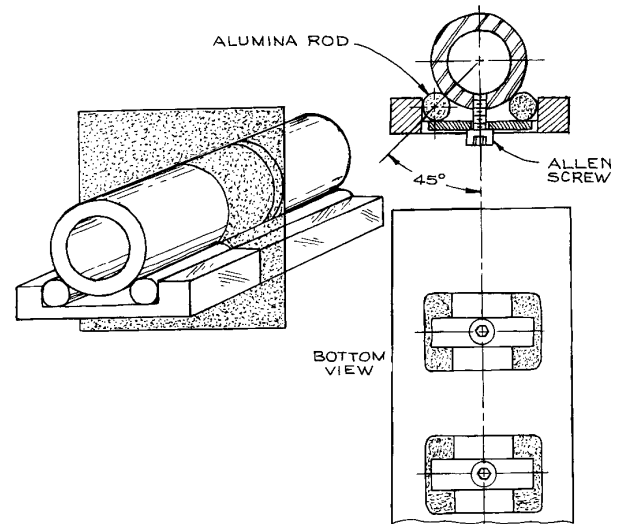
Most electron-optical lenses are created by the fields in the gap between coaxial cylindrically symmetric electrodes. These electrodes must be designed so that electric fields associated with the lens mounts and the vacuum-container walls do not penetrate the gap between electrodes. The involuted design illustrated in Figure 5.39 ensures that the lens gap is shielded. The step on the shoulder of the lens elements is designed so that the inner gap is of the correct size when the outer gap is adjusted to some standard width. In this way all the lenses in a system can be correctly positioned with the aid of a single gauge.



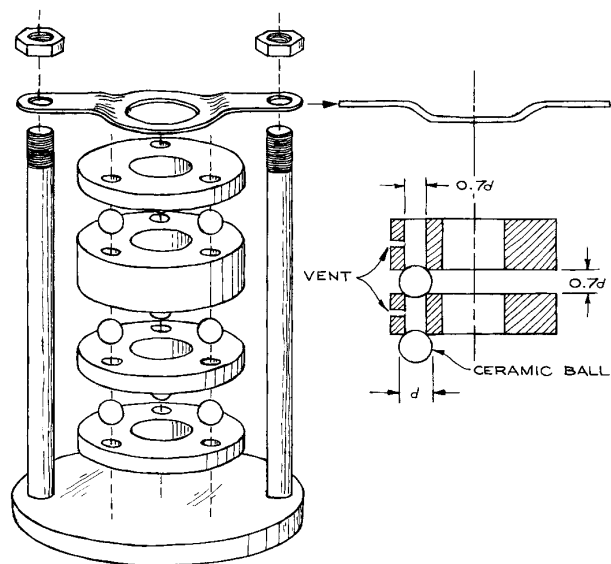
**Figure 5.39** Tube lens electrodes are constructed so that external fields cannot penetrate the lens gap. When mounting the lens elements, a gauge is inserted to ensure correct spacing at the invisible lens gap.

There are two widely used, semikinematic schemes for mounting lens elements. These are the rod mount and the ball mount, illustrated in Figures 5.40 and 5.41, respectively. In the *rod mount*, cylindrical lens elements rest on ceramic rods that insulate the elements from one another and from a grounded mounting plate. The critical dimensions are the diameters of the lens elements and the width of the channel in which the rods rest. In order for the lens elements to be coaxial, their outer diameters must be identical. This requirement is met by making all of the elements from a single piece of rod stock that has been carefully turned to a uniform diameter. To ensure that the elements are mounted coaxially, it is then only necessary that the sides of the channel in which the rods rest be parallel. This is easily accomplished in a milling operation. If the vertical position of the lens axis is important, then the width of the channel becomes a critical dimension. Of course, the diameter of the rods is important, but, as noted elsewhere (Section 1.3.6), alumina rod, centerless-ground to high precision, is commercially available. For maximum strength the rods should be about  $90^{\circ}$  apart around the circumference of the lens element.

In the *ball-mounting* scheme, lens elements are insulated from one another and positioned by ceramic balls



**Figure 5.40** Cylindrical lens elements mounted on ceramic rods.



**Figure 5.41** Lens elements mounted on ceramic balls.

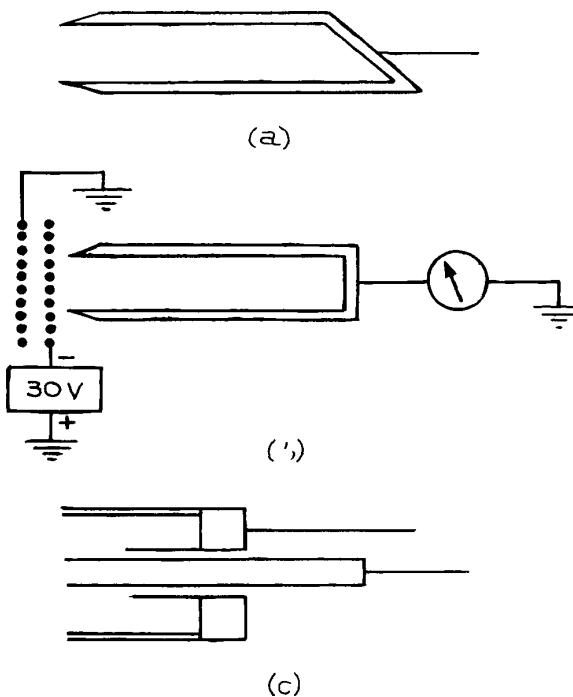
that rest in holes near the edges of the lenses. This system is preferred when mounting very thin elements, as in an aperture-lens system. Ideally only three balls should be used. The critical dimensions are the locations of the holes and their diameter. The holes should be bored, or drilled and reamed, in a milling machine or jig borer using a dividing head or a precision rotary table. For balls of diameter  $d$ , the hole diameter should be  $d \sin 45^\circ$ , in which case the spacing between lens elements will be  $d \cos 45^\circ$ . A lens system is assembled by stacking lens elements alternating with balls. The stack must be clamped. The clamp should have some spring and should bear on the topmost element only at one or two points near the center of the circle around which the balls are located. A rigid clamp will tend to drive the balls into their holes and may crush the edges of the holes. If the clamp bears too near the edge, there is a danger that the stack may become cocked off axis.

### 5.6.4 Charged-Particle Detection

For particle energies up to a few tens of keV, a stream of charged particles of sufficient intensity can be collected and measured directly as an electrical current. To achieve the greatest sensitivity, individual ions or electrons can be

detected with an electron multiplier. Charged particles can also be detected with photographic emulsions, scintillators, and various solid-state devices. Emulsions and scintillators have largely been supplanted by the methods mentioned above, and solid-state devices are only suitable for the detection of high-energy particles (above 30 keV).

The collector for the direct detection of a current of charged particles is called a *Faraday cup*. Typical designs are illustrated in Figure 5.42. These simple collectors are connected directly to a current-measuring device and are useful at currents down to the detection limits of modern electrometers about  $10^{-14}$  A. A properly-designed Faraday cup does not permit secondary electrons to escape. For a positive-ion collector, the loss of a secondary electron appears to the current-measuring instrument as an additional ion, while for an electron collector the loss of each secondary cancels the effect of an incident primary electron. To prevent the escape of secondary electrons, the depth of a



**Figure 5.42** Faraday-cup designs. Design (b) has a grid biased to suppress secondary electron emission. Design (c) is a double cup for aligning and focusing a beam.

Faraday cup should be at least five times its diameter. A suppressor aperture or grid biased to about  $-30$  V in front of a Faraday cup effectively prevents the escape of most secondaries. A grounded grid in front of the suppressor as illustrated in Figure 5.42(b) prevents field penetration from the suppressor in the direction of the incident current source. When a particle beam must be aligned and sharply focused, a concentric pair of collectors [Figure 5.42(c)] is useful. With this arrangement the beam-focusing elements are adjusted to maximize the ratio of current to the inner cup in relation to the current to the outer cup.

The use and application of electron multipliers is discussed in Chapter 7.

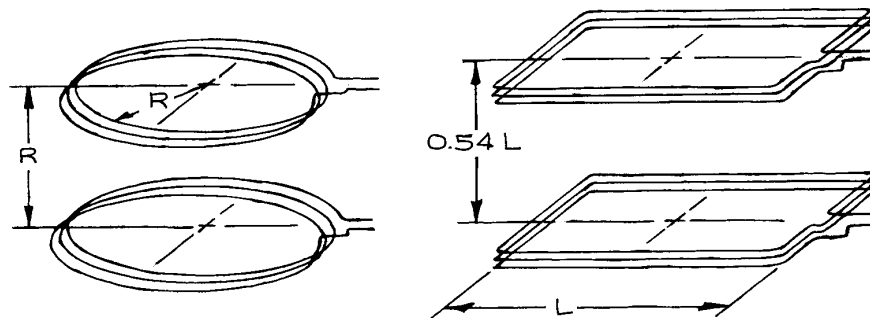
### 5.6.5 Magnetic-Field Control

A magnetic field may cause aberrations in a charged-particle optical system. The earth's magnetic field is the usual source of problems, but large concentrations of magnetic materials nearby or large currents, such as those associated with a cyclotron, may also produce magnetic fields of sufficient intensity to deflect the trajectories of electrons or ions in an optical system. The earth's magnetic field is about 0.6 gauss directed at an angle of elevation in the north-south plane roughly equal to the local latitude. Whether or not this is a problem depends on the energy and mass of the transmitted charged particles and the dimensions of lens elements and apertures through which the particles must pass. For example, the radius of curvature of the path of a 10 eV electron moving perpendicular to the earth's field is 15 cm; this amounts to a deflection of

1.5 mm along a 1 cm path. For a low-energy ( $<100$  eV) electron spectrometer, such considerations may lead to a residual-field tolerance of less than 1 milligauss.

There are two methods to eliminate magnetic fields: an opposing field can be applied to cancel the offending field throughout the volume occupied by the apparatus, or the apparatus can be enclosed in a shield of high permeability ferromagnetic metal that shunts the local field around the enclosure. An electromagnet constructed to make an opposing field is the least expensive, but also the most ungainly, approach; the dimensions of the coils may exceed by a factor of 10 the dimensions of the apparatus to be shielded. In addition, without an elaborate feedback control, an electromagnet can only cancel a constant (dipole) field. A magnetic shield need be only large enough to contain the apparatus, but the construction requires expensive fabrication and heat treatment of the shielding alloy that is itself relatively expensive.

Electromagnets for canceling the ambient field are assembled of two or more pairs of current-carrying loops. A continuous solenoid would be impractical, since access to the interior volume is limited. Caprari has surveyed optimal geometries for a range of possibilities.<sup>35</sup> The most common arrangement consists of a pair of identical circular magnet coils spaced apart by half the coil diameter. These are referred to as *Helmholtz pairs* or *Helmholtz coils*. The axis of the pair of coils must be oriented parallel to the field that is to be canceled. Alternatively, three pairs arranged orthogonally can be used with the current to each adjusted to give a net field that opposes the offending field. The Helmholtz coil geometry is illustrated in Figure 5.43.



**Figure 5.43** Round and square Helmholtz coils.

The field along the axis of a round Helmholtz pair, a distance  $z$  above the bottom coil is given by:

$$B = 0.32 \frac{NI}{R} \left\{ \left[ 1 + \left( \frac{z}{R} \right)^2 \right]^{-3/2} + \left[ 1 + \left( 1 - \frac{z}{R} \right)^2 \right]^{-3/2} \right\} \text{gauss} \quad (5.106)$$

where  $N$  is the total number of turns,  $I$  is in amperes, and  $R$  and  $Z$  are in cm. For square coils, a similar expression with  $R$  replaced by  $L/2$  applies. The orientation of the coils and the exact value of the current must be determined with the aid of a magnetometer. The field produced by a Helmholtz pair is of high uniformity only in a small volume midway between the coils and near the axis. For example, a uniformity of 0.1% is achieved in a central volume of  $2 \times 10^{-2}$  cubic radii, a uniformity of 1% in a volume of  $1.5 \times 10^{-1}$  cubic radii, and a uniformity of 5% in a volume of 0.6 cubic radii.<sup>34</sup> Thus, the coil dimensions must exceed those of the apparatus to be shielded by at least a factor of 10 in order to reduce the magnetic field to a few milligauss. In addition, it is important to recognize that a Helmholtz coil can only cancel the dipole component of the ambient field.

Shielding electron-optical devices from ambient magnetic fields can be accomplished with an enclosure fabricated of special high-permeability nickel-iron alloys.<sup>36</sup> These materials are available as sheet or tubing and thus as a practical matter are best formed into simple cylindrical shapes, either with or without endcaps, depending on requirements of access to the interior. Construction details are important in determining the performance of a magnetic shield. Removable partitions and endcaps should be designed to overlap the fixed pieces to which they are attached and they should fit tightly. Permanent joints should be spot-welded lap joints or arc-welded butt joints. Magnetic-shielding alloys only attain their high permeability after annealing in a hydrogen atmosphere following a precisely defined temperature schedule. Mechanical stress degrades the permeability of the material so the annealing is carried out after fabrication. It usually makes sense to have a shield fabricated by a shop that is prepared to do the annealing. Suppliers and fabricators of magnetic shielding materials include Amuneal Corp., Magnetic Metals Co., and the Magnetic Shield Division of Perfection Mica Co.

A magnetic shield must be handled carefully. Bending or sharp blows degrade the shielding properties, as will contact with magnetic materials; magnetized tools are frequently the culprits. The annealed material should not be heated above 400 °C. It must be emphasized that magnetic shielding cannot be accomplished casually; a few sheets of “μ-metal” wrapped around a vacuum system will often do more harm than good.

The performance of a magnetic shield is specified by an attenuation factor,  $A$ , given by the ratio of the magnetic field  $H_0$  at a point in the absence of the shield to the field strength  $H_s$  at the same point when it is surrounded by the shield:

$$A = \frac{H_0}{H_s} \quad (5.107)$$

The calculation of the attenuation factor is generally difficult, however, since a shield is usually constructed of sheet and tubing, a reasonable estimation can be made from approximations to the attenuation factor for simple geometric shapes. The attenuation of a long cylindrical shield can be approximated as that for an infinitely long cylinder and the attenuation of a closed box can be approximated as that for a sphere. The attenuation factor depends upon the thickness of the shielding material, the overall dimensions of the shield, and  $\mu$ , the permeability of the shielding material. Low-saturation alloys used to shield fields of the order of 1 gauss have permeabilities of about 40 000. From the formulae below it will be obvious that two or more nested shields of thin material are much preferable to a single thick shield.<sup>37</sup>

For a long cylinder, the attenuation factor is given approximately by:

$$A \approx \frac{\mu d}{2R} \quad (5.108)$$

and for a sphere by

$$A \approx \frac{2\mu d}{3R} \quad (5.109)$$

where  $R$  is the radius of the cylinder or sphere and  $d$  is the thickness of the material. For a nested pair of cylinders or spheres of radius  $R_1$  and  $R_2$  spaced apart by a distance  $\Delta = |R_1 - R_2|$ , the attenuation factor:

$$A \approx A_1 A_2 \frac{2\Delta}{R} \quad (5.110)$$



where  $A_1$  and  $A_2$  are the attenuation factors for each of the cylinders or each of the spheres alone and  $R$  is the mean radius of the two. Note that the attenuation increases with the spacing between the two shields.

A sample calculation illustrates the advantage of multiple shields: A single cylindrical shield, 20 cm in diameter of 1 mm thick material with a permeability of 40 000 gives an attenuation factor of 200. Nested shields 20 cm and 22 cm in diameter of the same material give an attenuation factor of 7000; for shielding from the earth's magnetic field, the double shield reduces the field well into the sub-milligauss range.

An opening in a magnetic shield is often unavoidable. For example a shield surrounding a vacuum system must be left open at one end for attachment of the pump and for electrical and mechanical feedthroughs. The external field penetrates into the opening. The attenuation near an opening reaches a value of about two-thirds of the maximum attenuation of the shield at a distance into the shield volume about equal to the mean radius of the opening.<sup>38</sup> A skirt of shielding alloy extending outward from an opening, thus effectively displacing the opening outward from the shielded volume, helps to alleviate the problem.

## Cited References

1. K. R. Spangenberg, *Vacuum Tubes*, McGraw-Hill, New York, 1948.
2. J. R. Pierce, *Theory and Design of Electron Beams*, 2nd edn., Van Nostrand, New York, 1954.
3. O. Klemperer and M. E. Barnett, *Electron Optics*, 3rd edn., Cambridge University Press, Cambridge, 1971.
4. V. E. Cosslett, *Introduction to Electron Optics*, Oxford University Press, Oxford, 1946.
5. D. Roy and J. D. Carette, "Design of Electron Spectrometers for Surface Analysis," in *Electron Spectroscopy*, H. Ibach, (Ed.), Topics in Current Physics, Springer, Berlin, 1977, Chapter 2.
6. V. W. Hughes and H. L. Schultz (Eds.), *Methods of Experimental Physics*, Vol. 4A, *Atomic Sources and Detectors* Academic Press, New York, 1967.
7. E. Harting and F. H. Read, *Electrostatic Lenses*, Elsevier, New York, 1976.
8. F. H. Read, *J. Phys.*, **E2**, 165, 1969; (b) F. H. Read, *J. Phys.*, **E2**, 679, 1969; (c) F. H. Read, A. Adams, and J. R. Sotomontiel, *J. Phys.*, **E4**, 625, 1971; (d) A. Adams and F. H. Read, *J. Phys.*, **E5**, 150, 1972; (e) A. Adams and F. H. Read, *J. Phys.*, **E5**, 156, 1972.
9. S. Natali, D. DiChio, E. Uva, and C. E. Kuyatt, *Rev. Sci. Instr.*, **43**, 80, 1972; D. DiChio, S. V. Natali, and C. E. Kuyatt, *Rev. Sci. Instr.*, **45**, 559, 1974.
10. D. W. O. Heddle, *Tables of Focal Properties of Three-Element Electrostatic Cylinder Lenses*, J.I.L.A. Report No. 104, University of Colorado, Boulder.
11. R. E. Collins, B. B. Aubrey, P. N. Eisner, and R. J. Celotta, *Rev. Sci. Instr.*, **41**, 1403, 1970.
12. D. DiChio, S.V. Natali, C. E. Kuyatt, and A. Galejs, *Rev. Sci. Instr.*, **45**, 566, 1974.
13. J. A. Simpson, "Electron Guns," in *Methods of Experimental Physics*, Vol. 4A, V. W. Hughes and H. L. Schultz (Eds.), Academic Press, New York, 1967, Section 1.15.
14. An estimation of image expansion at the space charge limit can be obtained from the data of W. Glaser, *Grundlagen der Elektronenoptik*, Springer, Vienna, 1952, p. 75.
15. R. G. Wilson and G. R. Brewer, *Ion Beams*, Wiley, New York, 1973.
16. C. D. Moak, H. E. Banta, J. N. Thurston, J. W. Johnson, and R.F. King, *Rev. Sci. Instr.*, **30**, 694, 1959; M. von Ardenne, *Tabellen der Electronenphysik Ionengrößen und Übermikroskopie*, Deutscher Verlag der Wissenschaften, Berlin, 1956.
17. W. Aberth and J. R. Peterson, *Rev. Sci. Instr.*, **38**, 745, 1967.
18. J. A. Simpson, *Rev. Sci. Instr.*, **32**, 1283, 1961.
19. M. E. Rudd, "Electrostatic Analyzers," in *Low Energy Electron Spectrometry*, by K. D. Sevier, Wiley-Interscience, New York, 1972, Chapter 2, Section 3; A. Poulin and D. Roy, *J. Phys.*, **E11**, 35, 1978.
20. T. S. Green and G. A. Proca, *Rev. Sci. Instr.*, **41**, 1409, 1970; G. A. Proca and T. S. Green, *Rev. Sci. Instr.*, **41**, 1778, 1970.
21. V. V. Zashkvara, M. I. Korsunskii, and O. S. Kosmachev, *Soviet Phys. Tech. Phys.*, **11**, 96, 1966; H. Sar-el, *Rev. Sci. Instr.*, **38**, 1210, 1967, and **39**, 533, 1968; J. S. Risley, *Rev. Sci. Instr.*, **43**, 95, 1972.
22. C. E. Kuyatt and J. A. Simpson, *Rev. Sci. Instr.*, **38**, 103, 1967.
23. C. E. Kuyatt, unpublished lecture notes; see also A. J. Williams, III, and J. P. Doering, *J. Chem. Phys.*, **51**, 2859, 1969; J. H. Moore, *J. Chem. Phys.*, **55**, 2760, 1971.
24. R. F. Herzog, *Z. Phys.*, **89**, 447, 1934; **97**, 596, 1935; *Phys. Z.*, **41**, 18, 1940.
25. H. Wollnik and H. Ewald, *Nucl. Instr. Meth.*, **36**, 93, 1965.
26. A. Stamatovic and G. J. Schulz, *Rev. Sci. Instr.*, **39**, 1752 (1968); A. Stamatovic and G. J. Schulz, *Rev. Sci. Instr.*, **41**, 423 (1970); L. Sanche and G. J. Schulz, *Phys. Rev. A*, **5**, 1672 (1972); D. Roy, *Rev. Sci. Instr.*, **43**, 535 (1972).

27. R. L. Seliger, *J. Appl. Phys.*, **43**, 2352, 1972.
28. E. Segre, *Experimental Nuclear Physics*, Vol. 1, Wiley, New York, 1953, part V.
29. H. Boersch, J. Geiger, and W. Stickel, *Z. Phys.*, **180**, 415 (1964).
30. E. W. Blauth, *Dynamic Mass Spectrometers*, Elsevier, Amsterdam, 1966; P. H. Dawson, *Quadrupole Mass Spectrometer*, Elsevier, Amsterdam, 1976.
31. W. D. Kingery, H. K. Bowen, and D. R. Uhlmann, *Introduction to Ceramics*, 2nd edn., John Wiley & Sons, Inc., New York, 1976, chapter 17.
32. J. Millman and H. Taub, *Pulse, Digital, and Switching Waveforms*, McGraw-Hill, New York, 1965, Chapter 3; C. N. Winningstad, *IRE Trans. Nucl. Sci.*, **NS43**, 26, 1959; C. L. Ruthroff, *Proc. IRE*, **47**, 1337, 1959.
33. L. J. Richter and W. Ho, *Rev. Sci. Instr.*, **57**, 1469(1986); R. M. Tromp, M. Copel, M. C. Reuter, M. Horn v. Hoegen, J. Speidell, and R. Koudijs, *Rev. Sci. Instr.*, **62**, 2679 (1991).
34. R. K. Cacak and J. R. Craig, *Rev. Sci. Instr.*, **40**, 1468, 1969.
35. R. Caprari, *Meas. Sci. Technol.*, **6**, 593 (1995).
36. *The Definitive Guide to Magnetic Shielding*, Amuneal Manufacturing Corp., 4737 Darrah Street, Philadelphia PA 19124.
37. V. Schmidt, *Electron Spectrometry of Atoms using Synchrotron Radiation*, Cambridge University Press, Cambridge, 1997, pp. 403–407.
38. W.G. Wadey, *Rev. Sci. Instr.*, **27**, 910, 1956.

# ELECTRONICS

---

This chapter discusses electronics at a level somewhere between that of a handbook, which consists essentially of charts, tables, and graphs, and a textbook, where the interesting, important, and useful conclusions come only after well-developed discussions with examples. The aim here is a presentation that has sufficient continuity and readability that individual sections can be profitably read without having to refer to preceding sections or other texts. On the other hand, it is important to have useful and frequently referenced material in the form of readily accessible tables, graphs, and diagrams that are sufficiently self-explanatory that very little reference to the text material is necessary. Another important goal is vocabulary. A large amount of jargon in electronics is meaningless to the uninitiated, but when it is necessary to understand the properties of an electronic device from a written technical description, when writing the specifications for electronic equipment, or when talking to an electronics engineer, salesman, or technician, this vocabulary is essential. With this in mind, terms not current outside of electronics are italicized.

To be used to best advantage, this chapter should be supplemented with manufacturers' catalogs, data books, applications texts, handbooks, and more specialized texts that treat the topic of interest in depth. Manufacturers of laboratory electronic equipment, discrete devices, and integrated circuits have publications that describe, in clear practical terms, the properties of their products and their applications to a wide variety of tasks. Much of this material is also available on the internet, and for this reason internet addresses are given when available.

The material has been organized and written as one explains it to a student or technician coming to work in a

laboratory for the first time. The complexity of modern electronics is such that the cut-and-try approach is too inefficient and costly in terms of material and time. There are just too many possibilities when connecting devices and multiple-component circuits, and it is important to establish a systematic approach based on a limited number of simple, well-understood principles. It is probably not reasonable in the laboratory to expect quick solutions to problems that are entirely outside one's previous experience. The number of really new situations that can arise is limited, however, most problems being variations on a few basic situations. The ability to recognize this and to isolate the source of difficulty comes with experience and mastery of basic principles. When confronted with a new situation involving rack upon rack of equipment, the tendency is to believe that an understanding of how everything works is beyond the capabilities of all but experienced electronics engineers. This is far from the truth. At the operational level, present-day electronics is the most reliable, easy-to-use, and easy-to-understand element of most experiments.

## 6.1 PRELIMINARIES

### 6.1.1 Circuit Theory

An understanding of elementary circuit theory and the accompanying vocabulary permits one to reduce complex circuits consisting of many elements to a few essentials, predict the behavior of complex circuits, specify the operation of components, and understand and use data sheets

and operation manuals. In routine laboratory work, it is not necessary to be skillful with circuit theory. It is necessary to be able to isolate the basic elements of a circuit and understand their behavior. With that ability, when circuits fail to operate correctly, the causes of the malfunction can be localized and repaired.

*Linear circuit theory* applies to devices whose output is directly proportional to the applied input. If one increases the current through a resistor by a factor of two, for example, the voltage across it will double. An example of a non-linear device is a switch that is either open or closed and whose state changes abruptly at a threshold. A nonlinear device can often be treated with linear theory by dividing the response of the device into separate regions over which it behaves in a quasi-linear manner. This is called *linearizing the response curve*. An example is the piecewise linearization of a diode's current-voltage response, as shown in Figure 6.1. The exponentially rising forward current and constant reverse current are represented by straight lines of slope  $1/R_f$  and  $1/R_r$ , which are joined at voltage  $V_T$ .

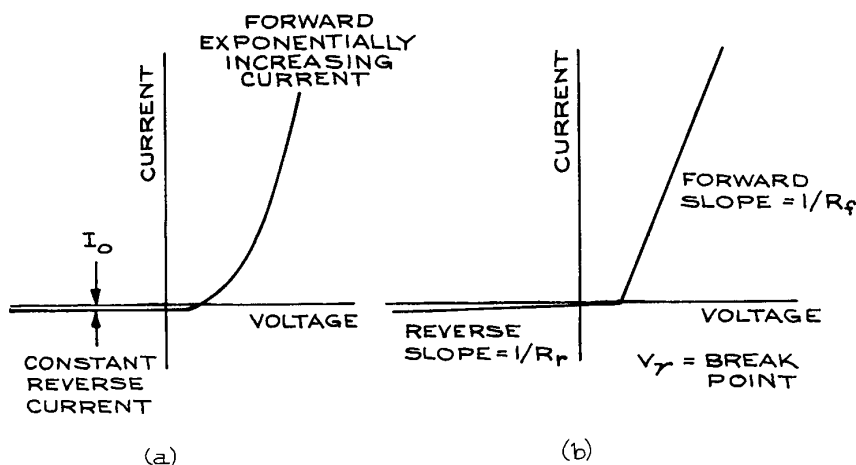
We begin by considering only *passive* linear devices; that is, devices that either dissipate energy (resistors) or store energy in electric (capacitors) or magnetic (inductors, transformers) fields. *Active* devices, such as transistors, can supply energy to a circuit when appropriately powered by external sources. The analysis of circuits with active devi-

ces is based on representations using equivalent circuits consisting of passive devices.

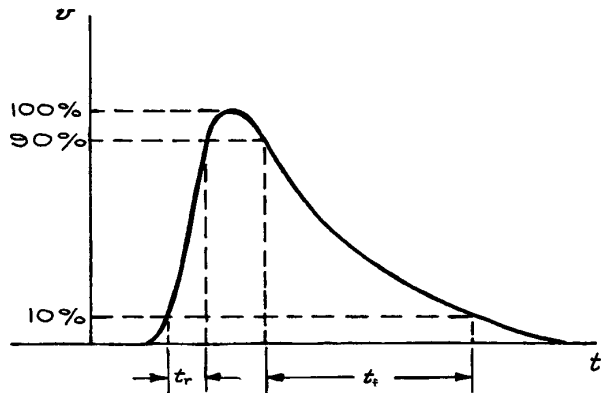
Conventional circuit analysis uses three *lumped* circuit elements – resistors ( $R$ ), capacitors ( $C$ ), and inductors ( $L$ ). This way of analyzing circuits is valid at signal frequencies  $f$  for which the wavelength  $\lambda$  is much larger than the physical dimensions of the circuit. Since  $\omega = c/f$ , where  $c$  is the speed of light, this means that analysis in terms of lumped elements is valid up to frequencies of a few hundred megahertz.

This frequency limitation also excludes waveforms with significant frequency components above a few hundred megahertz, even though the repetition rate of the waveform may be much less. A convenient way to estimate the frequency of the highest-frequency component of a nonsinusoidal waveform is to divide 0.3 by the *rise time* of the waveform,  $t_r$ , defined as the time between the 10% and 90% amplitude points on the leading edge of the waveform. A pulse with a 10 ns rise time, for example, has significant frequency components up to 30 MHz. The *fall time*,  $t_f$  of a waveform is the time between the 10% and 90% amplitude points on the trailing edge. Figure 6.2 illustrates *rise time* and *fall time*.

Even at low frequencies there are no ideal resistors, capacitors, or inductors. Actual resistors have some capacitance and inductance, while capacitors have



**Figure 6.1** (a) Real and (b) piecewise linear representation of the current-voltage characteristics of a diode.



**Figure 6.2** Rise time ( $t_r$ ) and fall time ( $t_f$ ) of a pulse.

resistance and inductance, and inductors have resistance and capacitance. These departures from ideality are largely a matter of construction.

At high frequencies, stray capacitances and inductances become significant, and one commonly speaks of *distributed* parameters, in contrast to the lumped parameters at low frequencies. Coaxial cable is an example of a type of distributed parameter circuit. The electrical properties of coaxial cable are normally given in terms of resistance and attenuation per unit length as a function of frequency. At high frequencies, the resistance of conductors (even connecting wires) increases due to what is termed *skin effect*. The magnitude of this effect for round cross-section wires is given in Table 6.1 as a function of frequency. High-frequency connections are best made with leads having a large surface area-to-volume ratio, with flat-ribbon geometry being the best.

Conventional circuit theory is based on a few laws, principles, and theorems. In the equations that follow,

**Table 6.1** Ratio of a.c. to d.c. wire resistance

Wire Gauge	$R_{ac}/R_{dc}$			
	$10^6$ Hz	$10^7$ Hz	$10^8$ HZ	$10^9$ HZ
#22	6.9	21.7	69	217
#18	10.9	34.5	109	345
#14	17.6	55.7	176	557
#10	27.6	87.3	276	873

lowercase letters represent instantaneous values of voltage and current, whereas uppercase letters indicate effective or d.c. values. It is also convenient to distinguish between root-mean-square (rms), peak-to-peak, and average values of voltage and current for sinusoidally varying voltages. If  $v = V \cos \omega t$ , where  $v$  is the instantaneous value of voltage and  $V$  is the peak value, the rms value is  $V/\sqrt{2}$ , the peak-to-peak value is  $2V$ , and the average value is clearly zero. This is illustrated in Figure 6.3. Common US line voltage is specified as 110 V a.c., which is the rms value. The peak voltage is 156 V, so the peak-to-peak voltage is 312 V.

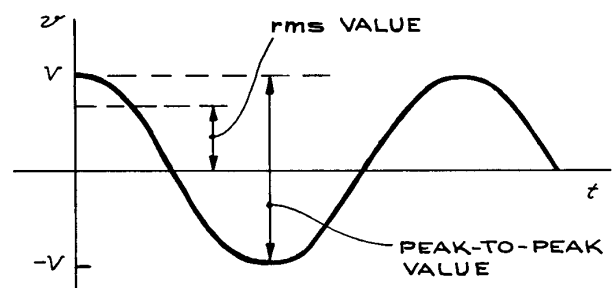
Under certain conditions, the rms value is not sufficient for specifying the output of an a.c. source. A source producing voltage spikes of large amplitude, but short duration superimposed on a small sinusoidally varying voltage will have an rms value very close to that without the spikes, but the spikes can have a large effect on circuits connected to the source. When specifying the output of a d.c. power supply, the magnitude, frequency, and duration of nonsinusoidal waveforms that appear at the output need to be specified, as well as the rms value of any a.c. component of the output.

## Laws.

**(i) Current–voltage relations.** For resistors, capacitors, and inductors we have:

$$v_R = iR, \quad v_C = \frac{1}{C} \int i dt, \quad v_L = L \frac{di}{dt} \quad (6.1)$$

respectively, where the resistance  $R$  is in ohms; the capacitance  $C$  in farads; and the inductance  $L$  in henrys.



**Figure 6.3** Relation of rms and peak-to-peak voltages for a sinusoidal waveform.

(ii) **Loops and nodes (Kirchhoff's laws).** In these, the sums are algebraic (signs taken into account):

- (1)  $\Sigma$  (voltage drops around a closed loop) =  $\Sigma$  (voltage sources).
- (2)  $\Sigma$  (current into a node) =  $\Sigma$  (current out of a node), where a node is a point where two or more elements have a common connection.

## (b) Theorems

(i) **Thevenin's theorem.** A real voltage source in a circuit can always be replaced by an ideal voltage source in series with a generalized resistance. An ideal voltage source is one that can maintain a constant voltage across its terminals regardless of the load across it. In other words, an ideal voltage source has zero internal resistance. An automobile battery, with an internal resistance of a few hundredths of an ohm, is a good approximation to an ideal voltage source at currents of a few amperes. Electronically regulated power supplies often have very low effective internal resistances when operated within their voltage and current ratings.

(ii) **Norton's theorem.** A real current source in a circuit can always be replaced by an ideal current source shunted by a generalized resistance. An ideal current source is one that supplies a constant current regardless of load – such a source has an infinite internal resistance. Photo-multiplier and electron-multiplier devices provide currents, albeit very low, that are almost independent of load, and they approximate ideal current sources.

## Superposition, Circuits with Multiple Sources.

For a circuit that contains several sources (voltage, current, or a combination of both) the contribution of each source to the voltage between any two points or the current past a point can be considered separately with all the other sources represented by their internal resistances. The total voltage or total current is then the algebraic sum of the separate contributions of each of the individual sources.

When connecting a source of current or voltage to a circuit, it is often important to know the internal resistance of the source. This can be determined by first measuring the open-circuit voltage of the source with a high internal-

resistance voltmeter and then connecting a variable resistance across the source and adjusting it until the voltmeter reading is one half the open-circuit value. The source resistance is then equal to the value of the variable-resistance setting. If the source has a very high internal resistance, a current measurement can be substituted for the voltage measurement. In this case, the output is shunted with an ammeter and the so-called *short-circuit current* is measured. A variable resistance is then placed in series with the ammeter and adjusted until the current through the ammeter is one half the short-circuit current. The value of the variable resistor at this point is equal to the internal resistance of the source. Analogous measurements can be made for a.c. sources by using either a.c. voltmeters and ammeters or an oscilloscope.

## 6.1.2 Circuit Analysis

For any given source, the choice of representation (Thevenin or Norton) is arbitrary and, in fact, the series resistance in the Thevenin representation is exactly equal to the parallel resistance in the Norton representation. Thevenin's and Norton's theorems simplify the application of the laws and principles discussed above.

The most general method for solving circuit problems is to apply Kirchhoff's laws using the appropriate current-voltage relations for each element in the circuit. This gives rise to one or more linear differential equations, which, when solved with the proper boundary conditions, give the general solution. This is illustrated for RC circuits in Section 6.1.3.

When dealing with sinusoidal sources of angular frequency  $\omega$ , circuit analysis can be greatly simplified when only the steady-state solution is required. In this case, circuit capacitances and inductances are replaced by *reactances*:

$$\begin{aligned} \text{capacitive reactance} &= jX_C, \quad \text{where } X_C = \frac{-1}{\omega C} \\ \text{inductive reactance} &= jX_L, \quad \text{where } X_L = \omega L \end{aligned} \quad (6.2)$$

$$j = \sqrt{-1}$$

The *impedance*  $Z$  of a circuit is obtained by combining reactances and resistances according to the formula

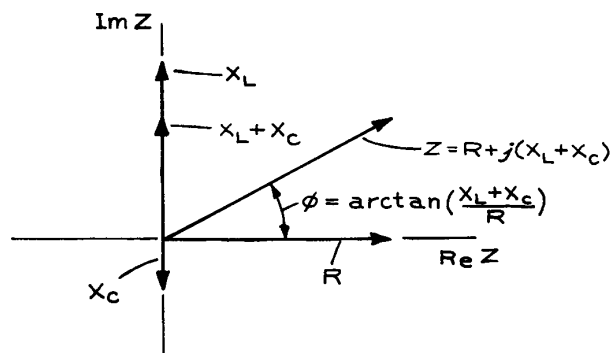
$Z = R + j(X_L + X_C)$ . These quantities can be represented in the complex plane by vectors (see Figure 6.4). The angle between  $Z$  and the real axis is  $\phi$ . By analogy with the  $I$ - $V$  relations for a pure resistance,  $X_C$  is the ratio of the a.c. voltage across a capacitor to the current through it,  $X_L$  is the ratio of the a.c. voltage across an inductor to the current through it, and  $Z$  is the net ratio of a.c. voltage to current in a circuit composed of resistors, capacitors, and inductors.

The fact that  $jX_C$  and  $jX_L$  are imaginary means that the voltage and current are  $90^\circ$  out of phase with each other. For a capacitor, the voltage lags the current by  $90^\circ$ , while for an inductor the voltage leads the current by  $90^\circ$ .

Another quantity that is occasionally useful in circuit analysis is the *complex admittance*  $Y$ , which is the reciprocal of the impedance. The SI unit of admittance is the *siemen*. The usefulness of the admittance arises in circuits with several parallel branches, where the net admittance is the sum of the admittances of the branches.

In carrying out circuit analysis, the following results of the above laws are useful:

**Series Circuits.** At any instant the current is the same everywhere in a series circuit, and the algebraic sum of the voltage drops around a circuit equals the algebraic sum of the sources. For circuit elements of impedance  $Z_1, Z_2, \dots, Z_N$  in an  $N$  element series circuit, the total impedance is  $Z = Z_1 + Z_2 + \dots + Z_N$ . If all the elements are resistors,



**Figure 6.4** Relations between reactance, resistance, impedance, and phase angle.

or inductors, or capacitors, the general expression reduces, respectively, to:

$$\begin{aligned} R &= R_1 + R_2 + \dots + R_N \\ \frac{1}{C} &= \frac{1}{C_1} + \frac{1}{C_2} + \dots + \frac{1}{C_N} \\ L &= L_1 + L_2 + \dots + L_N \end{aligned} \quad (6.3)$$

**Parallel Circuits.** For circuit elements in parallel, the voltage drop across each branch is the same while the current through each branch is inversely proportional to the impedance of the branch. The total current through all of the branches is the voltage across the network divided by the equivalent impedance for the network. The equivalent impedance  $Z$  and admittance  $Y$  for an  $N$  branch parallel circuit are:

$$\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2} + \dots + \frac{1}{Z_N} \quad (6.4)$$

and:

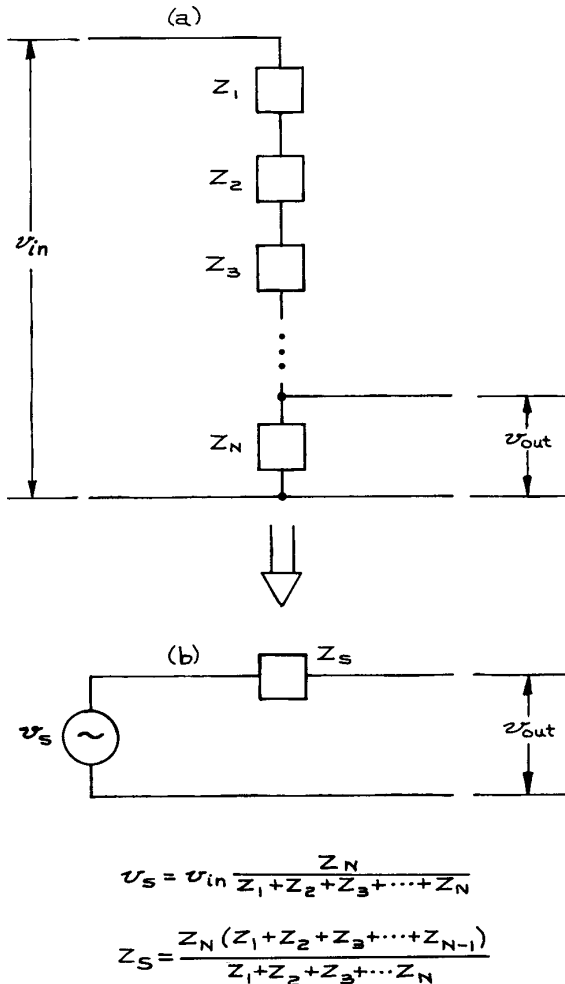
$$Y = Y_1 + Y_2 + \dots + Y_N \quad (6.5)$$

where  $Z_1, Z_2, \dots, Z_N$  are the impedances of the branches and  $Y_1, Y_2, \dots, Y_N$  are the admittances. In the special cases where all the circuit elements in the branches are of the same type:

$$\begin{aligned} \frac{1}{R} &= \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_N} \\ C &= C_1 + C_2 + \dots + C_N \\ \frac{1}{L} &= \frac{1}{L_1} + \frac{1}{L_2} + \dots + \frac{1}{L_N} \end{aligned} \quad (6.6)$$

where  $R, C,$  and  $L$  are the net resistance, capacitance, and inductance of the circuits.

**Voltage Dividers.** The *voltage divider*, illustrated in Figure 6.5(a), is a very common circuit element. The instantaneous voltage across  $Z_N$  is  $v_{in}[Z_N/(Z_1 + Z_2 + Z_3 + \dots + Z_N)]$ ; that is, the fraction of  $v_{in}$  that appears across any circuit element is the impedance of that element divided by the total impedance of the series circuit. Voltage dividers provide a convenient way to obtain a variable-voltage output from a fixed-voltage input, but there are



**Figure 6.5** (a) The voltage divider; (b) the Thevenin equivalent.

limitations. To avoid drawing too much current from the voltage source, the impedance of the voltage divider string should not be too small. If, in the interest of conserving power, however, the impedance is made large, the output impedance of the circuit will be large and  $v_{out}$  will depend critically on the load. This can be seen from the Thevenin equivalent of the circuit given in Figure 6.5(b) where  $v_s$  is the instantaneous voltage of the ideal voltage source. When  $Z_s$  is large, the voltage across a load will depend critically on the value of the load. Such *loading* of a divider

is to be avoided. For noncritical applications,  $Z_s$  should be at most 1/10 of any anticipated load.

Precision, highly linear, multiturn potentiometers are commonly used for position sensing. In this application, the shaft of the potentiometer is coupled mechanically to the moveable element whose position is to be determined, and a stable voltage source is connected across the ends of the potentiometer. The ratio of the voltage from the variable contact of the potentiometer to its end gives the angle through which the shaft has rotated.

**Equivalent Circuits.** Two circuits are *equivalent* if the relationships between the measurable currents and voltages are identical. As has been seen, a circuit with two external terminals can be replaced by its Thevenin or Norton equivalent. Common equivalence transformations for circuits with three terminals (*Miller* and *Y-Δ transformations*) are shown in Figures 6.6 and 6.7. In the first Miller transformation of Figure 6.6, it is necessary to know the ratio of the voltages at nodes 1 and 2; in the second it is necessary to know the ratio of the currents into nodes 1 and 2.

Discussions of amplifier circuits often refer to the *Miller effect*. This occurs when the input and output circuits of the amplifier are coupled by an impedance  $Z'$ . By using the transformation in Figure 6.6,  $Z'$  between input terminal 1 and output terminal 2 can be transformed into an equivalent circuit where  $Z'$  is replaced by  $Z_1$  and  $Z_2$  from terminals 1 and 2 to ground. The relations between  $Z_1$ ,  $Z_2$ , and  $Z'$  are given in the figure. When the coupling between terminals 1 and 2 is capacitive with  $Z' = 1/j\omega C'$ ,  $Z_1$  and  $Z_2$  are also capacitive impedances equal to  $1/j\omega C'(1 - K)$  and  $1/j\omega C'K/(K - 1)$ , where  $K$  is the voltage gain of the amplifier, negative for the example shown in Figure 6.6.

The Y-Δ transformation allows one to transform a circuit of three elements from a node to a loop configuration and from a loop to a node configuration.

### 6.1.3 High-Pass and Low-Pass Circuits

Analysis of the *high-pass* and *low-pass* circuits shown in Figure 6.8 illustrates some of the above circuit-analysis principles. The combination of  $v_s$  and  $R_s$  represents a real voltage source with instantaneous open-circuit voltage  $v_s$  and internal resistance  $R_s$ . For the high-pass or *differentiating circuit*, the output voltage  $v_o$  is across the resistor  $R$ ;



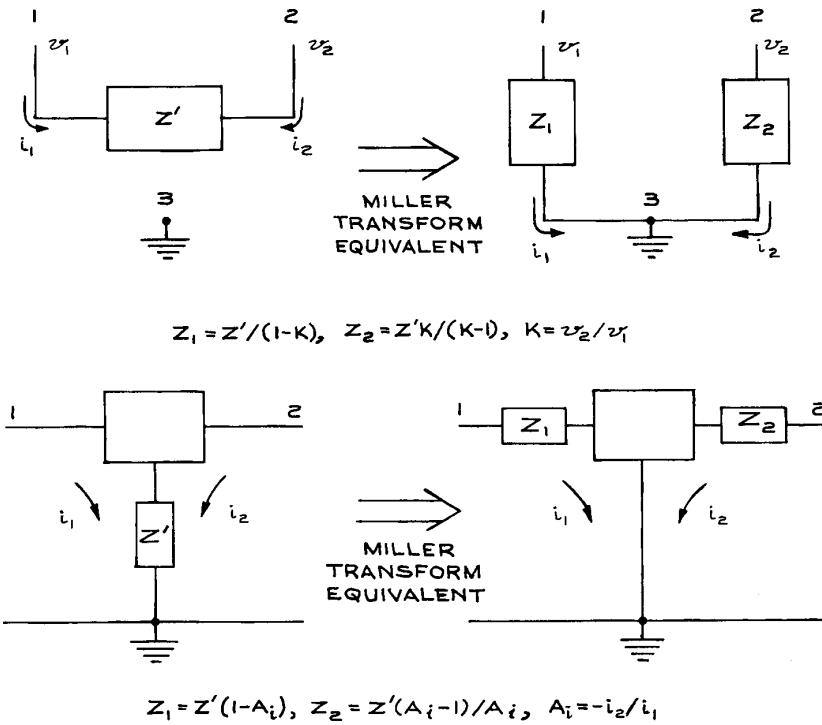


Figure 6.6 Miller transformations for circuits with three terminals.

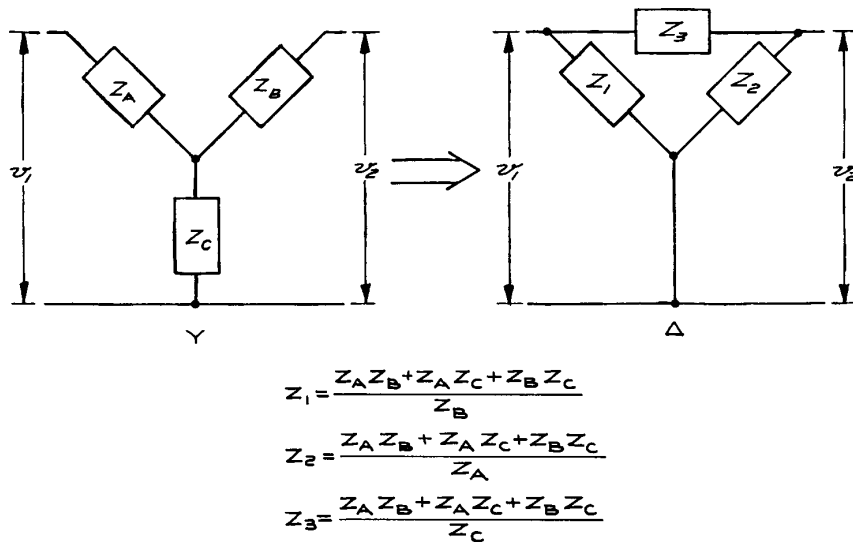
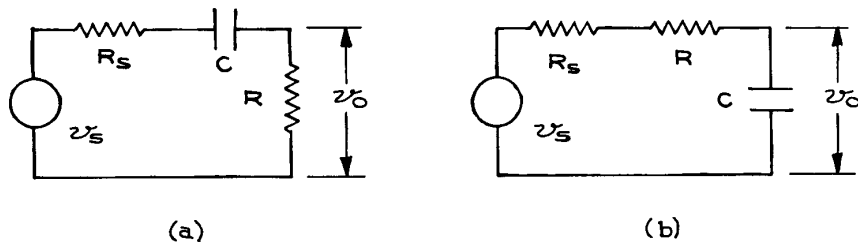


Figure 6.7 Y-Δ transformation for a circuit with three terminals.



**Figure 6.8** (a) High-pass or differentiating circuit; (b) low-pass or integrating circuit.

for the low-pass or *integrating circuit*, it is across the capacitor  $C$ . Very often, the essential properties of complex circuits can be understood in terms of one of these two circuits, so it is useful to be acquainted with their characteristics.

These circuits can be analyzed by the *differential-equation method*. For either circuit:

$$v_s(t) = iR_s + \frac{1}{C} \int i dt + iR \quad (6.7)$$

This equation uses the fact that the sum of the voltage drops in the circuit equals the sum of the voltage sources and the current is the same everywhere in a series circuit at any instant. Differentiating with respect to time:

$$\frac{dv_s}{dt} = \frac{di}{dt}(R_s + R) + \frac{i}{C} \quad (6.8)$$

The solution to the homogeneous equation ( $dv_s/dt = 0$ ) is:

$$i = Ae^{-t/R'C} \quad (6.9)$$

where  $R' = R_s + R$  and  $A$  is the integration constant determined from the initial conditions. The general solution requires that the functional form of  $v_s$  be known. Consider three cases:

(1) An a.c. voltage of amplitude  $V$ :

$$v_s = V \cos(\omega t + \phi) \quad (6.10)$$

(2) A step voltage of amplitude  $V$ :

$$v_s = \begin{cases} 0 & \text{for } t < 0 \\ V & \text{for } t > 0 \end{cases} \quad (6.11)$$

(3) A rectangular pulse of amplitude  $V$  and duration  $T$ :

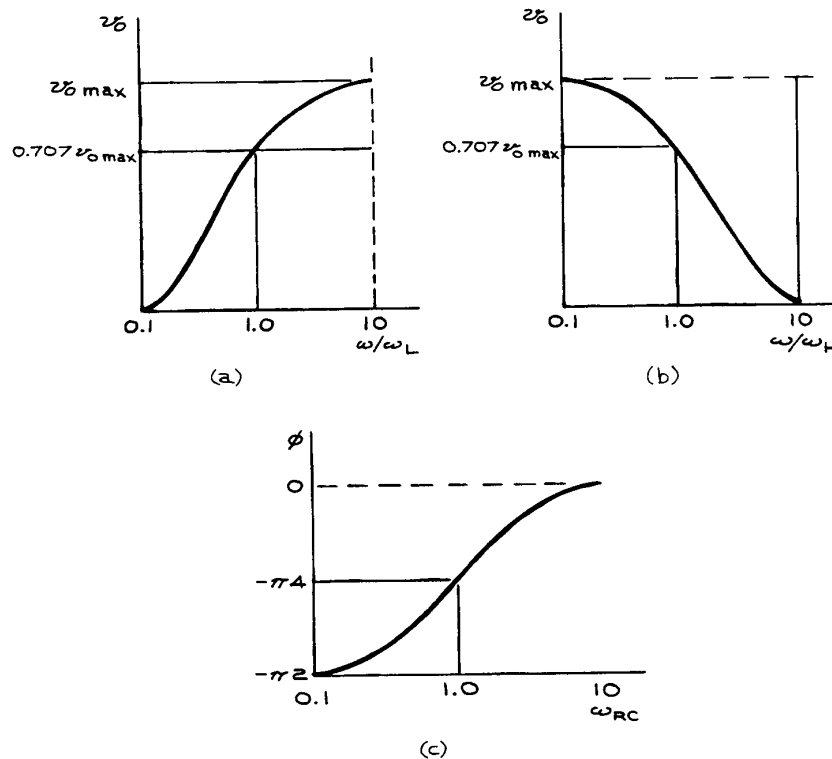
$$v_s = \begin{cases} V & \text{for } 0 \leq t \leq T \\ V & \text{for } t < 0, t > T \end{cases} \quad (6.12)$$

For case 1, the output voltage is sinusoidal at the same frequency as the input voltage. The ratio of  $v_o$  to  $v_s$  as a function of normalized frequency is shown in Figure 6.9(a). At the frequencies  $\omega = \omega_H$  and  $\omega = \omega_L$  for the two circuits,  $v_o$  is  $1/\sqrt{2}$  of the maximum value. These frequencies are called the *upper* and *lower corner frequencies*, respectively.

The maximum power that can be delivered to a load is proportional to the square of the output voltage, so that at  $\omega = \omega_H$  and  $\omega = \omega_L$ , the maximum power that the circuits can deliver to a constant load is one-half the maximum possible value. The usual way of expressing this is in terms of *decibels* dB, where:

$$\begin{aligned} \text{ratio in dB} &= 10 \log_{10} \left( \frac{\text{power out}}{\text{power in}} \right) \\ &= 10 \log_{10} \left( \frac{v_{\text{out}}^2/R_{\text{out}}}{v_{\text{in}}^2/R_{\text{in}}} \right) \end{aligned} \quad (6.13)$$

If  $R_{\text{out}} = R_{\text{in}}$ , which is often assumed, then (ratio in dB) =  $20 \log_{10}[v_{\text{out}}/v_{\text{in}}]$ . When  $v_{\text{out}}/v_{\text{in}} = 1/\sqrt{2}$ , this is approximately  $-3$ , so that  $-3$  dB represents a power reduction of a factor of two. Since the frequency response of amplifiers,



**Figure 6.9** Output voltage as a function of frequency for: (a) high-pass and (b) low-pass circuits; phase as a function of frequency for: (c) high-pass and (d) low-pass circuits.

filters, and transducers is routinely given in dB, it is important to keep in mind that the dB scale is logarithmic. Human sensory perception is approximately logarithmic, and a 3 dB change in sound level or light level is barely perceptible.

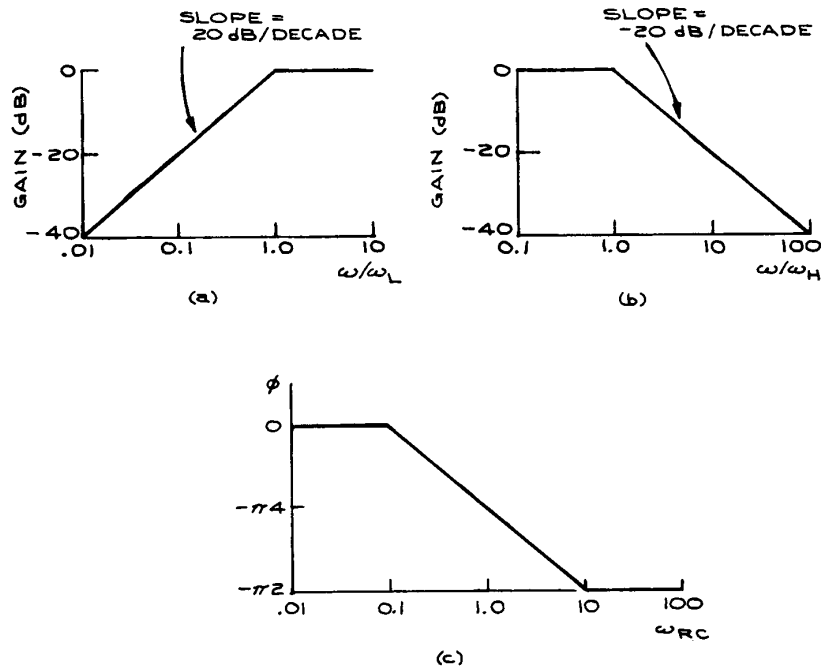
Because of the reactive element in  $RC$  circuits (the capacitor), the voltage is not in phase with the current, as illustrated in Figure 6.9(b). These plots of phase and log (output voltage) as a function of log(frequency) are called *Bode plots*, after H. W. Bode.<sup>1</sup>

It is often convenient to approximate frequency-response curves by a piecewise linear function. Such idealized Bode plots are shown in Figure 6.10(a) and (b). The *corner frequencies* are where  $\omega/\omega_L$  and  $\omega/\omega_H = 1.0$ . They correspond to the  $-3$  dB points on the unapproximated Bode plots. For most purposes, the simplified curves are an entirely satisfactory representation. From these

curves, every 10-fold reduction in frequency below  $\omega_L$  for the high-pass circuit decreases the output voltage by 20 dB, and every twofold reduction decreases it by 6 dB. The low-pass circuit has just the opposite properties: a 10-fold increase in frequency above  $\omega_H$  results in a 20 dB decrease in output voltage, and a twofold increase results in a 6 dB decrease. One often states these facts as *20 dB per decade* and *6 dB per octave*. The linearized phase-response curves are shown in Figure 6.10(c). The  $-3$  dB frequencies occur at a phase shift of  $-\pi/4$  ( $-45^\circ$ ) for the two circuits.

For the nonrepetitive input voltage waveforms of cases (2) and (3), the output waveforms are given in Figure 6.11.

The output, waveforms for the rectangular-wave input function can be used to determine the  $RC$  time constants for differentiating and integrating circuits. This is called *square-wave testing*. The  $RC$  time constant for the



**Figure 6.10** Idealized gain response for (a) high-pass and (b) low-pass circuits; idealized phase response for (c) high-pass and (d) low pass circuits

differentiating circuit is obtained by using a square-wave input with a rise time much smaller than  $RC$  and a period much larger than  $RC$ . For times small compared with  $RC$ , the tilt of the top edge of the output, as viewed with a fast-rise-time oscilloscope (see Figure 6.12) is directly related to  $RC$ . The fractional decrease in  $v_o$ , in time  $t_1$  is  $t_1/RC$ , which can be set equal to  $(V - V')/V$  and solved for  $RC$ . For the integrating circuit,  $RC$  is obtained by measuring the rise time of the output waveform on a fast rise time oscilloscope. Using the definition of the rise time  $t_r$ , as the time between the 10% and 90% points on the leading edge of the output waveform, one has the relation:

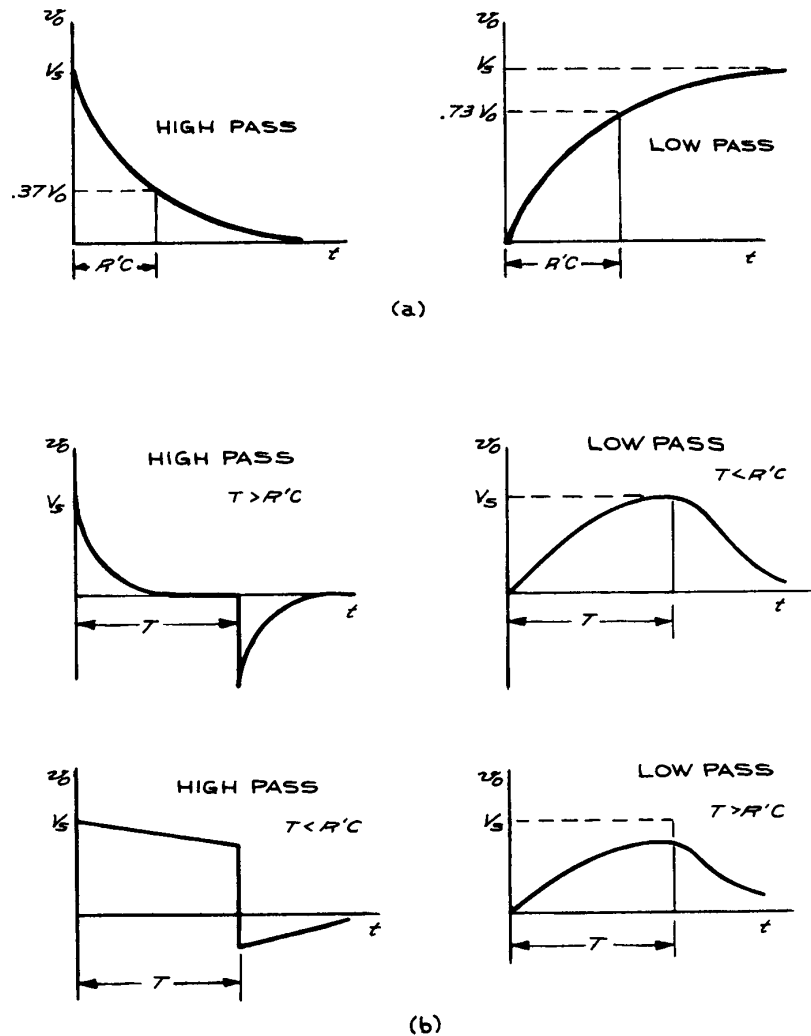
$$t_r = 2.2RC \quad (6.14)$$

### 6.1.4 Resonant Circuits

The voltages and currents in circuits with capacitors, inductors, and resistors show oscillatory properties much like those of mechanical oscillators. Electronic circuits have

natural frequencies of oscillation and can be critically damped, underdamped, or overdamped, depending on the relations between the values of the circuit parameters. Resonant circuits with ideal capacitors and inductors are of the series or parallel type shown in Figure 6.13. When driven by a sinusoidal input source, the capacitive reactance in the series circuit will cancel the inductive reactance at the resonant frequency  $\omega_o$ , where  $1/C\omega_o = \omega_o L$  and  $\omega_o = \sqrt{1/LC}$ . At  $\omega_o$  the impedance of the series circuit is a minimum and the current through it is a maximum.

For the parallel resonant circuit at low frequencies, the  $L$  branch will have a very low reactance and the current drawn from the source will flow almost entirely through that branch. At high frequencies, the current through the  $RC$  branch is limited by the value of  $R$ . The total impedance of the parallel circuit is therefore small at low and high frequencies, passing through a maximum at the frequency  $\omega_o = \sqrt{1/LC}$ , provided that  $R \ll \omega_o L$ . Graphs of the currents in the two circuits as a function of driving frequency are given in Figure 6.14. Real inductors have



**Figure 6.11** Response of high-pass and low-pass circuits to a step voltage (a) and a rectangular pulse of duration  $T$  (b).

an associated resistance, which generally must also be taken into account when analyzing circuits.

One measure of the resonance sharpness in the series and parallel circuits is the  $Q$  or *quality* of the circuit. For practical purposes,  $Q = \omega_0/\Delta\omega$  where  $\Delta\omega$  is the full width at half maximum of the peak or valley. In terms of the circuit parameters,  $1/Q = \omega_0 L/R = 1/\omega_0 RC$ . This is the ratio of the energy stored (in the capacitor or inductor) to

the energy dissipated in the resistor per cycle at resonance. Values of  $Q$  as large as 100 can be attained in electrical circuits while mechanical oscillators can attain values as high as  $10^6$ . The phase relationships between voltage and current in series and parallel resonant circuits are shown in Figure 6.15.

The behavior of an  $LRC$  circuit upon the application of a step or rectangular input is much like the response of a

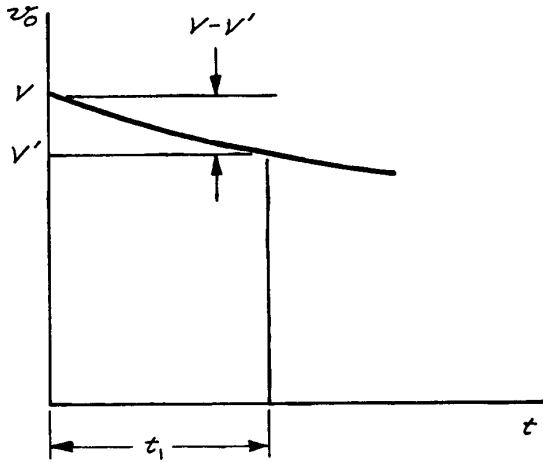


Figure 6.12 Square-wave testing of a high-pass circuit.

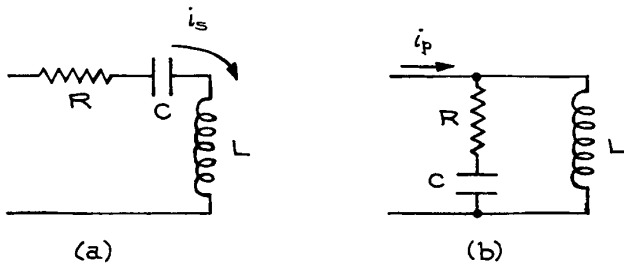


Figure 6.13 (a) Series resonant circuit; (b) parallel resonant circuit.

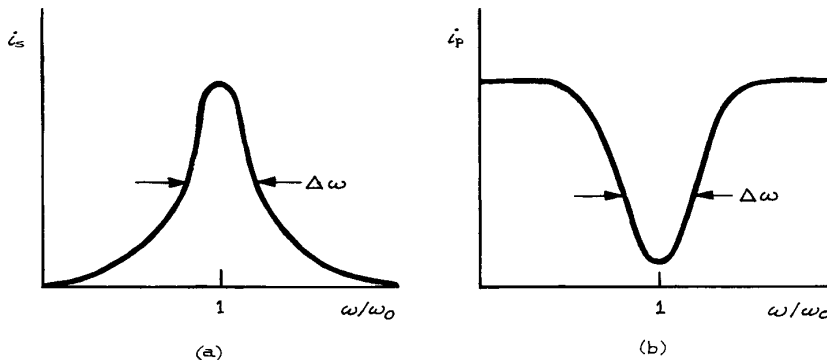


Figure 6.14 Current as a function of frequency for: (a) the series resonant circuit and (b) the parallel resonant circuit.

mechanical system to a sudden impulse. Critically damped, underdamped, and overdamped current flows result.

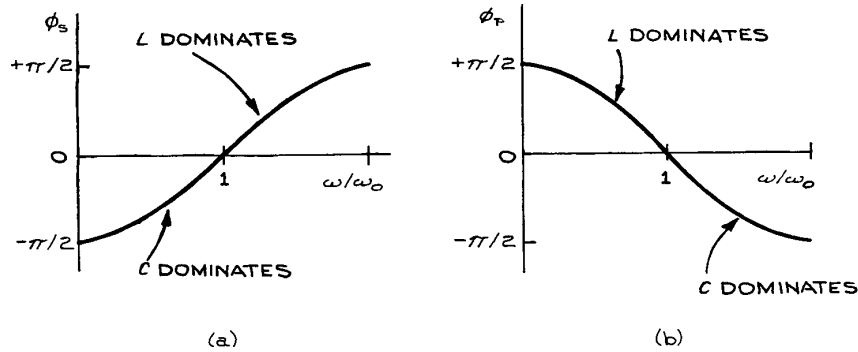
### 6.1.5 The Laplace-Transform Method

A general technique for analyzing circuits for arbitrary input voltage waveforms is the method of Laplace transforms. With this method it is possible to use only algebra and lists of transforms – such as those given in Table 6.2 – for the solution of differential equations and the evaluation of boundary conditions. The results of the method will be presented without any proofs. The vocabulary of Laplace transforms occurs in the discussion of circuits and is included in this chapter for that reason.

The method is based on an integral transform of the type:

$$\bar{f}(s) = \int_0^{\infty} f(t) e^{-st} dt \quad (6.14)$$

where  $\bar{f}(s)$  is the Laplace transform of  $f(t)$ , written as  $L[f(t)]$ . The function  $f(t)$  can involve integrals and differentials. When it is applied to the second-order differential equations that arise in circuit analysis, rather important simplifications occur and results can often be written down by inspection. The Laplace transform of the output voltage  $\bar{v}_o(s)$  of a circuit is the Laplace transform of the input voltage  $\bar{v}_i(s)$  times the Laplace transform of the transfer function  $\bar{T}(s)$  – the *transfer function* being the function relating the output to input. To obtain  $\bar{T}(s)$ , the values of all elements in the circuit are



**Figure 6.15** Phase relations between voltage and current in: (a) the series resonant circuit and (b) the parallel resonant circuit.

replaced by their transform equivalents according to the recipe  $R \rightarrow R$ ,  $C \rightarrow 1/sC$ , and  $L \rightarrow sL$ . In the simple case of the voltage divider, the transfer function is the ratio of the impedance of the output-circuit element to the total impedance of the circuit chain.  $T(s)$  is obtained in exactly the same way, using the equivalences for  $R$ ,  $C$ , and  $L$ . In general,  $\bar{T}(s)$  is in the form of a ratio of two functions of  $s$ ,  $G(s)$  and  $H(s)$ , which are polynomials in  $s$ :

$$\bar{T}(s) = \frac{G(s)}{H(s)} \quad (6.15)$$

The values of  $s$  for which  $G(s)$  is zero are called the *zeros* of  $\bar{T}(s)$ ; the values of  $s$  for which  $H(s)$  is zero are the locations of the *poles* of  $\bar{T}(s)$ . In the most general case, the zeros and poles are complex. The positions of the zeros and poles of  $\bar{T}(s)$  in the complex plane give important information on the properties of the circuit under analysis. When the poles are complex, they occur in pairs, while real-valued poles can occur singly or in pairs. The values of the real and imaginary components of the pole coordinates, usually labeled  $\sigma$  and  $\omega$ , have important physical meaning. The real component  $\sigma$  is a measure of the damping in the circuit while the imaginary part  $\omega$  is the natural frequency of oscillation. Negative values of  $\sigma$  give stable circuits in which transient signals all decay to zero with time. Circuits employing only passive elements behave in this way and are stable. Circuits with active elements can behave in such a way that the output increases with time in response to a transient input signal. Such circuits are unsta-

ble and have values of  $\sigma$  greater than zero. They are to be avoided, except in the case of oscillators, which must be unstable in order to function.

The Laplace-transform equivalents of the high-pass and low-pass circuits are shown in Figure 6.16. For the high-pass circuit, the output voltage across the resistor  $R$  for the transformed circuit is:

$$\bar{v}_o(s) = \bar{v}_s(s) \frac{sRC}{1 + sR'C} \quad (6.16)$$

where  $R' = (R_s + R)$  and  $\bar{T}(s)$ , which is the transfer function, has a zero at  $s = 0$  and a pole at  $s = -1/R'C = \omega_L$ . For the low-pass circuit, the output voltage across the capacitor for the transformed circuit is:

$$\bar{v}_o(s) = \bar{v}_s(s) \frac{1}{1 + sR'C} \quad (6.17)$$

$\bar{T}(s)$  has a pole at  $s = -1/R'C = -\omega_H$ . The steady-state frequency and phase response of the circuits are obtained from  $T(s)$  by replacing  $s$  with  $j\omega$ . The transfer functions are now, for the low-pass circuit:

$$T(\omega) = \frac{j\omega}{1 + j\omega/\omega_L} \quad (6.18)$$

and:

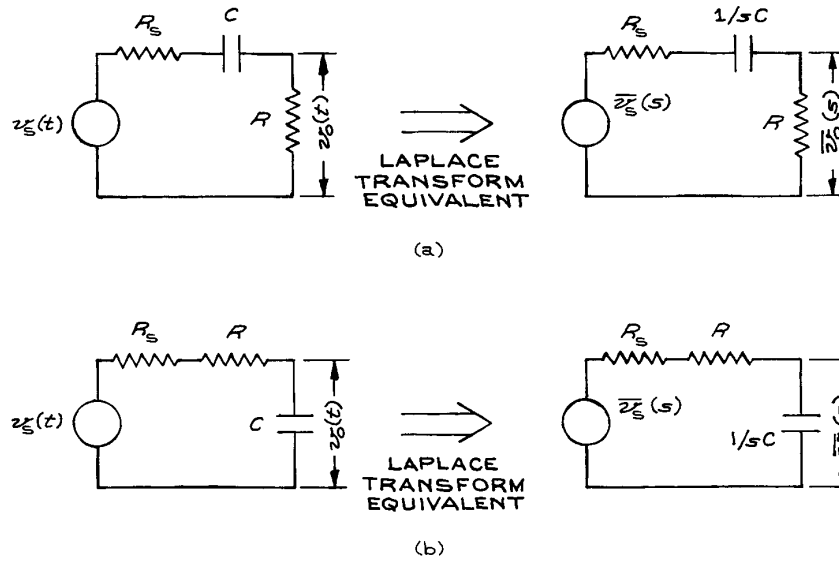
$$T(\omega) = \frac{1}{1 + j\omega/\omega_H} \quad (6.19)$$

for the high-pass circuit.

Table 6.2 Elementary laplace transforms

$f(t)(t > 0)$	$f(s)$
$\delta(t)$	1
1	1/s
$t^{n-1}/(n-1)!$	1/s <sup>n</sup> (n a positive integer)
$e^{at}$	$\frac{1}{s-a}$
$\sin at$	$\frac{a}{s^2+a^2}$
$\cos at$	$\frac{s}{s^2+a^2}$
$\sinh at$	$\frac{s}{s^2-a^2}$
$\cosh at$	$\frac{s}{s^2-a^2}$
$\frac{t}{2a} \sin at$	$\frac{s}{(s^2+a^2)^2}$
$\frac{1}{2a^3} (\sin at - at \cos at)$	$\frac{1}{(s^2+a^2)^2}$
$\frac{df(t)}{dt}$	$s\bar{f}(s) - f_0 \left[ f_0 = \lim_{t \rightarrow 0} f(t) \right]$
$\frac{d^2f(t)}{dt^2}$	$s^2\bar{f}(s) - sf_0 - f_1 \left[ f_1 = \lim_{t \rightarrow 0} \frac{df(t)}{dt} \right]$
$\int_0^t f(t') dt'$	$\frac{1}{s} \bar{f}(s)$
$\frac{1}{a-b} (e^{at} - e^{bt})$	$\frac{1}{(s-a)(s-b)}$
$\frac{1}{a-b} (ae^{at} - be^{bt})$	$\frac{s}{(s-a)(s-b)}$
$\frac{1}{a^2} (1 - \cos at)$	$\frac{1}{s(s^2+a^2)}$
$\frac{1}{a^3} (at - \sin at)$	$\frac{1}{s^2(s^2+a^2)}$
$\frac{1}{ab(b^2-a^2)} (b \sin at - a \sin bt)$	$\frac{1}{(s^2+a^2)(s^2+b^2)}$
$\frac{1}{b^2-a^2} (\cos at - \cos bt)$	$\frac{s}{(s^2+a^2)(s^2+b^2)}$
$\frac{1}{\alpha^2 + \beta^2} - \frac{e^{-\alpha t}}{\beta \sqrt{\alpha^2 + \beta^2}} \sin[\beta t + \arg(\alpha + i\beta)]$	$\frac{1}{s[(s+\alpha)^2 + \beta^2]}$





**Figure 6.16** Laplace-transform equivalents of: (a) the high-pass and (b) the low-pass circuit.

As seen previously,  $\omega_L$  and  $\omega_H$  are the corner frequencies of the circuits. By rationalizing the denominators of the transfer functions, one obtains the phase response. Since there are no inductive elements in the circuits, there is no natural frequency of oscillation. The poles lie on the negative real axis because there are no active elements in the circuit to cause sustained oscillations.

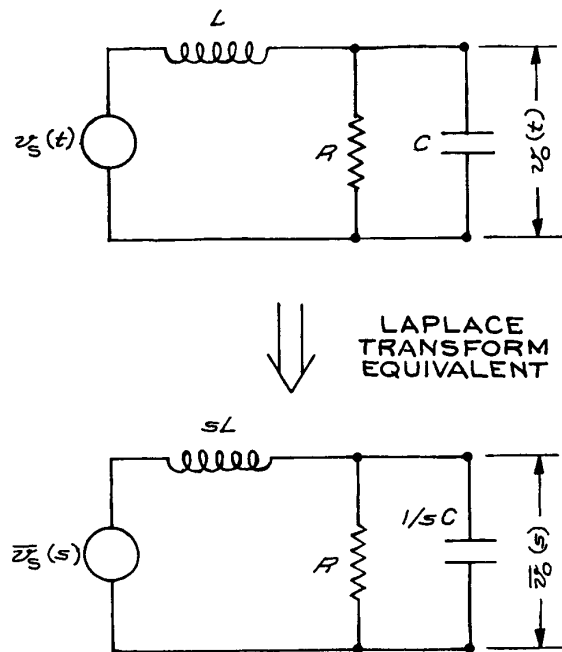
### 6.1.6 RLC Circuits

Consider the equivalent circuit and the Laplace transform given in Figure 6.17. To analyze the circuit, consider the parallel combination of  $R$  and  $1/sC$ , which is in series with  $sL$  in a voltage-divider configuration:

$$\frac{R/sC}{R + 1/sC} = \frac{R}{1 + sRC} \quad (6.20)$$

Thus:

$$\bar{v}_o(s) = \bar{v}_i(s) \frac{1}{1 + sL/R + s^2LC} \quad (6.21)$$



**Figure 6.17** An RLC circuit and the Laplace-transform equivalent.

The poles of  $\bar{T}(s)$  occur at:

$$s = \frac{-L/R \pm \sqrt{(L/R)^2 - 4LC}}{2LC} \quad (6.22)$$

Letting the natural frequency of oscillation of the circuit be  $\omega_0 = 1/\sqrt{LC}$ , we have  $Q = R/\omega_0 L$ , and the poles can be rewritten as:

$$s = \frac{-\omega_0}{2Q} \pm \frac{\omega_0}{2} \sqrt{1/Q^2 - 4} \quad (6.23)$$

There are three different possibilities for the roots of  $s$ :

- 1)  $1/Q^2 = 4$ : a single real root at  $s = -\omega_0/2Q$
- 2)  $1/Q^2 - 4 = m^2$  ( $m$  real): two real roots at  $s = -\omega_0/2Q \pm \omega_0 m/2$ .
- 3)  $1/Q^2 - 4 = -m^2$  ( $m$  real): two conjugate complex roots at  $s = -\omega_0/2Q \pm j\omega_0 m/2$ .

The magnitude of  $s$  in the  $\sigma - j\omega$  plane is  $\omega_0$ ; in geometric terms this means that the roots of  $s$  are confined to a semicircle of radius  $\omega_0$  in the left half of the complex plane (see Figure 6.18).

### 6.1.7 Transient Response of Resonant Circuits

The stability of a linear system subjected to a driving function depends on the system itself rather than the driv-

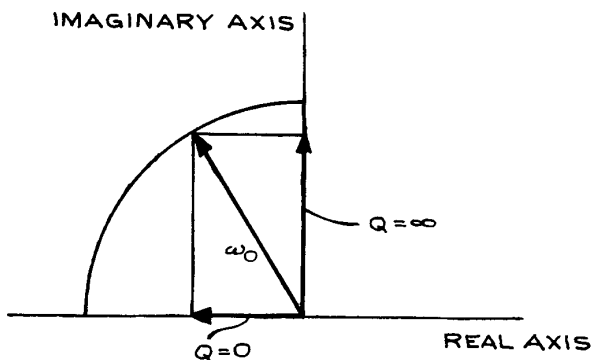


Figure 6.18 Pole trajectory for an RLC circuit.

ing function. For ease of analysis, consider the RLC circuit of Figure 6.17 to be driven by a unit step input voltage of the form  $v_i(t) = 1$  for  $t > 0$ . The Laplace transform of the output voltage is then the transform of the input voltage,  $\bar{v}_i(s)$ , multiplied by the transform of the transfer function  $\bar{T}(s)$ :

$$\bar{v}_o(s) = \bar{v}_i(s)\bar{T}(s) \quad (6.24)$$

For the unit step, input  $v_i(s) = 1/s$  from Table 6.2; therefore

$$\bar{v}_o(s) = \frac{1}{s(1 + sL/R + s^2LC)} \quad (6.25)$$

To find  $v_o(t)$  it is necessary to look up the inverse transform in Table 6.2. The result is:

$$v_o(t) = 1 - \frac{e^{-k\omega_0 t}}{\sqrt{1-k^2}} \sin \left[ \sqrt{1-k^2} \omega_0 t + \arctan \left( \frac{\sqrt{1-k^2}}{k} \right) \right] \quad (6.26)$$

where  $k = 1/2Q$ . Critical damping, overdamping, and underdamping correspond to  $k = 1$ ,  $k > 1$ , and  $k < 1$ , respectively. The normalized response of the circuit for various values of  $k$  is shown in Figure 6.19.

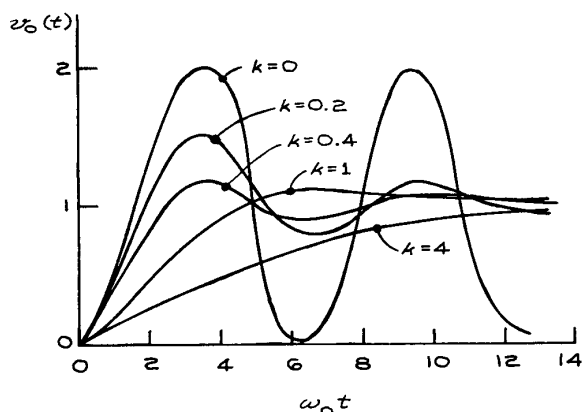


Figure 6.19 Response of a resonant circuit with no damping ( $k = 0$ ), underdamping ( $k = 0.2, 0.4$ ), critical damping ( $k = 1$ ), and overdamping ( $k = 4$ ).

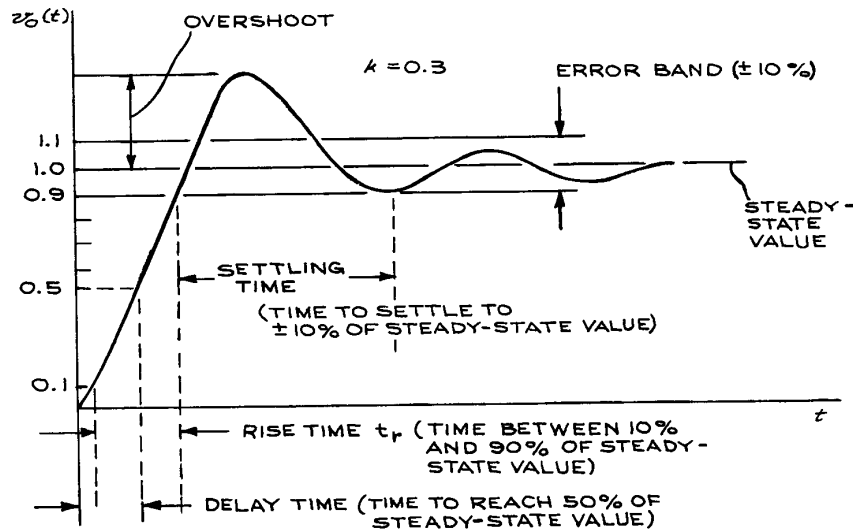


Figure 6.20 Output waveform in an underdamped resonant circuit.

As can be seen, the underdamped waveform overshoots the steady-state value of one and oscillates about it with decreasing amplitude. This is sometimes called *ringing*. As damping is increased, the output waveform oscillations decrease in amplitude until, for  $k = 1$  (critical damping), the oscillations are completely gone. Increasing the damping beyond  $k = 1$  only increases the time for the waveform to arrive at the steady-state value.

*Overshoot*, *rise time*, *settling time*, and *error band* are terms used to characterize output waveforms. They are illustrated in Figure 6.20. For critical damping,  $t_r = 3.3/\omega_0$ . By decreasing the damping so that  $k < 1$ ,  $t_r$  can be further decreased, but at the expense of ringing. A suitable compromise that is often used is  $k = 0.707$ , under which circumstance the overshoot is 4.3% and  $t_r$  is reduced to  $2.16/\omega_0$ . Figure 6.21 is a graph of the percentage overshoot as a function of  $k$ .

The above analysis is not restricted to passive circuits. Amplifiers with negative feedback can have response functions identical in form to those for *RLC* circuits. The response of such amplifiers to nonrepetitive waveforms is specified using the same terms as for the passive circuit above. In pulse amplifiers, the transient response is the fundamental limitation on the pulse

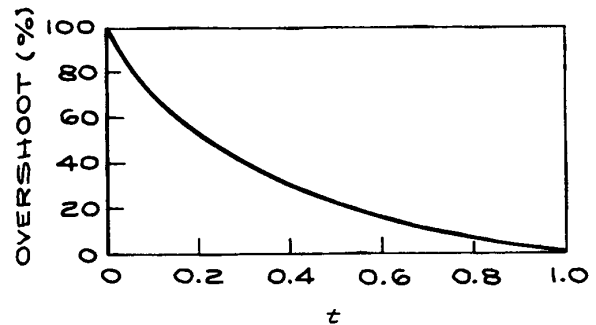


Figure 6.21 Percentage overshoot as a function of damping in a resonant circuit.

repetition rate. For critically damped amplifiers, the width  $T$  of the impulse response and the rise time  $t_r$  are approximately related by  $T = 1.5t_r$ , so that pulses cannot be sent at a rate greater than  $1/T$  per second.  $T$ , in this case, is generally taken to be the time between the 50% points on rising and falling edges of the output waveform.

For the precision control of temperature in a thermostat, three-term control is often used. With this method,

the power applied to the heater element is set equal to a constant background level plus a term proportional to the difference between the set-point temperature  $T_s$  and current temperature  $T$ , a term proportional to the time integral of  $T_s - T$ , and a term proportional to the time derivative of  $T_s - T$ . The resulting differential equation is, under certain circumstances, analogous to that for the  $RLC$  circuit. Experimental parameters are chosen to obtain the fastest time response consistent with overall system stability. This is discussed in further detail in Section 6.7.9 and in Chapter 7.

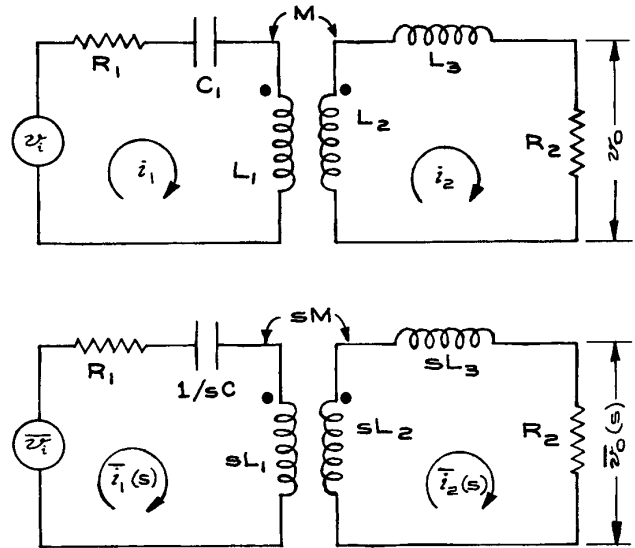
### 6.1.8 Transformers and Mutual Inductance

Transformers consist of primary and secondary windings coupled by a core, which can be of magnetic material or air. A voltage is induced in the secondary of the transformer when there is a change of current in the primary. The reverse also occurs – induced voltage in the primary due to a current change in the secondary. A transformer circuit and the Laplace-transform equivalent are illustrated in Figure 6.22. The dot associated with each winding of the transformer indicates the relative orientation of the windings. The convention is that  $M$ , the mutual inductance, is positive if the currents  $i_1$  and  $i_2$  both flow into or out of the dotted ends of the coils, and negative otherwise. The transfer function  $\bar{T}(s)$  is obtained by applying Kirchhoff's laws to the two loops:

$$\begin{aligned}\bar{v}_i(s) &= R_1 \bar{i}_1(s) + \frac{\bar{i}_1(s)}{sC} + \bar{i}_1(s)sL_1 - sM\bar{i}_2(s) \\ 0 &= \bar{i}_2(s)sL_2 - \bar{i}_1(s)sM + \bar{i}_2(s)sL_3 + \bar{i}_2(s)R_2 \quad (6.27) \\ \bar{v}_o(s) &= \bar{i}_2(s)R_2\end{aligned}$$

Eliminating  $\bar{i}_1(s)$  and substituting for  $\bar{i}_2(s)$  in the third equation gives the result:

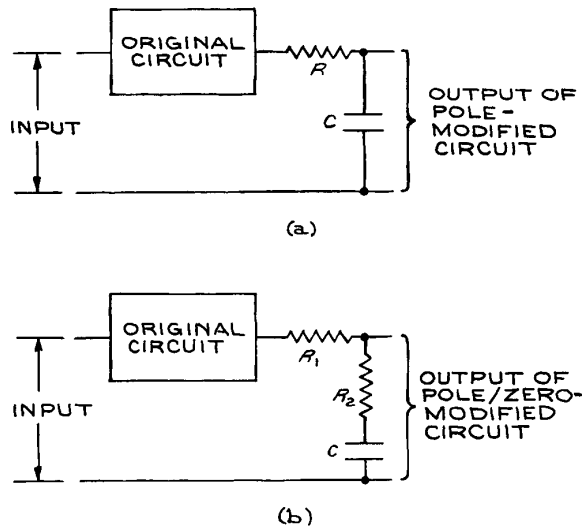
$$\bar{v}_o(s) = \bar{v}_i(s) \frac{R_2 M C s^2}{[cL_1(L_2 + L_3) - M^2]s^3 + [R_1(L_1 + L_2) + R_2L_1]Cs^2 + [L_2 + L_3 + R_1R_2C]R_2s} \quad (6.28)$$



**Figure 6.22** A transformer circuit and the Laplace-transform equivalent.  $M$ , the mutual inductance, is positive if  $i_1$  and  $i_2$  both flow into or out of the dotted ends of the coils.

### 6.1.9 Compensation

It is often desired to modify the frequency response of a given circuit. Two common methods are the addition of a pole that is much smaller than the other poles of the transfer function and the simultaneous addition of a pole and zero to the transfer function. The addition of a pole is accomplished with a resistor and capacitor in a low-pass configuration at the output of the circuit to be modified, as illustrated in Figure 6.23(a). If the original circuit had a pole  $\omega_1$  and the introduction of the  $RC$  circuit produced a pole at  $\omega_2$  where  $\omega_2 \ll \omega_1$ , the overall response of the circuit would be a 20 dB/decade decrease in gain from  $\omega_2$  to  $\omega_1$  followed by a 40 dB/decade decrease for frequencies greater than  $\omega_1$ .



**Figure 6.23** Compensation by: (a) the addition of a pole and (b) the addition of a pole and a zero to cancel the lowest-frequency pole of the original circuit.

The addition of a pole and a zero is accomplished with the circuit in Figure 6.23(b). The transfer function for the *pole-zero compensation* network is:

$$\frac{R_2 + 1/sC}{R_1 + R_2 + 1/sC} = \frac{1 + sR_2C}{1 + s(R_1 + R_2)C} \quad (6.29)$$

with the zero at  $s = -1/R_2C$  and the pole at  $s = -1/(R_1 + R_2)C$ . If the zero in the transfer function is chosen to cancel the smallest pole of the original circuit, the overall frequency response will be flat up to the frequency corresponding to the new pole. The gain will then decrease by 20 dB/decade with increasing frequency up to the second pole of the original circuit. The gain will then decrease by 40 dB/decade up to the next high-frequency pole, after which the decrease will be 60 dB/decade. The response of the pole-zero-compensated circuit is sharper than that of one with single-pole compensation.

### 6.1.10 Filters

Filters are generally classed as *low-pass*, *high-pass*, *band-pass*, and *band-reject*. *RC* circuits are examples of the first two kinds, while the series and parallel *RLC* circuits illus-

trate the last two. A series of band-pass or band-reject filters with pass bands at different, but closely spaced, frequencies is called a *comb filter*.

The ideal filter would pass unattenuated all frequencies in its pass band while completely rejecting frequencies outside the pass band. The simple circuits we have examined thus far are only approximations to the ideal. There are excellent approximations to the ideal filter that rely on judicious choices of the transfer function poles of *RC* and *RLC* circuits. There are four classes of filter design, each of which has its advantages:

- (1) *Maximally flat* has the flattest amplitude response within the pass band. The *Butterworth* filter is an example.
- (2) *Equal ripple* has fluctuations in the pass band but greater attenuation in the stop band than the maximally flat filter.
- (3) *Elliptic* has the maximum rate of attenuation between the pass band and stop band. The *Chebyshev* filter is an example of this type.
- (4) *Linear phase* has a much less sharp cutoff than the others, but maintains an almost linear phase response in the pass band. The *Bessel* filter is an example of this type.

Methods of filter design are well established, and there are many texts and handbooks that provide tables for filter synthesis. Before designing or specifying a filter, it is important to know, not only the cutoff and phase properties, but also the transient response and the input and output impedances.

Amplifiers with negative feedback provided by resistors and capacitors can have responses identical to *RLC* circuits. By appropriate choices of the elements in a feedback network, these amplifiers can be made into filters. These so-called *active filters* are especially useful at low frequencies, where large inductors with low resistance are heavy and expensive. There are many texts and handbooks on active-filter design. *Lancaster's Active Filter Cookbook* (D. Lancaster, 2nd ed., Butterworth-Heinemann, Oxford, 1996) is written from a practical point of view with a minimum of mathematics. Most active-filter designs are based on high-gain integrated-circuit operational amplifiers. With the frequency response of these amplifiers now extending well into the megahertz region, active filters are finding increasing application to high-frequency circuits. Stability is an important factor in filter design,

especially for large values of  $Q$ . If drift is to be minimized, highly stable passive components are required.

### 6.1.11 Computer-Aided Circuit Analysis

The *Simulation Program with Integrated-Circuit Emphasis* (SPICE) was originally intended for integrated circuit design where the very large number of circuit elements made traditional circuit analysis slow, cumbersome, and subject to errors. The SPICE Program was first limited to use on mainframe computers with network access, but various versions are now available for personal computers (PCs). One popular PC version of SPICE is PSPICE, now available from Cadence. Because SPICE and PSPICE figure prominently in electrical-engineering circuit-analysis courses, they are usually available over university networks and various versions, including student/faculty and evaluation versions, can be downloaded from the web. Analysis of digital circuits with SPICE requires detailed information about waveforms and delays within individual units. Digital circuits are increasingly assembled from programmable arrays and the manufacturers of the arrays generally supply the required information with their programming software.

The following procedures should be followed when using circuit analysis programs:

- (1) Create a schematic diagram of the circuit:
  - Assemble components from the program library
  - Import or create components unavailable in library
  - Assign values to resistors, capacitors, and inductors
  - Make interconnections
- (2) Create SPICE/PSPICE file:
  - Specify temperature
  - Specify sources
    - Independent
    - Controlled
  - Verify interconnections (nodes)
- (3) Analysis:
  - D.c., quiescent
  - Transient
  - A.c., small signal
- (4) Output:
  - Graphs
    - Bode plots

- Waveforms
  - Numerical tables
- (5) Adjust circuit to meet design criteria and reanalyze.

Analysis programs are extremely useful, but cannot replace an understanding of circuit theory. It is unrealistic to believe that one can begin with an approximate circuit and use a program to arrange the values of the components and the configuration until the desired result is attained. Even for a simple circuit of 10 components, the number of possibilities is so large that an uninformed attempt to arrive at a desired final result cannot be successful. On the other hand, an arbitrary circuit can be analyzed to determine whether it will perform according to the requirements. Most analysis programs come with libraries of common active devices (transistors, diodes, operational amplifiers) however, new devices appear on the market daily and older devices become obsolete so that it is impractical for any single library to contain all possible devices. Adding device specifications to analysis program libraries can be time consuming and it is often more efficient to use available devices with operating parameters close to those of the device in question. This requires additional research, but device specifications are now routinely available on the websites of the manufacturers. In some cases, manufacturers provide SPICE files for their devices for loading directly into libraries. An additional benefit of analysis programs is compatibility with schematic capture and printed-circuit-board layout software. This is discussed in more detail in Section 6.10.3 on printed circuit boards.

## 6.2 PASSIVE COMPONENTS

In discussing passive components, emphasis will be placed on choosing the correct type for the function to be performed. It is rarely sufficient to specify the nominal value of a component. This can be seen by looking through the catalog of any electronics supplier under *resistors* and finding perhaps 50 different resistor types, each with a range of resistance values. The following are some general considerations in the choice of passive components:

- (1) *Nominal value and tolerance.* An aspect of this is the coding applied to the component. Both color, letter, and number codes are used.

- (2) *Stability.* The nominal value as a function of temperature [the temperature coefficient (*tempco*) is often expressed in % or in parts per million change per °C or ppm/°C]. Age and environmental conditions such as humidity, vibration, or shock also effect component values. On the other hand, components can affect their environment – for example, the outgassing of a capacitor in vacuum or the heating of one component by another.
- (3) *Size and shape.* Along with the reduction in size that solid-state electronics has brought, there has been a similar reduction in the size of passive components. Generally, one uses the smallest practical size, but power-density considerations can limit the packing density of power-dissipating components. Also, components come in different shapes with different lead geometries depending on the intended mounting method. These are discussed in the section on hardware.
- (4) *Power dissipation and voltage rating.* A resistor has both a power rating and a voltage rating. For very large values of resistance, it is quite possible to operate well within the resistor's power rating, yet exceed the voltage rating. The result is electrical breakdown and destruction of the resistor. For moderate values of capacitance, capacitors dissipate very little power and power rating is of no importance. Very-high-capacitance capacitors with large effective surface areas, however, have non-negligible shunt conductances. The large currents that flow in such capacitors can cause considerable heat to be generated with potentially fatal (to the capacitor) results.
- (5) *Noise.* Passive components can introduce noise into an electronic circuit, and one must be particularly aware of this problem in low-level circuits and choose appropriate low-noise components.
- (6) *Frequency characteristics.* Pure resistors, capacitors, and inductors are practically unrealizable, and the electrical properties of real passive components depend on frequency in a nonideal way. This is often indicated by an equivalent circuit. Inductors and transformers with magnetic cores, however, are inherently nonlinear and can be treated only approximately in this way.
- (7) *Derating.* This is an engineering term related to the safety factors to be used when operating components under a variety of conditions. Good design requires

that components be operated at no more than 50% of the manufacturer's recommended maximum ratings, particularly with respect to power and voltage. At high ambient temperatures, components should be derated by even larger amounts. The resulting decrease in stress greatly increases component lifetime. In laboratory equipment, large safety factors should always be used for enhanced reliability. The abbreviation SOA, safe operating area, specifies the limits of voltage, current, and temperature for semiconductors.

- (8) *Cost.* For laboratory applications it is poor practice to economize on passive components by using the least expensive ones that will do the job. Laboratory work is labor intensive. The extra cost of top-grade components is greatly outweighed by the time saved in troubleshooting and repair.

### 6.2.1 Fixed Resistors and Capacitors

The more common types of fixed and variable resistors and fixed capacitors are listed in Tables 6.3 to 6.5. In Table 6.5, the *C* range and *V* range refer to the minimum and maximum values available for all types of a given dielectric. It is not possible to have the maximum capacitance at the maximum voltage. The larger capacitances have lower *working voltages* (WV), while the smaller ones have the higher WV. The temperature coefficients generally apply only to a limited temperature range within the larger operating temperature range.

High component-density circuits use *surface-mount technology* (SMT) chip resistors and capacitors. These are different from the more traditional wire lead types, both in their configuration and the way they are electrically and mechanically connected to circuits. Chip resistors and capacitors are common SMT components, but semiconductors and integrated circuits are also manufactured for surface mounting. While high packing density is not often a primary consideration in laboratory electronics, small component size can be an important advantage for some circuits where stray capacitance and inductance are factors. Surface-mount resistors and capacitors have metalized contact surfaces and are connected to circuit-board pads with solder that bridges the SMT metallization and circuit-board pads. Surface-mount technology resistors and capacitors can be little larger than

Table 6.3 Fixed resistors

<i>Carbon Composition</i>		<i>Precision Wire-Wound (bobbin)</i>	
R range:	2.7 $\Omega$ to 22 M $\Omega$	R range:	0.1 $\Omega$ to 18 M $\Omega$
Power range:	$\frac{1}{10}$ to 2W	Power range:	$\frac{1}{10}$ to 2W
Tolerances:	$\pm 5, \pm 10, \pm 20\%$	Tolerances:	$\pm 0.05$ to $\pm 5\%$
Temperature range:	$-55$ to $+150$ $^{\circ}\text{C}$	Temperature range:	$-55$ to $\pm 145$ $^{\circ}\text{C}$
Temperature coefficient	$\pm 0.1\%/^{\circ}\text{C}$	Temperature coefficient	$\pm 0.0001$ to $\pm 0.002\%/^{\circ}\text{C}$
Voltage coefficient	$\pm 0.03\%/V$ d.c.	Noise:	Low
Noise:	Highest	Frequency response:	To 25 kHz for high-resistance values unless noninductively wound
Typical capacitance:	0.25 pF, no inductance		
Working voltage:	150V $\left(\frac{1}{8}\text{W}\right)$ to 750V(2W)		
<i>Precision Carbon Film</i>		<i>Power Wire-Wound</i>	
R range:	1 $\Omega$ to 100 M $\Omega$	R range:	0.1 $\Omega$ to 1.3 M $\Omega$
Power range:	$\frac{1}{10}$ to 2W	Power range:	2 to 225 W
Tolerances:	$\pm 0.5, 1, 2\%$	Tolerances:	$\pm 5\%, \pm 10\%$
Temperature range:	$-55$ to $165$ $^{\circ}\text{C}$	Operating temperature:	$-75$ to $+350$ $^{\circ}\text{C}$
Temperature coefficient	0.02 to 0.05%/ $^{\circ}\text{C}$	Maximum rated-power temperature:	25 $^{\circ}\text{C}$
Noise:	Low	Temperature coefficient:	$\pm 0.026\%/^{\circ}\text{C}$
Frequency response:	To 1 MHz for $R > 1$ k $\Omega$ (spinal), to 100 MHz for $R < 1$ k $\Omega$	Frequency response:	To 25 kHz for high-resistance values unless noninductively wound
<i>Precision Metal Film</i>			
R range:	10 $\Omega$ to 200 M $\Omega$		
Power range:	$\frac{1}{8}$ to $\frac{1}{2}$ W		
Tolerances:	$\pm 0.1$ to $\pm 1\%$		
Temperature range:	$-55$ to $165$ $^{\circ}\text{C}$		
Temperature coefficient	0.0025 to 0.01%/ $^{\circ}\text{C}$		
Noise:	Lowest		

the end of a pencil lead and use special codes to identify their values. The resistor code is a three-digit code (for 2% and greater tolerances) and a tolerance character. The first two digits are the first two significant Figures of the resistance and the third digit is the power of ten multiplier. The code 563, for example, is 56 kohms. For resistances less than 100 ohms the letter R is used to represent a decimal point. For example 33R is a 33 ohm resistor. The tolerance is

specified by a single letter at the end of the numeric code. Table 6.8 lists SMT resistor and capacitor codes.

Tantalum is the most stable of all film-forming materials. Tantalum-foil electrolytic capacitors have similar characteristics to aluminum-foil electrolytic capacitors, but offer superior stability and freedom from leaking. The wet-slug type has the highest volumetric efficiency of any capacitor. Tantalum solid-electrolyte capacitors have



**Table 6.4 Variable resistors**

Resistive Element Material	R Range	Temperature Coefficient (PPm/°C)	Linearity (%)
Carbon composition	50 Ω to 1 MΩ	± 1000	5.0
Resistance wire	1 Ω to 100 kΩ	± 10	0.25
Conductive plastic	100 Ω to 5 MΩ	± 100	0.25
Cermet <sup>a</sup>	100 Ω to 5 MΩ	± 100	0.25

<sup>a</sup> Ceramic-metal hybrid.

much better stability, frequency, and temperature characteristics than liquid-electrolyte types.

The equivalent circuits for fixed resistors and capacitors are given in Figure 6.24. *Power factor (PF)*, the ratio of consumed power to apparent power, is a way of specifying the nonideal properties of resistors, capacitors, and inductors. The power factor of a pure resistance is one, while the power factor of a pure reactance (capacitor or inductor) is zero. For resistors and capacitors, power factor is specified at  $10^4$  to  $10^5$  Hz where  $X_L$  is generally much smaller than  $X_C$ . Figure 6.25 illustrates common capacitor types. Coding is given in Tables 6.6, 6.7, and 6.8.

## 6.2.2 Variable Resistors

In electronic equipment, variable resistors are generally called *potentiometers*. More specifically, they can be classed as potentiometers, trimmers, and rheostats. All are three-terminal devices with a terminal at each end of the resistive element and the third terminal attached to the sliding contact or *tap*. Potentiometers are designed for regular movement, trimmers for occasional changes, and rheostats for current limiting in high-power circuits. Some common types of variable resistors are shown in Figure 6.26. There are a number of different resistive-element materials for potentiometers. They are listed in Table 6.4.

General-purpose, single-turn potentiometers come in power ratings from 0.5 to 5 W and a variety of sizes and tapers (*taper* is the change in resistance as a function of shaft angle). Nonlinear tapers are often used in volume controls, where an approximately logarithmic response is desired. Some designs can be stacked on a common shaft, and it is often also possible to incorporate an on-off

switch. Shafts with locking nuts are very useful when it is necessary to fix a given setting. Precision potentiometers can be single-turn or 3-, 5-, or 10-turn types. The resistive elements are wire, conductive plastic, or a ceramic-metal substance called *cermet*. The rotating shaft is usually supported in a bushing, which is mounted with a threaded collar. Power ratings are from 0.25 to 2 W with a maximum operating temperature of 125 °C. Resistance values are from 50 ohms to 200 k ohms, depending on the model, and tolerances are +1 % to +5 %. Linearity is +0.25 % to +1.0 % and temperature coefficients are +20 ppm/°C. For the 10-turn models, turn-counting dials are available.

*Trimmers* are used to compensate for the tolerance and variation of fixed components. They are not designed for continuous adjustment, 200 cycles being their design life. The resistance elements are carbon composition, cermet, and wire. They can be single-turn or have multiturn lead screws or worm gears.

## 6.2.3 Transmission Lines

The most common transmission lines are those with two conductors. A number of common configurations are shown in Figure 6.27. Common VHF flat television antenna lead-in cable is an example of a two-wire transmission line. RG/U cables found in almost all laboratories are examples of coaxial transmission lines – such coaxial cable is also used for UHF and satellite television leads. The wire-and-plane configuration is less common. The ribbon-and-plane configuration occurs very often in printed circuit boards (PCBs), where the ribbon is the metal trace on one side of the board and the plane is a continuous conducting sheet on the opposite side, often called the *ground plane*. A transmission line of this type is called a *microstrip line*. The formulae for the characteristic impedance  $Z_0$  of the geometries in Figure 6.27 are given in Table 6.9.

The properties of a transmission line are given in terms of the series impedance  $Z$  and shunt admittance  $Y$  per unit length:

$$\begin{aligned} Z &= R + j\omega L \\ Y &= G + j\omega C \end{aligned} \quad (6.30)$$

where  $j = \sqrt{-1}$  and  $G$  is the shunt conductance, which is the reciprocal of the shunt resistance per unit length. The

Table 6.5 Fixed capacitors

<i>Air</i>			
C range:		<100pF	
Tolerances:		To 0.01%	
Stability:		High	
Temperature coefficient		20 ppm/°C	
Dissipation factor:		$10^{-5}$	
<i>Mica and Glass</i>			
C range:		pF to 1 μF	
Tolerance:		$\pm 1$ to $\pm 20\%$	
Operating Temperature:		-55 to + 70 °C, + 85 °C, +125 °C, or + 150 °C, depending on construction	
Temperature Coefficient:		-20 to + 100 ppm/°C	
V range:		100 to 8000 WV d.c.	
Dissipation factor:		$(1-7) \times 10^{-4}$	
<i>Ceramic Disc</i>			
C range:		1 pF to 1 μF	
Tolerance:		$\pm 1$ to $\pm 20\%$	
V range:		100 to 7500 WV d.c.	
Operating Temperature:		-55 to + 150 °C	
Temperature Coefficient:		0 ppm/°C for NPO to $\pm 750$ ppm/°C for N750	
Dissipation factor:		$5 \times 10^{-4}$	
<i>Film</i>			
Dielectric:	Polyester	Polycarbonate	Polystyrene
C range (μF):	0.0001 to 12 (metalized)	0.0001 to 50 (metalized)	0.0001 to 1 (metalized)
V range (WV d.c.):	50 to 1600	50 to 600	50 to 600
Operating temp. (°C)	-55 to + 125	-55 to + 125	-55 to + 85
Temp. coeff.:	+250 ppm/°C	Nonmonotonic, $\pm 1\%$ total	-50 to 100 ppm/°C
Tolerances:	$\pm 1$ to $\pm 20\%$	$\pm 1$ to $\pm 10\%$	$\pm 1$ to $\pm 10\%$
Dissipation factor:	$10^{-3}$	$3 \times 10^{-3}$	$10^{-4}$
<i>Electrolytic (polarized or unpolarized)</i>			
<i>Aluminum (general purpose)</i>			
C range:		1-50 000 μF	
Tolerance:		-10 to +75%	
V range:		3 to 450 WV d.c.	
Maximum operating temperature:		85 °C	
Direct current leakage:		High	
Low temperature stability:		Low	
Internal inductance limits high-frequency performance			
Prone to leaking			
<i>Tantalum (foil and wet slug)</i>			
Electrolyte:	Wet	Dry	
C range (μF):	15-2200	0.0047-1000	
V range (WV d.c.):	3-300	2-150	
Tolerance:	$\pm 10\%$ to -15 + 75%	$\pm 10$ to $\pm 40\%$	

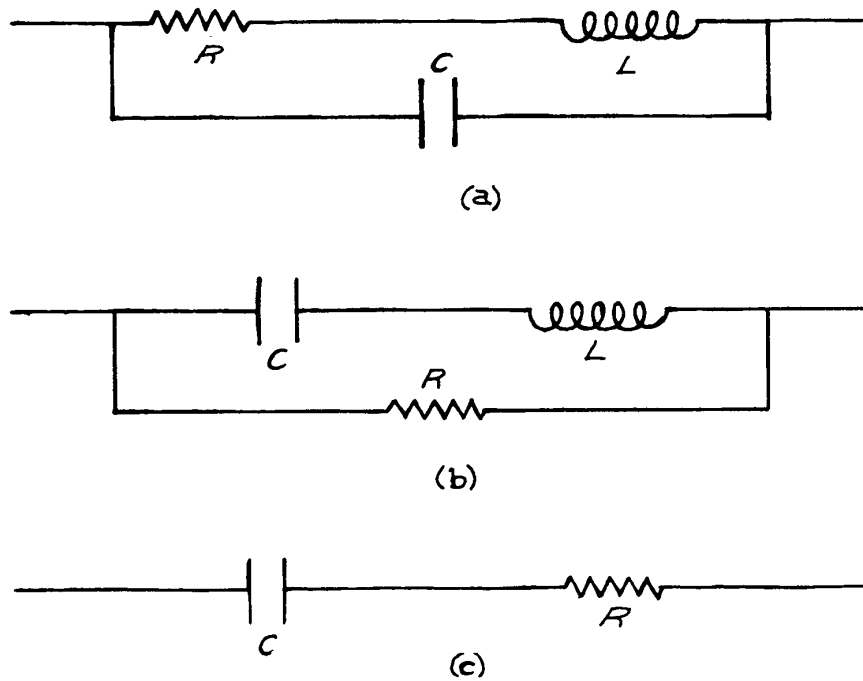


Figure 6.24 Equivalent circuits: (a) resistor; (b) capacitor.

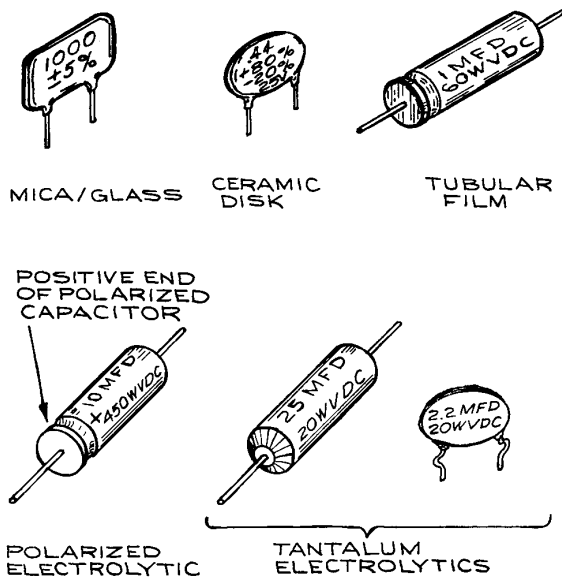


Figure 6.25 Some common capacitor types.

Table 6.6 Resistor coding

Color	Significant Digit	Multiplier	Tolerance
Black	0	1	—
Brown	1	10	—
Red	2	100	—
Orange	3	1 000	—
Yellow	4	10 000	—
Green	5	100 000	—
Blue	6	1 000 000	—
Violet	7	10 000 000	—
Gray	8	—	—
White	9	—	—
Gold	—	—	±5%
Silver	—	—	±10%
No color	—	—	±20%

**Table 6.7 Capacitor coding**

<i>Color</i>	<i>Significant Digits</i>	<i>Multiplier</i>	<i>Capacitance Tolerance</i>	<i>Characteristic</i>	<i>d.c. Working Voltage</i>	<i>Operating Temperature (°C)</i>	<i>EIA/Vibration</i>
Black	0	1	±20	—	—	−55 to +70	10 to 55 Hz
Brown	1	10	±1	B	100	—	—
Red	2	100	±2	C	—	−55 to +85	—
Orange	3	1 000	—	D	300	—	—
Yellow	4	10 000	—	E	—	−55 to +125	10 to 2000 Hz
Green	5	—	±5	F	500	—	—
Blue	6	—	—	—	—	−55 to +150	—
Violet	7	—	—	—	—	—	—
Gray	8	—	—	—	—	—	—
White	9	—	—	—	—	—	EIA
Gold	—	—	±0.5	—	1000	—	—
Silver	—	—	±10	—	—	—	—

Equivalences:

$10^{-6}$  F = 1 μF, 1 MFD.

$10^{-9}$  F = 0.001 μF, 1000 pF.

$10^{-12}$  F = 1 pF, 1 MMFD, 1 μμF.

*Six-Dot or Six-Band Code*

A and B (when marked) Temperature coefficient  
 C first significant figure  
 D second significant figure  
 E Decimal multiplier  
 F Military code number

*Five-Dot or Five-Band Code*

A and B (when marked) Temperature Coefficient  
 C first significant figure  
 D second significant figure  
 E Decimal multiplier

*Temperature Characteristics*

<i>A</i>	<i>B</i>	<i>Temp. Coeff.<sup>a</sup></i>	<i>A</i>	<i>B</i>	<i>Temp. Coeff.<sup>b</sup></i>	<i>Ppm/°C</i>
Gray	Black	Gen. Purpose	Black	—	NP0	0
Orange	Orange	N 1500	Brown	—	N030	−30
Yellow	Orange	N 2200	Red	—	N080	−80
Green	Orange	N 3300	Orange	—	N150	−150
Blue	Orange	N 4700	Yellow	—	N220	−220
Red	Violet	P100	Green	—	N330	−330
Green	Blue	P030	Blue	—	N470	−470
Gold	Orange	X5F	Violet	—	N750	−750
Brown	Orange	Z5F	Gold	—	P100	+100
Gold	Yellow	X5P	White	—	—	—
Brown	Yellow	Z5P	Gray	—	—	—
Gold	Blue	X5S	—	—	—	—
Brown	Blue	Z5S	—	—	—	—

Table 6.7. (contd.)

Temperature Characteristics						
A	B	Temp. Coeff. <sup>a</sup>	A	B	Temp. Coeff. <sup>b</sup>	Ppm/°C
Gold	Gray	X5U				
Brown	Gray	Z5U				

Capacitance					
Color	Digit (C&D)	Multiplier (E)	Nominal Capacitance		
			≤ 10 pF	≥ 10 pF	
Black	0	1	±2.0 pF	±20%	
Brown	1	10	±0.1 pF	±1%	
Red	2	100	—	±2%	
Orange	3	1 000	—	±3%	
Yellow	4	10 000	—	+100%, -0%	
Green	5	—	±0.5 pF	±5%	
Blue	6	—	—	—	
Violet	7	—	—	—	
Gray	8	0.01	±0.2 pF	+80%, to -20%	
White	9	0.1	±1.0 pF	±10%	

<sup>a</sup> Nominal capacitance code is EIA-RS 198.MIL-SPEC code not the same. Five- and six-digit codes are both used for radial-lead and axial-lead capacitors. Disc capacitors normally have typographical marking but may be color-coded.

<sup>b</sup> N designates a negative temperature coefficient; P designates a positive temperature coefficient.

characteristic impedance  $Z_0$  is  $\sqrt{Z/Y}$ . Voltage and current propagate along a transmission line as waves. The properties of the waves are determined by the propagation constant  $\gamma$  and phase velocity  $v_p$ . They are given by  $\gamma = \sqrt{ZY} = \alpha + j\beta$  and  $v_p = \omega/\beta$ , where  $\omega$  is the frequency of the wave. The wavelength in the transmission line is  $\lambda = 2\pi v_p/\omega = 2\pi/\beta$ . The attenuation of the wave is given by the factor  $e^{-\alpha}$  per unit length. These formulae are summarized in Table 6.10. For most practical transmission lines, the series resistance per unit length  $R$  is small compared to the inductive reactance  $\omega L$ , and the shunt conductance  $G$  is small compared to the reciprocal of the capacitance reactance  $1/\omega C$ . Under these conditions,  $Z_0$  is essentially  $\sqrt{L/C}$ ,  $\alpha$  becomes  $[R/Z_0 + GZ_0]/2$ , and  $\beta$  becomes  $\omega\sqrt{LC}$ . The phase velocity is independent of frequency and is the ratio of the speed of light in vacuum to the square root of the dielectric constant of the material separating the conductors.

For reasonable values of conductor geometry,  $Z_0$  lies approximately between 50 and 500 ohms. The attenuation  $\alpha$  is obtained from the shunt conductance  $G$  and series resistance  $R$ , once  $Z_0$  is known. For most dielectrics,  $G$  is very small (for air,  $G$  can be taken equal to zero) and  $R$  becomes the dominant factor. Because of the *skin effect*,  $R$  increases with frequency (see Table 6.1). The properties of common coaxial cables are given in Table 6.11. Attenuation ratings for RG/U cables as a function of frequency are given in Table 6.12, along with jacket information. As is to be expected, the smaller cables show greater attenuation due to the necessity of using a small-diameter inner conductor to maintain a reasonable ratio of outer to inner conductor diameter. The most common cable has a characteristic impedance of 50 ohms. Twisted pairs of single-conductor wire are used where immunity from noise is important. Unwanted signals induced in both conductors can be canceled electronically by differential amplification.

**Table 6.8 Surface mount (SMT) codes***SMT Resistor Tolerance Code*

<i>Letter</i>	<i>Tolerance</i>
D	±0.5%
F	±1.0%
G	±2.0%
J	±5.0%

*SMT Capacitor Significant Figure Code*

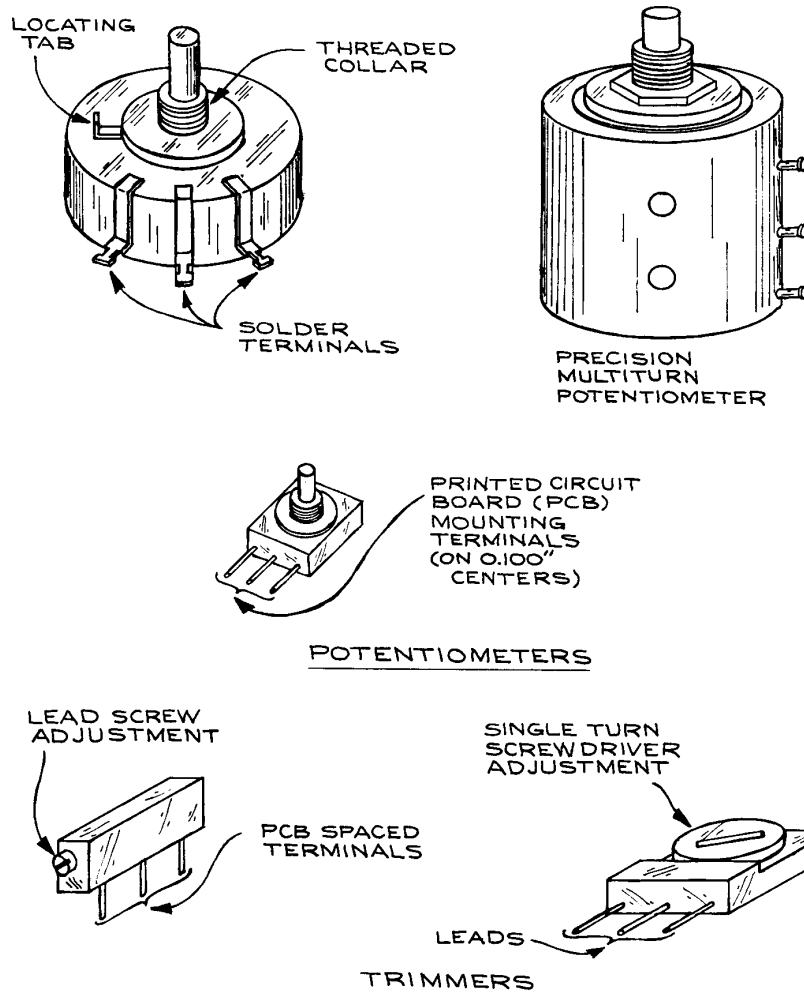
<i>Character</i>	<i>Figures</i>	<i>Character</i>	<i>Figures</i>	<i>Character</i>	<i>Figures</i>
A	1.0	N	3.3	a	2.5
B	1.1	P	3.6	b	3.5
C	1.2	Q	3.9	d	4.0
D	1.3	R	4.3	e	4.5
E	1.5	S	4.7	f	5.0
F	1.6	T	5.1	m	6.0
G	1.8	U	5.6	n	7.0
H	2.0	V	6.2	t	8.0
J	2.2	W	6.8	y	9.0
K	2.4	X	7.5		
L	2.7	Y	8.2		
M	3.0	Z	9.1		

*SMT Capacitor Code*

<i>Character</i>	<i>Multiplier</i>
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

As a general rule, at least one end of a transmission line should be terminated with a resistance equal to  $Z_0$  to prevent ringing and undesirable reflections if the line is used for sending signals over distances greater than  $\lambda/8$ . The transmission of pulses with nanosecond rise times over distances of a few centimeters on a printed circuit board,

for example, requires termination. In theory, termination of the source or load end of the line is sufficient. In practice, however, it is usually the load end that is terminated. Because 50 ohms cable is so common, electronic instruments designed to operate at high frequencies often have 50 ohms input and output impedances to facilitate



**Figure 6.26** Some common types of variable resistors.

connections. Such low-impedance instruments are not suitable for use in combination with lower-frequency devices, which may be seriously overloaded by 50 ohms.

The velocity of propagation  $v_p$  of an electrical signal through a transmission line is  $1/\sqrt{LC} = Z_0/L = 1/Z_0C$ . For 50 ohms cable with a polyethylene or Teflon™ insulation,  $v_p$  is of the order of  $2 \times 10^{10}$  m/s, which gives a delay of 5 ns/m. For delays up to a few hundred nanoseconds, high-quality, low-loss coaxial cable provides an excellent delay line, because no distortion of the propa-

gated signal occurs when it is properly terminated. The best cable to use for critical delay-line applications is rigid or semirigid air-core coaxial cable. With air as the dielectric, the shunt conductance is virtually zero. Semirigid construction assures that the coaxial geometry is maintained within tight tolerances, reducing distortion. Such cable is practical for use at frequencies in the microwave region.

For longer delays, lumped-element delays are the most practical. These delays consist of low-pass filter sections

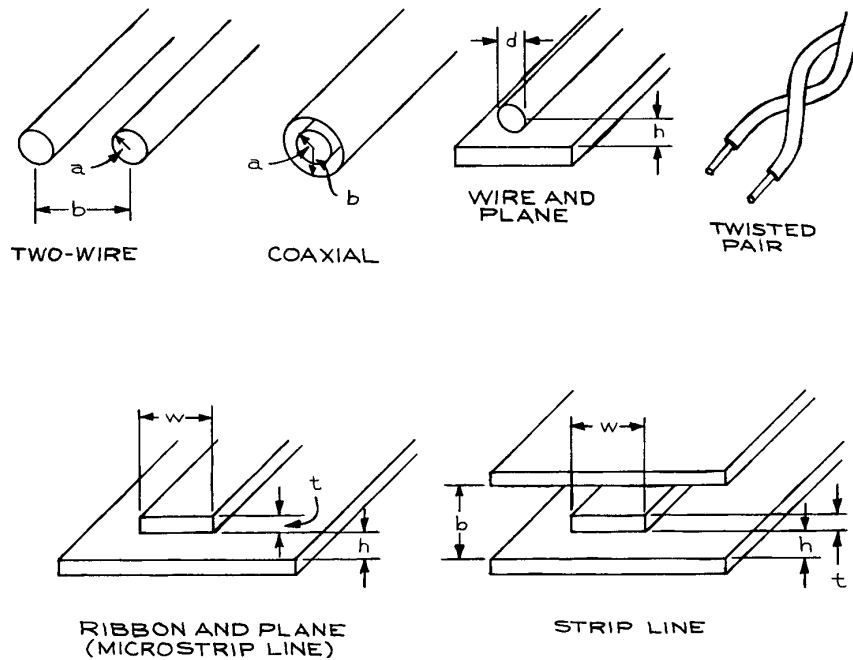


Figure 6.27 Fixed-geometry transmission-line configurations.

Table 6.9 Characteristic impedance of transmission-line geometries

Geometry	$Z_0(\Omega)^a$
Two-wire	$\frac{138}{\sqrt{\epsilon}} \ln \frac{b}{a}$
Coaxial	$\frac{276}{\sqrt{\epsilon}} \ln \frac{b}{a}$
Wire and plane	$\frac{60}{\sqrt{\epsilon}} \ln \frac{4h}{d}$
Ribbon and plane or Microstrip line <sup>b</sup>	$\frac{87}{\sqrt{\epsilon + 1.41}} \ln \frac{5.98h}{0.8w + t}$
Twisted pair (AWG 24–28, 30 turns/f.)	110
Strip line <sup>b</sup>	$\frac{60}{\sqrt{\epsilon}} \ln \frac{4b}{0.67\pi w(0.8 + tw)}$

<sup>a</sup>  $\epsilon = 1.0$  for air and 5.0 for G-10 fiberglass epoxy circuit board.

<sup>b</sup> Copper foil thickness is 0.001 in. for 1 oz cladding and 0.002 in. for 2 oz cladding.

Table 6.10 General transmission line properties

Quantity	Symbol	Relation to Fundamental Parameters
Shunt admittance	$Y$	$G + j\omega C$
Series impedance	$Z$	$R + j\omega L$
Characteristic impedance	$Z_0$	$\sqrt{Z/Y}$
Wavelength	$\lambda$	$2\pi/\beta$
Propagation velocity	$v_p$	$C/\sqrt{\epsilon}$
Attenuation constant	$\alpha$	
Propagation constant	$\gamma$	$\sqrt{ZY} = \alpha + j\beta$

connected in series. For frequencies below their corner frequency, they act as delay lines. The number of sections  $n$  required is related to the ratio of the delay time  $t_d$  to the rise time  $t_r$ , and is given by  $n = 1.5 t_d/t_r$ . The incorporation of multiple taps between the sections allows a choice of delays with a single unit.

For a line length less than a quarter wavelength, a transmission line behaves like a capacitor or an inductor,



Table 6.11 Properties of coaxial cable

<i>Military RG Number</i>	<i>Armor O.D. (in.)</i>	<i>Jacket O.D. (in.)</i>	<i>Jacket Type<sup>a</sup></i>	<i>Dielectric O.D. (in.)</i>	<i>Dielectric Type<sup>b</sup></i>	<i>Center Conductor<sup>c</sup></i>	<i>V.P.<sup>d</sup>(%)</i>	<i>Capacitance (pF/ft.)</i>	<i>Max. RMS Operating Voltage (V)</i>	<i>Nominal Impedance (Ω)</i>
5B	—	0.328	IIa	.181	P	16 S	65.9	28.5	3 000	50
6A/U	—	0.332	IIa	.185	P	21 CW	65.9	20	2 700	75
7A/U	—	0.405	I	.285	P	7/21 C	65.9	29.5	5 000	52
8A/U	—	0.405	IIa	.285	P	7/21 C	65.9	29.5	5 000	52
9	—	0.420	II Grey	.280	P	7/21 S	65.9	30	5 000	51
9A	—	0.420	II Grey	.280	P	7/21 S	65.9	30	5 000	51
9B/U	—	0.420	IIa	.280	P	7/21 S	65.9	30	5 000	50
10A/U	0.475	0.405	IIa	.285	P	7/21 C	65.9	29.5	5 000	52
11A	—	0.405	I	.285	P	7/26 TC	65.9	20.5	5 000	75
11A/U	—	0.405	IIa	.285	P	7/26 TC	65.9	20.5	5 000	75
12A/U	0.475	0.405	IIa	.285	P	7/26 TC	65.9	20.5	5 000	75
13	—	0.420	I	.280	P	7/26 TC	65.9	20.5	5 000	74
13A/U	—	0.420	IIa	.280	P	7/26 T	65.9	20.5	5 000	74
14A/U	—	0.545	IIa	.370	P	10 C	65.9	29.5	7 000	52
17A	—	0.870	IIa	.680	P	.188 C	65.9	29.5	11 000	52
17A/U	—	0.870	IIa	.680	P	.188 C	65.9	29.5	11 000	52
18A/U	0.945	0.870	IIa	.680	P	.188 C	65.9	29.5	11 000	52
19A/U	—	1.120	IIa	.910	P	.250 C	65.9	29.5	14 000	52
20A/U	1.195	1.120	IIa	.910	P	.250 C	65.9	29.5	14 000	52
21A	—	0.332	IIa	.185	P	16 N	65.9	29	2 700	53
22	—	0.405	I	.285	P	Two 7/.0152 C	65.9	16	1 000	95
22B/U	—	0.420	IIa	.285	P'	Two 7/.0152 C	65.9	16	1 000	95
34B/U	—	0.630	IIa	.460	P	7/.0249 C	65.9	21.5	6 500	75
35B/U	0.945	0-870	IIa	.680	P	.1045 C	65.9	21	10 000	75
55B/U	—	0.206	IIIa	.116	P	20 S	65.9	28.5	1 900	53.5
57A/U	—	0.625	IIa	.472	P	Two 7/21 C	65.9	16	3 900	95
58/U	—	0.195	I	.116	P	20 C	65.9	28.5	1 900	53.5
58A/U	—	0.195	I	.116	P	19/.0071 TC	65.9	30	1 900	50
58C/U	—	0.195	IIa	.116	P	19/.0071 TC	65.9	30	1 900	50
59/U	—	0.242	I	.146	P	22 CW	65.9	21.5	2 300	73
59B/U	—	0.242	IIa	.146	P	.023 CW	65.9	21	2 300	75
62/U	—	0.242	I	.146	SSP	22 CW	84	13.5	750	93
62A/U	—	0.242	IIa	.146	SSP	22 CW	84	13.5	750	93
62B/U	—	0.242	IIa	.146	SSP	7/32 CW	84	13.5	750	93
63/U	—	0.405	I	.285	SSP	22 CW	84	10	1 000	125

Table 6.11. (contd.)

Military RG Number	Armor O.D. (in.)	Jacket O.D. (in.)	Jacket Type <sup>a</sup>	Dielectric O.D. (in.)	Dielectric Type <sup>b</sup>	Center Conductor <sup>c</sup>	V.P. <sup>d</sup> (%)	Capacitance (pF/ft.)	Max. RMS Operating Voltage (V)	Nominal Impedance (Ω)
63B/U	—	0.405	IIa	.285	SSP	22 CW	84	10	1 000	125
71A	—	≤ 0.250	I	.146	SSP	22 CW	84	13.5	750	93
71B/U	—	0.250	IIIa	.146	SSP	22 CW	84	13.5	750	93
74A/U	0.615	0.545	IIa	.370	P	10 C	65.9	29.5	7 000	52
79B/U	0.475	0.405	IIa	.285	SSP	22 CW	84	10	1 000	125
87A/U	—	0.425	V	.280	TF	7/20 S	69.5	29.5	5 000	50
108A/U	—	0.235	IIa	.079 ea.	P	Two 7/28 TC	68	23.5	1 000	78
111A/U	0.490	0.420	IIa	.285	P	Two 7/.0152 C	65.9	16	1 000	95
114/U	—	0.405	I	.285	SSP	33 CW	88	6.5	1 000	185
114A/U	—	0.405	IIa	.285	SSP	33 CW	88	6.5	1 000	185
115	—	0.375	V	.250	'IT	7/21 S	70	29.5	5 000	50
115A/U	—	0.415	V	.250	'IT	7/21 S	70	29.5	4 000	50
116/U	0.475	0.425	V	.280	TF	7/20 S	69.5	29.5	5 000	50
122/U	—	0.160	IIa	.096	P	27/37 TC	65.9	29.5	1 900	50
140/U	—	0.233	V	.146	TF	.025 SCW	69.5	21	2 300	75
141/U	—	0.190	V	.116	TF	.0359 SCW	69.5	28.5	1 900	50
141A/U	—	0.190	V	.116	TF'	.039 SCW	69.5	28.5	1 900	50
142/U	—	0.206	V	.116	TF'	.0359 SCW	69.5	28.5	1 900	50
142/B	—	0.195	IX	.116	TF	.039 SCW	69.5	28.5	1 900	50
143	—	0.325	V	.185	TF	.057 SCW	69.5	28.5	3 000	50
142A/U	—	0.206	V	.116	TF	.039 SCW	69.5	28.5	1 900	50
143A/U	—	0.325	V	.185	TF	.059 SCW	69.5	28.5	3 000	50
149/U	—	0.405	I	.285	P	7/26 TC	65.9	20.5	5 000	75
164/U	—	0.870	IIa	.680	P	.1045 C	65.9	21	10 000	75
174/U	—	0.100	I	.060	P	7/.0063 CW	65.9	30	1 500	50
178B/U	—	0.075	VII	.034	TF	7/38 SCW	69.5	29.0	1 000	50
179B/U	—	0.105	VII	.063	TF'	7/38 SCW	69.5	19.5	1 200	75
180B/U	—	0.145	VII	.102	TF	7/38 SCW	69.5	15.0	1 500	95
187/U	—	0.110	VII	.063	TF	7/38 SCW	69.5	19.5	1 200	75
188/U	—	0.110	VII	.060	TF	7/.0067 SCW	69.5	29	1 200	50
195/U	—	0.155	VII	.102	TF	7/38 SCW	69.5	15	1 500	95
196/U	—	0.080	VII	.034	TF	7/38 SCW	69.5	29	1 000	50
209/U	—	0.745	VI	.500	SST	19/.0378 S	84	26.5	3 200	50
210/U	—	0.242	V	.146	SST	22 SCW	84	13.5	750	93
211A/U	—	0.730	V	.620	TF	.190 C	69.5	29.0	7 000	50

Table 6.11. (contd.)

Military RG Number	Armor O.D. (in.)	Jacket O.D. (in.)	Jacket Type <sup>a</sup>	Dielectric O.D. (in.)	Dielectric Type <sup>b</sup>	Center Conductor <sup>c</sup>	V.P. <sup>d</sup> (%)	Capacitance (pF/ft.)	Max. RMS Operating Voltage (V)	Nominal Impedance (Ω)
212/U	—	0.332	IIa	.185	P	.0556 S	65.9	28.5	3 000	50
213/U	—	0.405	IIa	.285	P	7/.0296 C	65.9	29.5	5 000	50
214/U	—	0.425	IIa	.285	P	7/.0296 S	65.9	30	5 000	50
215/U	0.475	0.405	IIa	.285	P	7/.0296 C	65.9	29.5	5 000	50
216/U	—	0.425	IIa	.285	P	7/26 TC	65.9	20.5	5 000	75
217/U	—	0.545	IIa	.370	P	.106 C	65.9	29.5	7 000	50
218/U	—	0.870	IIa	.680	P	.195 C	65.9	29.5	11 000	50
219/U	0.945	0.870	IIa	.680	P	.195 C	65.9	29.5	11 000	50
220/U	—	1.120	IIa	.910	P	.260 C	65.9	29.5	14 000	50
221/U	1.195	1.120	IIa	.910	P	.260 C	65.9	29.5	14 000	50
222/U	—	0.322	IIa	.185	P	.0556 N	65.9	29	2 700	50
223/U	—	0.216	IIa	.116	P	.035 S	65.9	29.5	1 900	50
225/U	—	0.430	V	.285	TF	7/.0312 S	69.5	29.5	5 000	50
227/U	0.490	0.430	V	.285	TF	7/.0312 S	69.5	29.5	5 000	50
228A/U	0.795	0.730	V	.620	TF	.190 C	69.5	29.0	7 000	50
264/U	—	0.750	PU	.176	P	Four 19/27 C	69.5	40.0	N.A.	40
280/U	—	0.480	IX	.327	TT	9C	80	27	3 000	50
281/U	—	0.750	VI	.500	TT	19/.0378 S	80	27	3 200	50
301/U	—	0.245	IX	.185	TF	7/.0203 K	69.5	28.5	3 000	50
302/U	—	0.206	IX	.146	TF	22 SCW	69.5	21	2 300	75
303/U	—	0.170	IX	.116	TF	.038 SCW	69.5	28.5	1 900	50
304/U	—	0.280	IX	.185	TF	.059 SCW	69.5	28.5	3 000	50
307A/U	—	0.270	IIIa	.146	SSP	19/.0058 S	80.0	17	N.A.	75
316/U	—	0.102	IX	.060	TF	7/.0067 SCW	69.5	29	1 200	50

<sup>a</sup> Designation listed at end of this table.

<sup>b</sup> P: polyethylene; SSP: semisolid polyethylene; TF: Teflon; TT: Teflon tape; SST: semisolid Teflon.

<sup>c</sup> Number of strands, gauge (B&S) or O.D., and material are specified. A single solid wire results in the lowest cable attenuation. Stranding increases flexibility. S: silvered copper; C: copper; TC: tinned copper; N: Nichrome; CW: Copperweld; SCW: silvered Copperweld; K: Karma.

<sup>d</sup> 100% x (velocity of propagation)/(velocity in vacuum).

depending on whether the end of the line is an open circuit or short circuit. A line that is an odd number of quarter-wavelengths long behaves like a parallel resonant *RLC* circuit when short-circuited at one end, and like a series resonant circuit when open-circuited. The *Q*s of such resonant circuits are much higher than

attainable with lumped elements. There are, however, an infinite number of resonances corresponding to frequencies where:

$$v/pf = (2n+1)(\lambda/4)$$

where *n* is an integer.

## 6.2.4 Coaxial Connectors

Probably more time is spent with connectors than with any other element of electronic hardware. When deciding on the method of connecting various pieces of electronic equipment, it is wise to be aware of the variety of connectors available and choose a family of connectors best suited to the task. In many cases, the choice of connector is dictated by the connectors at the outputs and inputs of existing equipment.

The most common connector is currently the BNC. It is a miniature bayonet type, but circuit density has increased enormously with the extensive use of integrated circuits

and the BNC connector is now being replaced by a variety of subminiature connectors – among which the LEMO™ has taken the lead for nuclear instrumentation. A list of common coaxial connectors is given in Table 6.13.

In addition to coaxial connector plugs, each series of connectors has a wide variety of jacks and adaptors (generally *plug* refers to the connector attached to the cable or wire, while *jack* refers to the connector to which the plug mates). Jacks for chassis mounting, bulkhead mounting, and printed circuit-board mounting are available, as well as hermetically sealed jacks for making coaxial connections through a vacuum wall. *Tees* and *barrels* allow one to

**TABLE 6.12 Properties of coaxial cable**  
*Attenuation ratings for RG/U cable*

Frequency (MHz):	Nominal Attenuation [dB/(100 ft)]									
	1.0	10	50	100	200	400	1000	3000	5000	10 000
5, 5A, 5B, 6, 6A, 212	0.26	0.83	1.9	2.7	4.1	5.9	9.8	23.0	32.0	56.0
7	0.18	0.64	1.6	2.4	3.5	5.2	9.0	18.0	25.0	43.0
8, 8A, 10, 10A, 213, 215	0.15	0.55	1.3	1.9	2.7	4.1	8.0	16.0	27.0	>100.0
9, 9A, 9B, 214	0.21	0.66	1.5	2.3	3.3	5.0	8.8	18.0	27.0	45.0
11, 11A, 12, 12A, 13, 13A, 216	0.19	0.66	1.6	2.3	3.3	4.8	7.8	16.5	26.5	> 100.0
14, 14A, 74, 74A, 217, 224	0.12	0.41	1.0	1.4	2.0	3.1	5.5	12.4	19.0	50.0
17, 17A, 18, 18A, 177, 218, 219	0.06	0.24	0.62	0.95	1.5	2.4	4.4	9.5	15.3	> 100.0
19, 19A, 20, 20A, 220, 221	0.04	0.17	0.45	0.69	1.12	1.85	3.6	7.7	11.5	> 100.0
21, 21A, 222	1.5	4.4	9.3	13.0	18.0	26.0	43.0	85.0	> 100.0	> 100.0
22, 22B, 111, 111A	0.24	0.80	2.0	3.0	4.5	6.8	12.0	25.0	> 100.0	> 100.0
29	0.32	1.20	2.95	4.4	6.5	9.6	16.2	30.0	44.0	> 100.0
34, 34A, 34B	0.08	0.32	0.85	1.4	2.1	3.3	5.8	16.0	28.0	> 100.0
35, 35A, 35B, 164 0.08	0.08	0.58	0.85	1.27	1.95	3.5	8.6	15.5	> 100.0	> 100.0
54, 54A	0.33	0.92	2.15	3.2	4.7	6.8	13.0	25.0	37.0	> 100.0
55, 55A, 55B, 223 0.03	0.30	1.2	3.2	4.8	7.0	10.0	16.5	30.5	46.0	> 100.0
57, 57A, 130, 131 0.18	0.65	1.6	2.4	3.5	5.4	9.8	21.0	> 100.0	> 100.0	> 100.0
58, 58B	0.33	1.25	3.15	4.6	6.9	10.5	17.5	37.5	60.0	> 100.0
58A, 58C	0.44	1.4	3.3	4.9	7.4	12.0	24.0	54.0	83.0	> 100.0
59, 59A, 59B	0.33	1.1	2.4	3.4	4.9	7.0	12.0	26.5	42.0	> 100.0
62, 62A, 71, 71A, 71B	0.25	0.85	1.9	2.7	3.8	5.3	8.7	18.5	30.0	83.0
62B	0.31	0.90	2.0	2.9	4.2	6.2	11.0	24.0	38.0	92.0
63, 63B, 79, 79B	0.19	0.52	1.1	1.5	2.3	3.4	5.8	12.0	20.5	44.0
87A, 116, 165, 166, 225, 227	0.18	0.60	1.4	2.1	3.0	4.5	7.6	15.0	21.5	36.5
94	0.15	0.60	1.6	2.2	3.3	5.0	7.0	16.0	25.0	60.0
94A, 226	0.15	0.55	1.2	1.7	2.5	3.5	6.6	15.0	23.0	50.0

**Table 6.12 (contd.)**

<i>Frequency (MHz):</i>	<i>Nominal Attenuation [dB/(100 ft)]</i>									
	1.0	10	50	100	200	400	1000	3000	5000	10 000
108, 108A	0.70	2.3	5.2	7.5	11.0	16.0	26.0	54.0	86.0	> 100.0
114, 114A	0.95	1.3	2.1	2.9	4.4	6.7	11.6	26.0	40.0	65.0
115, IISA, 235	0.17	0.60	1.4	2.0	2.9	4.2	7.0	13.0	20.0	33.0
117, 118, 211, 228	0.09	0.24	0.60	0.90	1.35	2.0	3.5	7.5	12.0	37.0
119, 120	0.12	0.43	1.0	1.5	2.2	3.3	5.5	12.0	17.5	54.0
122	0.40	1.7	4.5	7.0	11.0	16.5	29.0	57.0	87.0	> 100.0
125	0.17	0.50	1.1	1.6	2.3	3.5	6.0	13.5	23.0	> 100.0
140, 141, 141A	0.30	0.90	2.1	3.3	4.7	6.9	13.0	26.0	40.0	90.0
142, 142A, 142B	0.34	1.1	2.7	3.9	5.6	8.0	13.5	27.0	39.0	70.0
143, 143A	0.25	0.85	1.9	2.8	4.0	5.8	9.5	18.0	25.5	52.0
144	0.19	0.60	1.3	1.8	2.6	3.9	7.0	14.0	22.0	50.0
149, 150	0.24	0.88	2.3	3.5	5.4	8.5	16.0	38.0	65.0	> 100.0
161, 174	2.3	3.9	6.6	8.9	12.0	17.5	30.0	64.0	99.0	> 100.0
178, 178A, 196	2.6	5.6	10.5	14.0	19.0	28.0	46.0	85.0	> 100.0	> 100.0
179, 179A, 187	3.0	5.3	8.5	10.0	12.5	16.0	24.0	44.0	64.0	> 100.0
180, 180A, 195	2.4	3.3	4.6	5.7	7.6	10.8	17.0	35.0	50.0	88.0
188, 188A	3.1	6.0	9.6	11.4	14.2	16.7	31.0	60.0	82.0	> 100.0
209	0.08	0.27	0.68	1.0	1.6	2.5	4.4	9.5	15.0	48.0
281	0.09	0.32	0.78	1.1	1.7	2.5	4.5	9.0	13.0	24.0

*Jacket Type*

<i>Designation</i>	<i>Material</i>	<i>Temperature Limits (°C)</i>
Type I	Black polyvinyl chloride	-40 to +80
Type IIa	Black polyvinyl chloride (noncontaminating):	
	< 1/4 in.	-55 to +80
	> 1/4 in.	-40 to +80
Type IIIa	Black polyethylene	-55 to +85
Type IV	Black synthetic rubber:	
	< 1/2 in.	-55 to +80
	> 1/2 in.	-40 to +80
Type V	Fiber glass	-55 to +250
Type VI	Silicone rubber	-55 to +175
Type VII	Polytetrafluoroethylene (Teflon)	-55 to +200
Type IX	Fluorinated Ethylene Propylene (FEP)	-55 to +200
PU	Polyurethane (Estane)	-60 to +180

**Table 6.13 Common coaxial connector types**

<i>Size Classification</i>	<i>Connector Type</i>	<i>Attachment Method</i>	<i>Cable Size Range (in.)</i>	<i>Maximum Frequency (GHz)</i>	<i>RMS Working Voltage (V)</i>
Medium	UHF	Screw	$\frac{3}{16}$ to $\frac{7}{8}$	0.5	500
	N	Screw	$\frac{3}{8}$ to $\frac{7}{8}$	11	1000
	C	Bayonet			1500
	SC	Screw			1500
	7-mm precision	Sexless Screw	.141, .250, .325, semirigid; RG 214/U, RG 142B/U	18	1500
Miniature	BNC	Bayonet	$\frac{1}{8}$ to $\frac{3}{8}$	$\left\{ \begin{array}{l} 4 \\ 4 \\ 11 \end{array} \right.$	500
	MHV <sup>a</sup>	Bayonet			5000
	TNC	Screw			500
	SHV	Bayonet	0.080 to 0.420	—	10000 d.c. <sup>b</sup>
Subminiature	SMB	Snapon	$\frac{1}{16}$ to $\frac{1}{8}$	$\left\{ \begin{array}{l} 3 \\ 10 \\ 18 \end{array} \right.$	500
	SMC	Screw			500
	SMA	Screw			350
	LEMO <sup>c</sup>	Push On	0.100 to 0.195	4t	1500

<sup>a</sup> Higher-voltage version of the BNC connector, not mateable with BNC connectors.

<sup>b</sup> NIM high-voltage connector.

<sup>c</sup> 00 shell size. Manufactured in the United States as K-Loc by Kings Electronics Co.

Source: Reference 2.

connect cables together and there are *terminators* that can be attached directly to input or output jacks. In addition, each series has a certain number of interseries adaptors for connection of plugs or jacks to those of other series. Some of the different connectors are shown in Figure 6.28. It is advisable to reduce the number of adaptors to a minimum to maintain the transmission-line properties of the connections with as few reflection-producing discontinuities as possible.

A good connector should be easy to fit to the proper cable and should not introduce any discontinuities in the transmission-line properties of the cable. Needless to say, the characteristic impedance of the connector should be identical to that of the cable. The connector should be a

strong mechanical fit to the cable, and it should be easy to connect and disconnect from mating connectors.

Any given connector type comes in a variety of styles to accommodate different cable sizes. Configurations – such as straight and right angle – are available, and often there are different choices for the mechanical and electrical connections of the connector to the cable, the three most common being a ferrule clamp and screw, crimp, and solder. The procedure for stripping coaxial cable by hand is shown in Figure 6.29(a). Shielded lead wire extractors are available for removing insulated lead wire from braided shield. Similar to a syringe, the extractor removes the lead wire from the shield leaving the braid intact. Figure 6.29(b) shows the assembly of a

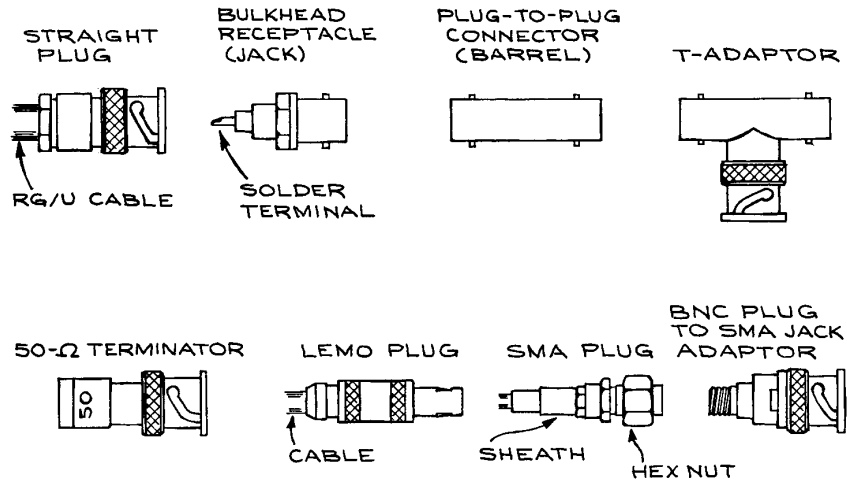


Figure 6.28 Some coaxial connectors.

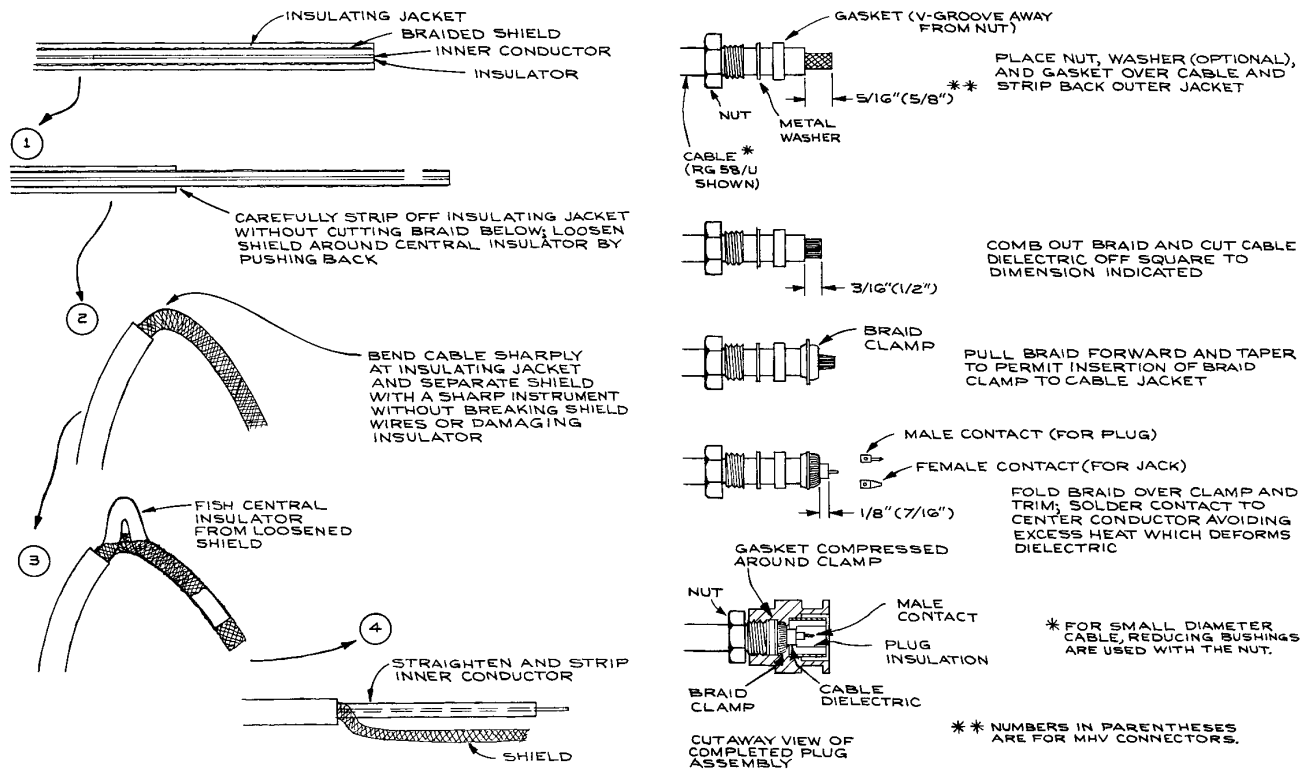


Figure 6.29 Procedure for stripping coaxial cables; Assembly procedure for BNC/MHV ferrule-clamp connector.

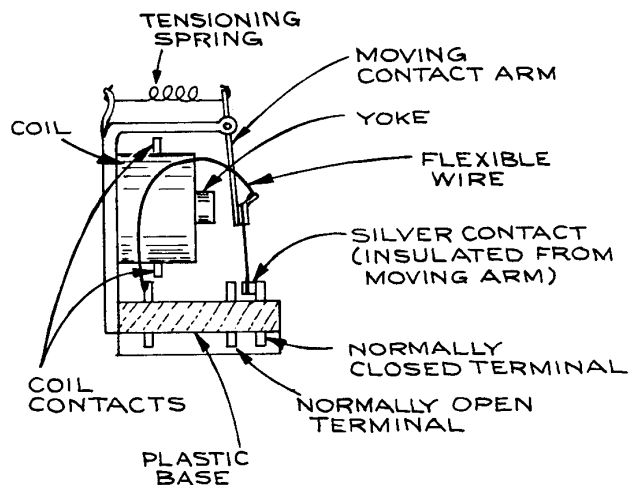
ferrule-clamp BNC/MHV connector. Crimp style connectors require the use of a special crimping tool and die. They are stronger, less bulky, more uniform, and electrically more robust than clamp and solder connectors. Crimping avoids the swelling or melting of the cable dielectric that can occur when solder-type connectors are assembled. When attaching connectors to cable, it is highly recommended that the manufacturer's cable-cutting procedures be followed so that the inner conductor, insulator, outer conductor (braid or foil), and outer insulator are of the proper length. In this way no discontinuities in the electrical properties of the cable are introduced at the cable–connector interface. Special cutting jigs are available to assure the proper geometry when preparing the cable for insertion into the connector.

Assembly instructions for several types of coaxial connectors are given in the Chapter 6 Appendix.

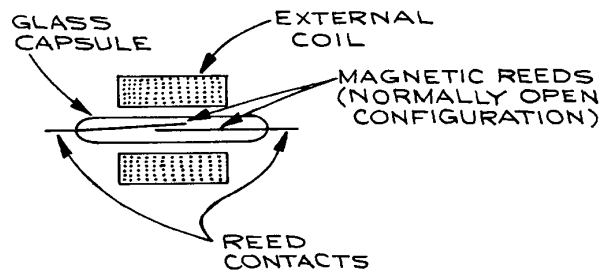
## 6.2.5 Relays

A relay is an electrically actuated switch. Relay action can be electromechanical or can be based on solid-state devices. A large number of contact configurations exist. Important parameters are the operating voltage and power rating of the relay, as well as the contact rating. The speed at which mechanical relays operate is generally in the 10 to 100 ms range. General-purpose electromagnetic relays operate at 6, 12, 24, 48, and 110 V d.c. and 6, 12, 24, 48, 110, and 220 V a.c. with coil power ratings of around 1 W [see Figure 6.30(a)]. Such relays have contacts capable of handling from 2 to 10 A. Sensitive mechanical relays use a large number of turns of fine wire on the electromagnet coil. They operate with 28 V d.c. and require only 1 to 40 mW of coil power. Contacts are rated from 0.5 to 2 A.

A *reed relay* is a glass-encapsulated switch having two flexible thin metal strips, or reeds, as contactors. The switch is placed inside an electromagnetic coil [Figure 6.30(b)]. The relay contacts can be dry or mercury-wetted. Standard coil voltages are 6, 12, and 24 V d.c., and the coil power is from 50 to 500 mW per reed-switch capsule. Contacts are rated at 10 mA to 3 A, depending on size. There is also a maximum open-circuit voltage that can be sustained by the opened contacts. The low mass of the reeds makes operating times of 0.2 to 2 ms possible. Small reed relays are made for PCB mounting.



(a)



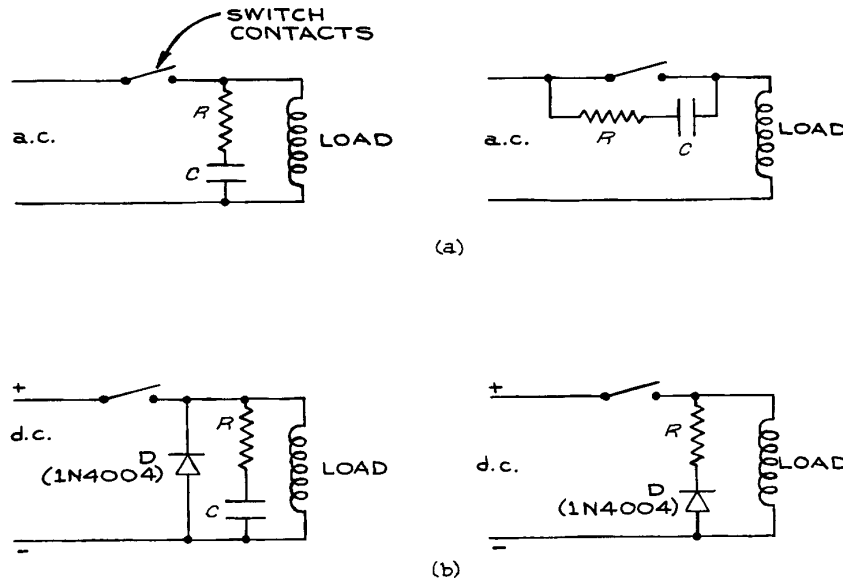
(b)

**Figure 6.30** Mechanical relays: (a) medium-power relay; (b) reed relay.

With inductive loads, the fast opening of relay contacts can cause a large voltage to appear across them. Some protective circuits for reducing this problem are illustrated in Figure 6.31.

Solid-state relays consist of a solid-state switching element driven by an appropriate amplifier. Because the input circuit is often optically isolated from the load, these relays introduce little or no noise into the circuit that drives them. The load current determines the switching element. Field-effect transistors (FETs) are used for low-level d.c., bipolar junction transistors (BJTs) for intermediate d.c. currents, and triacs and silicon-controlled rectifiers (SCRs) for large a.c. and d.c. currents. Section 6.3 has a discussion





**Figure 6.31** Contact-conditioning circuits for mechanical relays: (a) a.c. (for small loads driven from the power line,  $R = 100$  ohms and  $C = 0.05$   $\mu\text{F}$ ); (b) d.c. (diodes, D, reduce the peak voltage from the inductive load and should be able to handle the steady-state current through the inductor).

of transistors and solid-state switches. Though there is no contact wear with solid-state relays, their *on* resistance may be large compared with electromechanical devices and they require carefully conditioned activating signals. They are much less tolerant of overload than are electromechanical devices.

## 6.3 ACTIVE COMPONENTS

The overall properties of semiconductor diodes and transistors can be understood in terms of the properties of the  $p$ - $n$  junctions that form them. Without going into a discussion of solid-state physics, it is sufficient to know that  $p$ -type semiconductor material conducts electricity principally through the motion of positive charges (holes), while  $n$ -type material owes its conductivity mainly to electrons. Both  $p$ - and  $n$ -material can be produced from germanium or silicon by the addition of impurity atoms – a procedure called *doping*.

### 6.3.1 Diodes

Most semiconductor diodes consist of a  $p$ - $n$  junction as illustrated in Figure 6.32. The schematic representation is shown next to the simplified drawing of the diode. The diode is a unidirectional device: its d.c. current–voltage ( $I$ - $V$ ) characteristics depend on the applied potential on the *anode* with respect to the *cathode*. The  $I$ - $V$  characteristics of a typical small-signal silicon diode are drawn in Figure 6.33. When the anode is positive with respect to the cathode (*forward bias*), the diode conducts a large current. When the polarity of the applied potential is reversed (*reverse bias*), the current is very small. If the reverse voltage is increased beyond the *peak reverse voltage* (PRV), breakdown occurs and the diode conducts strongly again. Except in the case of specially designed Zener diodes, this usually destroys the diode. The forward-bias voltage at which the diode begins to conduct strongly is  $V_{\text{cutin}}$ . For germanium diodes,  $V_{\text{cutin}}$  is 0.2 V, while for

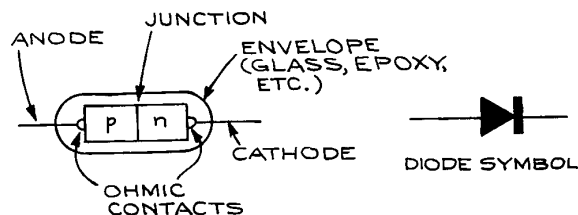


Figure 6.32 A diode and the corresponding symbol.

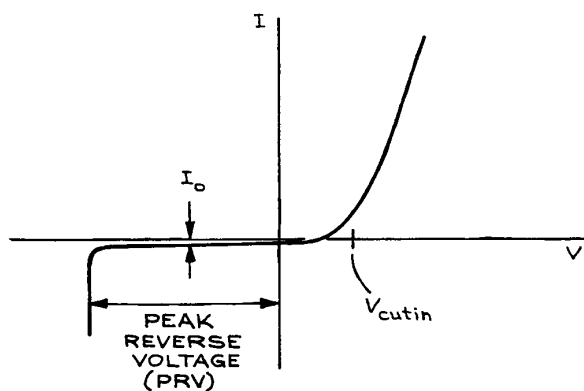


Figure 6.33 Typical current-voltage characteristics of a small-signal silicon diode. Note change of vertical and horizontal scales at the origin.

silicon diodes it is 0.6 V. The  $I$ - $V$  characteristics shown in the graph are accurately represented by the formula:

$$I_d = I_0 \left[ e^{V_d/\eta V_T} - 1 \right] \quad (6.31)$$

where  $I_0 = 10^{-9}$  A for silicon and  $10^{-6}$  A for germanium.  $I_0$  is temperature-dependent, increasing at a rate of 11%/°C for Ge and 8%/°C for Si.  $V_T$  has the value 0.026 V at 300 K and is directly proportional to the absolute temperature, and  $\eta$  is a unitless constant equal to one for germanium and two for silicon.

Diodes can be separated into two types according to the power they can dissipate. *Power diodes* are used in high-current applications, for example, as rectifiers in power supplies. Important specifications for power diodes are the *maximum forward current*, maximum PRV (peak

reverse voltage), and *effective forward-bias resistance*. Low-power diodes or *signal diodes* are used to rectify small signals, mix two frequencies to produce sum and difference frequencies, and switch low voltages and currents. Besides the current and voltage ratings, the speed with which diodes can be changed from the forward-bias to the reverse-bias condition and back is a factor in high-speed switching circuits. Switching speed is related to the effective capacitance of the  $p$ - $n$  junction. Typical specifications for a power and signal diode are given in Tables 6.14 and 6.15.

Many older diodes have a 1N-prefix designation. The two ends of a diode are usually distinguished from each other by a mark – the cathode end having a black band for glass-encapsulated diodes and a white band for black plastic-encapsulated diodes. Power diodes can dissipate large amounts of heat and are often constructed for ease of mounting to a heat sink for efficient heat dissipation. Some diode configurations are shown in Figure 6.34. In the case of the high-power diode, the cathode is a stud that can be attached directly to a metal heat sink. Generally, fiber or mica washers are used to isolate the cathode electrically from the metal heat sink, and thermal conductivity is enhanced with special silicone grease between the washers and heat sink. The mounting technique is shown in Figure 6.35.

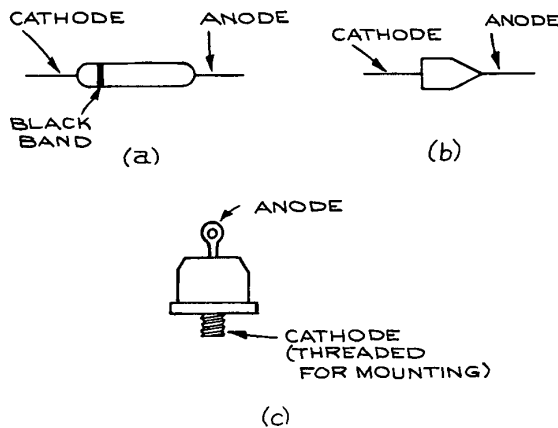
Because the current through a diode is a nonlinear function of the voltage across it, the effective resistance changes with diode voltage. This is shown in Figure 6.33, where the reciprocal of the slope of the curve passing through the origin and a point on the  $I$ - $V$  characteristic curve defines the *static resistance* at that point. The

Table 6.14 Properties of the 1n914 fast-switching diode

Maximum Ratings	
Peak reverse voltage	75 V
Reverse Current	25 nA
Average forward current	75 mA
Peak surge current (1 sec)	500 mA
Electrical Characteristics	
Junction capacitance	4 pF
Reverse recovery time	4–8 ns

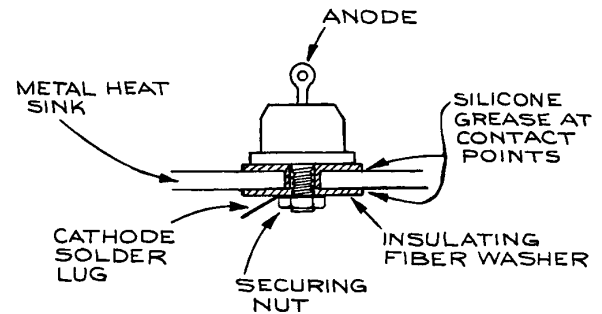
**Table 6.15 Properties of 1N4001-1N4007 1-ampere silicon rectifiers**

	Maximum Ratings						
	4001	4002	4003	4004	4005	4006	4007
Peak reverse voltage (V)	50	100	200	400	600	800	1000
Non-repetitive peak reverse voltage (V)	50	100	200	400	600	800	1000
Average forward current (A)				1			
Peak Surge current (1 cycle) (A)				30			
Operating temperature range (°C)	50			-65 to +175			



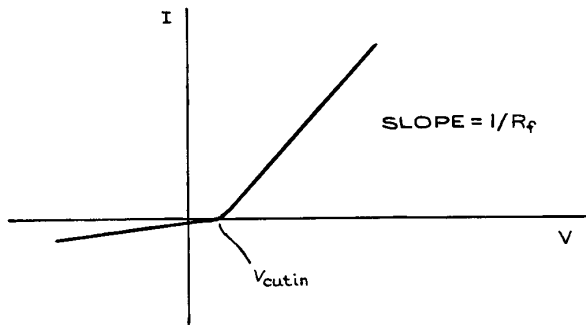
**Figure 6.34** Diode case configurations: (a) glass-encapsulated signal diode; (b) plastic-encapsulated medium-power diode; (c) high-power diode.

*dynamic resistance* is the reciprocal of the slope of the tangent to the curve at the point. The change in current through a diode for small changes in voltage across it requires knowledge of both static and dynamic resistance. For the purposes of simplified circuit analysis, it is useful to represent the  $I$ - $V$  characteristic curve of a diode by a so-called *piecewise linear model*. With this model, the forward- and reverse-bias regions are represented by straight lines with the transition occurring at the cutin voltage. The piecewise linear curve is shown in Figure 6.36. The slope of the curve at voltages above  $V_{\text{cutin}}$  is the

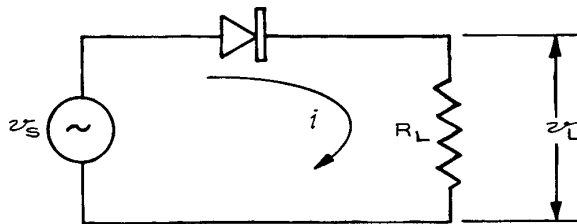


**Figure 6.35** Power-diode mount that provides electrical insulation with good heat dissipation.

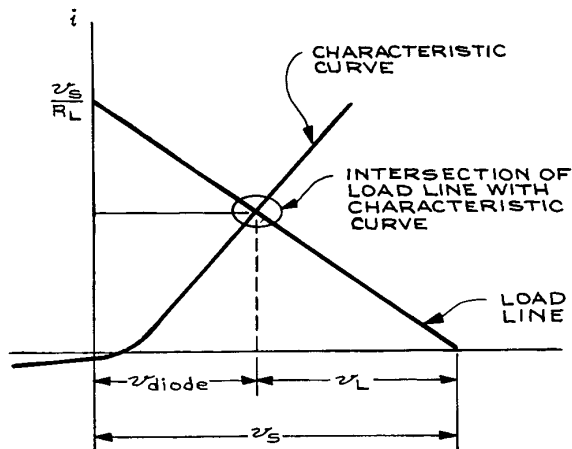
reciprocal of the average forward resistance. With this model it is a simple matter to calculate the effect of the diode in a circuit, since it behaves like a resistor in the forward-bias condition with resistance  $R_f$ , and in the reverse-bias condition it acts essentially like an open circuit. A graphical representation of this can be obtained by what is called *load-line analysis*. Consider the circuit with an a.c. voltage source shown in Figure 6.37. The current through  $R_L$  can be determined in the following manner: given a value of  $V_s$ , the voltage across the diode,  $V_d$ , and the current through it,  $I_d$ , will be determined by the intersection of the straight line with coordinates  $(V_s, 0)$  and  $(0, V_s/R_L)$ , as in Figure 6.38. Different values of  $V_s$  will give a family of parallel straight lines. From the intersection points, one constructs a graph of  $V_L$  as a function of  $V_s$ .



**Figure 6.36** Linearized current–voltage characteristic of a diode.



**Figure 6.37** Circuit for load-line analysis of a diode.



**Figure 6.38** Graphical construction for determining the dynamic transfer curve from the load line and the static characteristic curve.

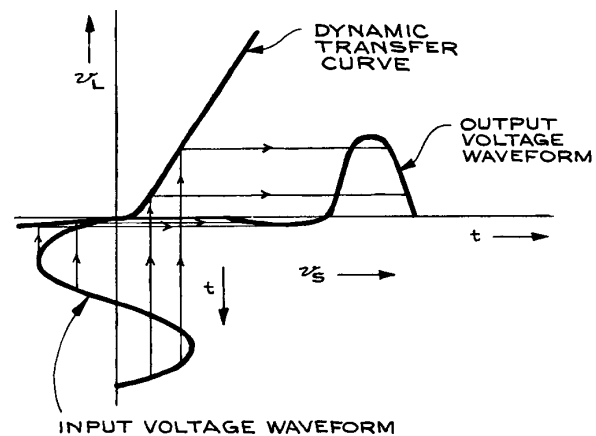
This *dynamic curve*, which is shown in Figure 6.39, has much the same appearance as the original *static curve* from which it is derived. If the input waveform is drawn below the horizontal axis, the reflection of it on the dynamic curve will give the output waveform as shown.

## 6.3.2 Transistors

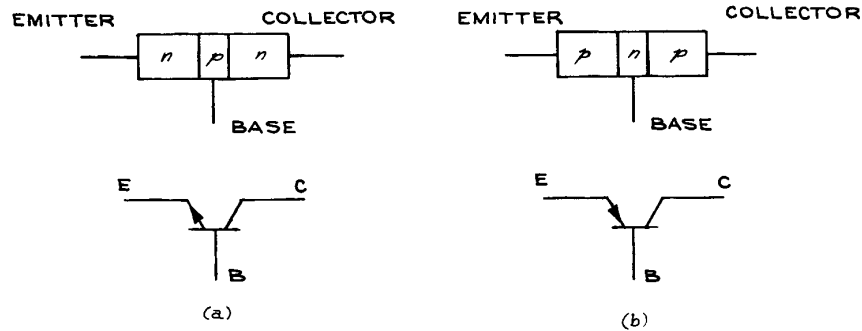
Transistors are three-terminal devices. Any one of the three terminals can be used as an input with a second terminal as output and the third providing the common connection between the input and output circuits.

**Bipolar Junction Transistors (BJTs).** These transistors consist of an *emitter*, a *base*, and a *collector*. They are designated *npn* or *pnp*, depending on the materials used for the elements. As can be seen from Figure 6.40, a transistor is two back-to-back diodes with the base element in common. The major current flow in a BJT transistor is from the emitter to the collector across the base. Relatively little current flows through the base lead. The arrow on the emitter shows the direction of the current flow.

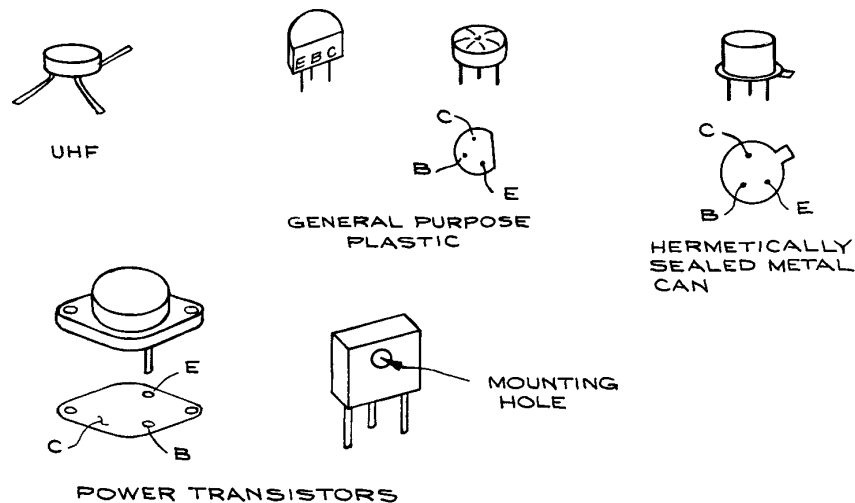
Some common BJT transistor case styles are shown in Figure 6.41. To identify the leads of high frequency transistors, manufacturers' base diagrams should be consulted. With plastic cases, there is a flat on one side of the case to help identify the leads. Leads are sometimes identified by



**Figure 6.39** Graphical construction of output waveform from input waveform and the dynamic transfer curve.



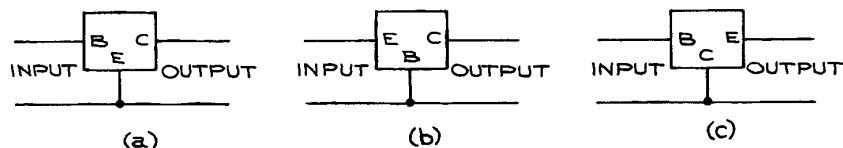
**Figure 6.40** Bipolar junction transistors (BJTs) and the corresponding symbols.



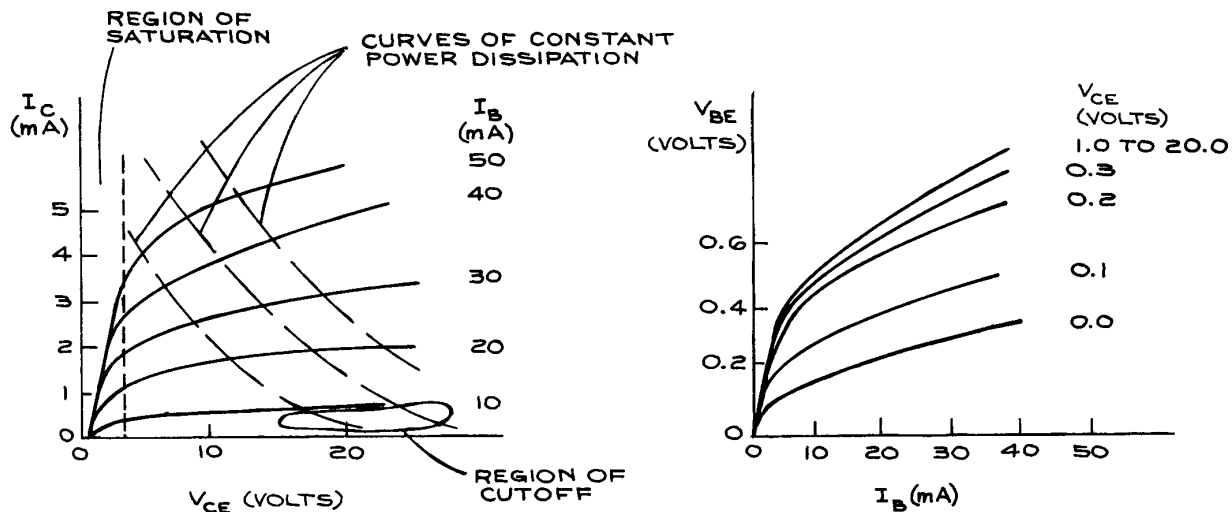
**Figure 6.41** Transistor case styles.

letters, or are in the following order – emitter, base, collector – starting at the flat or tab of sealed metal cans and proceeding clockwise as viewed from the lead side of the transistors. In power transistors, most of the power is dissipated in the collector and the collector terminal is the one with provision for attachment to a heat sink. Standard transistor types have code designations beginning with 2N. Several manufacturers often make the same transistors. Proprietary code designations are also used by manufacturers to identify their products.

Since BJT transistors consist of two  $p-n$  junctions, a simple ohmmeter test is a good way to verify that they are in working condition once the transistor is isolated from the circuit. The ohmmeter must provide at least 0.6 V, enough to forward-bias a silicon  $p-n$  junction. The resistance between the emitter and base should be low, of the order of a few hundred ohms or less when the ohmmeter leads are attached, so as to forward-bias the junction. When the leads are reversed, the resistance should be of the order of megohms. The same should be true of the



**Figure 6.42** (a) Common-emitter (CE), (b) common-base (CB), and (c) common-collector (CC) configurations.



**Figure 6.43** Current-voltage characteristics of a BJT transistor in the common emitter (CE) configuration: (a) output circuit, and (b) input circuit.

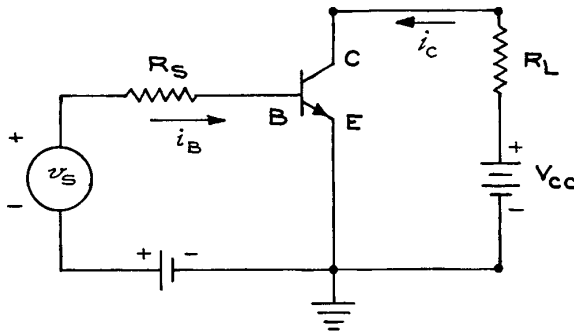
collector–base junction. The polarity of the ohmmeter leads can be checked with a separate voltmeter.

In normal operation, the emitter–base junction of a transistor is forward-biased, while the base–collector junction is reverse-biased. The three usual configurations, *common emitter* (CE), *common base* (CB), and *common collector* (CC), are shown schematically in Figure 6.42. The  $I$ - $V$  characteristics of each configuration consist of two sets of curves, one for the input and the other for the output. The  $I$ - $V$  curves for a typical general-purpose transistor in the CE configuration, for example, are shown in Figure 6.43. There are a set of curves rather than a single one for the input and output circuits because the two circuits interact with each other. When measuring the  $I$ - $V$  characteristics of either the input or output circuit, the condition

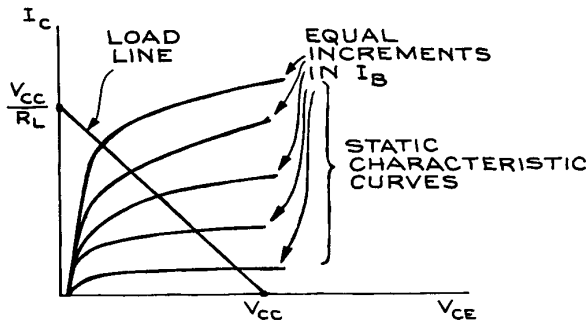
of the other circuit must also be specified. Figure 6.44 shows a circuit for measuring the output characteristics of a CE configuration.

For a fixed series of  $I_B$  values, set by adjusting  $V_{BB}$ , the current into the collector  $I_C$  is measured as a function of  $V_{CE}$ . The variation in  $V_{CE}$  is obtained by changing  $V_{CC}$ . Transistor operation can be understood in terms of the characteristic curves. For a fixed value of  $I_B$ ,  $I_C$  increases rapidly with  $V_{CE}$  until a plateau is reached.  $I_C$  increases thereafter much more slowly with  $V_{CE}$ . In the plateau region, the value of  $I_C$  depends on  $I_B$  and is almost independent of  $V_{CE}$ . This can be made more quantitative by using load-line analysis. Two points are sufficient to establish the load line on the output curves – they are points corresponding to  $I_C = 0$  where  $V_{CE} = V_{CC}$  and to  $V_{CE} = 0$

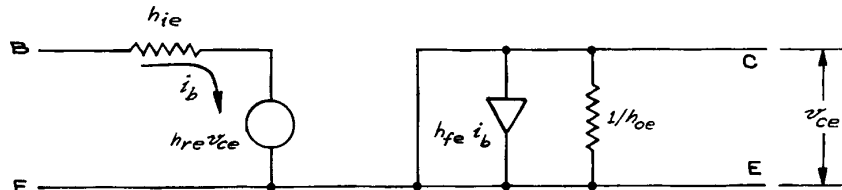
where  $I_C = V_{CC}/R_L$ . The output curves with a superimposed load line are shown in Figure 6.45. The intersections of the load line with the characteristic curves establish the operating voltages and currents. It is possible to predict quantitatively how the current through  $R_L$  varies with input base current from these curves. For linear operation, conditions are adjusted so that the transistor operates



**Figure 6.44** Biasing circuit for the BJT common-emitter (CE) configuration.



**Figure 6.45** Relation between the static characteristic curves and the load line.



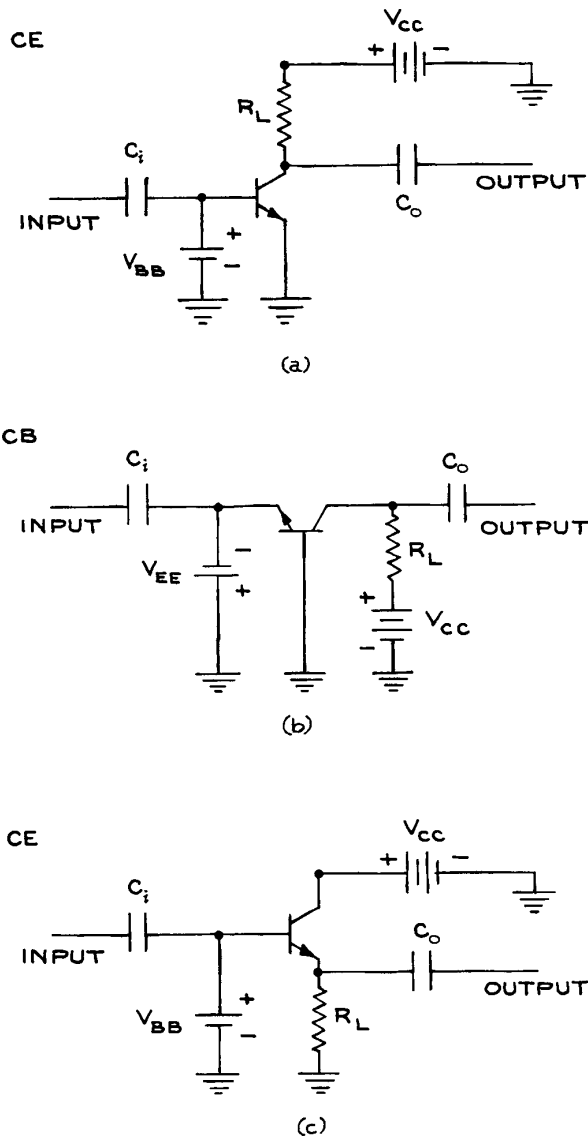
**Figure 6.46** Small-signal model of a BJT.

somewhere in the middle of the characteristic output curves, where they are approximately spaced equally and parallel for equal base-current increments.

Another way of analyzing transistor operation is by constructing an equivalent circuit with passive elements and sources. Such a circuit can then be analyzed by standard methods. Because the equivalent circuit contains only linear elements, it can represent the properties of the transistor only within a small region of the characteristic curves about given nominal values of the d.c., steady-state, or quiescent voltages and currents. For this reason such a model is called a small-signal model. A simplified low-frequency, small-signal circuit model of the BJT transistor in the CE configuration treats the output circuit as a controlled current source with  $i_c = \beta i_b$  in parallel with output resistance  $r_o$ . Here  $i_b$  is the base current and  $\beta$ , the ratio of collector current to base current, is the current gain, a constant. The input circuit can be modeled as a forward biased diode with  $r_\pi$  the diode effective resistance. The cutin voltage of the diode is assumed to be constant and for this reason does not appear in the small-signal model. This is shown in Figure 6.45(a).

It is necessary to keep in mind that the linear circuit is a good representation of the transistor only for small excursions of the input and output voltages and currents about specified quiescent d.c. values. The equivalent circuit does not show the biasing network necessary to establish the quiescent operating point. Practical circuits using separate power supplies for biasing voltages are shown in Figure 6.47. The output signal is developed across the load resistor  $R_L$ , while capacitors  $C_i$  and  $C_o$  isolate the a.c. input and output signals from the d.c. biasing levels.

Transistor data sheets give the values of the  $\beta$ , the current gain, and  $r_\pi$ , the input resistance, for different operating points. The parameters for the CB and CC



**Figure 6.47** Practical transistor circuits using separate power supplies for biasing voltages: (a) common emitter; (b) common base; (c) common collector.

configurations can be calculated from those of the CE with standard formulae.<sup>3</sup> By applying the standard methods of circuit analysis to the equivalent circuit, the current, voltage, and power gains can be calculated, as well as the input

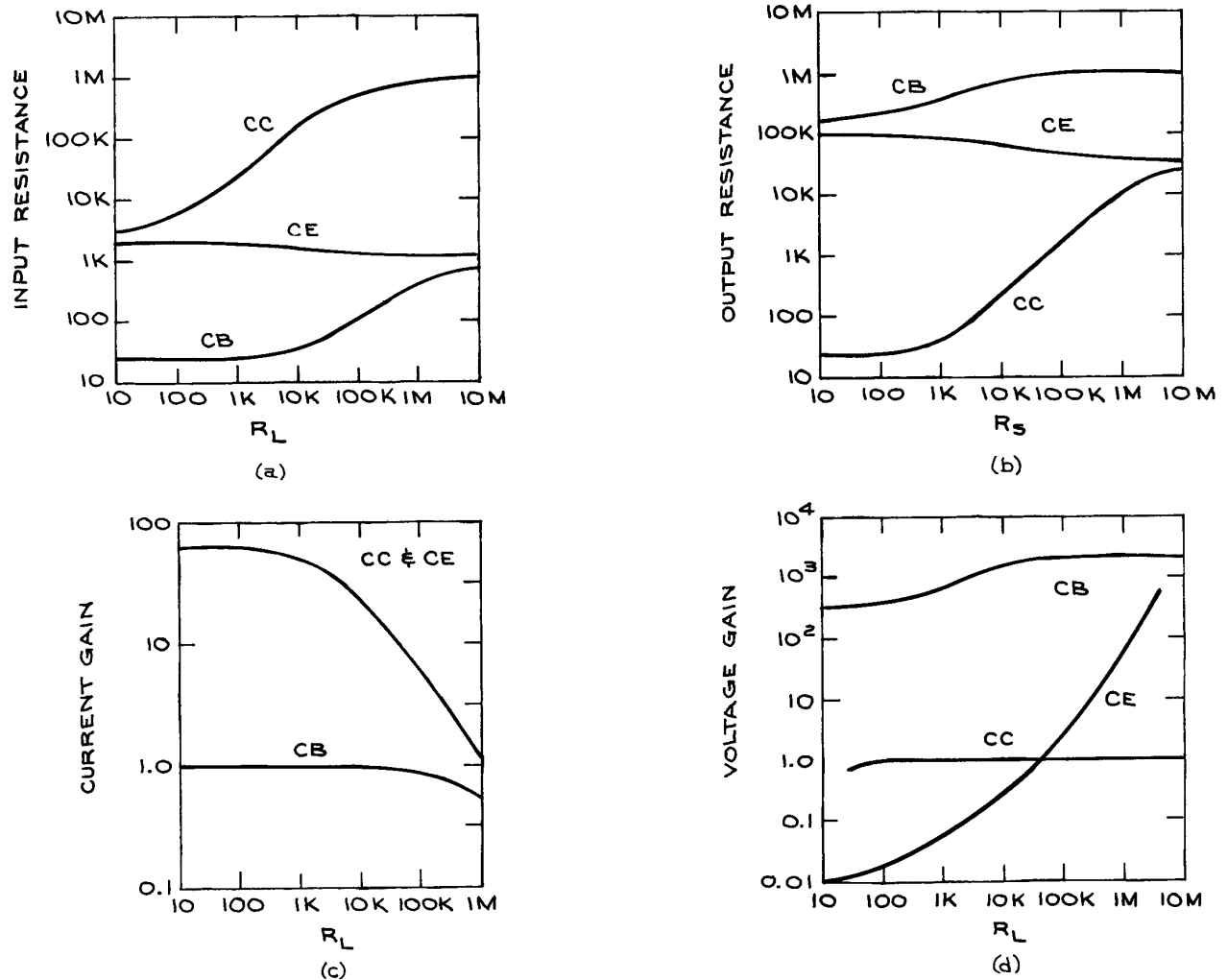
and output impedances under small-signal conditions. The general properties of the three configurations are given in the graphs of Figure 6.48. Because the charge carriers in the  $p$ - and  $n$ -materials that make up the transistor cannot respond instantaneously to changes in voltage across the junctions, the simple three-parameter ( $r_{\pi}$ ,  $\beta$ ,  $r_o$ ) model is approximately valid only at low frequencies. The response of the transistor at high frequencies can be very different, and equivalent-circuit models exist for the high-frequency response. In these models, the effective capacitances between the base-emitter, base-collector, and collector-emitter are represented by capacitors  $C_{be}$ ,  $C_{bc}$ , and  $C_{ce}$  shown in Figure 6.49.

A BJT can also be operated as a switch. If the emitter-base junction of a transistor is strongly forward-biased (0.8 V for silicon), the total current through the transistor will be limited by the external resistance in the circuit and the voltage between emitter and collector will be of the order of 0.2 V. In this condition, the transistor is said to be *saturated* or *on* and the base current necessary to sustain the current flow from emitter to collector is  $I_C/\beta_F$ , where  $I_C$  is the total current in the collector circuit and  $\beta_F$  is the large signal current gain. If the emitter-base junction is reverse-biased, no base current will flow and no collector current will flow. Under this condition, the transistor acts as an open circuit and is said to be *cut off*. In the saturated or the cut off condition, a transistor is similar to a closed or an open switch. Transistor switches can operate at high frequencies, unlike mechanical switches, but the saturated resistance is never zero, and there is always at least a 0.2 V drop from the emitter to the collector.

The operating conditions for bipolar junction transistors are summarized in Table 6.16. They determine whether a transistor is operating in the linear region (*active*), is *saturated*, or is *cut off*. When reading the specifications for a 2N3904 transistor in Tables 6.16 and 6.17, the following conventions are to be noted:

- (1) Voltages between two terminals are specified by either  $V$  or  $v$  with two or three subscripts. The first two subscripts indicate the terminals between which the voltage occurs, and the third subscript, when present, indicates the condition of the third terminal. Thus  $V_{CBO}$  indicates a d.c. voltage between collector and base with the emitter open. Subscripts



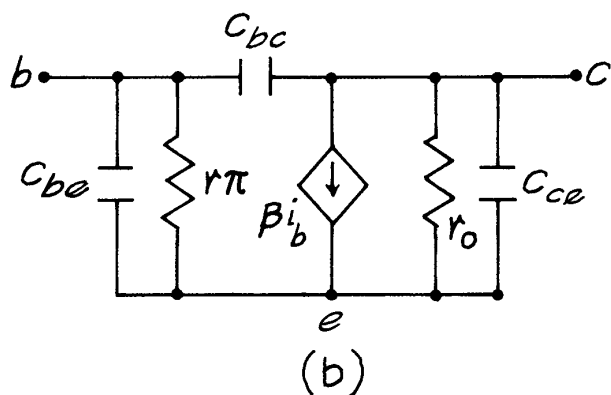
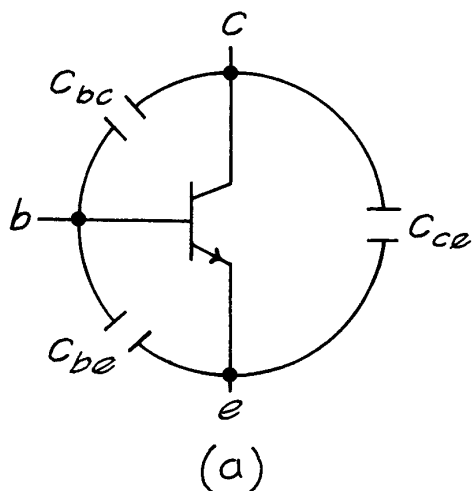


**Figure 6.48** Variation of the properties of the CE, CB, and CC transistor configurations with various input parameters: (a) input resistance as a function of load resistance; (b) output resistance as a function of source resistance; (c) current gain as a function of load resistance; (d) voltage gain as a function of load resistance.

- with repeated letters indicate power-supply voltages. For example,  $V_{CC}$  is the power-supply voltage to the collector, and  $V_{BB}$  is the power-supply voltage to the base.
- (2) By convention, all currents are taken to be positive when they flow into the leads of the transistor and negative when they flow out of the leads.

- (3) Currents are specified by  $I$  or  $i$  with a subscript indicating the terminal involved.  $I_B$  is therefore a d.c. base current.

Transistors are destroyed if their maximum power capabilities are exceeded, or if, when reverse-biasing the junctions, the maximum reverse voltages are exceeded, resulting in electrical breakdown. The maximum power



**Figure 6.49** BJT showing: (a) Junction capacitances and (b) small signal model with junction capacitances.

dissipation depends on temperature, and a transistor fixed to a heat sink can dissipate considerably more power than one standing alone in free air. In pulse operation, the maximum steady-state power levels can be momentarily exceeded without damage.

**Field-Effect Transistors (FETs).** The *junction FET* or *JFET* is the simplest FET. It has three terminals – a *source*, a *gate*, and a *drain*. An *n*-channel device is shown in Figure 6.50. Current flow is from source to drain with the gate reverse-biased with respect to the channel. Because of this

**Table 6.16** 2N3904 Transistor parameters

Parameter	Condition of Transistor		
	Active	Saturated	Cut-Off
$V_{BE}$	0.6 V	0.8 V	< 0.0 V
$V_{BC}$	5–10 V (reverse bias)	0.6 V (forward bias)	$\approx V_{CC}$ (reverse bias)
$V_{EC}$	5–10 V	0.2 V	$V_{CC}$
$I_B$ ( $i_b$ )	$i_c$ $\beta$	$I_c$ $\beta$	$\approx 0$
$I_C$	$\frac{V_{CC}-V_{CE}}{R_L}$	$\frac{V_{CC}}{R_L}$	$\approx 0$

reverse bias, negligible current flows into the gate. The  $I$ - $V$  characteristics of a small-signal JFET are shown in Figure 6.51. From the curves it can be seen that there is a region where the drain current decreases with increasing reverse bias from the gate to source, independent of the drain-source voltage. This is the region where the transistor can be operated as an amplifier. About the region  $V_{DS} = 0$ ,  $I_D$  increases linearly with  $V_{DS}$  for fixed  $V_{GS}$ , with the slope equal to the reciprocal of the channel resistance  $R_{channel}$ . As  $V_{GS}$  becomes more negative,  $R_{channel}$  increases. Transistor operation in this region is that of a *voltage-controlled resistor*.

The FET can also be used as a switch, with the advantage of no offset voltage from source to drain and almost perfect electrical isolation of the gate signal from the output circuit. Disadvantages are rather high *on* resistances (of the order of hundreds of ohms) and low *off* resistances. Also, switching times are greater than for BJTs.

The simple structure of FETs results in low noise. As a consequence, FETs are often used as the active elements in low-level signal amplifiers. Field-effect transistors with gates insulated by an oxide layer from the conducting channel are called *MOSFETs* (metal oxide-semiconductor FETs). The oxide gate insulation effectively blocks any d.c. gate current flow. The conducting channel in a MOSFET can be either *p*-doped (PMOS) or *n*-doped (NMOS). Moreover PMOS and NMOS transistors can be of the *depletion mode* or *enhancement mode*. For depletion-mode transistors, the maximum drain current corresponds to  $V_{GS} = 0$ . For both PMOS and NMOS enhancement-mode transistors, there is no drain current at  $V_{GS} = 0$ . Only when

TABLE 6.17 Properties of the 2N3904 N-P-N silicon transistor

Condition	Parameter	Value	Conditions
Absolute Maximum Ratings (25°C in Free Air)	$V_{CBO}$	60 V	$I_C = 10 \mu\text{A}, I_E = 0 \text{ mA}$
	$V_{CEO}$	40 V	$I_C = 1 \text{ mA}, I_B = 0 \mu\text{A}$
	$V_{EBO}$	6 V	$I_E = 10 \mu\text{A}, I_C = 0 \text{ mA}$
$I_C$ (continuous) = 200 mA			
Maximum Power Dissipation = 310 mW			
“On” Characteristics	$\beta_F$	40	$V_{CE} = 1 \text{ V}, I_C = 100 \mu\text{A}$
		70	$V_{CE} = 1 \text{ V}, I_C = 1 \text{ mA}$
	$V_{BE}^a$	0.65 to 0.85 V	$I_B = 1 \text{ mA}, I_C = 10 \text{ mA}$
		0.95 V	$I_B = 5 \text{ mA}, I_C = 50 \text{ mA}$
	$V_{CE}^a$	0.2	$I_B = 1 \text{ mA}, I_C = 10 \text{ mA}$
0.3		$I_B = 5 \text{ mA}, I_C = 50 \text{ mA}$	
Small Signal Characteristics	$r_\pi$	1 k $\Omega$ to 10 k $\Omega$	
	$\beta$	100–400	$V_{CE} = 10 \text{ V}, I_C = 1 \text{ mA}$ ,
	$r_o$	25 k $\Omega$ to 1 M $\Omega$	
	$f_T^b$	300 MHz	$V_{CE} = 20 \text{ V}, I_C = 10 \text{ mA}$
	$C_{obo}^c$	4 pF	$V_{CB} = 5 \text{ V}, I_E = 0 \text{ mA}$
	$C_{ibo}^d$	8 pF	$V_{EB} = 0.5 \text{ V}, I_C = 0 \text{ mA}$

<sup>a</sup> Saturation.

<sup>b</sup> Transition frequency for  $\beta=1$ .

<sup>c</sup> Common-base open-circuit output capacitance.

<sup>d</sup> Common-base open-circuit input capacitance.

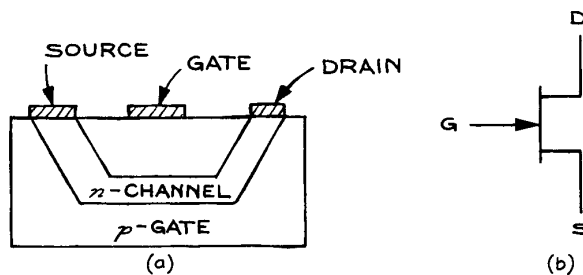


Figure 6.50 N-channel junction field-effect transistor (JFET): (a) Construction; (b) Symbol.

$V_{GS}$  reaches the threshold value  $V_T$ , will drain current flow. For NMOS enhancement  $V_T$  is approximately +2 V, for PMOS enhancement  $V_T$  is approximately -2 V. The construction of PMOS and NMOS depletion-and enhancement-mode transistors along with their schematic symbols are shown in Figure. 6.52. Typical characteristic output curves are given in Figure 6.53.

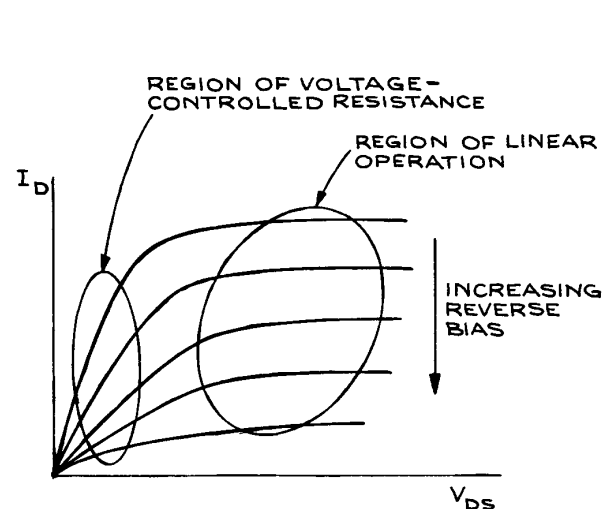


Figure 6.51 Current-voltage characteristics of a small-signal, junction-field-effect transistor (JFET).

Depletion-mode MOSFETs are less common than enhancement-mode devices because of their relatively low transconductance. Because of the higher mobility of electrons compared to holes, NMOS transistors have properties that are superior to PMOS transistors. The most common application of PMOS transistors is in CMOS logic circuits (see Section 6.6.10). Low power consumption, ease of fabrication, and high packing densities have made MOSFETs extremely common in *large-scale* (LSI) and *very large-scale* (VLSI) *integrated circuits*.

Because of the gate insulation, static charge can accumulate on the gate and voltage can build to levels sufficient to punch through the insulation. To avoid this, some MOSFET gates are protected with reverse-biased Zener diodes. MOSFET devices are normally packaged in conducting foam or with metal protecting rings that are removed just prior to insertion into the circuit.

For switching high currents the bipolar junction transistor was the device of choice until the development of the power MOSFET and the insulated gate bipolar transistor (IGBT). Because the BJT is current controlled and the MOSFET is voltage controlled, the MOSFET has a number of significant advantages at moderately high voltages and currents and especially at high switching speeds. The current-carrying capacity of a small-signal MOSFET is limited by the size of the conducting channel. One way to increase current capacity is with a short and wide channel that has the form of a trench or groove. This is shown in Figure 6.54 for an enhancement  $n$ -channel power MOSFET. The materials are labeled  $p$ ,  $n^+$ , and  $n^-$  indicating  $p$ -doped material, lightly  $n$ -doped material, and heavily  $n$ -doped material. With the drain biased positive with respect to the source and the gate at the source potential, no current flows between drain and source because the  $p$ - $n$  junctions are reversed biased. As the potential of the gate is increased with respect to the source, an  $n$ -channel is formed around the gate, resulting in a conduction path from source to drain. The  $I$ - $V$  static characteristics of a typical  $n$ -channel enhancement power MOSFET, the VN66AFD, is shown in Figure 6.55. The main characteristics to note are:

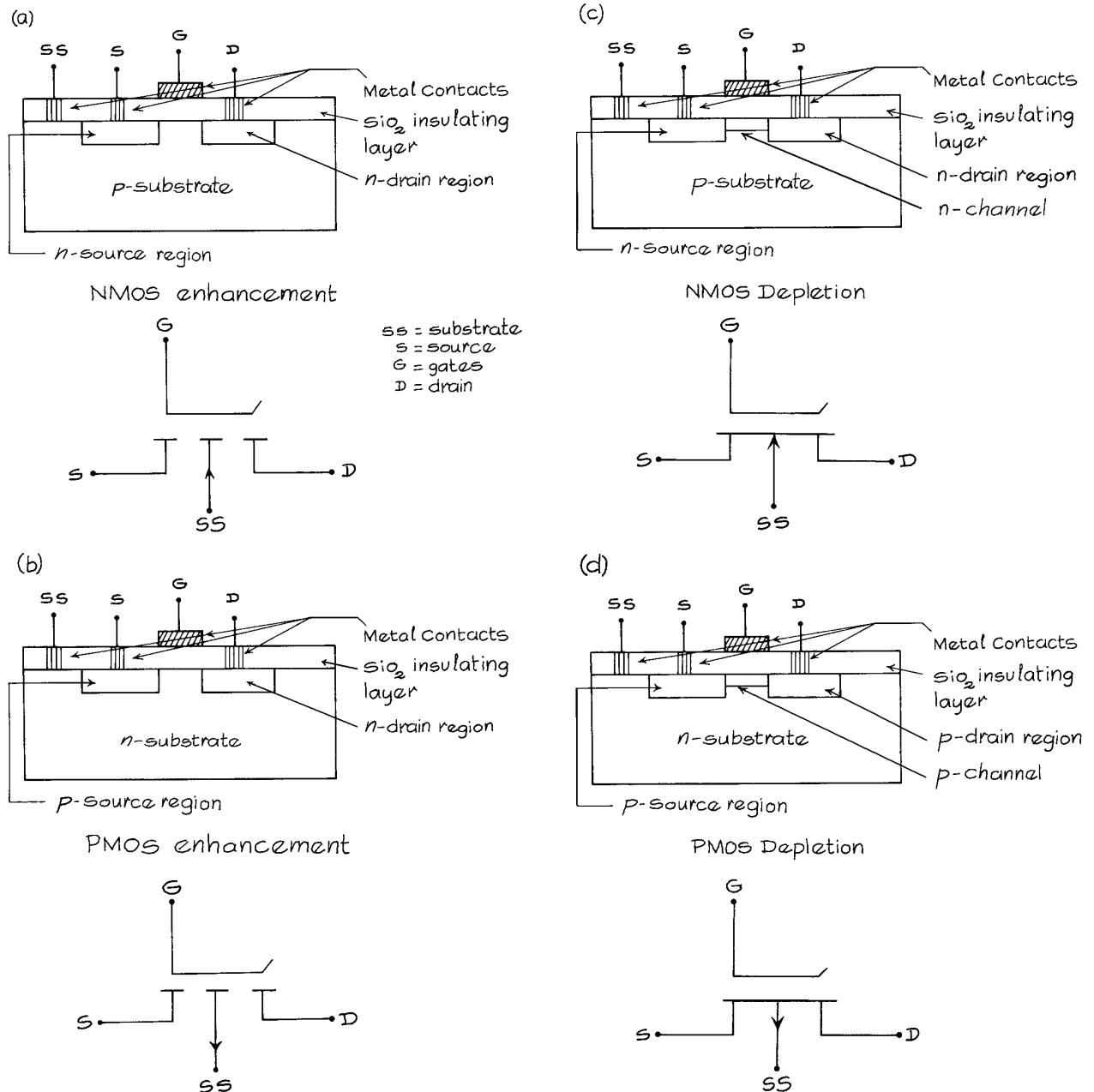
- (1) Drain current is essentially zero for  $V_{GS}$  less than 2 V. This is the threshold voltage and varies from 0.5 to 2.5 V for power MOSFETs.
- (2) Between  $V_{DS} = 2$  V and  $V_{DS} = 4$  V,  $I_D$  is a nonlinear function of  $V_{GS}$ , increasing quadratically with  $V_{GS}$ . For  $V_{GS}$  greater than 4 V,  $I_D$  is a linear function of  $V_{GS}$ .
- (3)  $I_D$  depends almost entirely on  $V_{GS}$  and is nearly independent of  $V_{DS}$ .

Because the gate is insulated by an oxide layer from the underlying  $p$  and  $n$  materials, there is no current flow into the gate under steady-state conditions. The input impedance is essentially infinite. However, because the gate acts as one plate of a capacitor, with the oxide insulating layer as the dielectric, current can flow into the gate when the voltage on the gate is changed. Typical gate capacitances are in the tens of picofarads. As a result, switching times of nanoseconds are possible when the gate drive voltage is derived from a low-impedance source.

The power MOSFET is an ideal voltage-controlled switch. For the VN66AFD, the drain to source resistance  $R_{DS}$  varies from 2  $\Omega$  at  $V_{GS} = 4$  V and  $I_D = 0.1$  A to less than 1  $\Omega$  at  $V_{GS} = 10$  V and  $I_D = 0.1$  A. For  $V_{GS} = 2$  V,  $R_{DS}$  is of the order of megohms. Power MOSFETs can be switched with TTL or CMOS gates. Sample interface circuits using the VN66AFD are shown in Figure 6.56(a) and (b). With 5 V on the gate of the MOSFET, the saturation drain current is 0.8 A. Higher drain currents can be achieved by increasing the MOSFET gate voltage.

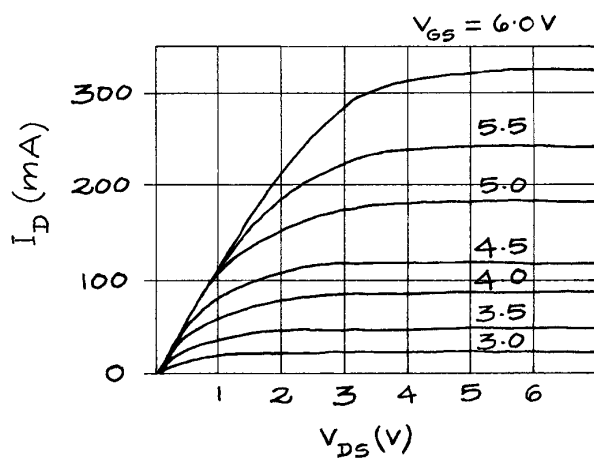
While power MOSFETs are mostly used in switching applications, they can also be used as power amplifiers by biasing them into the linear portion of the characteristic curve. This is shown in Figure 6.56(c). With this arrangement the quiescent drain current is 0.675 A with  $V_{GS} = 8.6$  V. An input voltage swing of  $\pm 1$  V results in a change of current of  $\pm 0.17$  A about the quiescent value. While inefficient, the circuit is simple and has less than 1% harmonic distortion at 1 W.

Commercial applications for power MOSFETs are in switch mode power supplies, motor controllers, lamp drivers, and automotive ignition systems. While trench and V power MOSFETs are common, there are other MOSFET configurations based on laterally diffused (LDMOS) geometries and parallel arrays of MOSFETs (HEXFETs). An important property of enhancement MOSFETs that allows them to be connected in parallel is the drain-current negative-temperature coefficient. This means that as the temperature of the device increases, the drain current will



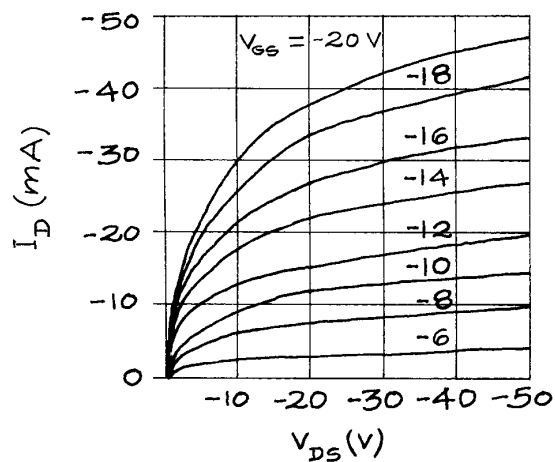
**Figure 6.52** MOSFET structures and schematic diagrams. (a) NMOS enhancement; (b) PMOS enhancement; (c) NMOS depletion; (d) PMOS depletion.

NMOS Enhancement



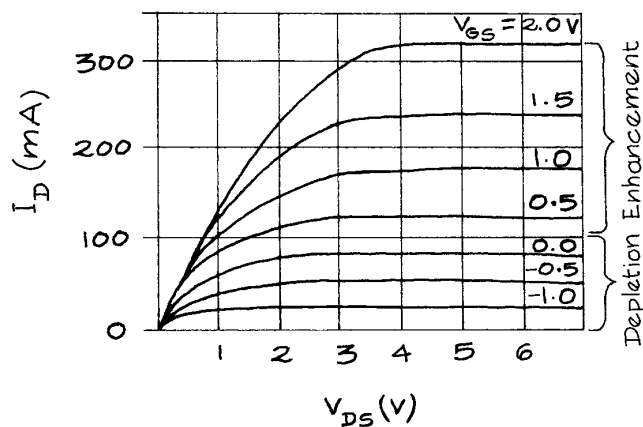
(a)

PMOS Enhancement



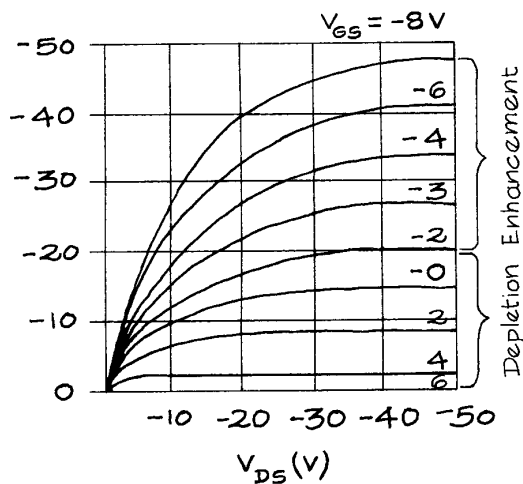
(b)

NMOS Depletion



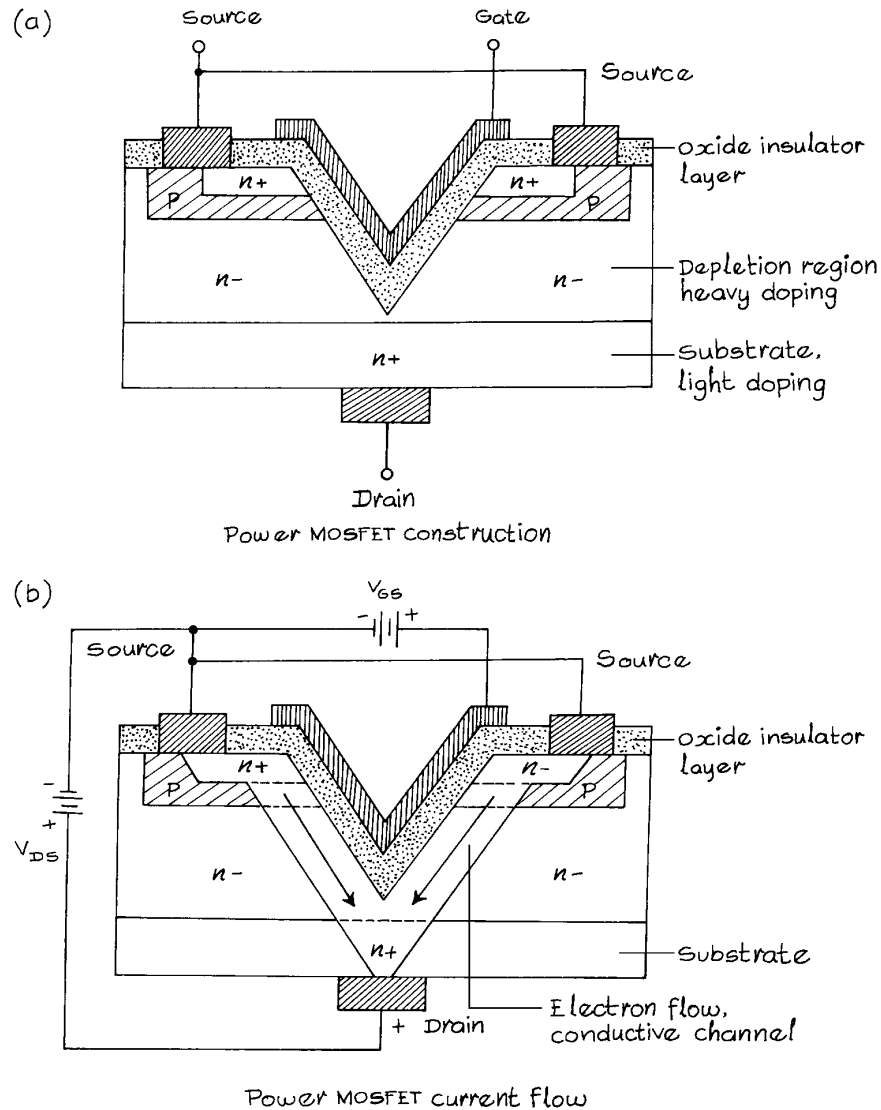
(c)

PMOS Depletion



(d)

**Figure 6.53** MOSFET output IV characteristics: (a) NMOS enhancement; (b) PMOS enhancement; (c) NMOS depletion; (d) PMOS depletion.

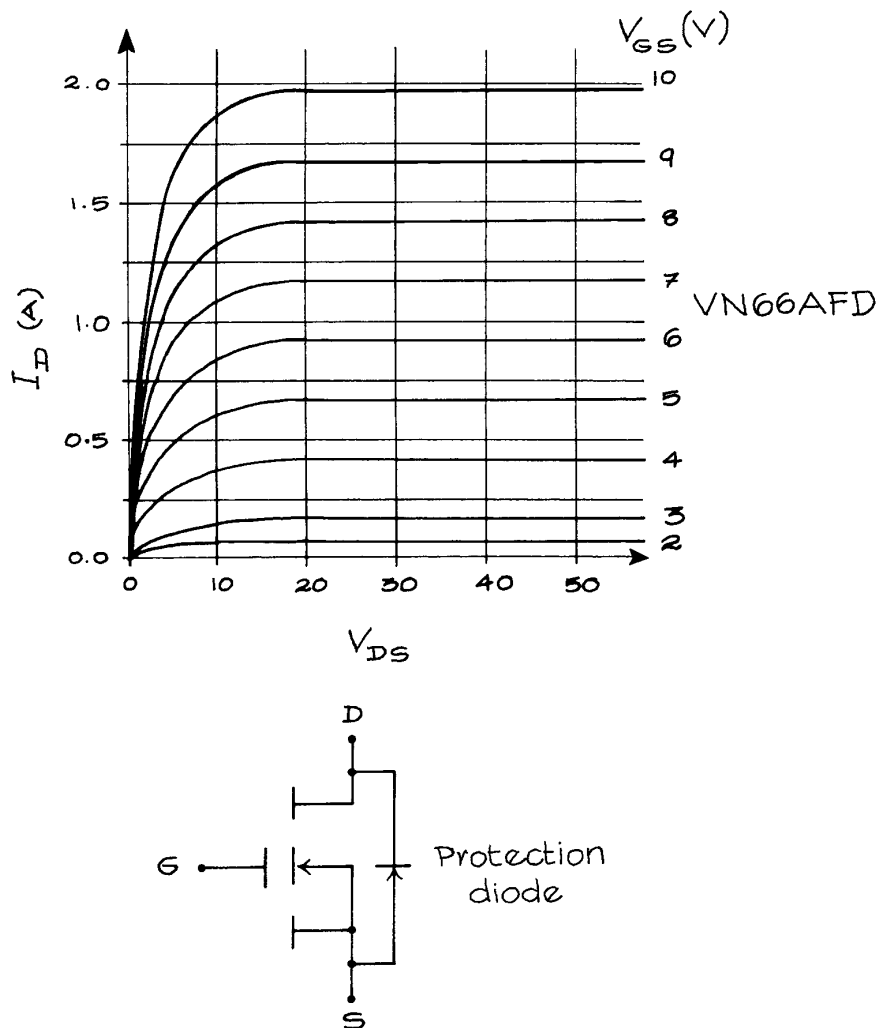


**Figure 6.54** Power MOSFET: (a) Construction; (b) Current flow.

decrease. This is a consequence of majority carrier conduction in the drain-source channel. By contrast, the collector-current temperature coefficient in BJTs is positive; collector current increases with temperature, leading to the possibility of thermal runaway, in which increasing temperature leads to increasing collector current that increases power dissipation and higher temperatures, etc.

Power MOSFETs can operate at frequencies well in excess of 100 kHz. While maximum voltages as high as 1000 V are available, most power MOSFETs are designed for operation below 300 V. Current handling capabilities extend to 100 A.

The IGBT is a device that combines the high input impedance of a power MOSFET with the high current



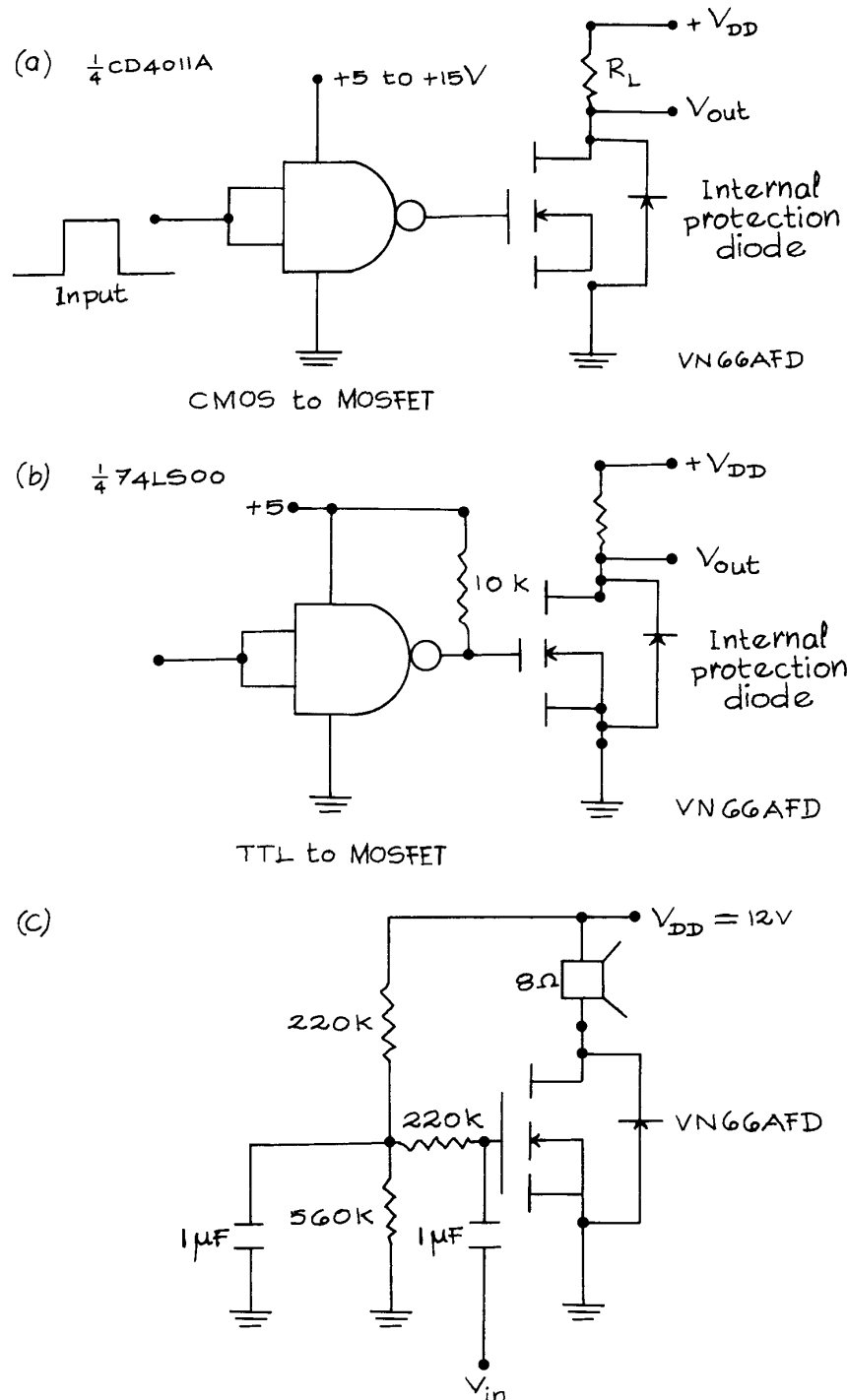
**Figure 6.55** VN66AFD power MOSFET output IV characteristics and schematic diagram.

capabilities of a power BJT. The schematic symbol for a IGBT is given in Figure 6.57(a) along with a simplified equivalent circuit in Figure 6.57(b). The device is “on” when the base (B) of the  $p-n-p$  output transistor is negative with respect to the collector (C) by 0.6 to 0.8 V. The total voltage drop from C to E is equal to the voltage drop across the CB diode plus the drop across the input MOSFET. This latter voltage drop is a function of the gate-to-source volt-

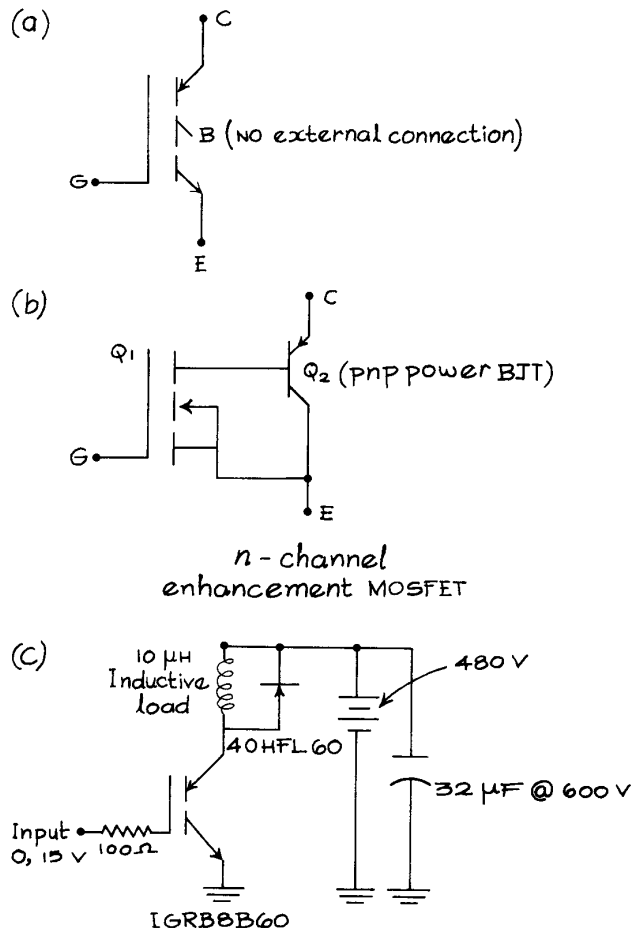
age,  $V_{GS}$ . The use of IGBTs is restricted to switching applications. In Table 6.18 the switching characteristics of power MOSFETs, IGBTs, and power BJTs are compared.

The IGBT offers clear advantages over power MOSFETs at voltages above 300 V and at high currents and speeds from 5 to 50 kHz. They are to be preferred in high-power applications with voltages greater than 1000 V provided the frequency requirements are less than





**Figure 6.56** Power MOSFET circuits: (a) CMOS to MOSFET; (b) TTL to MOSFET; (c) MOSFET power amplifier.



**Figure 6.57** IGBT: (a) schematic diagram; (b) internal connections; (c) inductive load switching at high voltage.

20 kHz, the duty cycle is low, and variations in load and line are small. A representative IGBT is the IRGB860. It is available in TO-220AB, TO262, and D<sup>2</sup>Pak packages. The maximum collector-to-emitter voltage is 600 V and the maximum continuous collector current is 28 A at 25 °C, falling to 19 A at 100 °C. The maximum value of  $V_{GE}$  is  $\pm 20$  V. Exceeding this limit subjects the device to permanent damage through destruction of the insulating layer between gate and substrate. When “on” ( $V_{GE} = 15$  V and  $I_C = 8$  A),  $V_{CE}$  is approximately 2 V. Switching rise times are 20 to 30 ns and fall times are 30 to 40 ns. Max-

imum power dissipation is 167 W at 25 °C and 83 W at 100 °C. It is worth noting that  $V_{CE}$  decreases with temperature so that there is the possibility of thermal runaway in high-power applications. This is not the case for power MOSFETs. A circuit for switching an inductive load with an IRGB860 is shown in Figure 6.57(c). The diode suppresses inductive voltage transients associated with the switching of currents through the inductor. This protects the IGBT and results in cleaner switching waveforms.

**Active Loads and Current Sources.** Both BJT and FET transistors can be used as resistors with static resistances of kilohms and dynamic resistances of tens and even hundreds of kilohms. Active loads are commonly used in integrated circuits where resistors are difficult to fabricate. The high dynamic resistance ( $dv/di$ ) of BJTs and FETs in their active region is also used for high output-impedance current-source circuits. Examples of BJT current-source circuits are given in Figure 6.58. An NMOS depletion-mode transistor can also be converted to a resistor by connecting the drain to the gate. The nonlinear  $I$ - $V$  characteristics of this configuration can be found from the output characteristics of the transistor with  $V_{GS} = V_{DS}$ .

### 6.3.3 Silicon-Controlled Rectifiers

The *silicon-controlled rectifier* (SCR) is a switching device with only two states – *on* and *off*. In the *on* state, the SCR has a low resistance, and some models are capable of passing over 100 A. In the *off* state, the resistance is in the megohm range. The SCR has three terminals: an *anode*, a *cathode*, and a *gate*. The transition from *off* to *on* occurs when the voltage on the anode is positive with respect to the cathode and a pulse of the correct polarity and magnitude is applied to the gate. Once the SCR is *on*, the gate loses all control over the functioning of the device. The only way to turn it *off* is to reduce the anode–cathode voltage to zero. If a sinusoidal voltage is applied to the SCR anode, the device will be turned *off* once each half cycle. By controlling the point in each positive half cycle when the trigger pulse is applied, the average current through the SCR can be varied over wide limits (see Figure 6.59). This way of regulating the power to a load is very efficient because very little power is dissipated in the SCR. When the SCR is *off*, there is no

Table 6.18 Comparison of the properties of power MOSFET, IGBT and power BJT transistors

	Power MOSFET	IGBT	Power BJT
Input	Voltage	Voltage	Current
Input Power	Very Low	Very Low	Moderate to High
Output Current	High at low voltages, low at high voltages	Very High, almost independent of switching speed	Moderate, strong function of switching speed
Switching Losses	Very low	Low to moderate	Moderate to high

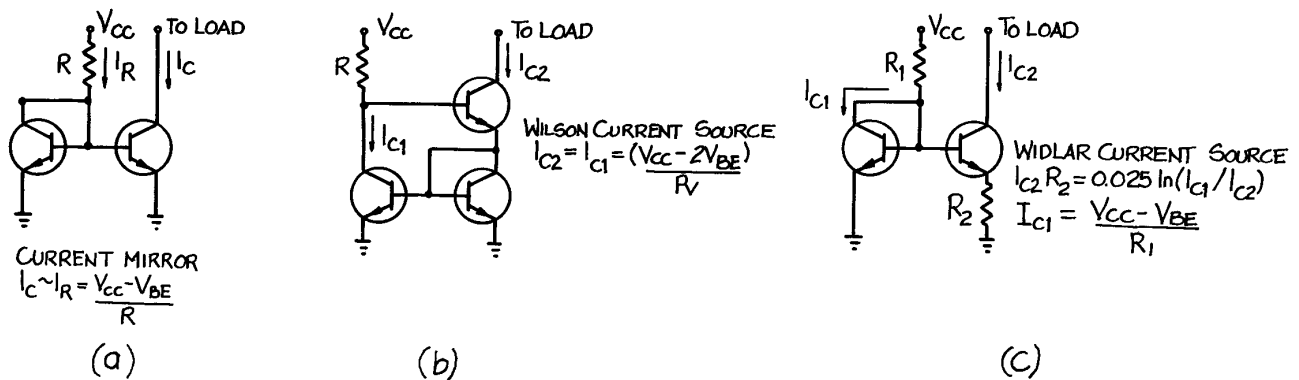


Figure 6.58 Current sources.

current through it and the power dissipation is zero. When the current through the SCR is at a maximum, the voltage across it is low, resulting in low power dissipation. Bilateral triggering can be accomplished with a bilateral switch or *triac*. A common application of this kind of power control is the incandescent light dimmer. High-power switching devices like the SCR and triac produce *radio-frequency interference* (rfi) during switching when there is a voltage across them. *Zero crossing* switching avoids rfi by timing the switching to those instants when there is no voltage across the device. Silicon-controlled switches (SCSs) are similar to SCRs, but have the advantage of being able to be turned off with a pulse.

### 6.3.4 Unijunction Transistors

The *unijunction transistor* (UJT) is a single bar of *n*-material to the middle of which *p*-material is attached, forming

a *p-n* junction. This is shown schematically in Figure 6.60. The two ends of the bar are bases 1 and 2 (*B1* and *B2*), while the *p*-material is the emitter (*E*). When *B2* is made positive with respect to *B1*, there will be a potential gradient along the bar and the potential at the point of attachment of the emitter will be a fixed fraction of the potential across the bar. As long as the emitter is less positive than the bar at its point of attachment, the emitter junction will be reverse-biased and no current will flow through it. When the emitter potential is increased sufficiently to forward-bias the junction, current will flow in the *E-B1* circuit and  $V_{EB1}$  will fall. This is shown in the characteristic curve (see Figure 6.61). A common application of the UJT is in relaxation oscillators where a capacitor is charged through a resistor to the peak potential of the UJT and then discharged. The UJT is also used in trigger circuits, pulse generators, and frequency dividers. The stable peak voltage, which is a fixed fraction of the interbase voltage, and

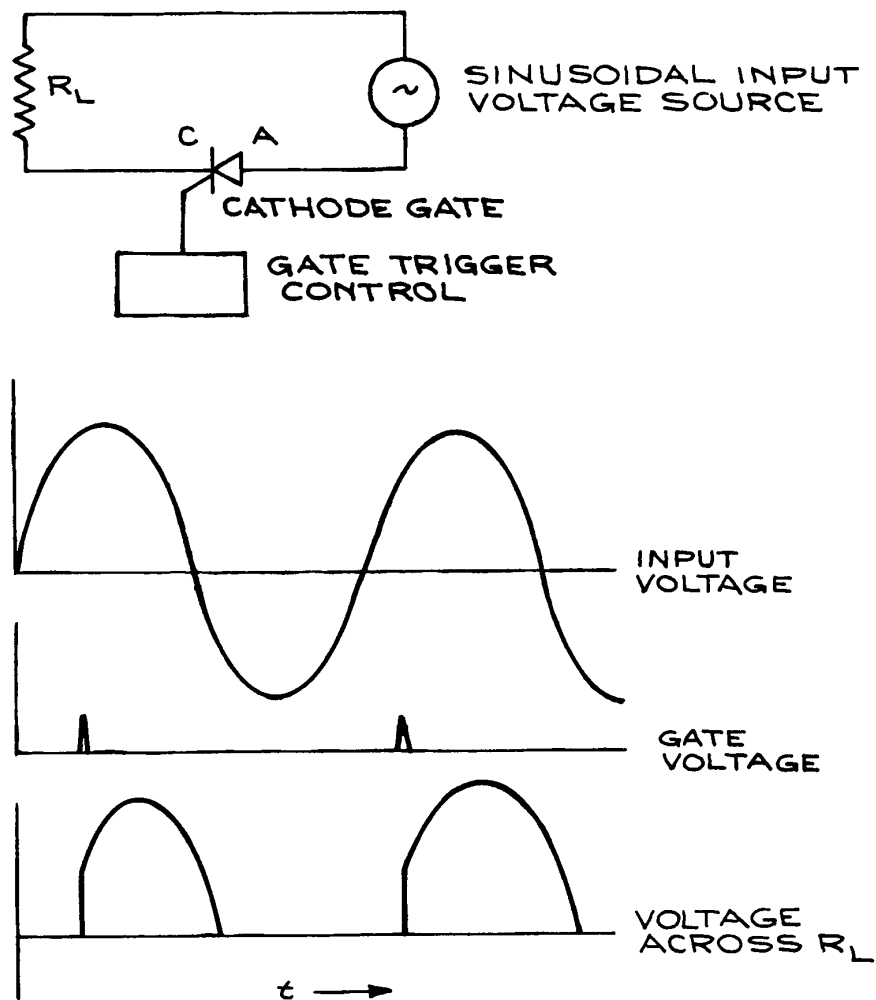


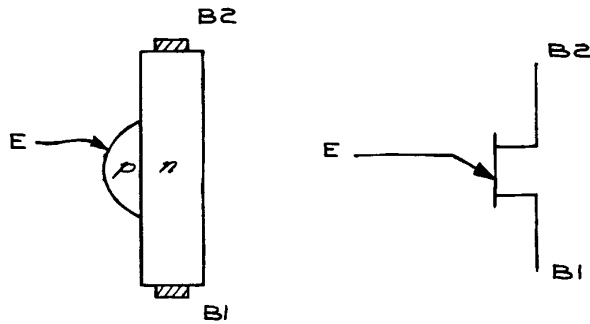
Figure 6.59 Current regulation with a silicon controlled rectifier (SCR).

high pulse-current capability make the UJT very useful for SCR control circuits.

### 6.3.5 Thyratrons

For fast switching of high currents at high voltages, gas-filled tubes called *thyratrons* are used. They are analogs of SCRs and consist of a heated cathode, an anode, and a grid sealed into a glass, gas-filled envelope. An arc can be struck between the cathode and anode, with the initiation

of the arc controlled by the potential on the grid. The grid is usually a cylindrical structure surrounding the anode and cathode from which a baffle or series of baffles with small holes extends between anode and cathode. Since the anode and cathode are almost completely shielded from each other, only a small grid potential is needed to overcome the field at the cathode resulting from the large anode potential. Once the arc has been initiated, the grid loses control over the arc. Grid control is reestablished only when the anode potential falls below the level necessary



**Figure 6.60** Unijunction transistor (UJT) and the corresponding schematic symbol.

to sustain the arc. Deuterium-filled thyratrons can switch hundreds of amperes at tens of kilovolts in microseconds or less and are commonly used in the high-voltage discharge sources of lasers. SCRs are often used to supply the trigger pulse to the grid of the thyatron.

## 6.4 AMPLIFIERS AND PULSE ELECTRONICS

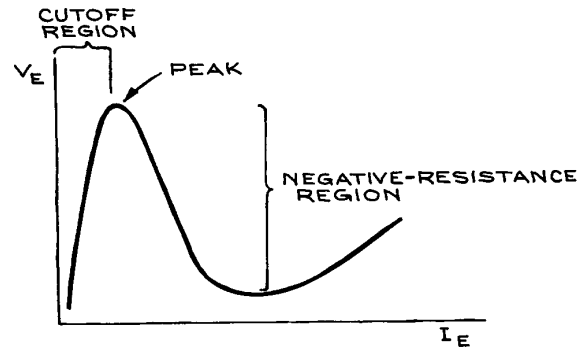
### 6.4.1 Definition of Terms

In laboratory applications, amplifiers are used to transform low-level signals to levels sufficient for observation and recording, or for operation of electronic or electromechanical devices. When choosing an amplifier for a given task, there are a number of considerations that require the definition and explanation of several special terms. The goal is to provide information for the experimentalist who must make decisions regarding the use and specifications of amplifiers.

Amplifiers can be classified in a number of ways, among them the following:

- (1) By input and output variables
- (2) By frequency domain
- (3) By power levels.

Each of these classifications will be treated in turn, and any given amplifier will have its place somewhere within each of the classifications. The two most common electrical input and output variables are current and voltage. There are therefore four different possible types of ampli-



**Figure 6.61** Current-voltage characteristic curve for a unijunction transistor showing the region of negative resistance.

**Table 6.19** Types of amplifier

<i>Input Variable</i>	<i>Output Variable</i>	<i>Amplifier Type</i>
Voltage	Voltage	Voltage
Voltage	Current	Transconductance
Current	Voltage	Transresistance
Current	Current	Current

fier, corresponding to the two input-variable possibilities and the two output-variable possibilities. These are listed in Table 6.19. The *gain* of an amplifier is the ratio of the output variable to the input variable. This is the equivalent of the transfer function for the passive circuits already considered. For voltage and current amplifiers, the gain is unitless. For the transconductance and transresistance amplifiers, however, their gains have units of siemens ( $\text{ohm}^{-1}$  or sometimes mho) and ohm, respectively.

There are also amplifiers for which the input variable is the time integral of the current or the charge. Such *charge-sensitive* amplifiers have a voltage as an output, and the gain is expressed in volts per coulomb or inverse farads. Amplifiers with the time derivative of current as an input variable are also possible.

The amplifiers in Table 6.19 have very different input and output properties. If the input is a current, the amplifier should have a low input impedance so as to affect the source as little as possible. On the other hand, if the input is a voltage the input impedance should be as large as

possible to avoid drawing current from the source and changing the source output voltage. The opposite considerations apply to the outputs of the amplifiers. If the output is a current, it is desirable that the output be close to that of an ideal current source – that is, a very high output impedance. If the output is a voltage, the output impedance should be low to approximate an ideal voltage source as closely as possible. These considerations of input and output impedance (see Table 6.20) are of fundamental importance when matching an amplifier to a detector or transducer at the input and a load at the output. If the input device is a current source, one should generally choose a current, transresistance, or charge-sensitive amplifier; if the input device is a voltage source, a voltage or transconductance amplifier is required. This assumes that amplification with minimum loading of the input circuit is desired. When efficient power transfer is desired, the impedances of the source and load should match the input and output impedances of the amplifier as closely as possible.

Amplifiers can be divided into groups according to the frequency domain in which they are designed to operate (see Table 6.21). Consider the Bode plots of gain and phase shift shown in Figure 6.62 for a capacitance-coupled inverting (output  $180^\circ$  out of phase with input) amplifier. The curves can be understood qualitatively in the following ways – at low frequencies the coupling capacitors (capacitors in series with input and output to block d.c. levels) reduce the low-frequency gain just like a high-pass filter. At high frequencies the reduced gain of the active devices (transistors) has the effect of shunt capacitances across both the input and output circuits of the amplifier, resulting in properties similar to that of a low-pass filter. The frequency difference between the corner frequencies (3 dB points of the gain curve) is the *bandwidth* of the amplifier. Because the low corner frequency is usually orders of magnitude smaller than the high corner frequency (except in the case of tuned amplifiers that operate over a very small frequency range), the bandwidth can be considered equal to the upper corner frequency. An often-used Figure of merit for amplifiers is the *gain-bandwidth product* (GBWP), which is the midband gain times the bandwidth. By this criterion, a low-gain, wide-bandwidth amplifier is equivalent to a high-gain, narrow-bandwidth amplifier.

If it is possible for the output of an amplifier to be coupled back to the input (capacitance coupling is

---

**TABLE 6.20 Amplifier impedances**


---

<i>Amplifier Type</i>	<i>Input Impedance</i>	<i>Output Impedance</i>
Voltage	High	Low
Transconductance	High	High
Transresistance	Low	Low
Current	Low	High
Charge-sensitive	Low	Low

---



---

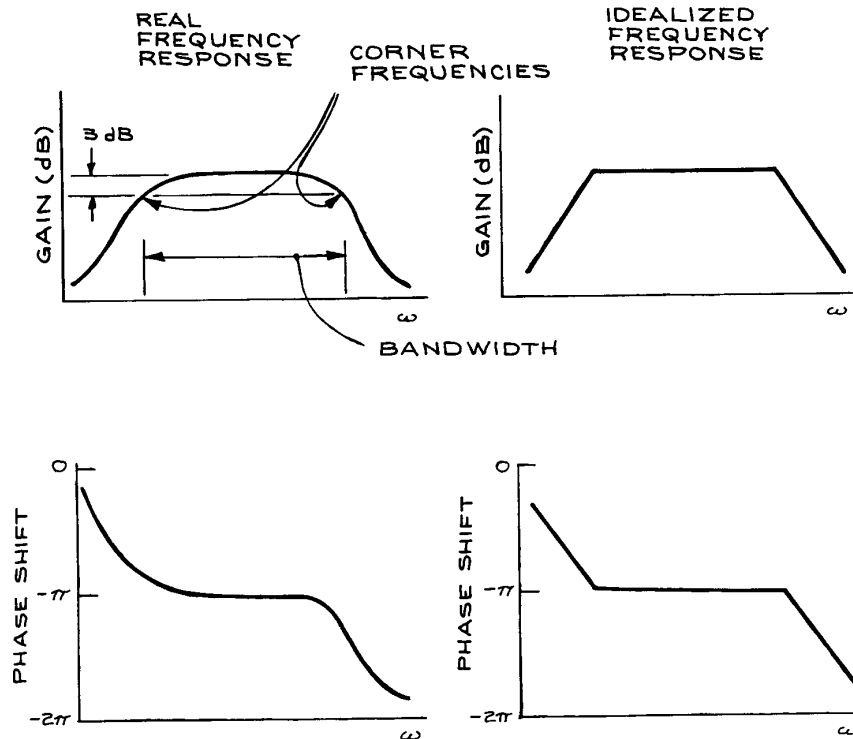
**TABLE 6.21 Amplifier frequency ranges**


---

<i>Amplifier Type</i>	<i>Frequency Range</i>
d.c.	0 to 10 Hz
Audio	10 Hz to 10 kHz
r.f.	100 kHz to 1MHz
Video	30 to 1000 MHz
VHF	30 to 300 MHz
UHF	300 to 1000 MHz
Microwave	1000 MHz to 50 GHz

---

always present to some degree because of stray capacitances) in such a way as to be in phase with the signal at the input, a *positive feedback* or *regenerative* situation will occur, giving rise to oscillations and unstable behavior. This is the origin of the familiar squawking and whistling that occurs in public address systems when the output from the loudspeakers finds its way back to the microphone. The stability of an amplifier against such oscillations is expressed in terms of *gain* and *phase margins*. If the value of the gain of an amplifier (in dB) is negative at frequencies for which the phase of the output is equal to that of the input, oscillations cannot occur. Correspondingly, if the output is out of phase with the input at all frequencies for which the gain is greater than unity, oscillations cannot occur. The amount of negative gain at zero phase is the *gain margin*; the phase difference at 0 dB gain is the *phase margin* (see Figure 6.63). The larger these margins, the more stable the amplifier. Wide-band video amplifiers are particularly susceptible to unstable behavior because even small stray capacitances are sufficient to provide low-impedance paths from the output to the input for high frequencies. Because of the large power consumption in the positive feedback



**Figure 6.62** Bode plots for a capacitance-coupled inverting amplifier.

(regenerative) mode, prolonged unstable operation of an amplifier can destroy it. This is especially true in the case of fast-pulse amplifiers, which must have a very large bandwidth in order to accurately amplify pulses with short rise times. Compensation by the addition of poles and zeros to the transfer function can remedy this at the expense of reduced bandwidth.

In classifying amplifiers by power rating, the distinctions are rather arbitrary. Generally, *preamplifiers* are designed to operate with input voltage levels of millivolts and less, input currents of microamperes and less, and input charge of picocoulombs and less. Gain is not so important in preamplifiers as is amplifier noise, which limits ultimate sensitivity. In some applications, a unit-gain preamplifier is used as an impedance-matching device. *Power amplifiers* are designed to furnish anywhere from a few to several hundred watts to a matched load. Audio amplifiers are examples. Intermediate between preamplifiers and power amplifiers are amplifiers that operate with input signal levels from

preamplifier outputs and produce outputs from a few to several hundred milliwatts. Gains of such amplifiers are high, and noise considerations are much less important than overload recovery. Frequency- and gain-compensating circuits are often incorporated in such amplifiers.

It has been assumed thus far that the output of an amplifier is a linear function of the input at any frequency. Because the active elements (transistors) in solid-state amplifiers have nonlinear characteristics, it follows that amplifiers can only operate in a linear way within a limited range of input conditions. For all amplifiers, there is an input signal level beyond which the output signal is severely distorted. Clipping of the output waveform is an obvious form of distortion. Less obvious distortions can be quantified by a Fourier analysis of the output waveform for a sinusoidal input. The distortion in the output can be described in terms of the coefficients of the harmonics. This *harmonic distortion* is important in critical audio-amplifier applications, but not generally in the laboratory.

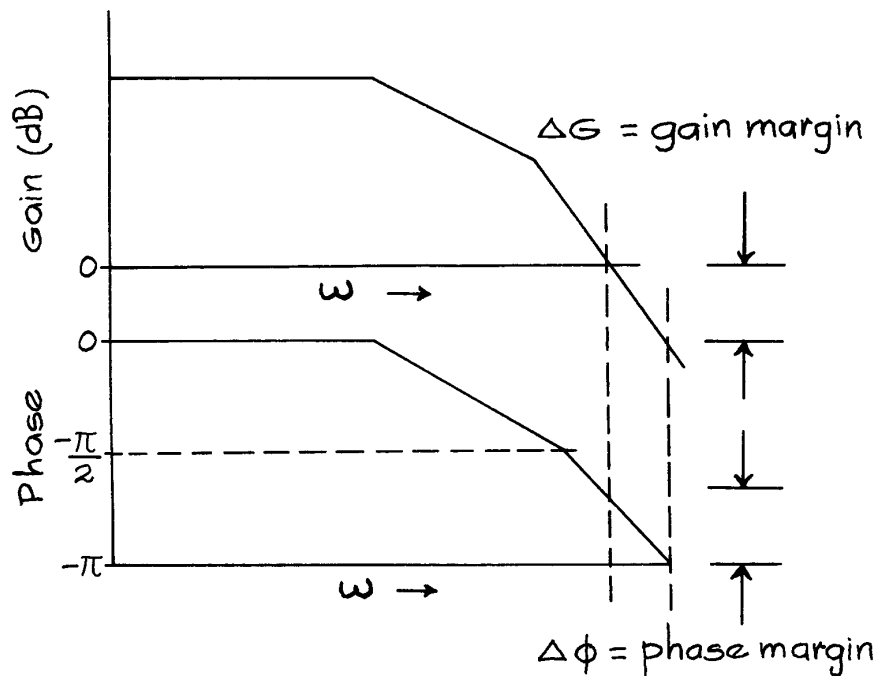


Figure 6.63 Gain and phase margins.

The properties of solid-state amplifiers may also be temperature-dependent, and such factors as gain and distortion, though acceptable at room temperature, may become unacceptable at high or low temperatures.

### 6.4.2 General Transistor-Amplifier Operating Principles

The active elements in solid-state amplifiers can be  $n$ - $p$ - $n$  or  $p$ - $n$ - $p$  bipolar junction transistors or field-effect transistors. Field-effect transistors can be  $n$ -channel or  $p$ -channel and can be operated in the depletion or the enhancement mode. They can have gate structures in direct contact with the conducting channel (JFETs) or gates that are insulated from the conducting channel by an oxide layer (MOSFETs).

To use an inherently nonlinear transistor to make a linear amplifier requires ingenuity. Standard methods involve operating the transistors over a limited range of voltage and current where their nonlinear characteristics can be

approximated by linear functions, using compensating circuits to cancel the nonlinearities with devices that have opposite characteristics, or applying negative feedback.

Transistors must be properly biased if they are to work correctly. This means supplying, from external sources such as batteries or d.c. power supply circuits, the correct potential differences across the terminals and the correct currents into them. Even in the absence of an external signal, transistors dissipate power. The values of the d.c. bias (or quiescent) voltages and currents define the *operating point* of the transistor. Signals from an outside source or previous stage are then superimposed on the quiescent values. As long as the signal does not represent too large a deviation from these values, the transistor will operate linearly. If the external signal is sufficiently large, however, it can *cut off* the transistor (that is, reduce all the currents through it to zero) or *saturate* it (that is, cause the transistor to act as a short circuit).

Since the bias voltages and currents are d.c. and the signal voltages and currents are a.c., it is possible to



separate them. Capacitors or transformers can be used to couple a.c. signals without disturbing the quiescent d.c. values. When capacitance coupling is used, the capacitors are called *blocking* capacitors because they block d.c. voltages and currents. Reasonably sized blocking capacitors usually limit the low-frequency response of amplifiers to  $> 1$  Hz (capacitor lead inductance and stray shunt capacitances to ground can also limit the high-frequency response). In many applications, the poor transient response of the high-pass circuits formed by the blocking capacitors cannot be tolerated. Direct-coupled amplifiers – where the bias levels of each transistor stage are designed to be compatible with the preceding and succeeding stages – provide high gain to 0 Hz. Because of the need to match active components and resistor ratios very precisely, such amplifiers are often *integrated-circuit (IC)* amplifiers, where all the elements of the circuit are manufactured simultaneously on a single chip. Direct-coupled designs are, in fact, particularly suited to the IC manufacturing process, which favors the production of transistors and diodes, but requires quite special techniques for the production of resistors and capacitors. *Chopping* is another method of achieving good low-frequency response in amplifiers. With this method, the input signal is chopped at a frequency much higher than the highest characteristic frequency of the input signal. The resulting a.c. signal can be amplified by conventional a.c.-coupled amplifier stages. Rectification of the output of the final stage restores the input waveform. The advantages of chopper over direct-coupled amplifiers are lower uncompensated input currents and voltages, and comparative freedom from drifts with temperature. These advantages have been largely offset by IC direct-coupled amplifier designs with FET input stages.

Amplifier properties and specifications are best understood by considering as an example the general-purpose 741 IC operational amplifier. The name *operational amplifier (op amp)* was originally applied to amplifiers used in analog-computer circuits, which performed a wide variety of mathematical operations. The term is now more generally applied to any high-gain, differential-input amplifier.

*Differential-input amplifiers* have two input terminals that are electrically isolated from ground and each other. They are called the *non-inverting (+)* and *inverting (–)* terminals. The output signal is proportional to the

difference between the signals at the two input terminals. *Single-ended* inputs are those with one input terminal and a ground terminal. Most operational amplifier outputs are single-ended – that is, the output is developed between a single output terminal and ground. The ideal operational amplifier has the following characteristics:

- Infinite input impedance
- Zero output impedance
- Infinite voltage gain,  $180^\circ$  out of phase with the inverting input terminal
- Infinite bandwidth
- Zero output voltage for identical input voltages at the two differential input terminals
- Properties independent of temperature, voltage levels, and frequency.

Real operational amplifiers depart from the ideal. Table 6.22 compares the properties of the common 741 IC operational amplifier and the premium OP07. The operational amplifier symbol is shown in Figure 6.64. The supply voltages  $V_+$  and  $V_-$  are applied to the indicated terminals, not to be confused with the signal input terminals. The output is at the apex of the triangle and the terminals marked *null* are for an external potentiometer to cancel the offset voltage in critical applications.

Common operational-amplifier terms include:

*Common-mode rejection ratio (CMRR)*. This is the ratio of the differential gain  $A_D$  to the common-mode gain  $A_{CM}$  (often expressed in dB). The output of an operational amplifier is  $A_D$  times the differential voltage plus  $A_{CM}$  times the common mode voltage. For arbitrary voltages  $v_1$  and  $v_2$  at the two amplifier inputs, the differential voltage is  $(v_1 - v_2)/2$  and the common mode voltage is  $(v_1 + v_2)/2$ . For the 741, the 90 dB value of the CMRR means that a 1 V signal at each input terminal will result in a 32  $\mu$ V output signal. This is shown in the following calculation:

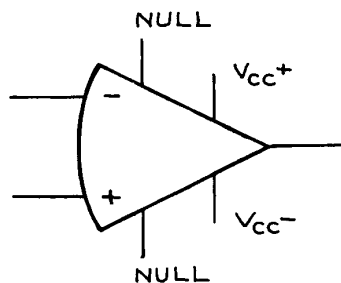
$$90 \text{ dB} = 20 \log \frac{A_D}{A_{CM}} \quad \frac{A_D}{A_{CM}} = 10^{4.5} = 3.2 \times 10^4 \quad (6.32)$$

Since the differential voltage is 0 V, the output voltage will be the common-mode voltage multiplied by the common-mode gain  $A_{CM}$ . For a common-mode

**Table 6.22 Comparison of Operational Amplifier Parameters<sup>a</sup>**

	741	OP07
CMRR (dB)	90	120
Open-loop d.c. gain	200 000	500 000
GBWP (MHz)	1.5	0.6
Slew rate (V/ $\mu$ s)	0.5	0.3
Input resistance (M $\Omega$ )	2	80
Output resistance ( $\Omega$ )	75	60
Input offset voltage (mV)	5	0.010
Input offset current (nA)	20	0.3
Input bias current (nA)	80	0.7
Input voltage range (V)	$\pm 13$	$\pm 13$
Maximum peak-to-peak output voltage swing (V)	28	25
Power supply rejection ratio ( $\mu$ V/V)	150	5
Power consumption (mW)	50	75

<sup>a</sup> Typical characteristics at 25 °C with power-supply voltages of  $\pm 15$  V.



**Figure 6.64** Schematic symbol for the operational amplifier, showing input, output, power supply, and null terminals.

voltage of 1.0 V, the output voltage is then  $1.0/3.2 \times 10^4 = 32 \times 10^{-6}$  V.

**Open-loop d.c. gain.** This is the gain in the absence of a connection from output to input – that is, in the

absence of any *feedback*. The gain of the amplifier in the presence of a feedback network is called the *closed-loop gain*. With open-loop gains of the order of 200 000 or more, it can be seen that the region of linear operation applies to only very small input signals. At a gain of 200 000, an input signal of  $+75 \mu$ V is sufficient to drive the output to  $+15$  V, the maximum possible level.

**Frequency response.** This is given in terms of either the GBWP or the frequency at which the gain equals 0 dB (unit gain). The two methods give nearly the same numbers for a single-pole transfer function. For the 741, at 1 MHz the gain has fallen by over five orders of magnitude from the d.c. value. It should be kept in mind that these frequency-response data are based on so-called *small-signal* conditions, where the input is small enough so that the output stages are operating well within their linear region.

**Slew rate.** This gives the time response of the amplifier to large input signals. For this specification the amplifier is connected in a unit-gain (closed-loop) configuration with negative feedback, and an input voltage step is applied. The slew rate is then taken as the ratio of the amplitude of the input voltage to the rise-time for the output voltage waveform. For the 741, the slew rate is 0.5 V/ $\mu$ s.

**Input resistance.** This is the resistance  $R_i$  between the input terminals with one terminal grounded.

**Output resistance.** This is the resistance  $R_o$  (seen by the load) between the output terminal and ground. It determines whether the amplifier approximates a voltage or a current source.

**Input offset voltage.** This is the d.c. voltage  $V_{IO}$  that must be applied across the input terminals to bring the d.c. output to zero. For the 741 it is 0.8 mV and can be compensated for. The 741 is not a good choice for the amplification of millivolt d.c. levels.

**Input offset current ( $I_{IO}$ ).** This is the difference between the currents into the input terminals necessary to bring the output voltage to zero.

**Input bias current ( $I_{IB}$ ).** This is the average of the currents into the input terminals with the output voltage at zero. It should be noted that these bias currents will flow through circuit elements (generally resistors) attached to the input terminals. If the

resistances are not identical, voltage differences will be developed across the input terminals and amplified along with the input signal.

*Input voltage range ( $V_I$ ).* The maximum allowed input voltage for which proper operation can be maintained.

*Input voltage range ( $V_I$ ).* The maximum allowed input voltage for which proper operation can be maintained.

*Maximum peak-to-peak output voltage swing.* The maximum peak-to-peak output voltage that can be obtained without clipping about a quiescent output voltage of zero.

*Power-supply rejection ratio.* The change in input offset voltage for a given change in power-supply voltage. Operational amplifiers are particularly good in this regard and do not require highly regulated power supplies.

*Temperature effects.* The various offset and bias currents and voltages are temperature dependent. The variations of these parameters with temperature are specified by temperature coefficients.

The 741 is internally compensated with a capacitor at the high-gain second stage to produce a pole in the transfer function at 6 Hz. This ensures stable operation under even 100% feedback conditions (of course, this gain in stability reduces the high-frequency response of the amplifier). Two extra terminals are available for connection to an external 10 k $\Omega$  potentiometer, the wiper arm of which is connected to the negative power supply. By adjusting the potentiometer, the input offset voltage can be eliminated. Additional features include short-circuit output protection and input overload protection. Single amplifiers come in at least four different packages: an 8 lead metal can, a 14 lead dual in-line package (DIP), an 8 lead mini-DIP, and a 10 lead *Flat Pak*. There is a commercial model – the 741C – that operates over a restricted temperature range, and a military model – the 741A – that operates over an extended temperature range. The 741 sells for about the price of a few 1/4 W composition resistors.

The OP07 is a premium grade operational amplifier with a very low input offset voltage, obtained by trimming thin-film resistors in the input circuit. The low offset voltages usually eliminate the need for external null circuits. The OP07 also has low input bias current and high open-loop gain. It is a direct pin-for-pin replacement for the 741.

**Table 6.23 Some special operational amplifiers**

Low offset voltage	OP07
Low bias current	(Harris) HA-5180
High voltage	(Harris) HA-2640/45
Low power consumption	(National) LF441
Single-power-supply operation	(National) LM10
Wide bandwidth	(National) LM318 (Burr-Brown) OPA606
Video preamplifier	733

Some examples of more specialized operational amplifiers are given in Table 6.23.

Because of their high gain, wide bandwidth, and small offsets, operational amplifiers with suitable feedback networks can perform a wide range of electronic functions. There are many excellent texts devoted to operational-amplifier applications.<sup>4</sup> Here we give only a list of the basic configurations. In all of these it is important to keep in mind the following design rules:

- (1) The offset-voltage temperature coefficient must produce voltages much less than the input signal throughout the anticipated temperature range. The offset voltage can be eliminated with an external nulling circuit at any given temperature, but variations with temperature cannot be easily accommodated.
- (2) The voltages created by the bias currents flowing through resistances attached to the input terminals must be less than the signal voltage. For high output-impedance sources this may create a problem, but it can be solved by introducing an intermediate buffer stage with a high input impedance (so as not to load the source) and a low output impedance (to minimize the effects of input bias current). The temperature coefficient of the offset current must produce offset voltages much less than the signal voltage over the anticipated temperature range. To minimize bias-current effects, the resistances to ground from both input terminals should be identical.
- (3) Frequency response and compensation must prevent the amplifier from breaking into oscillation. The 741 is internally compensated. High-frequency amplifiers

lack internal compensation, but have extra terminals for external compensation circuits.

- (4) The maximum rate of change of a sinusoidal output voltage  $V_0 \sin \omega t$  is  $V_0 \omega$ . If this exceeds the slew rate of the amplifier, distortion will result. Slew rate is directly related to frequency compensation, and high slew rates are obtained with externally compensated amplifiers using the minimum compensation capacitance consistent with the particular configuration. Data sheets should be consulted.

### 6.4.3 Operational-Amplifier Circuit Analysis

If ideal behavior can be assumed (as is reasonable for circuits where the open-loop gain is much larger than the closed-loop gain), only the following two rules need to be used to obtain the transfer function of an operational-amplifier circuit, as shown schematically in Figure 6.65:

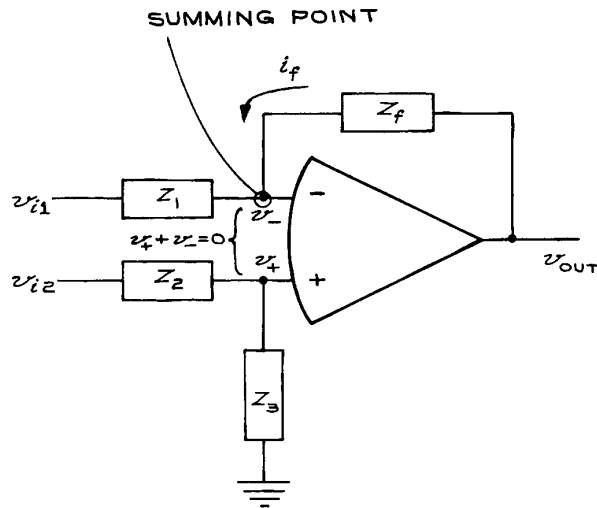
- (1) The voltage across the input terminals is zero.
- (2) The currents into the input terminals are zero (the inverting terminal is sometimes called the *summing point*).

Applying these to the generalized circuit of Figure 6.65, we see that the voltage at the + input is  $v_{i2}Z_3/(Z_1 + Z_2)$ , since there is no current flowing into the + input and  $Z_2, Z_3$  form a voltage divider. From Rule 1, the voltage at the - input must also equal  $v_{i2}Z_3/(Z_1 + Z_2)$ . The current through  $Z_1$  is then the potential difference across it divided by  $Z_1$ , or:

$$\left[ v_{i1} - v_{i2} \frac{Z_3}{(Z_2 + Z_3)} \right] / Z_1 \quad (6.33)$$

This current must be equal in magnitude and opposite in sign to the current from the output through the feedback element  $Z_f$ . This current,  $i_f$ , is given by the potential difference across  $Z_f$  divided by  $Z_f$ :

$$\begin{aligned} i_f &= \left[ v_{out} - v_{i2} \frac{Z_3}{(Z_2 + Z_3)} \right] / Z_f \\ &= - \left[ v_{i1} - v_{i2} \frac{Z_3}{(Z_2 + Z_3)} \right] / Z_f \end{aligned} \quad (6.34)$$



**Figure 6.65** The general operational-amplifier circuit. Power supply and null terminals not shown.

Solving for  $v_{out}$ , we obtain:

$$-v_{out} = -v_{i1} \frac{Z_f}{Z_1} + v_{i2} \left( \frac{Z_3}{Z_2 + Z_3} \right) \left( 1 + \frac{Z_f}{Z_1} \right) \quad (6.35)$$

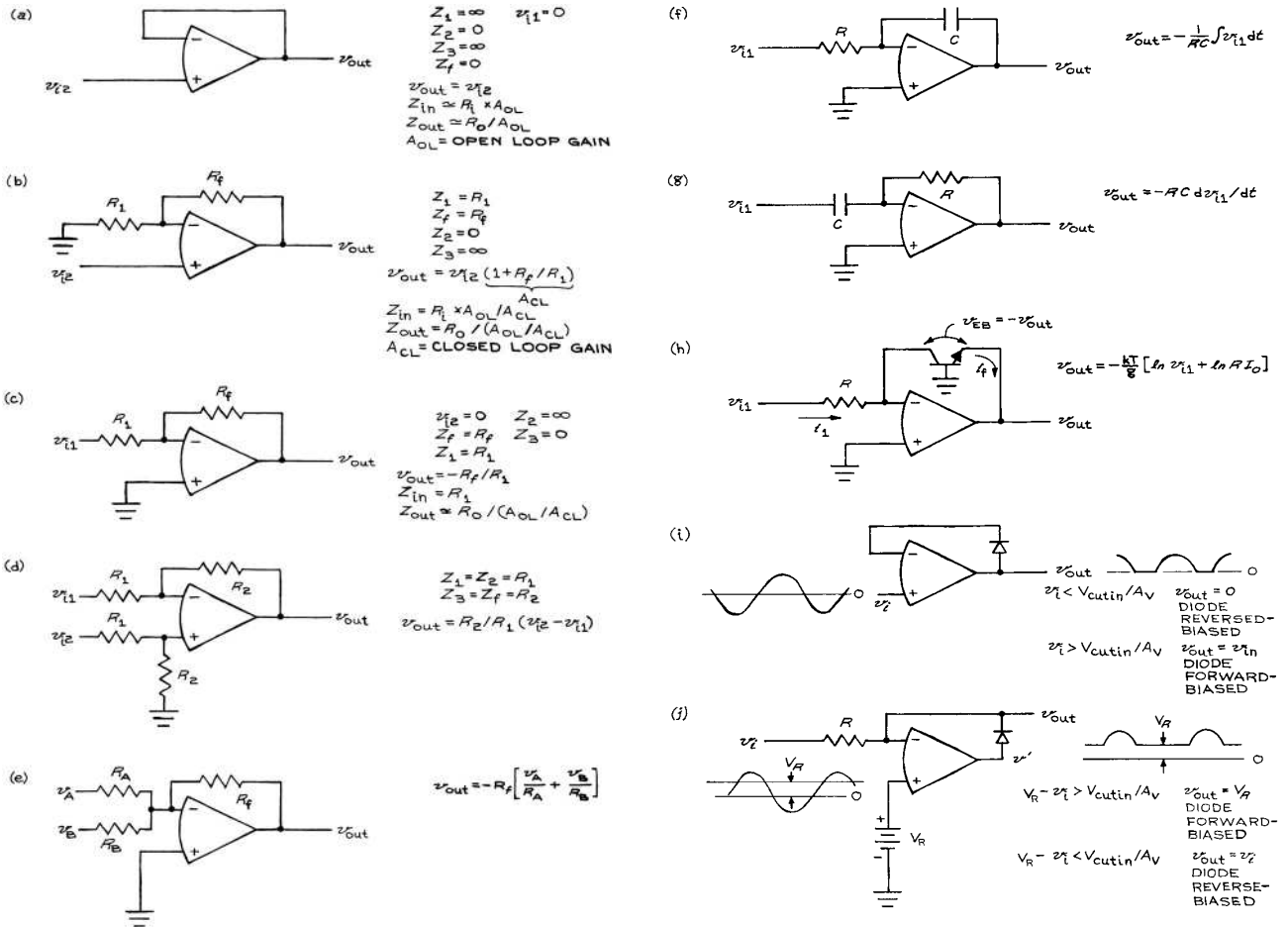
Some useful configurations are illustrated in Figure 6.66.

The *summer* circuit [Figure 6.66(e)] is an elaboration of the inverting amplifier, with the current into the summing point coming from two external sources ( $v_A$  and  $v_B$ ) canceled by the current from the feedback loop,  $v_{out}/R_f$ . If  $R_A = R_B = R$ , then  $v_{out} = -(R_f/R)(v_A + v_B)$ . If  $R_A = R_B$ , then  $v_{out}$  is equal to the negative of the weighted sum of voltages  $v_A$  and  $v_B$ , that is:

$$v_{out} = - \left[ \left( \frac{R_f}{R_A} \right) v_A + \left( \frac{R_f}{R_B} \right) v_B \right] \quad (6.36)$$

Any number of external signal sources can be summed in this way.

A sinusoidal input to the *low-pass filter* [Figure 6.66(f)] gives  $v_{out} = (j/\omega RC)v_{i1}$ . For a nonsinusoidal input, the current into the summing point from  $v_{i1}$  is  $v_{i1}/R$ ,



**Figure 6.66** Operational-amplifier configurations: (a) follower; (b) follower with gain; (c) inverting amplifier; (d) subtractor; (e) summer; (f) low-pass filter (integrator); (g) high-pass filter (differentiator); (h) logarithmic amplifier; (i) precision rectifier; (j) clamp. The parameters are defined in Figure 6.59.

because the inverting input is at ground (condition 1). Often one speaks of the inverting input being at *virtual ground*. Although there is no direct connection from the inverting input to ground, the condition in which both inputs are at the same potential results in the inverting input assuming a potential of zero when the noninverting input is at ground potential. The current through the feedback capacitor of this circuit is  $i_f$ , and the voltage across the capacitor,  $v_{out}$  is equal to

$(1/C) \int i_f dt$ . Since  $i_f$  and  $v_{i1}/R$  are equal in magnitude and opposite in sign:

$$v_{out} = (1/C) \int i_f dt = -(1/RC) \int v_{i1} dt. \quad (6.37)$$

and the circuit acts as an *integrator* of the input voltage. Integration times of several minutes or even hours are possible with high quality, low-leakage capacitors and operational amplifiers with small offset voltages and bias currents.

For the high-pass *filter* [Figure 6.66(g)],  $v_{\text{out}} = -RC(dv_{\text{in}}/dt)$  and the circuit acts as a differentiator.

The integrator and differentiator circuits are simple forms of active filters. More complex networks involving only capacitors and resistors can be used to obtain poles in the left half of the complex  $s$ -plane, which in the past were only obtained with inductors in passive filter networks. The operational amplifier allows the use of reasonable resistor and capacitor values even at frequencies as low as a few hertz. An additional advantage is the high isolation of input from output due to the low output impedance of most operational-amplifier circuits. The limitations of active filters are directly related to the properties of the operational amplifier. Inputs and outputs are usually single-ended and cannot be floated as passive filters can. The input and output voltage ranges are limited, as is the output current. Offset currents and voltages, bias currents, and temperature drifts can all affect active filter performance. A good introduction to active filters is the *Active Filter Cookbook* by D. Lancaster.

The circuits in Figure 6.66(b) to (g) are examples of the use of operational amplifiers in analog computation. With these circuits, the mathematical operations of addition, subtraction, multiplication by a constant, integration, and differentiation can be performed. Multiplication or division of two voltages is accomplished by a logarithmic amplifier, an adder (for multiplication) or subtractor (for division), and an exponential amplifier.

Both the *logarithmic* and *exponential* amplifiers rely on the exponential  $I$ - $V$  characteristics of a  $p$ - $n$  junction. When this junction is the emitter-base junction of a transistor, the collector current  $I_C$  with zero collector-base voltage is:

$$I_C = I_0[\exp(qV_{\text{EB}}/kT) - 1] \quad (6.38)$$

where  $I_0$  is a constant for all transistors of a given type and  $V_{\text{EB}}$  is the emitter-base voltage. Typically,  $I_0$  is 10 to 15 nA for silicon planar transistors.

For  $I_C \geq 10^{-8}$  A, the equation reduces to  $I_C = I_0 \exp(qV_{\text{EB}}/kT)$ . For the configuration shown in Figure 6.66(h),  $i_1 = v_{\text{in}}/R = I_0 \exp(qv_{\text{out}}/kT)$  and  $v_{\text{out}} = -(kT/q)(\ln v_{\text{in}} + \ln RI_0)$ . Since  $k$ ,  $T$ ,  $q$ ,  $R$ , and  $I_0$  are constants,  $v_{\text{out}}$  will be proportional to  $v_{\text{in}}$ . Practical logarithmic amplifiers can operate over three decades of input voltages. They do, however, need temperature-compensating circuits that generally require the use

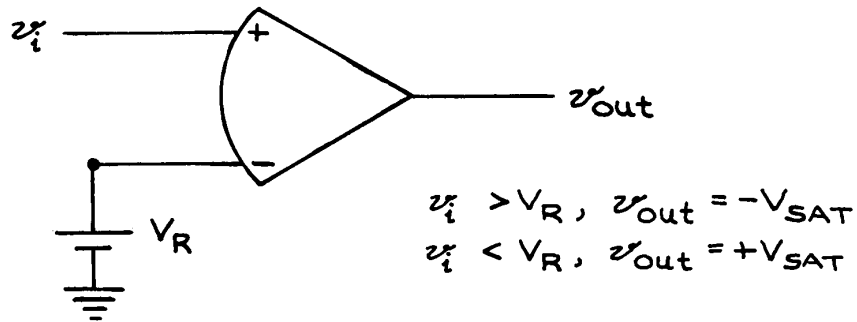
of matched pairs of transistors. In addition to their arithmetic use, logarithmic amplifiers are frequently used for transforming signal amplitudes that cover orders of magnitude to a linear scale. Exponential (antilogarithmic) amplifier circuits are logarithmic circuits with the input and feedback circuit elements interchanged.

*Logarithmic* and *exponential* amplifiers are examples of nonlinear circuits – the output is not linearly related to the input. Operational amplifiers are used extensively in nonlinear circuits. Because the cutin voltage of simple diodes is a few tenths of a volt (0.2 V for Ge and 0.6 V for Si), they cannot be used to rectify millivolt-level a.c. voltages. A diode in the feedback loop of an operational amplifier [see Figure 6.66(i)] results in a rectifier with a cutin voltage equal to the diode cutin voltage divided by the open-loop gain of the amplifier. With a slight modification, the rectifier circuit can be converted to a *clamp* [see Figure 6.66(j)] in which the output follows the input for voltages greater than a reference voltage  $V_R$ , but equals  $V_R$  for input voltages less than  $V_R$ .

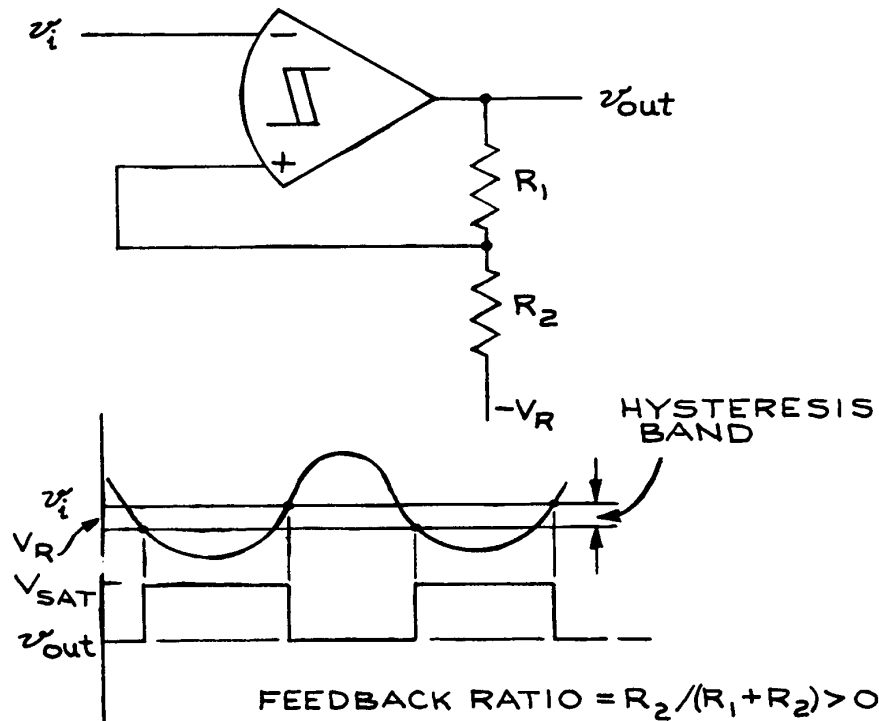
*Comparator* circuits are used to compare an input signal with a reference voltage level. The output is driven to  $V^+$  or  $V^-$  depending on whether the input is less than or greater than the reference level [see Figure 6.67(a)]. Using an amplifier in such a saturated mode is very poor practice and practical comparators with internally clamped outputs less than  $|V^+|$  and  $|V^-|$  are available. The transition between the two output states can be accelerated by the use of positive feedback [Figure 6.67(b)]. Such a circuit is called a *Schmitt trigger* and finds wide application in signal conditioning when it is necessary to convert a slowly changing input voltage into an output waveform with a very steep edge. The price paid for the fast transition is *hysteresis* – where the threshold voltage for a positive transition of the output is different from the threshold voltage for a negative transition. Hysteresis inhibits the output from oscillating between the two output states when the input is very close to the threshold level.

#### 6.4.4 Instrumentation and Isolation Amplifiers

*Instrumentation amplifiers* are used for amplifying low-level signals in the presence of large common-mode voltages, as, for example, in thermocouple and bridge



(a)



(b)

**Figure 6.67** (a) Simple comparator ( $-V_{sat}$  and  $+V_{sat}$  are the minimum and maximum voltages the amplifier can deliver;  $V_R$  is a constant reference voltage); (b) Schmitt trigger (comparator with positive feedback).

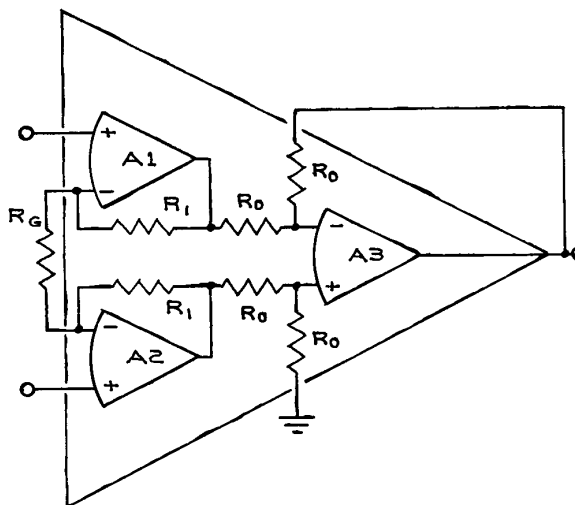
circuits. The instrumentation amplifier is different from the operational amplifier; the feedback network of the instrumentation amplifier is integral to it and a single external resistor sets the gain. The output of the instrumentation amplifier is usually single-ended with respect to ground and is equal to the gain multiplied by the differential input voltage. Compared to an operational amplifier in the differential configuration with equivalent gain, the instrumentation amplifier has a higher common-mode rejection ratio and higher input impedance.

*Instrumentation amplifiers* are often made up of three operational amplifiers, as shown in Figure 6.68. Amplifiers A1 and A2 together provide a differential gain of  $1 + 2R_1/R_G$  and a common-mode gain of unity. Amplifier A3 is a unit-gain differential amplifier. Amplifiers A1 and A2 operate as followers with gain, and resistors  $R_1$  do not affect the input impedances. Because amplifier A3 is driven by the low output impedances of A1 and A2, the resistors  $R_o$  can have relatively small values in order to optimize CMRR and frequency response while minimizing the effects of input offset currents.

The three-amplifier configuration is limited to common-mode voltages of  $\pm 8$  V in most applications. Typical high-performance instrumentation amplifiers are the INA110 (Burr-Brown), LH0038 (National Semiconductor), and AD624 (Analog Devices). Typical specifications for high-performance amplifiers are given in Table 6.24.

Instrumentation amplifiers cannot be used when the common-mode voltage exceeds the power-supply voltage. In these cases, an *isolation amplifier* is required. Isolation amplifiers are also useful when the input and output signals of an amplifier are referenced to different common voltage levels. It is often necessary, for example, to impose a low-voltage waveform on a d.c. high voltage. With an isolation amplifier, the modulation voltage can be derived from a source referenced to ground and the amplifier output can be referenced to the d.c. high voltage. The internal construction of the amplifier isolates the high-voltage output from the low-voltage input.

The basic isolation-amplifier configuration is shown in Figure 6.69. For the above example,  $V_{IN}$  is the modulation voltage referred to ground ( $V_{CM} = 0$ ), and  $V_{ISO}$  is the high voltage modulated by  $V_{IN}$ . In cases where  $V_{CM}$  exceeds  $\pm 10$  V, the input common terminal is not grounded and the common-mode voltage is referenced



**Figure 6.68** Three-amplifier instrumentation amplifier.

to the output common terminal. This effectively transfers  $V_{CM}$  across the isolation barrier and allows for  $V_{CM}$  voltages up to the maximum isolation rating of the amplifier.

The *isolation-mode rejection ratio (IMRR)* and CMRR are critical isolation-amplifier parameters because of the large common-mode and isolation voltages present. To calculate the effect of common-mode and isolation

**Table 6.24** Typical high-performance instrumentation amplifiers specifications

	INA 110	LH0038	AD6243
Gain range	1–500	1–2000	1–1000
Gain nonlinearity (%)	0.001–0.01	0.001	0.001
Input impedance ( $\Omega$ )	$5 \times 10^{12}$	$5 \times 10^6$	$10^9$
CMRR (dB)	90–110	120	130
Input offset-voltage drift ( $\mu\text{V}/^\circ\text{C}$ )	2	0.25	0.25
Input bias current (nA)	0.050	50	50
Bandwidth (MHz)	2.5	2	25
Noise, referred to input ( $\text{nV}/\sqrt{\text{Hz}}$ )	10	6	4



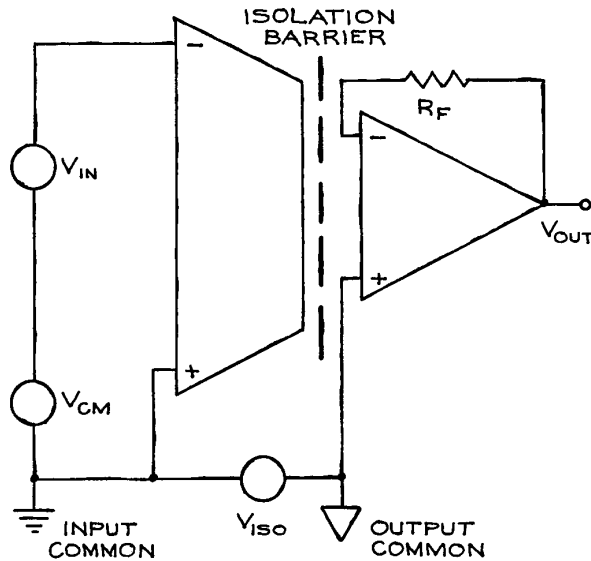


Figure 6.69 Isolation amplifier.

voltages on amplifier performance, it is necessary to add the product of IMRR and  $V_{ISO}$ , and that of CMRR and  $V_{CM}$  to  $V_{IN}$ . Typical values of IMRR and CMRR are 140 and 90 dB, respectively. For an IMRR of 140 dB, a 1 kV isolation voltage produces the equivalent of 0.1 mV at the input terminals of the amplifier. Other important parameters specific to isolation amplifiers are the maximum isolation voltage, isolation resistance, and isolation capacitance.

Isolation between input and output is achieved with optical couplers or transformers. The latter are more expensive, but have higher isolation voltage ratings and IMRR. An additional advantage of the transformer-coupled amplifiers is their built-in isolated power supply. The input and output stages of the optically coupled amplifiers must have separate power supplies. This means that the output-stage power supplies must be floated at the isolation voltage. Burr-Brown manufactures a full line of isolation amplifiers for medical and industrial applications. The 3450 and 3656 series are transformer-coupled, while the 3650 and 150100 series are optically coupled. Prices range from tens to hundreds of dollars, depending on specifications. Analog Devices produces a premium trans-

former-coupled isolation amplifier, the AD295, as well as the less expensive AD202.

### 6.4.5 Stability and Oscillators

For circuits composed only of passive elements, the poles of the transfer function will always lie on the left-hand side of the complex  $s = \sigma + j\omega$  plane, that is, will always be less than zero. For circuits with active elements, this is not necessarily the case. If, for example, a fraction of the output signal of an amplifier is returned to the input in phase with the original input signal, an unstable situation will result and the output of the amplifier will increase with time. Analysis of such a circuit will show at least one pole to lie in the positive half of the complex  $s$ -plane. Unwanted oscillations in amplifiers are due to the regenerative coupling of a fraction of the output signal back to the input (see Figure 6.70). Stray capacitances between input and output may be sufficient at high frequencies to couple the output signal back to the input. This can be minimized by having the output and input physically separated as far apart as possible. A copper shield separating the output and input is often also effective, and high-frequency circuits built on printed circuit boards should have generous ground-plane areas.

While the lack of stability is undesirable in amplifiers, sinusoidal *oscillators* depend on such instabilities for their operation. Consider an amplifier with gain  $A_o$  (it can be any one of the four already considered). If  $X_o$  represents the output signal (a current or voltage) and  $X_i$  the input signal (also a current or voltage),  $X_o = A_o X_i$ . For a simple amplifier  $X_i = X_s$  (the source signal).

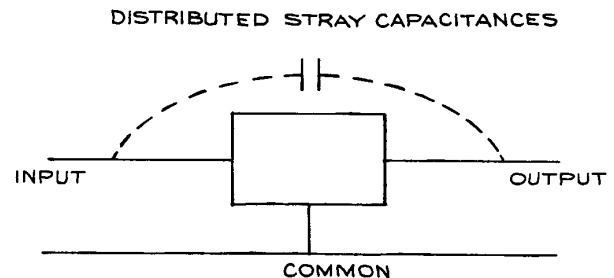


Figure 6.70 Unwanted feedback from stray capacitance.

If a fraction  $\beta$  of the output is sent back to the input in such a way as to cancel part of the source signal, the new signal  $X_i$  is  $X_s - \beta X_o$ , and the gain of the amplifier with feedback is the output signal divided by the source signal:

$$A_f = \frac{X_o}{X_s} = \frac{X_o}{X_i + \beta X_o} \quad (6.39)$$

$$A_f = \frac{A_o}{1 + \beta A_o}$$

For  $\beta > 0$ , the feedback signal acts to cancel the signal from the source and  $A_f < A_o$ . This is called *negative* or *degenerative feedback* and is frequently used to stabilize amplifier gain and improve linearity and noise. For  $\beta < 0$ , the feedback signal adds to the source signal at the input and  $A_f > A_o$ . This is *positive* or *regenerative feedback* and is sometimes used to increase the gain of amplifiers at the expense of stability. If  $\beta A_o = -1$ ,  $A_f = \infty$  and the amplifier is unstable, with an output that breaks into oscillation or saturates.

In practice, both  $A_o$  and  $\beta$  are frequency dependent and the response of such feedback amplifiers can be calculated. If only stability criteria are desired, it is only necessary to determine whether any poles of the transfer function lie in the right half of the complex plane.

One will have an oscillator if  $\beta A_o$  can be made equal to  $-1$  for only a single frequency, by means of a frequency selective feedback network. There are a large number of frequency-selective circuits based on *LC* and *RC* networks. Quartz crystal oscillators are particularly stable and free of temperature effects. These devices work by the application of a potential difference across the faces of a quartz crystal that becomes deformed and oscillates at a resonant frequency determined by the size and orientation of the crystal planes with respect to the electrodes. Crystal oscillators with  $Q$  values of  $10^4$  to  $10^6$  at frequencies from tens of kHz to over 100 MHz are commercially available in DIP and SMT configurations.

### 6.4.6 Detecting and Processing Pulses

There is a large class of experiments, especially in nuclear and particle physics, that deals with the detection and analysis of single events. With the advent of detectors

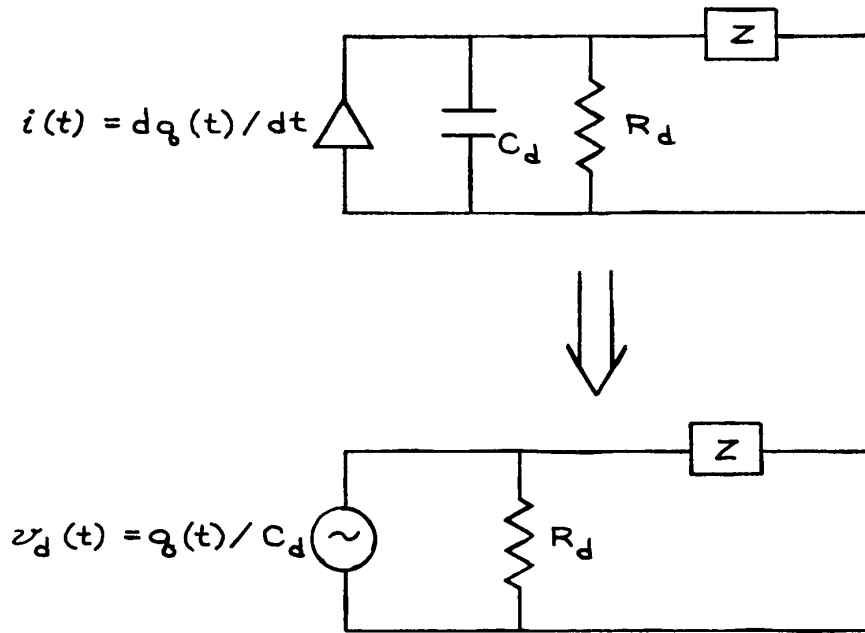
for low-energy photons, electrons, and ions, these techniques have spread to the fields of atomic and molecular physics, chemistry, and biology. Commercial instrumentation is widely available, but to specify and use the units to best advantage a knowledge of the basic properties of the detectors, amplifiers, and associated processing circuits is required.

Table 6.25 summarizes the properties of the more common types of detectors. The first part of the table allows one to estimate the output of a given detector for an arbitrary input. A 5 MeV alpha particle depositing all its energy in a silicon detector, for example, will create  $1.4 \times 10^6$  electron-hole pairs. With a suitable bias across the detector, the charge can be collected in 0.1 to 10 ns. Similar considerations allow one to calculate the outputs of other detectors. The detectors can be represented electrically by the equivalent circuit of Figure 6.71. The choice of preamplifier, amplifier, and shaping and analysis circuits now depends on the information to be obtained from the pulse. Figure 6.72 shows the arrangement of three general categories of experiments – counting, timing, and pulse-height analysis. Each one has different electronics requirements, but before considering each of these systems in detail it is worthwhile to examine the circuits commonly available for amplification and shaping.

**Table 6.25 Particle Detectors**

Particle Detector	Output Signal Level	Charge-Collection Time
Semiconductor:		
Ge	2.8 eV/electron hole pair	0.1–10 ns <sup>a</sup>
Si	3.5 eV/electron hole pair	
Photomultiplier	$10^6$ – $10^7$ electrons/ photoelectron	1 ns
Electron multiplier	$10^6$ – $10^8$ electrons/incident electron	0.5–1 ns
Microchannel plate	$10^3$ – $10^4$ electrons/incident electron	0.1 ns
Scintillator:		
Plastic } Alkali halide }	3 keV/photon	{ 0.1–10 <sup>o</sup> ns 1 <sup>o</sup> us
Gas-filled tube	25–35 eV/ion pair	0.01–5 $\mu$ s

<sup>a</sup> For narrow depletion depth.



**Figure 6.71** Equivalent circuit of a detector.

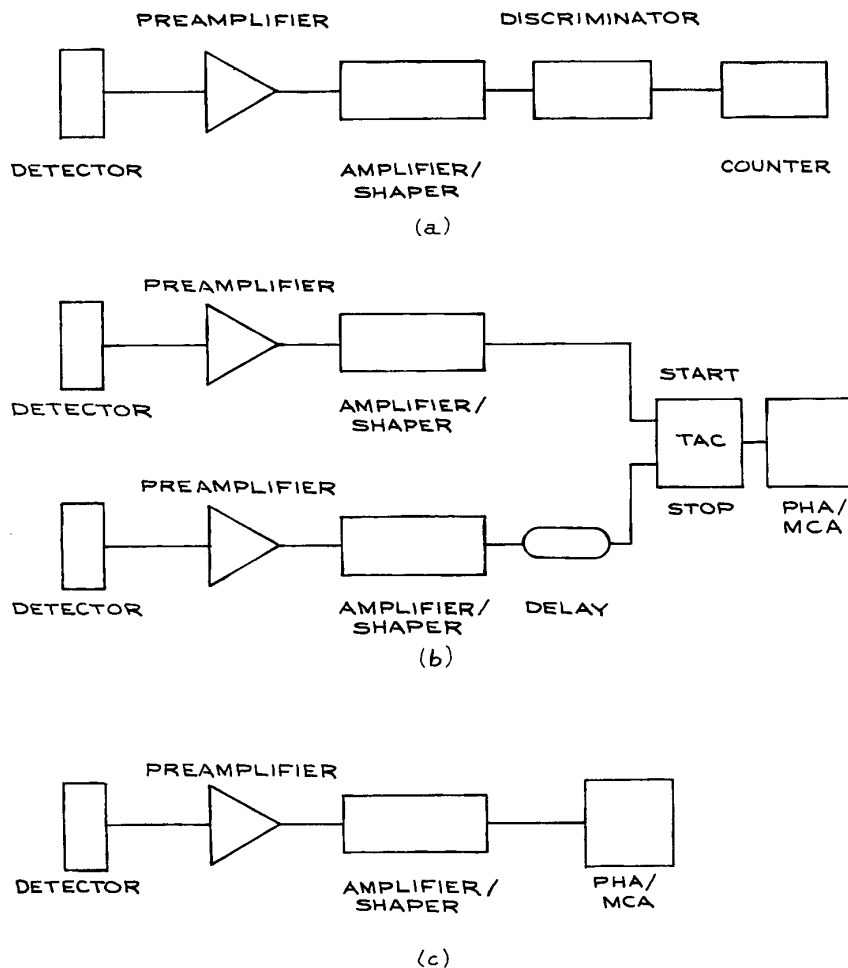
Preamplifiers can be voltage, current, or charge-sensitive. Voltage preamplifiers are not generally used, because the voltage generated by most detectors depends on the capacitance of the detector (Figure 6.73) that can change with bias voltage and other circuit parameters. Current preamplifiers have low input impedances and are designed to convert the fast current pulse from a photomultiplier or electron multiplier to a voltage pulse. Their sensitivity is specified as output millivolts per input milliamper. Charge-sensitive preamplifiers are the most commonly used preamplifiers because their gain is independent of the capacitance of the detector. Their sensitivity is specified in output millivolts per unit input charge or in equivalent output millivolts per MeV deposited in a given solid-state detector.

A schematic diagram of a combined detector and charge-sensitive preamplifier is shown in Figure 6.73. Neglecting the effect of the coupling capacitor, the Laplace transform of the output voltage  $v_o(s)$  is:

$$\bar{v}_o(s) = \frac{\bar{Q}(s)}{C_f} \frac{1}{1 + 1/R_f C_f s} \quad (6.40)$$

The function has a pole at  $s = -1/R_f C_f$ ; the Bode plot is given in Figure 6.74. The high-frequency pole at  $\omega_2$  is a result of the preamplifier transfer function. The output waveform for a rectangular input pulse of duration  $\tau$  will be a *tail pulse* of rise time  $\tau$ , amplitude  $Q/C_f$ , and decay time  $R_f C_f$ .

For very small values of  $\tau$ , the rise time of the output pulse will be determined by the characteristics of the amplifier itself. Practical amplifiers use high-stability NPO capacitors for  $C_f$  with values from 0.1 to 5 pF; the feedback resistor  $R_f$  is made as large as is consistent with the signal rate and leakage current. Noise in charge-sensitive preamplifiers is specified in terms of the energy resolution FWHM in keV. This can be converted to an equivalent charge noise with the factors in Table 6.25. To test charge-sensitive preamplifiers, the circuit of Figure 6.75 can be used. An input step voltage applied to the capacitor  $C_T$  results in an amount of charge equal to  $C_T V_T$  deposited on the input of the preamplifier. For  $10^{-15}$  C with  $C_T = 10$  pF, the amplitude of the input voltage pulse must be 10 mV. The 50 Ohm resistor in the circuit is for termination purposes. The duration of



**Figure 6.72** Three categories of pulse experiments: (a) counting, (b) coincidence, and (c) pulse-height analysis.

the test input current will be equal to the rise time of the voltage input.

The output pulses from the preamplifier are generally very different in shape from the input signals. This is the case even with fast-current preamplifiers. An important function of the amplifier and shaping circuits is to increase the amplitude of the preamplifier output signal and change its shape to minimize pulse overlap and optimize the signal-to-noise ratio. To minimize pulse overlap, the duration of the pulses should be short compared with the average

time between them. In more quantitative terms, the rms voltage level of the pulse train to be processed should be much less than the required voltage resolution of the system. If  $n$  is the pulse rate and  $v(t)$  is the functional form of the pulse, then:

$$v_{\text{rms}} = \left[ n \int_0^{\infty} v(t)^2 dt \right]^{1/2} \quad (6.41)$$

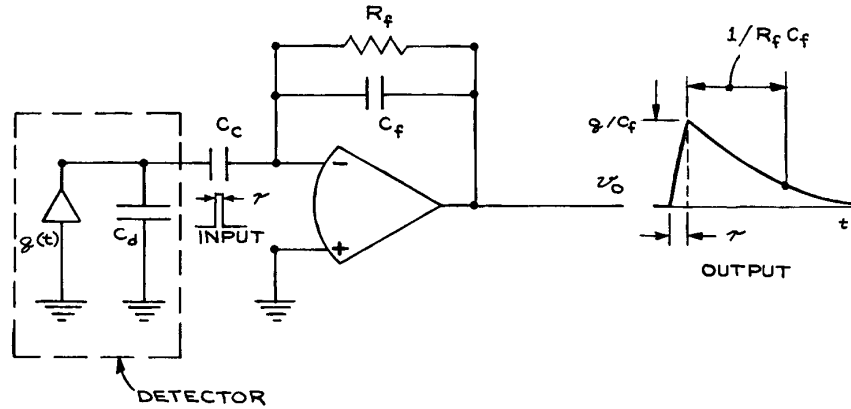


Figure 6.73 Combined detector and charge-sensitive preamplifier.

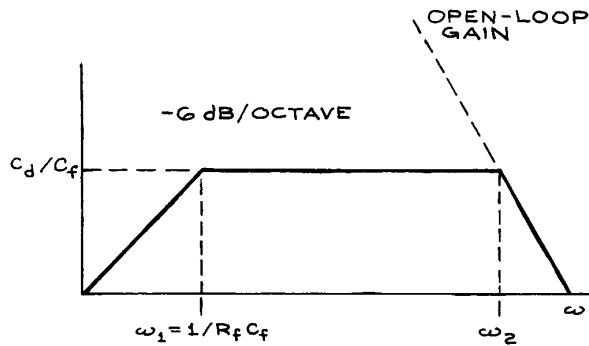


Figure 6.74 Bode plot for a charge-sensitive preamplifier.

For rectangular pulses of unit height and width  $\tau$ :

$$v_{\text{rms}} = n^{1/2} \tau^{1/2} \quad (6.42)$$

The long-tail pulse from a charge-sensitive preamplifier can be shortened with a differentiating circuit [Figure 6.76(a)]. The circuit reduces the fall time from  $R_f C_f$  to  $RC$ . A disadvantage of simple differentiation is undershoot, which appears on the trailing edge of the output pulse. This can be substantially eliminated by modifying the transfer function of the differentiating circuit to include a zero that exactly cancels the pole in the preamplifier

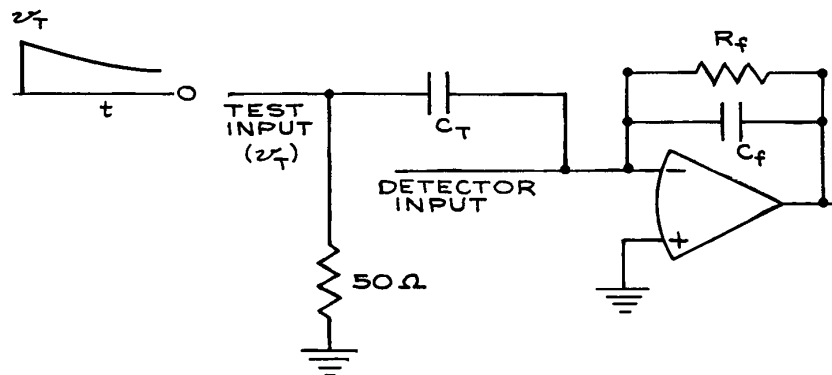
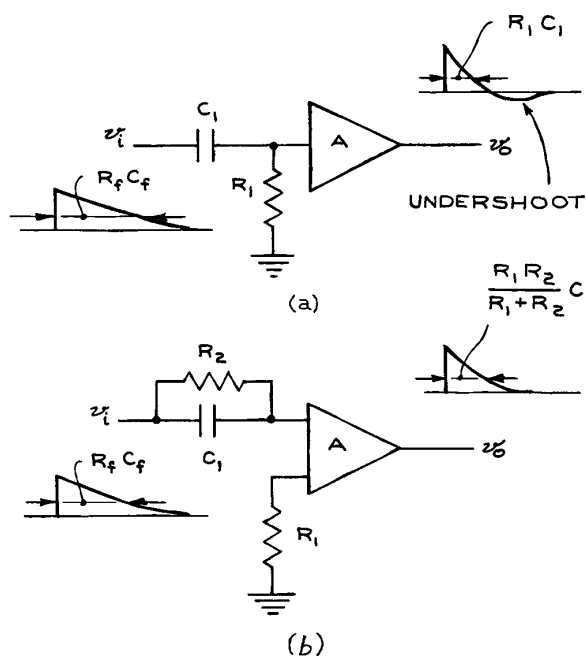


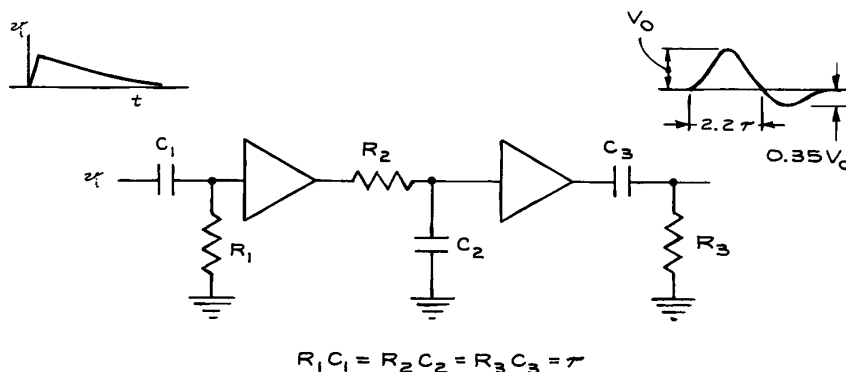
Figure 6.75 Test circuit for charge-sensitive preamplifiers.



**Figure 6.76** Differentiating circuits: (a) without pole-zero compensation; (b) with pole-zero compensation.

transfer function [Figure 6.76(b)]. The Laplace transform of the transfer function of Figure 6.76(b) is:

$$\frac{s + 1/R_2 C_1}{s + (R_1 + R_2)/(R_1 R_2 C_1)} \quad (6.43)$$

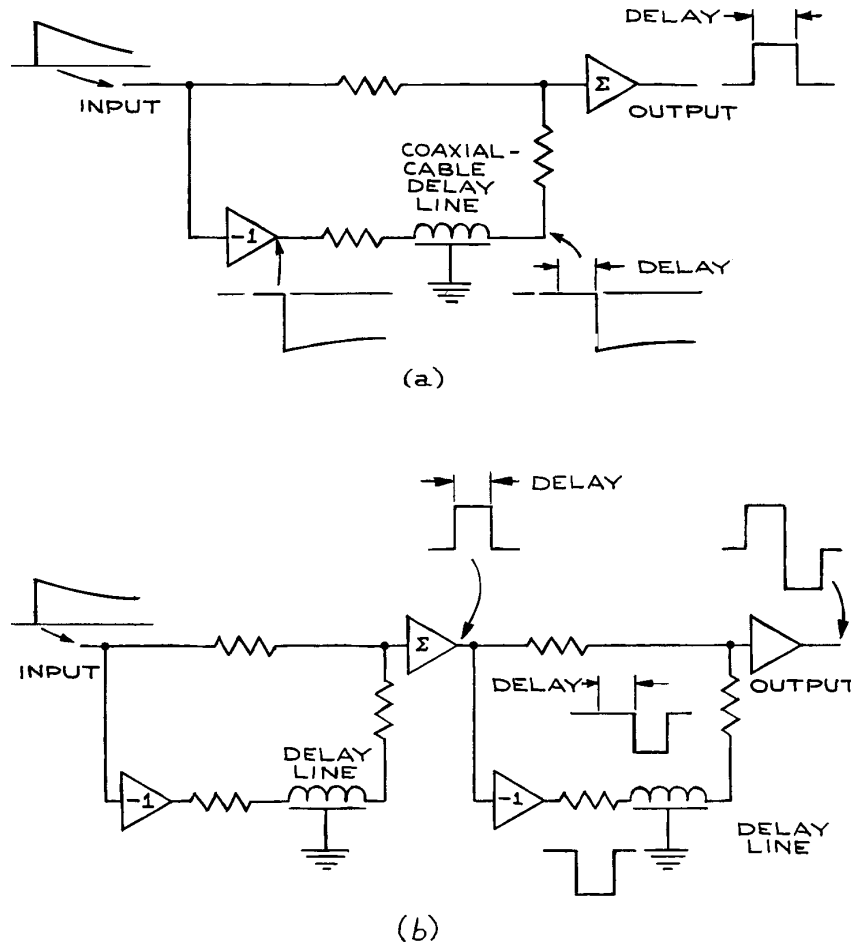


**Figure 6.77** Bipolar pulse shaping with CR-RC-CR circuits.

If  $R_2 C_1$  is set equal to  $R_f C_f$  from the preamplifier, undershoot can be eliminated. This technique of *pole-zero compensation* is commonly used in pulse shaping.

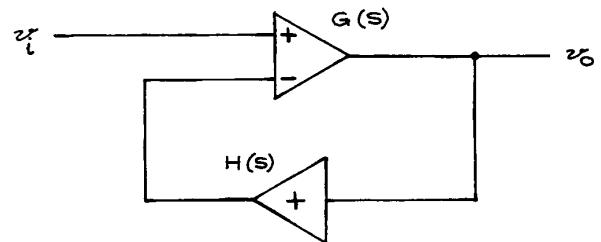
For pulse-height analysis, a Gaussian waveform has the best noise characteristics.<sup>5</sup> Such a shape can be obtained from a tail pulse by a single differentiation and an infinite number of cascaded integrations with integration time constants all equal to the differentiation constant. In practice, four integrations give a shape very close to Gaussian. Bipolar pulses can be obtained by double differentiation of a tail pulse with an intermediate integration (Figure 6.77). An important property of the bipolar pulse is that the zero crossing always occurs at a fixed time after its start, determined by the time constant, and is independent of pulse height. When used with zero-crossing discriminators, such pulses permit reliable timing to be accomplished with long-decay-time pulses. Bipolar pulses, however, have poorer noise characteristics than unipolar Gaussian pulses.

An alternative to RC shaping is delay-line shaping. The equivalent of unipolar and bipolar pulses can be obtained with single- and double-delay-line circuits (Figure 6.78). Advantages are the preservation of fast rise times, sharp trailing-edge clipping, and greatly reduced base-line shifts. These advantages are offset by the poor noise properties of the pulses compared with RC shaping. Delay-line shaping is mostly confined to timing applications.



**Figure 6.78** (a) Unipolar and (b) bipolar pulse shaping with delay lines.

The shaping circuits discussed in the previous paragraph are generally incorporated in the variable-gain main amplifier that follows the preamplifier. While noise is a consideration with such amplifiers, it is not so critical as with preamplifiers because of the larger signal levels. Important considerations are linearity and fast overload recovery. To reduce base-line shifts from pulse pileup, main amplifiers often incorporate a *baseline restorer circuit*. The principle of operation is illustrated in Figure 6.79. A noninverting amplifier with gain  $\bar{H}(s)$  is placed in the negative-feedback loop of an amplifier with gain



**Figure 6.79** A baseline restorer circuit.

$\bar{G}(s)$ . The Laplace transform of the transfer function for the system is:

$$\frac{\bar{G}(s)}{1 + \bar{G}(s)\bar{H}(s)} \quad (6.44)$$

If both  $\bar{G}(s)$  and  $\bar{H}(s)$  are themselves transforms of single-pole transfer functions, given by:

$$\bar{G}(s) = \frac{A}{1 + T_1s} \quad \text{and} \quad \bar{H}(s) = \frac{K}{1 + T_2s} \quad (6.45)$$

where  $A$  and  $K$  are the mid-band gains of the amplifiers, the overall transfer function is:

$$\frac{A(1 + T_2s)}{AK + (1 + T_1s)(1 + T_2s)} \quad (6.46)$$

At high frequencies ( $s \gg 1$ ) the transfer function reduces to  $\bar{G}(s)$ , while for d.c. ( $s = 0$ ) the function is  $1/K$ . The overall effect of the circuit is to amplify high-frequency pulses while attenuating low-frequency baseline shifts.

For counting applications, the principal considerations are pulse rate and discriminator setting. The shaped pulses from the amplifier should be sufficiently short in duration to avoid pileup and baseline shifts. A leading edge discriminator can be used with a unipolar pulse. Since such units are based on regenerative (positive-feedback) comparator circuits that have hysteresis, the decay of the input pulse must not pass through the recovery voltage level until the triggered state is attained. When bipolar pulses are to be counted, zero-crossing discriminators should be used.

Timing systems require that the leading-edge information from the detector pulse be faithfully preserved. Timing errors can arise from due time variations in the processed pulses relative to the input pulse, long-term drift from component aging, and noise. Wide-band current preamplifiers and fast timing amplifiers with short integrating and differentiating shaping circuits or delay-line shaping should be used. When using wide-band amplifiers, precautions should be taken to electrically terminate all connections properly and avoid stray capacitances that can form positive-feedback paths and

produce instability. When the detector signal itself is of sufficient amplitude – as it is for the output of a photomultiplier tube used with a scintillation detector – a timing signal can be taken directly from the detector without further amplification or shaping. Leading-edge and constant-fraction discriminators are most commonly used for timing. The *constant-fraction discriminator* provides the better resolution times. With this discriminator, the input signal is delayed and combined with a fixed fraction of the undelayed signal. A bipolar pulse with a zero crossing independent of rise time and amplitude results can be used to activate a zero-crossing detector.

In Figure 6.72(b), the delay inserted in the *stop* line of the system is to ensure that the *stop* pulse to the time-to-amplitude converter (TAC) will always follow the *start* pulse. The TAC converts the time difference between the *start* and *stop* input pulses to a single pulse of an amplitude proportional to the time difference. High quality, low-attenuation coaxial cable makes an excellent delay line, and lumped-parameter delay lines are also commercially available. Timing information is usually obtained by processing the output of the TAC with a pulse-height analyzer (PHA). A more direct method is with a time-to-digital converter that converts the time difference between the *start* and *stop* pulses to a digital word for processing. Depending on detector characteristics, time resolution of pico-seconds can now be obtained.

Particle-spectroscopy experiments rely on the preservation of the pulse-amplitude information from the detector because the amplitude is a direct measure of the particle energy. Optimum amplitude resolution is obtained by operating the preamplifier and amplifier in their linear ranges, reducing sources of noise and baseline shifts from pileup, and using semi-Gaussian pulse shapes as inputs to the PHA. The function of the PHA is to convert the pulse amplitude to a binary number. This number can then be used to address a memory location in a multichannel analyzer (MCA) or PC. The contents of the memory location specified by the encoded pulse height is then increased by one. Nonlinearities in the preamplifier and amplifier can be both differential and integral, illustrated in Figure 6.80. The *differential non-linearity* is the ratio of the slope of the amplifier gain curve, at a specified input level, to the slope, at a reference input level. The *integral nonlinearity* is the



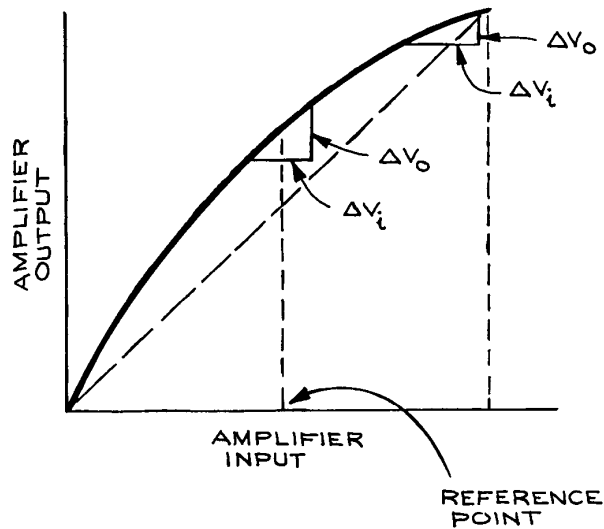


Figure 6.80 Amplifier linearity.

maximum vertical deviation of the real gain curve from the ideal straight-line gain curve, expressed as a percentage of maximum output. Such nonlinearities also occur in the circuits of the PHA and must be taken into consideration for high-resolution work.

The manufacturers of pulse-processing equipment include application notes and product guides with the catalog descriptions of their instruments. These are generally very helpful and should be read thoroughly before choosing a system and selecting components.

## 6.5 POWER SUPPLIES

The design of regulated power supplies follows a standard configuration (see Figure 6.81). Raw a.c. from the outlet is converted, usually with a transformer, to the desired level. This converted a.c. is rectified to produce unregulated d.c.. The unregulated d.c. then becomes the input to a regulator circuit, whose output is the desired d.c. output voltage.

### 6.5.1 Power-Supply Specifications

Power supplies are specified by the maximum current, maximum voltage, and maximum power they can

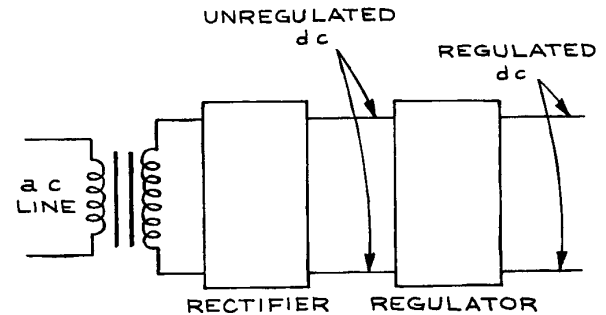


Figure 6.81 Block diagram of a regulated power supply.

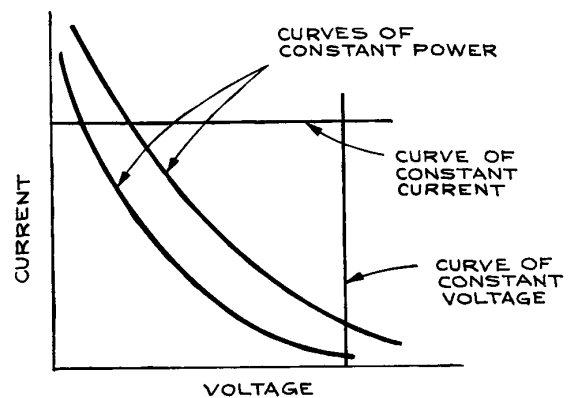


Figure 6.82 Current, voltage, and power relations for a power supply.

deliver. It is not always the case that maximum current can be delivered over the complete voltage range of the supply, nor can the maximum voltage be obtained at all current levels. This situation is illustrated in Figure 6.82. Maximum current can only be obtained over a limited range of voltage without exceeding the power rating of the supply.

The change in output voltage per unit change in input a.c. voltage is called the *line regulation*, while the change in output voltage per unit change in load or output current is called the *load regulation*. Both are generally specified as a percentage. Many power supplies can be operated in a *constant-current mode* where the output is a current

essentially independent of input and output voltage. Regulation specifications then apply to the output current. Load regulation can be translated into an equivalent dynamic-output resistance, since the relative change in load is the negative of the relative change in current ( $\Delta R/R = -\Delta I/I$ ):

$$\begin{aligned} \text{load regulation} &= \frac{\Delta V/V}{\Delta R/R} = -\frac{\Delta V/V}{\Delta I/I} \\ &= \frac{|\Delta V/\Delta I|}{V/I} = \frac{R_{\text{dynamic output}}}{R_{\text{load}}} \end{aligned} \quad (6.47)$$

As an example, if a filament power supply produces 3 A at 5 V, and the load regulation is 0.1%, then the dynamic output resistance is:

$$\begin{aligned} R_{\text{dynamic output}} &= \frac{1 \times 10^{-3} \times 5 \text{ V}}{3 \text{ A}} \\ &= 1.67 \times 10^{-3} \Omega \end{aligned} \quad (6.48)$$

*Transient response* and *recovery time* relate to the ability of the power supply to recover from sudden changes in load or line voltage at the stated operating point. If the line and load fluctuations are rapid, the power-supply regulating circuits may not be able to keep up with them and the resulting regulation will be considerably worse than specified. Generally, the better the regulation, the slower the response to changes of line and load conditions. The regulation may also depend on the output power level, a point worth noting in the specifications.

As well as maintaining a constant average d.c. level, instantaneous values of the output voltage should not deviate appreciably from the average. Deviations come from ripple at twice the line frequency for dissipative regulators, at the switching frequency for switching regulators, and from regulating circuit transients. Specification of the rms ripple gives an indication of the effectiveness of the filtering circuits. It gives no indication, however, of the presence of short-duration, large-amplitude voltage spikes at the output. A maximum peak-to-peak noise specification is used to describe this kind of instantaneous voltage deviation. As with regulation, ripple and noise may depend upon the power level.

Even with a constant line voltage and constant load, the output voltage can change if the temperature changes. This

is generally indicated by a *temperature-coefficient* specification, which gives the percentage change in output voltage per degree change in temperature, about a specified temperature. For laboratory applications, this is usually not a critical specification. When working in extreme environments, however, it may be significant (for compact units that depend on forced-air cooling or free-convection cooling, adequate space must be provided in the mounting to prevent overheating).

Power supplies that convert a d.c. input voltage to a different d.c. output voltage are particularly useful when d.c. rather than a.c. voltages are available. Voltage conversion with these supplies is accomplished by changing the d.c. input to a.c. at frequencies of 20 kHz and above, changing the amplitude of the high-frequency a.c. with a transformer, and rectifying and regulating the output of the transformer. The ratio of the maximum output power to maximum input power is the efficiency of conversion. d.c. to d.c. converters have efficiencies that exceed 70 to 80% and are available in a wide range of input and output voltages and powers. When high voltage at low current is required, as, for example, for a photomultiplier tube or electron multiplier, these units are particularly convenient.

Though often not included in the specification list, radio frequency (r.f.) noise can be an important consideration for supplies used in the vicinity of sensitive wide-band amplifiers and when trying to extract a weak signal from noise. Some types of power supplies produce more r.f. noise than others because of the design of the regulating circuit. Switching regulators with SCRs are particularly poor in this regard.

Other factors to consider in a power supply are *bipolar operation* (can both the negative and positive terminals be grounded to give positive and negative output voltages, respectively?) and insulation from ground if the supply is to be floated on another supply. Operation is simplified by front-panel meters indicating both output voltage and current, overload reset switches, and a convenient output terminal configuration and location.

A single output power supply can never be used to supply two voltages, one negative and one positive with respect to ground. When it is necessary to have such voltages as, for example, with operational-amplifiers, two separate supplies must be used.

## 6.5.2 Regulator Circuits and Programmable Power Supplies

A block diagram of a dissipative regulator circuit in a power supply is illustrated in Figure 6.83. The sensing resistors  $R_1$  and  $R_2$  are across the output of the power supply. A fraction  $R_2/(R_1 + R_2)$  of the output voltage is sent to a difference amplifier, where it is compared with a reference voltage, and the amplified difference used to control the *pass element*, a transistor in the common-collector configuration in this example. Depending on the output of the difference amplifier, the pass-element output voltage will adjust itself to a value just sufficient to sustain the required input from the difference amplifier. If, for some reason, the output voltage rises, the output of the difference amplifier will decrease and the pass-element output will decrease. The opposite occurs for a falling output. The series dissipative regulator is inefficient because the difference in voltage between the raw d.c. input and regulated d.c. output falls across the pass element, which must then dissipate power equal to the product of this voltage difference and the output current.

More efficient power supplies use circuits to control the *duty cycle* of the pass element. In these circuits, the pass element (a transistor, SCR, power MOSFET, IGBT, or other solid-state device) acts as a switch and is either *on*

(no voltage drop across it) or *off* (no current through it). The output voltage from such a regulator is:

$$V_{\text{out}} = V_{\text{in}} \frac{t_{\text{on}}}{t_{\text{on}} + t_{\text{off}}} \quad (6.49)$$

where  $t_{\text{on}}$  is the time for which the pass element is on and  $t_{\text{off}}$  is the time for which it is off in each cycle. The optimum switching frequency for switching regulators is between 20 and 100 kHz. At these high frequencies, r.f.i. is present and can cause problems with sensitive electronic circuits.

In practice, regulator circuits can be very complex, with high-gain difference amplifiers or comparator preregulators, multiple pass elements, overload and overvoltage protection, and temperature compensation. The above general power-supply description, however, is sufficient for an understanding of remote programmable power supplies. If  $R_1$  and  $R_2$  in Figure 6.83 are replaced by a potentiometer, a change in its setting will result in a change in the regulated output voltage. This is what occurs when one changes the dial settings on the front panel of a variable power supply. The internal potentiometer can clearly be replaced by an external potentiometer (*resistance programming*) – or even more directly – by an external voltage source (*voltage programming*). The addition of such features requires very little modification to the basic power-supply

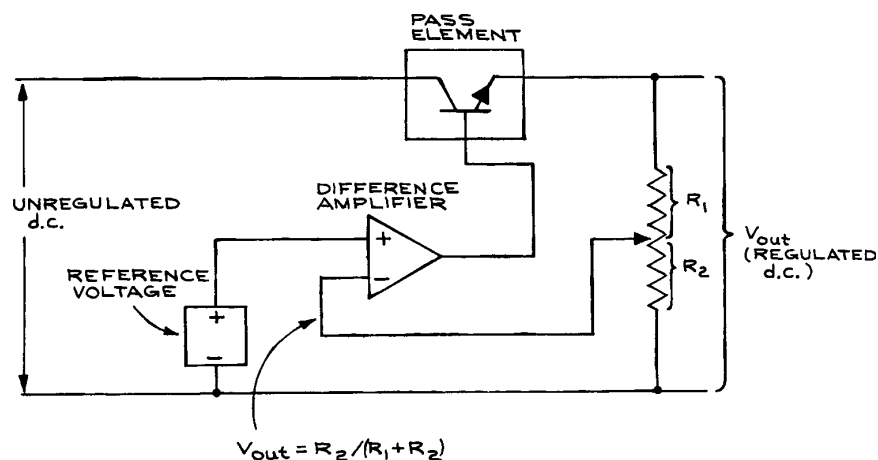


Figure 6.83 Block diagram of the regulator circuit in a power supply.

regulator circuit, and they are often available at no cost or as low-cost options. Remote programming can be particularly useful in control operations, where a digital code from a computer is changed to a voltage by a digital-to-analog-converter (DAC), and this voltage is used to set the output of a power supply.

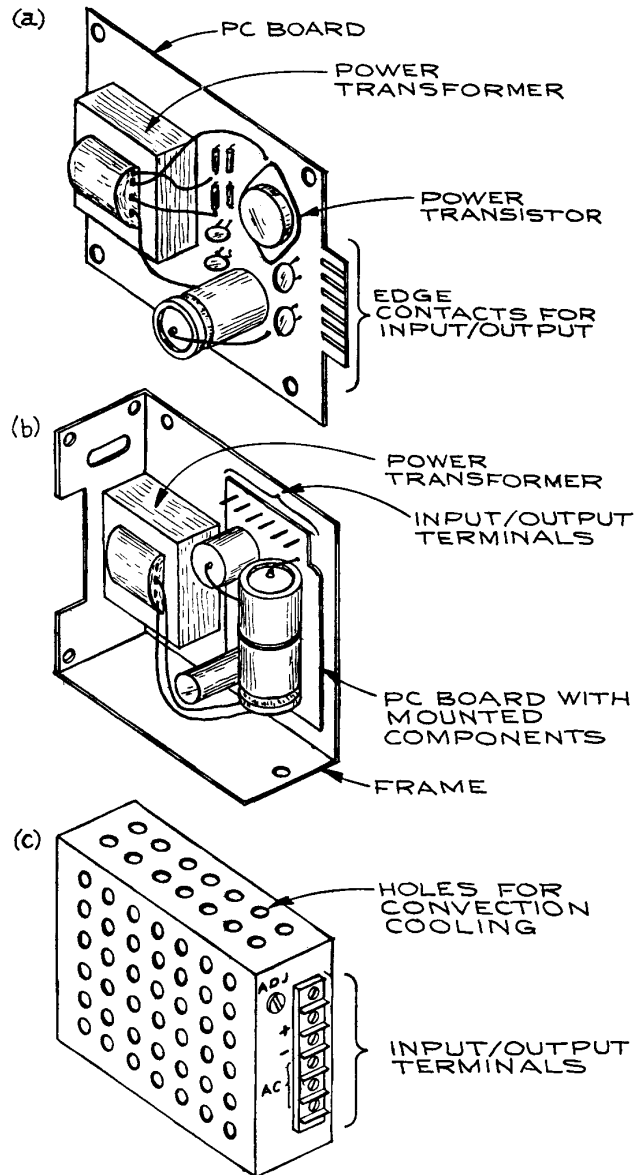
*Overvoltage protection* is an important feature in a power supply. With such protection, one is assured that no failure will result in a high voltage across the output terminals of the supply. *Crowbar* circuits using SCR switches are often used for overvoltage protection. Such circuits short the output terminals of the supply when the output voltage exceeds a preset value. The resulting short-circuit current causes a fuse to blow or a circuit breaker to open, shutting down the power supply until the fault that caused the overvoltage is corrected. It is essential that the protection circuit respond quickly to overload conditions to avoid damage to the power supply. Response speed should be included in the specifications.

Another built-in safety feature found on power supplies is *foldback current limiting*. When the output current exceeds a specified limit, it is automatically reduced by such circuits to a low level and maintained there for a specified time before being reset. Current-limiting circuits are triggered by temperature-sensitive or power-dissipating elements.

In addition to variable output power supplies there are a large number of fixed-voltage power supplies on the market. When working with logic circuits and operational amplifier circuits, only fixed power-supply voltages are required.

Most fixed-voltage supplies incorporated in other units are designed to operate from the a.c. line, although there are a number of d.c.-to-d.c. converters. The least expensive configuration is the card power supply [Figure 6.84(a)]. All components are on a single printed circuit board with edge contacts. The a.c. input must be brought to one set of contacts and the d.c. output taken from another. For routine use, the card supply should be mounted in a case or on a rack panel with an on-off switch, a pilot light, a fuse, and output terminals. This requires additional time and expense, which may cancel the initial cost savings from purchasing such a supply.

Frame-mounted supplies [see Figure 6.84(b)] are similar to the card supplies in that all external connections, switches, and output terminals are missing. The metal



**Figure 6.84** Fixed-voltage power supplies; (a) card supply; (b) frame supply; (c) enclosed case.

frame facilitates mounting, however, and affords some protection to the components. Rather than PCB contacts, frame supplies usually have screw-type barrier connectors.

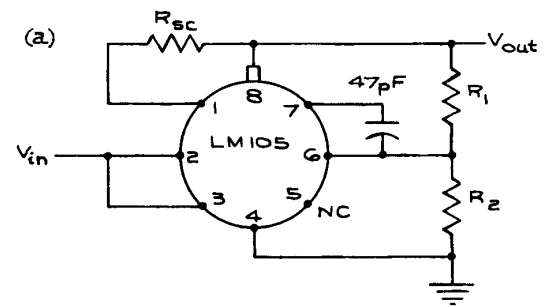
Fixed-voltage supplies also come in enclosed cases and in encapsulated units suitable for direct connection to a printed circuit board [see Figure 6.84(c)]. Many of the so-called fixed-voltage supplies can be adjusted over a 5–10 % range about the nominal output voltage value. Multiple fixed-voltage units with two and even three voltages are also common. It is often useful to have  $\pm 15$  V and +5 V available from a single power supply when both linear and digital circuits are used together.

Manufacturers produce a wide range of supplies with different voltages, power ratings, and mechanical configurations. It is therefore rarely worthwhile to construct a power supply. A possible exception lies in the use of integrated-circuit modules. There are two general types: fixed and variable output. The fixed-output units are three-terminal devices – unregulated d.c. is applied across the input and common terminals, and regulated d.c. at the appropriate voltage appears between the output and common terminals. A representative unit of this type is the LM109 5 V, 1.5 A monolithic regulator. The regulator combines on-chip thermal shutdown with current limiting, an on-chip series pass transistor, and a stable internal voltage reference. Further details about the LM109 and its use in other applications are available from the manufacturer.<sup>6</sup>

The LM105 is an example of a variable-voltage integrated-circuit regulator. For this unit, the output is set by the ratio of two external resistances. The maximum output voltage is 40 V, and the maximum output current is 20 mA. When an external pass transistor is added to the basic regulator circuit, the load current can be increased to 500 mA. The LM105 can also be used as a switching regulator. Data and application sheets should be consulted when using such regulators.

Useful power-supply circuits with integrated regulators are illustrated in Figure 6.85. Fixed-voltage units are made by several manufacturers at several fixed voltages from 4 to 20 V. Variable units are available for both positive and negative voltages.

Though not strictly power supplies, integrated-circuit voltage references provide precise output voltages for d.c. input voltages that can span a relatively wide range. Output voltages are from 1 to 10 V with initial accuracies from 0.01 to 1.2%. Change in output voltage with temperature for these units varies between 1 and 50 parts per

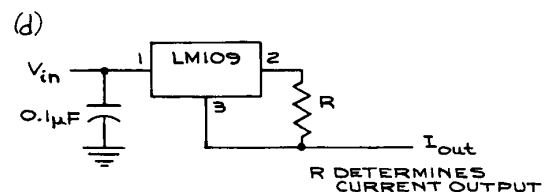
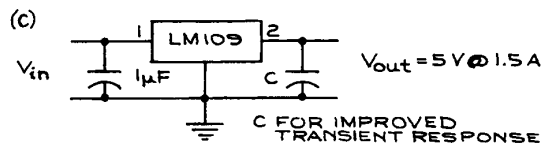
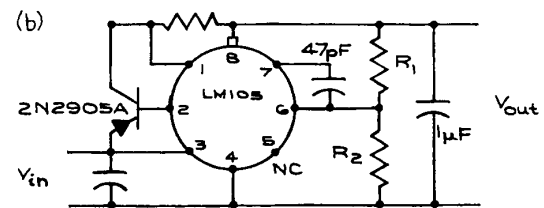


$$\frac{R_2}{R_1 + R_2} V_{out} = 1.8$$

$$\frac{R_1 R_2}{R_1 + R_2} = 2 \text{ k}\Omega$$

$$R_{sc} = \frac{325}{I_{sc}}$$

- 1 CURRENT LIMIT
- 2 BOOSTER OUTPUT
- 3 UNREGULATED INPUT
- 4 GROUND
- 5 REFERENCE BYPASS
- 6 FEEDBACK
- 7 COMPENSATION SHUTDOWN
- 8 REGULATED OUTPUT



**Figure 6.85** Examples of circuits with the LM109 fixed-voltage regulator and the LM105 variable-voltage regulator.

million per degree Celsius (ppm/°C). Output noise is exceptionally low, varying from 1 to 50  $\mu\text{V}$  peak-to-peak ( $\mu\text{V}$  p to p) over a bandwidth of 0.1 to 10 Hz. Output currents are low, 5 to 100 mA.

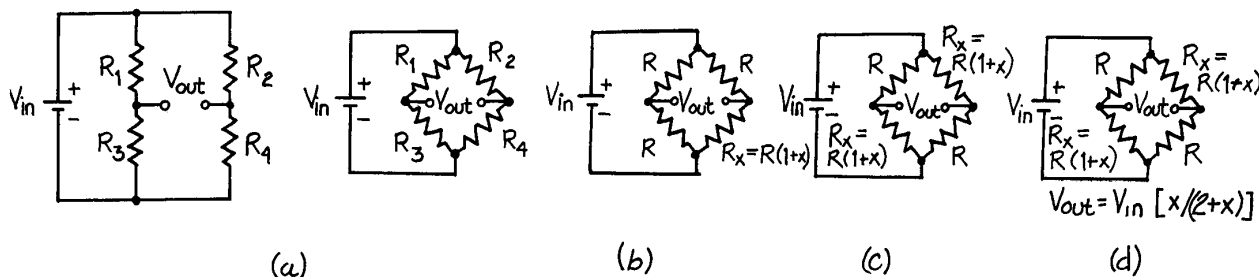
### 6.5.3 Bridges

Bridges are used to measure the electrical properties of a circuit element by comparison with a similar element. The Wheatstone bridge, developed in the early nineteenth century, compares the potentials from two voltage dividers connected in parallel across a voltage source, as shown in Figure 6.86(a).

Taking the difference between the potentials at the junction of  $R_1$  and  $R_2$  and  $R_3$  and  $R_4$ ,  $V_{\text{out}} = V_{\text{in}}[R_3/R_1 + R_2 - R_4/R_2 + R_4]$ . When  $V_{\text{out}} = 0$ ,  $R_1/R_2 = R_3/R_4$ , the null condition. In the standard configuration,  $R_1$  and  $R_2$  are fixed,  $R_4$  is the resistance to be measured, and  $R_3$  is varied until a null or balance condition is achieved. At null the relation between the resistances is independent of  $V_{\text{in}}$ . Null-type measurements require a means for varying one element of the bridge, and this is usually accomplished with a system that incorporates feedback.

Bridges in most transducer applications are operated *off-null*, or out of balance. For the simple bridge in Figure 6.86(b), three of the resistors are fixed at value  $R$  while the one to be measured,  $R_x$ , deviates from  $R$  by  $xR$  – where  $x$  is a unitless variable. The output voltage  $V_{\text{out}}$  is given by:

$$V_{\text{out}} = V_{\text{in}} \left[ \frac{R}{2R} - \frac{R_x}{R + R_x} \right] = -\frac{V_{\text{in}}}{4} \left[ \frac{x}{1 + x/2} \right] \quad (6.50)$$



**Figure 6.86** (a) Wheatstone bridge; (b) off-null bridge with one variable element; (b) off-null bridge with two variable elements; (c) off-null bridge with two compensating variable elements.

For  $x/2 \ll 1$ ,  $V_{\text{out}} \approx -V_{\text{in}}(x/4)$ . As an example, let  $x = 0.01$ , then  $V_{\text{out}} = -V_{\text{in}}(0.01/4)$ , so that a 1% change in  $R_x$  gives a 0.25% change in  $V_{\text{out}}$ . The sensitivity of the bridge, defined as the change in output voltage for a given change in input voltage, is 2.5 mV/V. Bridge sensitivity can be doubled by adding a second identical variable element opposite the first, as shown in Figure 6.86(c). While the output of such a configuration is doubled, the nonlinearity remains. The addition of two more variable elements,  $R_C = R(1 - x)$ , with resistances that change in opposite ways to the original elements, results in a bridge with output equal to the fractional change in resistance times the reference voltage. This is shown in Figure 6.86(d). Bridges of this type are used with two-element strain gauges and piezoresistive transducers. The transducers are arranged so that there are complementary changes in resistance for sensors in adjacent bridge arms.

Another way to linearize the output of a bridge circuit is by inserting an operational amplifier to null the bridge by adding a voltage in series with the variable element  $R_x$  that is equal in magnitude and opposite in sign to the incremental voltage across  $R_x$ . This gives an output that is linear in  $x$  for large values of  $x$ .

In bridge circuits that are operated out of balance, the output is proportional to the reference voltage  $V_{\text{in}}$ . A stable reference voltage can be derived from an IC reference source followed by an operational amplifier stage. Analog Devices and Maxim Integrated Products manufacture signal-conditioning power supplies that have programmable outputs and include amplification and filtering. Because neither side of the bridge from which the voltage is

measured is at ground, instrumentation amplifiers are the preferred devices for reading bridge outputs. These amplifiers have the necessary differential inputs, high common-mode rejection ratios, high input impedances, and resistance programmable gain.

## 6.6 DIGITAL ELECTRONICS

Digital systems are based on circuit elements (usually transistors) operated in such a way that they exist in only one of two states. This is in contrast to analog systems, where the outputs are continuous functions of the input variables. Combinations of two-state, or *binary*, devices can perform arithmetic and logic operations of arbitrary degrees of complexity.

### 6.6.1 Binary Counting

With a binary system, it is usual to call the states 0 and 1. Combinations of 0s and 1s can then be used for counting. Some schemes are given in Table 6.26. Note that the *least significant bit* (LSB) or column is  $2^0$ , the next  $2^1$ , and the last or *most significant bit* (MSB) is  $2^4$ . In other words, the

Decimal	Binary	Octal	Hexadecimal
0	0000	00	0
1	0001	01	1
2	0010	02	2
3	0011	03	3
4	0100	04	4
5	0101	05	5
6	0110	06	6
7	0111	07	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F

binary number system is merely a base 2 system, while the decimal system is base 10. Other systems derived from the binary system are the octal (base 8) and the hexadecimal (base 16), as detailed in Table 6.26. In the binary system, the number 2 does not appear, and in the octal system 8 does not appear. The hexadecimal system substitutes the letters A, B, C, D, E, and F for the decimal numbers 10, 11, 12, 13, 14, and 15, for which a single symbol does not exist.

The state of the output of a logic device can be represented by a voltage level or the presence or absence of a signal pulse within a given time window. A voltage-level system in which the most positive voltage level corresponds to logical 1 is a *positive-logic* system. One in which the least positive level corresponds to logical 1 is a *negative-logic* system.

### 6.6.2 Elementary Functions

The basic logic gates, along with the corresponding relationships between input and output variables, are given in Table 6.27 in *truth-table* form along with the schematic symbols.

### 6.6.3 Boolean Algebra

The rules for manipulating logic expressions based on binary logic were formulated in the nineteenth century by George Boole. The laws and identities are given in Table 6.28. Boolean algebra is useful for reducing complex truth tables to the fewest possible gates. Consider, for example, the three-variable ( $A, B, C$ ), two-function ( $X, Y$ ) truth table at the top of Figure 6.87. The Boolean algebra expressions for  $X$  and  $Y$  are obtained by identifying the states of each of the variables ( $A, B, C$ ) for which the function under consideration ( $X$  or  $Y$ ) is in the 1 state. For example,  $X = 1$  when  $A = 0, B = 0$ , and  $C = 1$ . By convention, a bar over a variable symbol indicates negation, so that if  $A = 0, \bar{A} = 1$ . The coincidence of 1s at  $\bar{A}, \bar{B}$ , and  $C$  to give  $X = 1$  requires an AND operation, which is the first term in the full expression for  $X$  in Figure 6.87. The other terms are derived from the truth table in the same way, and they are all connected together by OR operations. After simplification by application of the laws in Table 6.28, the expressions can be implemented with elementary gates. The gates for the function  $X$  are shown at the bottom of Figure 6.87.

**Table 6.27 Logic gates**

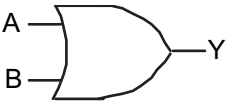
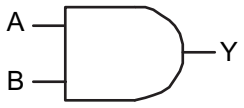
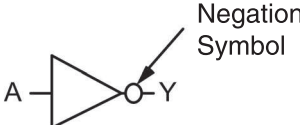
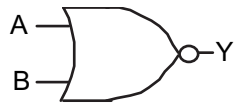
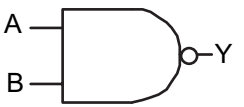
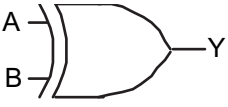
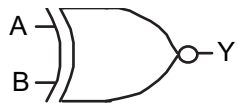
<i>Name</i>	<i>Symbol</i>	<i>Truth Table</i>	<i>Boolean Expression</i>															
OR		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td></tr> <tr><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>1</td></tr> </table>	A	B	Y	0	0	0	0	1	1	1	0	1	1	1	1	$Y = A + B$
A	B	Y																
0	0	0																
0	1	1																
1	0	1																
1	1	1																
AND (coincidence)		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>1</td></tr> </table>	A	B	Y	0	0	0	0	1	0	1	0	0	1	1	1	$Y = A \cdot B$
A	B	Y																
0	0	0																
0	1	0																
1	0	0																
1	1	1																
NOT (inverter)		<table border="1"> <tr><td>A</td><td>Y</td></tr> <tr><td>0</td><td>1</td></tr> <tr><td>1</td><td>0</td></tr> </table>	A	Y	0	1	1	0	$Y = \bar{A}$									
A	Y																	
0	1																	
1	0																	
NOR		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td></tr> </table>	A	B	Y	0	0	1	0	1	0	1	0	0	1	1	0	$Y = \overline{A + B}$
A	B	Y																
0	0	1																
0	1	0																
1	0	0																
1	1	0																
NAND		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>1</td><td>1</td></tr> <tr><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>0</td></tr> </table>	A	B	Y	0	0	1	0	1	1	1	0	1	1	1	0	$Y = \overline{A \cdot B}$
A	B	Y																
0	0	1																
0	1	1																
1	0	1																
1	1	0																
XOR (exclusive OR, nticoincidence)		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td></tr> <tr><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>0</td></tr> </table>	A	B	Y	0	0	0	0	1	1	1	0	1	1	1	0	$Y = (B \cdot \bar{A}) + (\bar{B} \cdot A)$
A	B	Y																
0	0	0																
0	1	1																
1	0	1																
1	1	0																
XNOR (exclusive, NOR, equivalence)		<table border="1"> <tr><td>A</td><td>B</td><td>Y</td></tr> <tr><td>0</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>1</td></tr> </table>	A	B	Y	0	0	1	0	1	0	1	0	0	1	1	1	$Y = \bar{A} \cdot \bar{B} + AB$
A	B	Y																
0	0	1																
0	1	0																
1	0	0																
1	1	1																



Table 6.28 Boolean algebra	
<i>Laws</i>	<i>Identities</i>
Commutative	$A + A = A$
$A + B = B + A$	$A \cdot A = A$
$A \cdot B = B \cdot A$	$A + 1 = 1$
	$A \cdot 1 = A$
Distributive	$A + 0 = A$
$A \cdot (B + C) = A \cdot B + A \cdot C$	$A \cdot 0 = 0$
De Morgan	
$\overline{A \cdot B \cdot C} = \overline{A} + \overline{B} + \overline{C}$	
$\overline{A + B + C} = \overline{A} \cdot \overline{B} \cdot \overline{C}$	

Another way of simplifying truth tables is by making a logic map called a Karnaugh map. The map for the truth table of Figure 6.87 is shown in Figure 6.88. The squares corresponding to the variables for which X and Y are 1 are shaded. Contiguous shaded squares corresponding to 0 and 1 for a single variable, say C, with the other variables constant, show that the value of the function is independent of the state of C. In the map for Y, it can be seen that the squares for  $\overline{A} \cdot B$  are shaded for C and  $\overline{C}$ . It therefore follows that the state of Y is not dependent on the variable C for these cases.

### 6.6.4 Arithmetic Units

After elementary gates, the next simplest logic units are those that perform elementary binary arithmetic operations. Some common arithmetic units are given in Table 6.29. The truth table for the *half adder* follows the rules for binary addition with C the low-order bit of the sum and D the high-order bit. The *full adder* is an elaboration of the half adder with an extra input. When adding two binary numbers of more than one bit each, provision must be made for the carry operation. This is the function of the  $C_{n-1}$  terminal, that accepts the carry bit from the result of adding the preceding lower-order bits in the string (thus the designation  $C_{n-1}$ ). The results from the addition of the n-order bits appear at terminals  $C_n$  and  $S_n$ . The *digital comparator* compares the magnitude of the digital signals at input  $A_n$  and  $B_n$  and activates output  $C_n$ ,  $D_n$ , or  $E_n$  depending on whether  $A_n > B_n$ ,  $A_n < B_n$ , or  $A_n = B_n$ . Such elements

A	B	C	X	Y
0	0	0	0	1
0	0	1	1	0
0	1	0	1	1
0	1	1	0	1
1	0	0	1	0
1	0	1	0	0
1	1	0	0	1
1	1	1	1	0

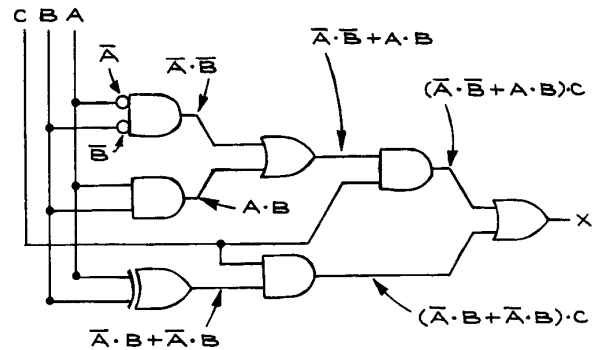
$$X = (\overline{A} \cdot \overline{B} \cdot C) + (\overline{A} \cdot B \cdot \overline{C}) + (A \cdot \overline{B} \cdot \overline{C}) + (A \cdot B \cdot C)$$

$$Y = (\overline{A} \cdot \overline{B} \cdot \overline{C}) + (\overline{A} \cdot B \cdot \overline{C}) + (\overline{A} \cdot B \cdot C) + (A \cdot B \cdot \overline{C})$$

WHICH SIMPLIFY TO:

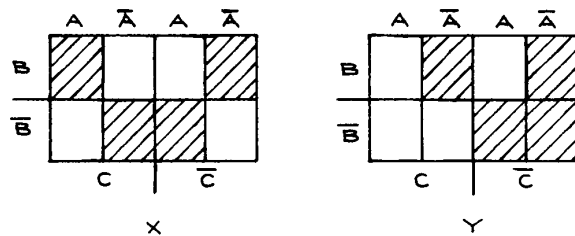
$$X = (\overline{A} \cdot \overline{B} + A \cdot B) \cdot C + (\overline{A} \cdot B + A \cdot \overline{B}) \cdot \overline{C}$$

$$Y = \overline{A} \cdot B + (\overline{A} \cdot \overline{B} + A \cdot B) \cdot \overline{C}$$



**Figure 6.87** Example of a truth table, the equivalent Boolean algebraic expression, and implementation with elementary gates.

can be connected together to compare two n-bit binary numbers. The parity of a binary number is defined as odd (0) if there is an even number of 1s and even (1) if there is an odd number of 1s. The parity code is often used to check for transmission errors in bit strings. By sending a parity bit with an n-bit word and testing the received word against the parity bit, single bit errors can be detected. This is illustrated for a four-bit word with the *parity generator/checker* in Table 6.29.



**Figure 6.88** Karnaugh map for the truth table of Figure 6.87.

### 6.6.5 Data Units

For data acquisition and transmission, *decoders*, *encoders*, *multiplexers*, and *demultiplexers* are commonly used. The *read-only memory* (ROM) is a combination of decoder and encoder and is frequently used to implement complex truth tables. *Programmable* and *erasable* ROMs, called PROMs and EPROMs, can have their internal logic modified for the particular application of the user. Table 6.30 gives the truth tables and input-output terminal arrangements for some simple data units.

### 6.6.6 Dynamic Systems

The logic units described so far are considered to be *static* or *combinatorial*, because the output levels depend only on the current input levels. The previous history of the signals at the different inputs is irrelevant. With *dynamic* or *sequential systems*, output signal levels depend on the history of the signal levels at the input terminals. The truth tables for dynamic systems must specify the previous states of the inputs and outputs in order for the state of a particular output to be fully determined. Subscripts are used to differentiate the time sequence of the output states. Table 6.31 lists a number of *flip-flops* that are the basic building blocks in dynamic systems. Flip-flops can be strung together to make *shift registers* and *counters*. A shift register is a series of *J/K flip-flops* with the  $Q$  and  $\bar{Q}$  outputs of each stage connected to the  $J$  and  $K$  inputs of the next stage (see Figure 6.89). Data are entered serially in the first unit, connected as a *D flip-flop*, and move from unit to unit at each clock transition. Shift registers can be used as counters, serial-to-parallel data converters using the paral-

lel outputs, and parallel-to-serial data converters using the parallel inputs.

The basic serial or *ripple* binary counter is shown in Figure 6.90. Three *J/K* flip-flops can count to  $2^3$  in this arrangement. In the shift-register configuration, eight flip-flops are needed to count to  $2^3$ .

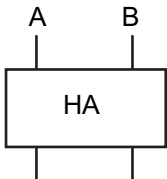
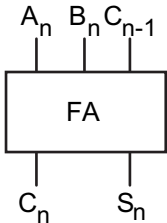
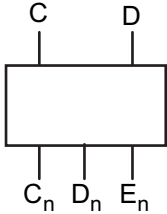
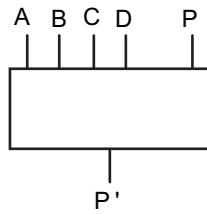
**Modulo- $N$  counters with  $N \neq 2$ .** These counters can be made by decoding the counter outputs with gates and resetting the flip-flops after  $N$  counts. Shift registers and counters are manufactured in a variety of configurations in single 14- and 16-pin DIP and SMT packages.

### 6.6.7 Digital-to-Analog Conversion

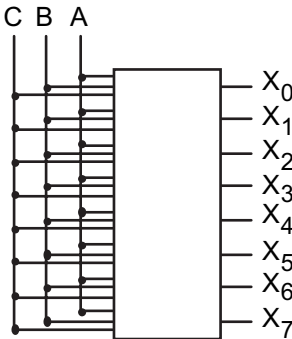
Whenever it is necessary to connect a digital system to an external device that operates in an analog mode, converters are needed – either *digital-to-analog* (DAC) or *analog-to-digital* (ADC). Of the two, the DAC is the more fundamental, since many ADCs use DACs in their construction.

A DAC produces an output current or voltage proportional to the magnitude of the binary number at its input terminals (see Figure 6.91). In one design, the binary inputs to a DAC activate FET analog switches that connect the properly weighted current or voltage from a reference source to the output. In the case of voltage DACs, an operational amplifier in the summing mode adds the weighted input voltages. Electrical specifications to consider with a DAC are *resolution*, the number of input bits or steps (an 8-bit DAC has 256 steps); *linearity*, the maximum deviation from the best straight line drawn through the graph of output versus digital input; and *accuracy*, the deviation of the output from that computed on the basis of the digital input. Linearity primarily depends on the resistors in the DAC circuit and the voltage drops across the internal transistor switches. For these reasons, linearity is temperature dependent. Accuracy depends on the factors affecting linearity and also on the reference voltage, either internal or external. In applications where speed of conversion is important, *settling time* must be considered. This is the time necessary for the analog output to stabilize to a level that is the equivalent of  $\pm 1/2$  of the least significant bit of the binary input. Other considerations are power-supply sensitivity and compatibility with input and output logic levels. If a

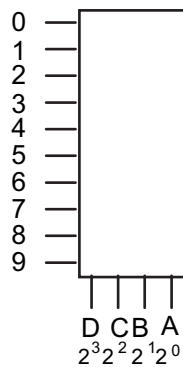
**Table 6.29 Arithmetic Units**

Name	Symbol	Truth Table																																																																																																												
Half adder		<table border="1"> <thead> <tr> <th>A</th> <th>B</th> <th>Sum</th> <th>C</th> <th>D</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>00</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>01</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>0</td><td>01</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>10</td><td>1</td><td>0</td></tr> </tbody> </table>	A	B	Sum	C	D	0	0	00	0	0	0	1	01	0	1	1	0	01	0	1	1	1	10	1	0																																																																																			
A	B	Sum	C	D																																																																																																										
0	0	00	0	0																																																																																																										
0	1	01	0	1																																																																																																										
1	0	01	0	1																																																																																																										
1	1	10	1	0																																																																																																										
Full adder		<table border="1"> <thead> <tr> <th><math>A_n</math></th> <th><math>B_n</math></th> <th><math>C_{n-1}</math></th> <th><math>S_n</math> (sum)</th> <th><math>C_n</math> (carry)</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> </tbody> </table>	$A_n$	$B_n$	$C_{n-1}$	$S_n$ (sum)	$C_n$ (carry)	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	0	1	1	0	1	1	0	1	0	1	1	1	1	1	1																																																															
$A_n$	$B_n$	$C_{n-1}$	$S_n$ (sum)	$C_n$ (carry)																																																																																																										
0	0	0	0	0																																																																																																										
0	1	0	1	0																																																																																																										
1	0	0	1	0																																																																																																										
1	1	0	0	1																																																																																																										
0	0	1	1	0																																																																																																										
0	1	1	0	1																																																																																																										
1	0	1	0	1																																																																																																										
1	1	1	1	1																																																																																																										
Digital comparator		<table border="1"> <thead> <tr> <th><math>A_n</math></th> <th><math>B_n</math></th> <th><math>C_n (A_n &gt; B_n)</math></th> <th><math>D_n (B_n &gt; A_n)</math></th> <th><math>E_n (A_n = B_n)</math></th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td></tr> </tbody> </table>	$A_n$	$B_n$	$C_n (A_n > B_n)$	$D_n (B_n > A_n)$	$E_n (A_n = B_n)$	0	0	0	0	1	0	1	0	1	0	1	0	1	0	0	1	1	0	0	1																																																																																			
$A_n$	$B_n$	$C_n (A_n > B_n)$	$D_n (B_n > A_n)$	$E_n (A_n = B_n)$																																																																																																										
0	0	0	0	1																																																																																																										
0	1	0	1	0																																																																																																										
1	0	1	0	0																																																																																																										
1	1	0	0	1																																																																																																										
Parity generator/checker		<table border="1"> <thead> <tr> <th colspan="4">4-bit word</th> <th colspan="2">Parity bit</th> </tr> <tr> <th>A</th> <th>B</th> <th>C</th> <th>D</th> <th>P</th> <th>P'</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>1</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td></tr> </tbody> </table>	4-bit word				Parity bit		A	B	C	D	P	P'	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	1	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	1	1	1	0	1	1	0	0	0	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1	1	0	0
4-bit word				Parity bit																																																																																																										
A	B	C	D	P	P'																																																																																																									
0	0	0	0	0	0																																																																																																									
0	0	0	1	1	0																																																																																																									
0	0	1	0	1	0																																																																																																									
0	0	1	1	0	0																																																																																																									
0	1	0	0	1	0																																																																																																									
0	1	0	1	0	0																																																																																																									
0	1	1	0	0	0																																																																																																									
0	1	1	1	1	0																																																																																																									
1	0	0	0	1	0																																																																																																									
1	0	0	1	0	0																																																																																																									
1	0	1	0	0	0																																																																																																									
1	0	1	1	1	0																																																																																																									
1	1	0	0	0	0																																																																																																									
1	1	0	1	1	0																																																																																																									
1	1	1	0	1	0																																																																																																									
1	1	1	1	0	0																																																																																																									

**Table 6.30 Data units**

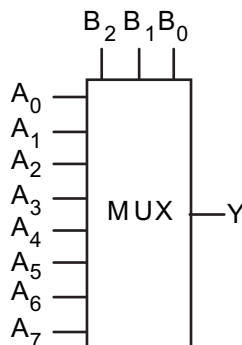
Unit	Symbol	Truth Table																																																																																																			
<p><i>Decoder:</i> Activates one of <math>2^n</math> outputs according to an <math>n</math>-bit code.</p>		<table border="1"> <thead> <tr> <th>C</th> <th>B</th> <th>A</th> <th><math>X_0</math></th> <th><math>X_1</math></th> <th><math>X_2</math></th> <th><math>X_3</math></th> <th><math>X_4</math></th> <th><math>X_5</math></th> <th><math>X_6</math></th> <th><math>X_7</math></th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> </tbody> </table>	C	B	A	$X_0$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1
		C	B	A	$X_0$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$																																																																																									
		0	0	0	1	0	0	0	0	0	0	0																																																																																									
		0	0	1	0	1	0	0	0	0	0	0																																																																																									
		0	1	0	0	0	1	0	0	0	0	0																																																																																									
		0	1	1	0	0	0	1	0	0	0	0																																																																																									
		1	0	0	0	0	0	0	1	0	0	0																																																																																									
		1	0	1	0	0	0	0	0	1	0	0																																																																																									
1	1	0	0	0	0	0	0	0	1	0																																																																																											
1	1	1	0	0	0	0	0	0	0	1																																																																																											

*Encoder:*



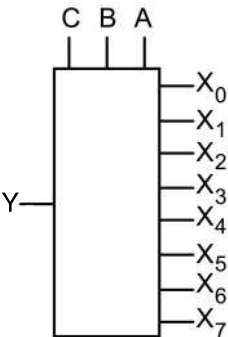
Input	D	C	B	A
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1

*Multiplexer:* According to an  $n$ -bit code, one of  $2^n$  signal input lines is connected to a single output line.

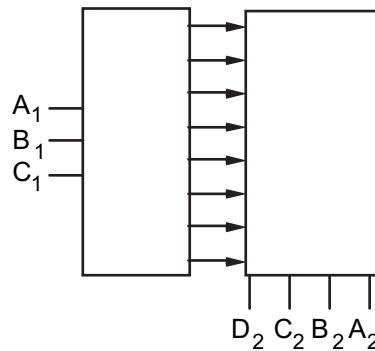


8-to-1 line multiplexer			
$B_2$	$B_1$	$B_0$	Y
0	0	0	$A_0$
0	0	1	$A_1$
0	1	0	$A_2$
0	1	1	$A_3$
1	0	0	$A_4$
1	0	1	$A_5$
1	1	0	$A_6$
1	1	1	$A_7$

Table 6.30. (contd.)

Unit	Symbol	Truth Table																																				
<p><i>Demultiplexer:</i> According to an <math>n</math>-bit code, a single input line is routed to one of <math>2^n</math> output lines.</p>		<table border="1"> <thead> <tr> <th>C</th> <th>B</th> <th>A</th> <th>Y Connected to</th> </tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>0</td><td><math>X_0</math></td></tr> <tr><td>0</td><td>0</td><td>1</td><td><math>X_1</math></td></tr> <tr><td>0</td><td>1</td><td>0</td><td><math>X_2</math></td></tr> <tr><td>0</td><td>1</td><td>1</td><td><math>X_3</math></td></tr> <tr><td>1</td><td>0</td><td>0</td><td><math>X_4</math></td></tr> <tr><td>1</td><td>0</td><td>1</td><td><math>X_5</math></td></tr> <tr><td>1</td><td>1</td><td>0</td><td><math>X_6</math></td></tr> <tr><td>1</td><td>1</td><td>1</td><td><math>X_7</math></td></tr> </tbody> </table>	C	B	A	Y Connected to	0	0	0	$X_0$	0	0	1	$X_1$	0	1	0	$X_2$	0	1	1	$X_3$	1	0	0	$X_4$	1	0	1	$X_5$	1	1	0	$X_6$	1	1	1	$X_7$
		C	B	A	Y Connected to																																	
		0	0	0	$X_0$																																	
		0	0	1	$X_1$																																	
		0	1	0	$X_2$																																	
		0	1	1	$X_3$																																	
		1	0	0	$X_4$																																	
		1	0	1	$X_5$																																	
1	1	0	$X_6$																																			
1	1	1	$X_7$																																			

*Read-only memory (ROM):* Combination of a decoder and encoder to convert an  $n$ -bit code to an  $m$ -bit code, where  $n$  and  $m$  are not necessarily equal.



current DAC is used in an application where the output must be a voltage, a current-to-voltage converter is needed. For low-voltage applications, this can be a resistor. When higher voltages are required, the operational-amplifier circuits shown in Figure 6.92 can be used. The characteristics of the amplifier should not degrade the expected performance of the DAC with respect to linearity, accuracy, and settling time.

A variation of the DAC is the *multiplying DAC*. Since the output of a DAC is directly proportional to the binary input and reference voltage, varying the reference voltage has the effect of multiplying the output by a constant.

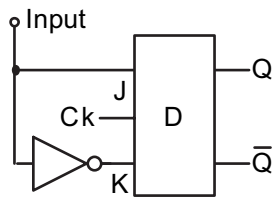
Digitally modulated waveforms can be easily produced with such devices.

There is a larger variety of ADCs than DACs because the complexity of analog-to-digital conversion lends itself to a variety of techniques. Three ADC schemes are shown in Figure 6.93. With the *counter ADC* [see Figure 6.93(a)], pulses from an oscillator or *clock* are routed to a counter through control logic, which is activated by the output of a comparator. The counter output goes to a DAC, the output of which is compared to the input signal. When the DAC output equals the input voltage, the comparator changes state and disconnects the clock from the counter. The

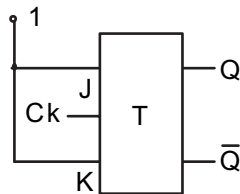
**Table 6.31 Flip-flops**

Unit	Symbol	Truth Table			
<p><i>R/S (reset-set) flip-flop:</i> Basic digital memory unit.</p>		<i>S</i>	<i>R</i>	<i>Q</i>	$\overline{Q}$
		0	1	0	1
		1	0	1	0
		0	0	Indeterminate	
		1	1	Indeterminate	
<p><i>J/K flip-flop:</i> An elaboration of the R/S avoiding the indeterminate 1, 1 and 0, 0 states, with separate preset (<i>Pr</i>) and clear (<i>Cr</i>) inputs, and a clock (<i>Ck</i>) input that must be activated for the outputs to follow input conditions.</p>		Upon the application of a <i>Ck</i> pulse: <sup>a</sup>			
		<i>J</i>	<i>K</i>	<i>Q</i> <sub><i>n</i>+1</sub>	$\overline{Q}$ <sub><i>n</i>+1</sub>
		0	0	<i>Q</i> <sub><i>n</i></sub>	$\overline{Q}$ <sub><i>n</i></sub>
		0	1	0	1
		1	0	1	0
		1	1	$\overline{Q}$ <sub><i>n</i></sub>	<i>Q</i> <sub><i>n</i></sub>
		Direct inputs:		<i>Q</i>	$\overline{Q}$
		<i>P<sub>r</sub></i>	<i>C<sub>r</sub></i>	Disallowed	
		0	0	1	0
		1	0	0	1
1	1	Normal clocked Operation			

*D (delay) flip-flop:* Input state appears at the output after a clock pulse. Constructed from a *J/K* flip-flop with an inverter from *J* to *K*, signal applied to *J*.



*T (toggle) flip-flop:* Output changes state at each clock transition. Used as a binary divider. Constructed from a *J/K* flip-flop by fixing *J* and *K* at logic 1.



<sup>a</sup> The subscript *n* + 1 designates the (*n* + 1)th state as distinct from the previous *n*th state. When *R* and *S* are 0, the output of *Q* remains constant, equal to its previous value.

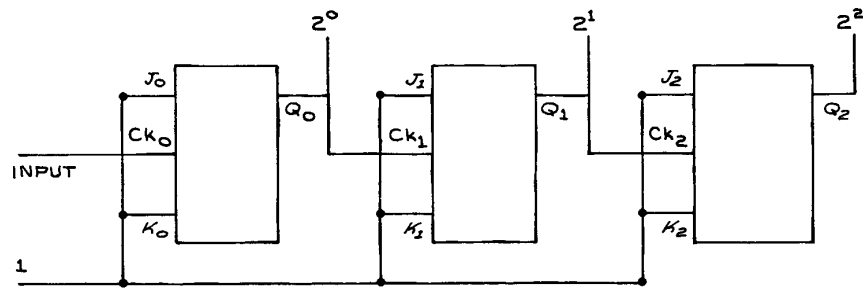


Figure 6.89 J/K flip-flops arranged as a 3-bit shift register.

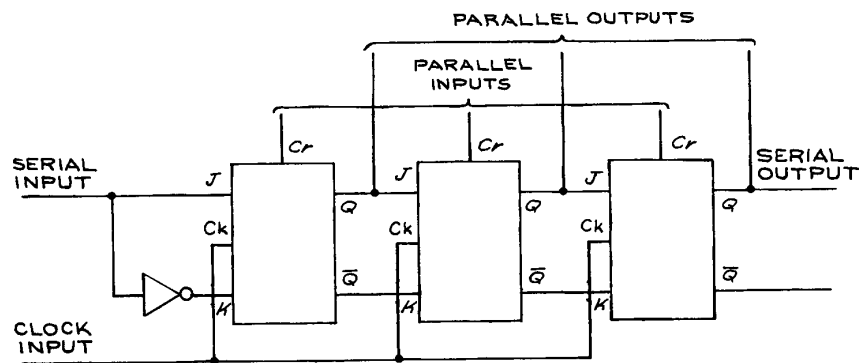


Figure 6.90 J/K flip-flops arranged as an 8-bit binary ripple counter.

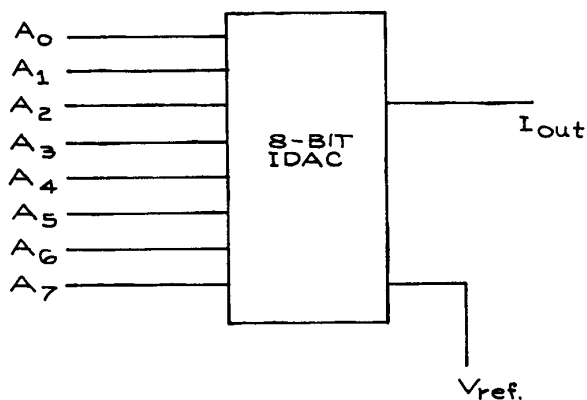
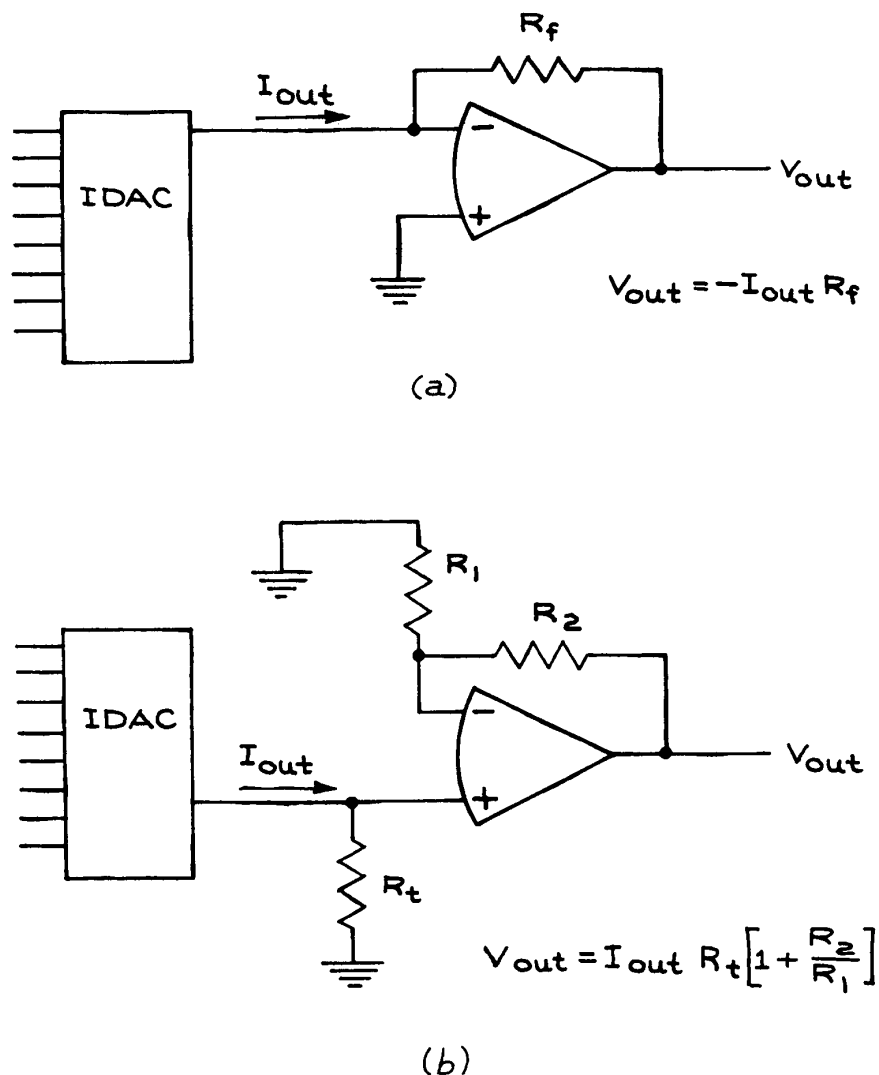


Figure 6.91 An 8-bit current digital-to-analog converter (IDAC).

binary number at the counter outputs is then proportional to the analog input voltage.

The *dual-slope ADC* [see Figure 6.93(b)] converts voltage to time by using the input voltage to charge a capacitor to a predetermined voltage and then discharging the capacitor by connecting it to a reference voltage of the opposite polarity. The input voltage is proportional to the ratio of the charge to the discharge times, which is recorded by a counter driven by a clock through control logic. As long as the capacitor, clock, and reference voltage are stable during the conversion interval, the resolution will depend only on the resolution of the capacitor circuit. This method is the one most commonly used in digital multimeters.

The *successive-approximation ADC* [see Figure 6.93(c)] works like the counter ADC except a programmer, rather



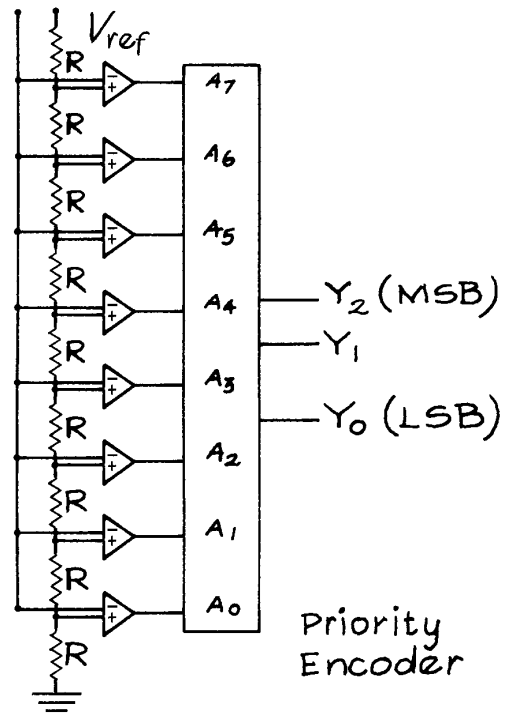
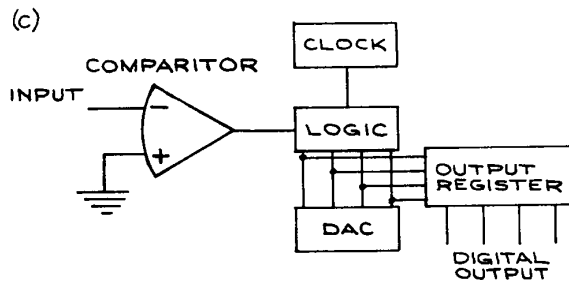
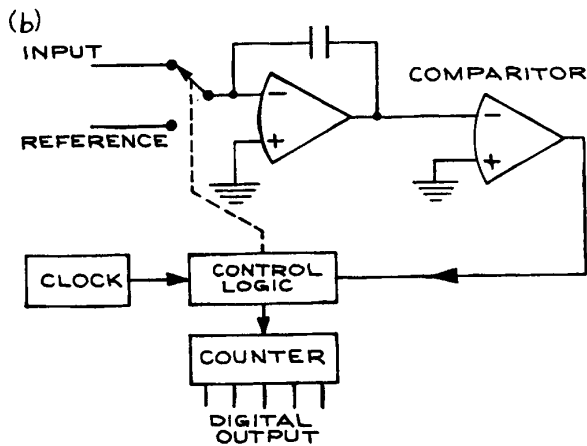
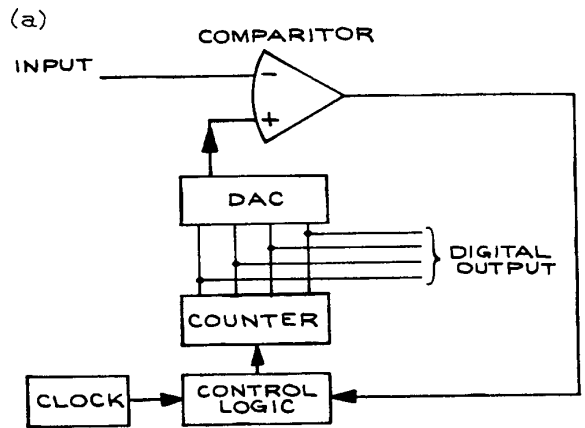
**Figure 6.92** Conversion of the current output of an IDAC to a voltage output: (a) inverting; (b) noninverting.

than a clock and counter, drives the DAC. The programmer starts by activating the most significant bit of the DAC. If the DAC output is greater than the input, the programmer turns the MSB off and the next MSB on, and the comparison is done again. The process of comparison continues to the LSB, after which the conversion

is complete. This method is fast and has high resolution. Resolution, linearity, accuracy, and settling time are important parameters used for characterizing the operation of ADCs.

The fastest ADCs are flash converters. These use  $2^N$  comparators, a voltage divider with  $N$  resistors, and a





$A_7$	$A_6$	$A_5$	$A_4$	$A_3$	$A_2$	$A_1$	$A_0$	$Y_2$	$Y_1$	$Y_0$
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	X	0	0	1
0	0	0	0	0	1	X	X	0	1	0
0	0	0	0	1	X	X	X	0	1	1
0	0	0	1	X	X	X	X	1	0	0
0	0	1	X	X	X	X	X	1	0	1
0	1	X	X	X	X	X	X	1	1	0
1	X	X	X	X	X	X	X	1	1	1

Priority Encoder Truth Table

Figure 6.93 Four analog-to-digital converters: (a) counter; (b) dual slope; (c) successive approximation; (d) flash.

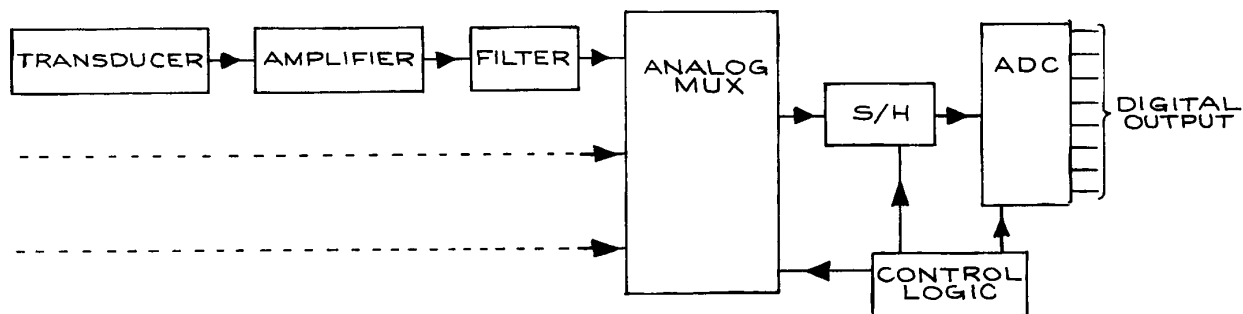
priority encoder to produce an  $N$ -bit binary output. The speed of conversion depends primarily on the speed of the comparators. eight-bit ADCs that can convert at a rate of  $10^9$  samples per second (Gsp/s) are widely available. A simple 3-bit flash converter circuit is shown in Figure 6.93(d) along with the truth table for the priority encoder.

Modular data-acquisition systems or *data loggers* with an analog multiplexer (MUX) *sample-and-hold* (S/H) circuit, and ADC are available from many sources. A block diagram of such a system is shown in Figure 6.94. Each of the transducer-amplifier-filter circuits is connected to the sample-and-hold circuit for a fixed time through the multiplexer. The output of the sample-and-hold circuit is then digitally encoded by the ADC. The frequency at which the various inputs are sampled depends on the rate at which the signals are varying. An important theorem of

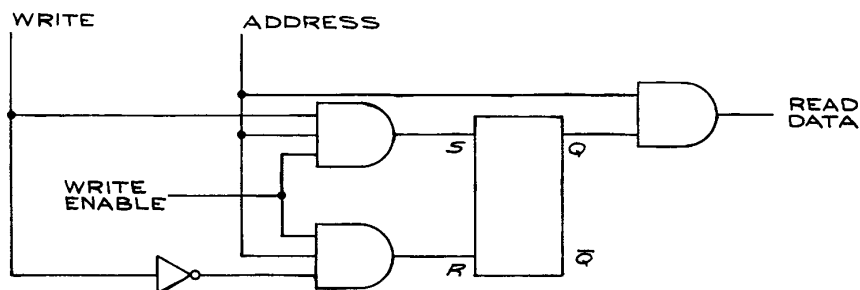
sampling theory states that if a sampled signal contains no Fourier components of higher than  $f_{\max}$ , the signal can be recovered with no distortion, if sampled at a frequency of twice  $f_{\max}$ .

### 6.6.8 Memories

*Large-scale integration* (LSI) and *very large-scale integration* (VLSI) have made possible the production of high-density, solid-state memories. The typical one-bit memory cell is an R/S flip-flop with supplementary gates, shown in Figure 6.95. When the *address* and *write enable* lines are activated, the signal on the *write* line is registered on the  $Q$  line. It will remain there as long as the *write enable* line is not subsequently activated. To read the contents of the memory cell, the *address* line has only to be



**Figure 6.94** Multiple-input data-acquisition system with an analog multiplexer (MUX), sample-and-hold (S/H) circuit, and ADC.



**Figure 6.95** Basic memory cell.

activated. Large-scale memories are composed of billions of such elementary cells. Because the contents of any single cell can be read by activating the appropriate address line, such memories are called random-access memories (RAMs). There are several different ways of arranging the basic memory cells for addressing purposes. For example, a 1024-bit RAM can be arranged as 1024 one-bit words or 256 four-bit words. The addressing scheme in the first case requires 10 address lines, one *data out* line, and one *data in* line. For the  $256 \times 4$  memory, eight address lines are needed as well as four *data in* and four *data out* lines. In addition to the address and data lines, a *read/not write* ( $R/\overline{W}$ ) line and a *chip enable* (CE) line are needed. The state of the  $R/\overline{W}$  line determines whether data are to be read into or out of the memory; the CE line is activated when reading or writing data. *Dynamic* RAMs require an additional input, called the *refresh*, to hold the contents of the memory. Refreshing is required continuously with the refresh signal supplied from a separate clock. *Static* RAMs require no such refreshing. Solid-state RAMs may be volatile, meaning that their contents are destroyed when the power is interrupted. Battery backup can be provided for these memories in applications where the memory contents must be retained in the event of a power failure. For modern memories based on MOS technology, external power sources are not necessary to retain memory contents. Examples are USB memory sticks and flash memory cards used in digital cameras and recorders.

Important parameters to consider for RAMs are power dissipation, access time, and cycle time. The *access time* is the time required after the activation of the address lines to obtain valid data at the output. The *cycle times* are the times required to complete a read- or write-data procedure. The access and cycle times are related to the power dissipation – generally, the lower the power dissipation, the slower the memory.

Other considerations are power-supply voltages and input and output signal levels. RAM must be able to accept data from external devices and, in turn, produce output levels capable of driving the circuits connected to its outputs.

ROMs are storage devices with permanent, unalterable patterns of 0s and 1s in the individual cells. ROMs are used extensively in computers for storage of the character-generation bit patterns, lookup tables for routine code con-

versions, and control programs, such as those used in calculators.

Programmable read-only memories (PROMs) are ROMs whose bit pattern can be altered. They are programmed with a PROM *programmer* that permanently *burns* the desired bit pattern into the PROM. The bit pattern can be entered manually (memory location by memory location), serially, or in blocks under the control of a computer connected to the programmer.

Even more useful devices are erasable PROMs, or EPROMs. These are programmed in the usual way, but can also be erased and reprogrammed for other applications. Erasing is usually accomplished with ultraviolet light in an EPROM eraser. This can take several minutes. The EEPROM or electrically erasable PROM, can be very quickly erased with electrical signals.

An example of the more common series of EPROMs is the 2700 series, manufactured by INTEL, Advanced Micro Devices, and others. Members of the series are the 2708, 2716, 2732, 2764, 27128, 27256, and 27512. The digits following the 27 give the number of kilobits in the EPROM. For example, the 2732 has 32 kilobits of memory arranged in a  $4096 \times 8$  pattern – that is, 4 kilobytes. A 12-bit address ( $2^{12} = 4096$ ) is necessary to address 4 kilobytes. The chip also has eight output pins – one pin each for the +5 V power, the ground, and the 25 V programming voltage, and a chip-enable pin. The 2732 chip has 24 pins and is programmed by entering 0s into the desired memory locations since all the bits are 1s upon delivery or after erasure. Programming is accomplished by applying +25 V to the OE/VPP programming-voltage pin, while the address of the byte to be programmed is applied to the appropriate address pins at the same time that the desired 8-bit pattern is placed on the output pins. When all voltage levels are stable, a 50 ms TTL low-level pulse is applied to the CE/PGM chip-enable pin.

Programmable PROMs are designed to apply the proper voltages in the required sequence to the PROMs. Programmable PROMs in kit form cost less than \$100. Fully assembled programmers cost from \$200 to \$500 – the more costly ones have the capability to program a wide variety of EPROMs as well as copy and verify EPROM codes. The programmers generally are controlled through the serial port of a personal computer, and software is supplied with the programmer and loaded into the

computer. The PROMs are useful in experimental work because they can be programmed to reproduce fairly complex truth tables, eliminating the need for several different logic chips. The truth table of Table 6.34, for example, can be fully implemented with a PROM having as few as  $32 \times 4$  bits.

The implementation of truth tables with PROMs is inefficient because, in practice, only a limited number of variable combinations are actually used. With a PROM, all possible combinations of all the input variables are represented, resulting in a large bit count, most of which is never used. Programmable array logic (PAL), programmable logic arrays (PLAs), and field programmable gate arrays (FPGAs) permit the economical realization of logic expressions with a limited number of inputs and terms. The units perform the functions of a custom IC. Of the three devices, the PLAs and FPGAs are the most flexible. Programmers for these devices are more complicated than PROM programmers and cost, with software, several hundred dollars for professional models and a few hundred dollars for student models. XILINX, Altera, and Actel are manufacturers of programmable logic devices. Currently FPGAs with as many as  $10^6$  gates are available for a few tens of dollars. The two most used languages for programming FPGAs are VHDL and Verilog. The acronym VHDL stands for **V**HSIC **H**ardware **D**escription **L**anguage, and VHSIC refers to the **V**ery **H**igh **S**peed **I**ntegrated **C**ircuit program originally sponsored by the Department of Defense for the development of high-speed integrated circuits. A working knowledge of VHDL is beyond the scope of the average scientist, however it is quite practical to implement rather complex codes using the very large variety of VHDL templates supplied by the manufacturers of FPGAs. The software necessary to implement VHDL codes is supplied by the manufacturer of the FPGAs and includes all steps including logic checking and simulation for implementing the code on an FPGA device. Communication with the FPGA is through a USB link to a PC.

### 6.6.9 Logic and Function

To illustrate the ideas in the preceding sections on digital logic, we now discuss a method for designing an interlock circuit for a vacuum system.

High-vacuum systems often have a vacuum chamber connected to a turbomolecular pump through a gate valve (see Section 3.6.1). The turbo-pump has a high-pressure outlet connected to a mechanical rotary pump or diaphragm pump via a foreline and foreline valve. A bypass roughing line with a roughing valve may be used for preliminary evacuation of the chamber. Before this is done, however, the foreline valve is closed. When the pressure in the chamber is sufficiently low, the roughing line is closed, the foreline valve is opened, and the chamber is turbo-pumped through the gate valve.

To guard against accidents, it is worthwhile to have a system of interlocks that activates valves in the proper sequence and turns the turbo-pump off in case of an accident. The interlock system consists of *sensors* that continuously sample the variables in the vacuum system, such as cooling-water temperature, cooling-water pressure, foreline pressure, vacuum-chamber pressure, and roughing-line pressure. The security of the system depends on whether the physical property sampled is above or below a predetermined threshold. The second section of the interlock system is the *logic*, which, with combinations of elementary gates, establishes the relationships between the input variables from the sensors and the output functions that control the various vacuum-system operations – such as the opening and closing of the various valves and the application of voltage to the turbo-pump. Because the output functions only assume one of two values, the interlock system can be analyzed using either Boolean algebra or Karnaugh maps. Examples of increasing complexity are given below. A way of implementing the system with a *data selector* is also discussed.

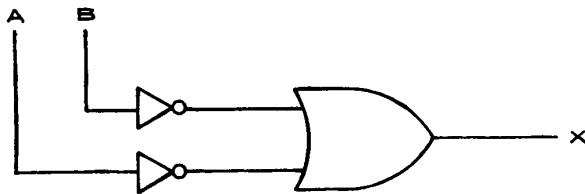
The simplest practical interlock system monitors the cooling-water temperature and pressure and turns off the voltage to the turbo-pump if the water temperature is too high, the water pressure is too low, or both. A truth table relating the variables to the function is shown in Table 6.32. The letters and numbers in parentheses are logic symbols assigned to the variables ( $A$ ,  $B$ ), function ( $X$ ), and states of the variables (0, 1). The Boolean logic expression for the truth table is:

$$X = \bar{A} \cdot \bar{B} = \overline{A + B} \quad (6.51)$$

This is obtained by noting the condition under which  $X$  is in the logical 1 state. Had there been more than one

**Table 6.32 Truth table**

Variables		Function
Water Pressure (A)	Water Temperature (B)	Diffusion Pump Heater (X)
Low (1)	Low (0)	Off (0)
Low (1)	High (1)	Off (0)
High (0)	Low (0)	On (1)
High (0)	High (1)	Off (0)

**Figure 6.96** Gate arrangement corresponding to Table 6.32.

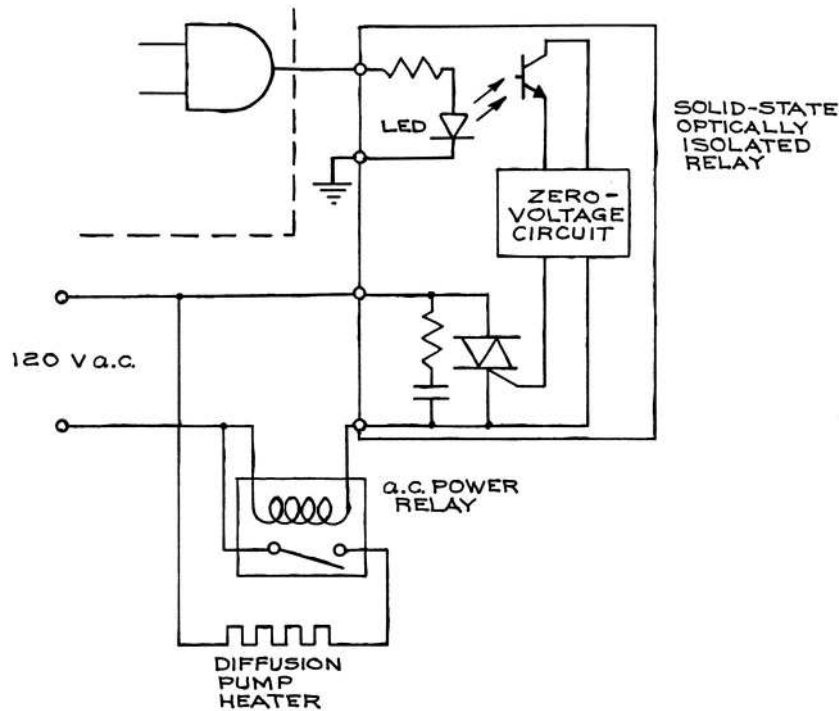
occurrence of 1 for X, the expression would have additional OR terms. The Boolean expression can be implemented with the inverter-and-gate arrangement of Figure 6.96. A bimetallic-switch temperature sensor can provide the water-temperature variable, and a diaphragm-type pressure sensor can provide the water-pressure variable. Power to the turbo-pump is controlled by a heavy-duty relay actuated by a solid-state, optically isolated relay driven by the output of the logic circuit. A practical circuit is shown in Figure 6.97. It is assumed that logical 0 is less than 0.8 V and logical 1 is greater than 2.4 V, consistent with TTL logic. Optical isolators provide an effective interface when controlling power circuits from logic circuits.

An optical isolator contains both an infrared-emitting diode and a photodetector arranged in a single package, so that the light from the diode is efficiently transmitted to the detector through a transparent dielectric medium that provides electrical isolation. The diode detector pair is shielded from ambient light by opaque encapsulating material. Since only infrared radiation is used to communicate between diode and detector, there is no electrical connection between the two. The communication is also

unidirectional, because the detector cannot provide a signal that can couple back to the diode. Typical applications for isolators are elimination of common mode input signals and ground loops, level shifting between two circuits, logic control of power circuits, and isolation of logic circuits from noise. The last application commonly occurs when using the output logic levels on a computer to control external devices. Infrared emitting diodes and photodetectors are inherently nonlinear devices, but some isolators are specially made for linear analog circuits. Photodiodes, phototransistors, photo-Darlington transistor pairs, and phototriacs are the detector units in isolators. Important parameters to consider when choosing an isolator are the isolation voltage, ratio of output current to light-emitting-diode input current, and output power-handling capabilities. Transistor isolators have transfer ratios from 2 to over 50%, while the Darlington transistor pairs have ratios from 50 to 500%. Isolators with triac outputs require input trigger currents in the 10 to 30 mA range in order to switch 120 V a.c. at currents up to a few hundred milliamperes. This is sufficient to drive high-power triacs, solenoids, small motors, and mechanical relays. For proper isolator operation, the characteristics of the output stage must be matched to the load. Data sheets should be consulted. A common package for low-power isolators is a six-pin DIP. A list of suppliers is given at the end of this chapter.

When switching transformer loads, solenoid valves and contactors, lamp loads, and heaters, output current rating should exceed the load rating by at least a factor of two. Solid-state relays provide isolation, and an important application of solid-state relays is the interface between computers and power devices. OPTO 22 makes a line of solid-state relays that can be turned on and off with standard TTL signal levels. The 120D10 can be used to switch up to 10 A of a.c. current at 120 V, with a maximum voltage drop of 1.6 V. The control circuit and switching circuits are electrically isolated from each other up to 4 kV, and switching of the load occurs at zero voltage in the a.c. cycle, thereby eliminating switching transients. Because of the isolation between control signal and load, the circuit supplying the control signal is protected from voltage spikes. The relays are 1.2 in. high, 2.4 in. long and 1.8 in. wide.

A more useful interlock system using three variables and two functions has the truth table of Table 6.33. The variables monitored are water pressure and temperature and



**Figure 6.97** Practical circuit for a vacuum-system interlock according to the truth table of Table Figure 6.29.

turbo-pump foreline pressure. The functions are the turbo-pump (on or off) and the gate valve (open or closed). The Boolean expression and logic circuit are also given in the table, while the Karnaugh map is shown in Figure 6.98. In the map for  $Y$ , the 1s are adjacent in the 10 column, so that  $Y$  is independent of the state of  $C$ .

As a final example of logic design, consider increasing the number of variables to five by including the roughing-line pressure and chamber pressure and adding the roughing valve and foreline valve to the functions. With five variables there will be  $2^5$  (or 32) combinations. Systematic design of the interlock circuit will ensure that none of them is left out. The truth table is given in Table 6.34. Rather than obtain the Boolean expressions for the truth table or make a map, we can use a *data selector* to establish the relationships among the variables and functions. A 5-bit data selector has 5 input address lines, 1 output line, and 32 data lines. Depending on the 5-bit input address, the state

of the addressed input line is transferred to the output line. Implementation of the expression for  $Y$  is shown in Figure 6.99(a). A 4-bit data selector [see Figure 6.99(b)] can also be used by employing a foldback arrangement. Looking at the  $A$ ,  $B$ ,  $C$ , and  $D$  bits, one sees that there are 16 pairs of combinations of  $A$ ,  $B$ ,  $C$ , and  $D$  that are independent of whether  $E$  is 1 or 0. For each of the 16 pairs, the function  $Y$  is independent of the state of  $E$ , equals  $E$ , or equals  $\bar{E}$ . The 16 input lines of the 4-bit data selector will each have 0, 1,  $E$ , or  $\bar{E}$  attached to them.

Even more complexity can be built into the circuit by having time delays on the functions so that one valve is not opening while the other is closing. Extra functions, such as turning on and off and venting the mechanical pump, can also be added. This systematic method of analysis ensures that nothing is overlooked. By noting when two functions change state upon the change of state of a variable, decisions about sequencing can be made.

**Table 6.33 Three-variable, two-function interlock**

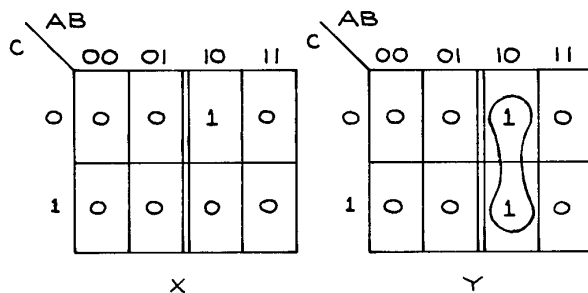
Truth Table			
Variables		Function	
Water Temperature (B)	Foreline Pressure (C)	Diffusion-Pump Heater (X)	Gate Valve (Y)
low (0)	low (0)	off (0)	closed (0)
low (0)	high (1)	off (0)	closed (0)
high (1)	low (0)	off (0)	closed (0)
high (1)	high (1)	off (0)	closed (0)
low (0)	low (0)	on (1)	open (1)
low (0)	high (1)	off (0)	open (1)
high (1)	low (0)	off (0)	closed (0)
high (1)	high (1)	off (0)	closed (0)

**Boolean Expressions**

$$X = A \cdot \bar{B} \cdot \bar{C}$$

$$Y = A \cdot \bar{B} \cdot \bar{C} + A \cdot \bar{B} \cdot C$$

$$Y = A \cdot \bar{B}(C + \bar{C})$$

$$Y = A \cdot \bar{B}$$
**Figure 6.98** Karnaugh map for the truth table of Table 6.33.

### 6.6.10 Implementing Logic Functions

Implementing functions with transistor gates requires that the 0 and 1 states be associated with voltage or current levels. There are a large number of logic families, all of which have internally consistent 0 and 1 levels, but which

are not necessarily mutually compatible. The choice of a logic family for the implementation of a given function depends on a number of factors – among them are speed, power dissipation, immunity to noise, number of available functions, cost, and compatibility with external circuits and other logic families. A summary of the properties of the most common logic families is given in Table 6.35. As can be seen from the table, there is a tradeoff between speed and power dissipation – the fastest logic dissipates the most power per gate and the slowest dissipates the least. In an effort to reduce power dissipation while maintaining speed, nominal voltage levels have been reduced to 3.5 and as low as 2.5 V for some logic families from the nominal TTL 5 V. *Noise immunity* is measured by *noise margin*, the meaning of which is illustrated in Figure 6.100. In general, logic families are not directly compatible – translator chips exist for interfamily conversion, and some simple conversion circuits are given in Figure 6.101. The abbreviations SSI, MSI, LSI, and VLSI stand for small-scale integration, medium-scale integration, large-scale integration, and very large-scale integration, and are used to specify the number of gates on a chip. SSI is less than 12 gates per chip, MSI is between 12 and 100, LSI is more than 100, and VLSI is more than 1000. Simple logic functions employ SSI and MSI.

The most common logic family has been TTL (transistor-transistor logic) and its variations, but CMOS logic (complementary symmetry, metal-oxide semiconductor) has now largely displaced TTL because of its low power consumption and wide range of usable power-supply voltages. Emitter-coupled logic (ECL) is only used when the highest speeds are required. The ECL 10000 series is a slower version of the ECL II series and is not so sensitive to the quality of the interconnections. With the II series, a circuit board with a ground plane must be used and the interconnections carefully laid out to minimize cross-coupling.

Important changes in the semiconductor electronics industry in the last few years reflect new technologies as well as new philosophies in electronics. Whole new families of logic have been born – based on CMOS, Schottky, and gallium arsenide (GaAs) technologies. Portable electronic devices from laptop computers to cell phones and personal digital assistants require high-speed and low-power electronics. The inherently lowest-power technology is CMOS, but it has suffered in the past from low

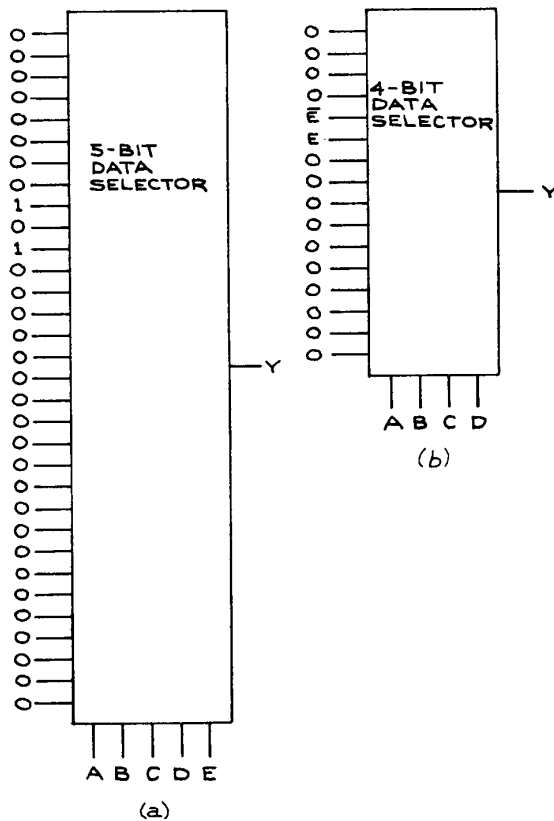
Table 6.34 Five-variable, four-function interlock truth table

$T_{H_2O}$ (A)	$P_{H_2O}$ (B)	$P_{foreline}$ (C)	$P_{roughing}$ (D)	$P_{chamber}$ (E)	Diffusion Pump (X)	Gate Valve (Y)	Roughing Valve (R)	Foreline Valve (W)
0 (low)	0 (low)	0 (low)	0 (low)	0 (low)	0 (off)	0 (closed)	0 (closed)	0 (closed)
0	0	0	0	1 (high)	0	0	0	0
0	0	0	1 (high)	0	0	0	0	0
0	0	0	1	1	0	0	0	0
0	0	1 (high)	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0
0	0	1	1	1	0	0	0	0
0	1 (high)	0	0	0	1 (on)	1 (open)	0	1 (open)
0	1	0	0	1	0	0	1 (open)	0
0	1	0	1	0	1	1	0	1
0	1	0	1	1	0	0	1	0
0	1	1	0	0	0	0	0	0
0	1	1	0	1	0	0	1	0
0	1	1	1	0	0	0	0	1
0	1	1	1	1	0	0	1	0
1 (high)	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	1	1	0	0	0	0
1	0	1	0	0	0	0	0	0
1	0	1	0	1	0	0	0	0
1	0	1	1	0	0	0	0	0
1	0	1	1	1	0	0	0	0
1	1	0	0	0	0	0	0	0
1	1	0	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0
1	1	0	1	1	0	0	0	0
1	1	1	0	0	0	0	0	0
1	1	1	0	1	0	0	0	0
1	1	1	1	0	0	0	0	0
1	1	1	1	1	0	0	0	0

speed. This has largely been overcome with the HC and HCT logic lines. Both are directly compatible with the older TTL LS families, but offer greatly improved speed-to-power ratios. Because TTL high levels can be as low as 2.2 V, there can be difficulties driving HC devices from

TTL outputs. This is overcome with the HCT family, which is completely compatible with TTL over the full range of TTL operating parameters. The HC logic has now replaced the TTL LS family. The 4000 D series of CMOS can operate with supply voltages from +5 to +15 V





**Figure 6.99** Implementation of the Y-function of the truth table of Table Figure 6.30 with (a) a 5-bit data selector and (b) a 4-bit data selector using a fold-back arrangement.

and is used in conjunction with higher-voltage analog circuits. FAST<sup>TM</sup> logic is based on Schottky technology and approaches ECL 10 000 logic in speed, but with much lower power consumption. FAST devices bear the designation 54/74F and have the advantage of TTL logic levels while avoiding the smaller voltage swings and lower noise margins of ECL.<sup>7</sup> FAST is a trademark of Fairchild Camera and Instruments Corporation and is an acronym for Fairchild Advanced Schottky TTL.

Because TTL is still used in some applications, a few simple TTL circuits are given in Figure 6.102. The mechanical contact conditioner [Figure 6.102(a)] ensures

that the output from a mechanical switch will be a single transition rather than the series of voltage spikes normally resulting from the mechanical bouncing of the contacts. Transistor-transistor logic gates are either *edge* or *level triggered*. Edge triggering refers to the transition between the 0 and 1 levels (positive edge) or 1 and 0 levels (negative edge). Level triggering occurs when the 0 or 1 is attained. Data sheets specify the type of triggering each logic unit requires.

With most logic families, there is the tendency to generate current spikes during switching; this creates power-supply noise. To reduce this noise, the power supply must be decoupled with capacitors from the logic circuits. A 0.01  $\mu\text{F}$  ceramic capacitor for each gate and an additional 0.1  $\mu\text{F}$  capacitor for each 20 gates are generally sufficient. Counters and shift registers are especially sensitive to power-supply noise and should be decoupled with a 0.1  $\mu\text{F}$  capacitor for every two devices. For optimum performance, unused inputs should be set to logic level 0 by direct connection to ground or logic level 1 by connection to the power-supply voltage through a 1 to 10  $\text{k}\Omega$  resistor.

Some additional features of TTL are three-state outputs and open-collector outputs. A *three-state output* has the standard 0 and 1 levels and also a high-impedance, inactive state. In this state, the output assumes the level of any active output connected to it. The three-state outputs are designed to be connected together in a common line or bus structure so that when any single output is active, all the others assume its level. The *open-collector output* requires a collector resistor to the power supply for operation. When connected together, open-collector outputs allow one to implement an output AND function (often incorrectly termed wired-OR) that is, when any output is 0, all become 0. Open-collector outputs have the disadvantage of reduced transition speed and high output impedance in the 1 state. They are usually used as line drivers.

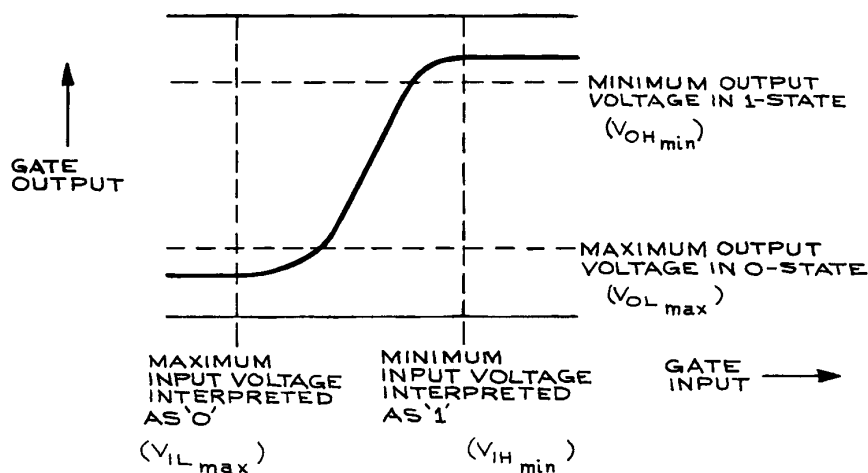
## 6.7 DATA ACQUISITION

### 6.7.1 Data Rates

When designing an experiment, one important consideration is the rate at which data are collected. It is usually useful to think of an experiment in terms of sampling rate, sampling time, and measurement precision. Consider a

Table 6.35 Logic families

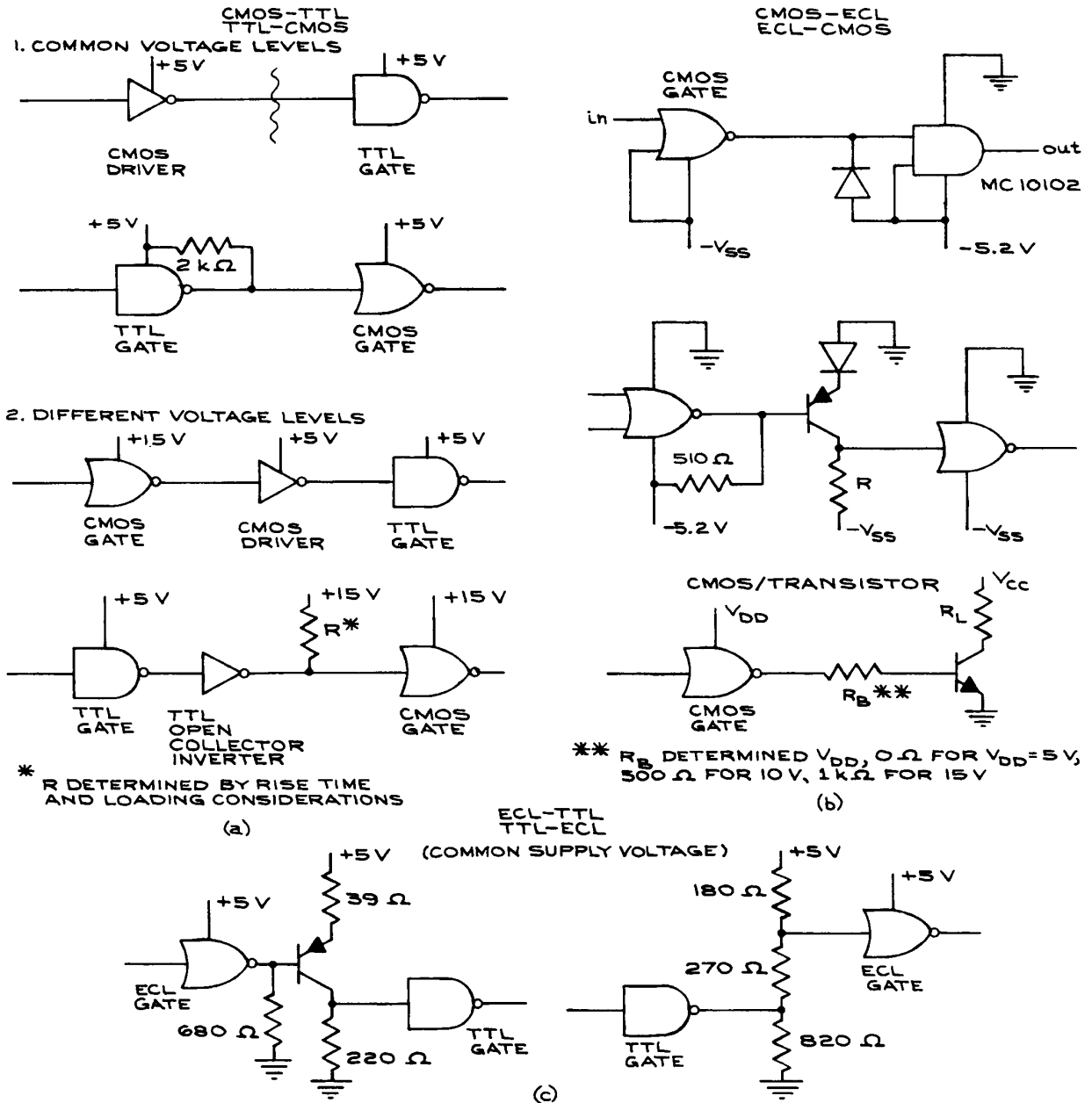
Logic Family	Propagation Delay (ns)	Power Dissipation (mW/gate)	0-Level (V)	1-Level (V)	Noise Margin, Low/High (V)	Fanout	Power Supply (V)
TTL:							
54/74	10	10	0.2	3.9	0.3/0.7	10	+5
54/74LS	10	2	0.2	3.9	0.3/0.7	20	+5
54/74F	4	4	0.2	3.9	0.3/0.7	33 (50LS)	+5
CMOS:							
54/74C	30	0	0.2	4.8	0/7/0.6	>50 (10LS)	+5
54/74HC	10	0	0.2	4.8	0.7/0.6	>50 (10LS)	+5
54/74HCT	8	0.010	0.2	4.8	0.7/0.6	>50 (10LS)	+5
4000 CD							
$V_{DD}$ 5 to 15 V	115	0.005	0.05	4.95	1.0	50	+5
ECL:							
10, 000	3	24	-1.75	-0.90	0.3/0.3	10	-5.2
100, 000	0.8	40	-1.75	-0.90	0.3/0.3	10	-4.5



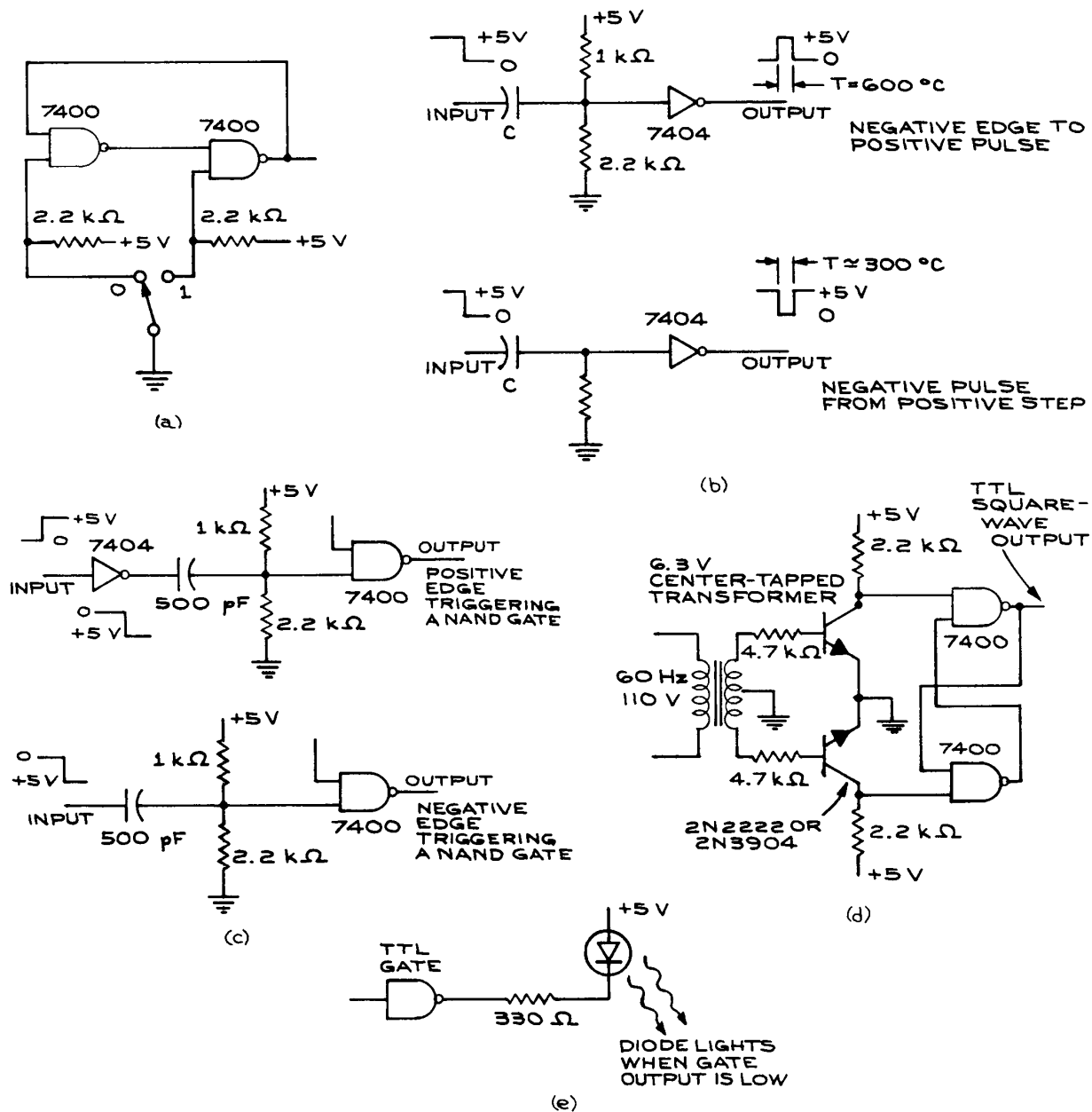
**Figure 6.100** Illustration of “noise margin.” Noise margin for 0-state =  $V_{OH_{min}} - V_{IH_{min}}$ ; noise margin for 1-state =  $V_{IL_{max}} - V_{OL_{max}}$ .

spectroscopy experiment where the output of a monochromator is incident on a photomultiplier tube – the output of which are pulses of electrons that are recorded with a counter. The monochromator is stepped at a regular rate and predetermined step size over the wavelength interval

of interest. The time spent at each step combined with the number of steps determines the total number of counts recorded. If the total number of signal counts in any interval is  $C_S$  and the number of noise counts is  $C_N$ , the total number of counts  $C_T$  is  $C_S + C_N$ . The statistical



**Figure 6.101** TTL, CMOS, ECL, and discrete transistor interface circuits. (a) CMOS-TTL and TTL-CMOS (the value of R is determined by rise-time and loading considerations). From *McMOS Integrated Circuits Data Book*, Motorola, Inc, 1973. Copyright of Motorola, Inc. Used with permission. (b) CMOS-ECL and ECL-CMOS (R<sub>B</sub> is determined by V<sub>DD</sub>: 0 Ω for V<sub>DD</sub> = 5 V, 500 Ω for V<sub>DD</sub> = 10 V, 1 k Ω for V<sub>DD</sub> = 15 V); (c) ECL-TTL and TTL-ECL (from *MECL System Design Handbook*, 2nd edition, W. R. Blood, Jr., and E. C. Tynan, Eds., Motorola Semiconductor Products, 1972. Copyright of Motorola, Inc. Used with permission.)



**Figure 6.102** TTL signal-conditioning circuits: (a) mechanical-contact conditioner; (b) pulses from step inputs (step duration must exceed output-pulse width); (c) edge triggering; (d) 60-Hz TTL square wave from the power line; (e) TTL to LED.

uncertainty in  $C_T$  is  $\sqrt{C_T}$ , and that in  $C_N$  is  $\sqrt{C_N}$ . The uncertainty in determining  $C_S$  is therefore  $\sqrt{C_T + C_N}$ , so the relative error is  $\sqrt{C_T + C_N}/(C_T - C_N)$ . The counts increase linearly with the time of measurement, but the relative uncertainty only decreases as the inverse square root of the time of the measurement. The *data rate* is the number of counts in a measurement interval divided by the interval time.

*Bits, bytes, words, and characters* are all used to specify the amount of information. The number of *bits* necessary to specify a given integer decimal  $D$  is given by  $N = 3.32 \log_{10} D$ . If  $N$  is a noninteger, it is rounded to the next higher integer. If the decimal number is signed (that is, plus or minus), another bit is added. Eight bits make a *byte*, and two bytes (16 bits) commonly, but not always, make a *word*.

A *character* is the equivalent of one word. If data are to be transferred from the counter to a printer or other display or storage device, the data acceptance rate of the device must match the data output rate of the counter. Data rates are often specified as a *baud rate*, which is the number of times per second there is a transition on the signal line from 0 to 1 or from 1 to 0.

For two devices to be compatible, a number of parameters must match. These include frequency response, voltage levels, timing, format, code, hand-shaking protocols, and housekeeping procedures.

## 6.7.2 Voltage Levels and Timing

The most positive and most negative voltages corresponding to the two logic states must match. For example, TTL 0 and 1 logic levels are +3.2 and +0.2 V, while ECL 0 and 1 levels are -0.75 and -1.55 V. Furthermore, the sign of the logic must be identical – that is, positive or negative. *Positive logic* means that the more positive voltage level is taken as a logical 1, and the more negative as logical 0. The reverse is true for *negative logic*. If the signs do not match and everything else is satisfactory, only inversion of the input signal will be required.

Often a digital input or output circuit is compatible with a logic family of which it is not a member. The 54/74C, HC, and HCT families are *TTL-compatible*, for example, which means that the voltage levels are compatible. The sinking and sourcing capabilities may be very different,

however, for so-called compatible logic families. *Sinking* has to do with the number of gates that can be attached to the output of a single gate and still maintain reliable operation. With TTL logic, when an output is at logical 0 it must be able to hold all inputs connected to it at logical 0. Since this condition is a consequence of the input transistors being in saturation, the output must be able to sink all the saturation currents and still keep the inputs at their required levels. A TTL output at logic level 1 attached to a TTL input will turn off the input transistor, and very little current will flow in the input circuit. This is the *sourcing* specification. *Fan-in* and *fan-out* are directly related to the sinking and sourcing capabilities of a gate. Fan-in is the number of logic inputs to a gate, while fan-out is the number of logic outputs it can drive. These specifications should be carefully considered when connecting two different logic families. Timing is also important – the required rise and fall times, as well as pulse widths, must be compatible. Timing diagrams show the relationship between signals and should be consulted. Converter chips that change the signals from one logic family directly to another are available.

## 6.7.3 Format

A *serial format* has all data on a single line in a sequential pattern. This is economical as far as wiring is concerned, but slower than a *parallel format* where there can be as many data lines as bits, and all of the bits in a word are transferred at the same time. This is the fastest way to transmit data, but the least economical.

There are also *parallel-serial* formats where the data are transmitted in sequential parallel codes. An eight-bit byte, for example, could be transmitted with four lines as two four-bit words. A synchronization code is necessary with such an arrangement to be able to distinguish between the beginning and end of a single byte string. It is possible to convert from a serial code to a parallel code and vice versa with converters that use shift registers. For serial-to-parallel conversion, data are sent to the registers from one end and then read from the outputs of each stage. For parallel-to-serial conversion, data are put into each register simultaneously and then extracted from the last register sequentially, a bit at a time. Data transmission in one direction over a single line is called *simplex*, in two

directions over a single line is called *half duplex*, and in two directions simultaneously, with a single line and two coding frequencies or two separate lines, is *full duplex*. Data conversion ICs called universal asynchronous receiver-transmitters (UARTs) take a wide variety of data formats and convert them to standard codes, one of which is ASCII (American Standard Computer Information Interchange). The ASCII computer code is given in Table 6.36.

A major difficulty in setting up data-handling equipment is the interface. The RS 232C serial interface can handle data rates to 20 kbps – it has a high input impedance (3 to 5 k $\Omega$ ), uses relatively high voltage (+25 V), and is unbalanced and therefore noise-susceptible. The pin assignments for the standard 25-pin connector used with this interface are given in Figure 6.103. More advanced interfaces such as the RS 485 use balanced circuitry and can handle data rates well in excess of 10 Mbps (megabits per second). There are computer interfaces designed to handle data rates to hundreds of Mbps.

*Modems* (modulator-demodulators) are devices for converting a binary code to audio frequencies (300 to 3300 Hz) for transmission over telephone lines, and reconverting the audio frequencies to binary. They are separate units to which both a telephone receiver and a data terminal can be attached.

### 6.7.4 System Overhead

In all data-transmission schemes there are additional signals that are necessary to assure that the data transfer occurs correctly. These are sometimes called *housekeeping* signals. In the case where such information is exchanged back and forth between the sending unit and the receiving unit, they are called *hand-shaking*. These additional signals contribute to what is called the *system overhead* – that is, additional data capacity which must exist independent of the useful information transmitted. When data are transferred at a regular rate between two units, they must be kept in step. A *sync* bit is often used to

Table 6.36 ASCII computer code

$b_4$	$b_3$	$b_2$	$b_1$	0	0	0	0	1	1	1	$1 \leftarrow b_7$
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	0	0	1	1	0	0	1	$1 \leftarrow b_6$
				0	1	0	1	0	1	0	$1 \leftarrow b_5$
0	0	0	0	NUL	DLE	SP	0	@	P	\	p
0	0	0	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	STX	DC2	"	2	B	R	b	r
0	0	1	1	ETX	DC3	#	3	C	S	c	s
0	1	0	0	EOT	DC4	\$	4	D	T	d	t
0	1	0	1	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	ACK	SYN	&	6	F	V	f	v
0	1	1	1	BEL	ETB	'	7	G	W	g	w
1	0	0	0	BS	CAN	(	8	H	X	h	x
1	0	0	1	HT	EM	)	9	I	Y	i	y
1	0	1	0	LF	SUB	*	:	J	Z	j	z
1	0	1	1	VT	ESC	+	;	K	[	k	{
1	1	0	0	FF	FS	,	<	L	\	l	
1	1	0	1	CR	GS	-	=	M	]	m	}
1	1	1	0	SO	RS	.	>	N	^	n	~
1	1	1	1	SI	US	/	?	O	_	o	DEL

7-bit code gives  $2^7 = 128$  different words. Columns 1 and 2 are machine commands such as carriage return (CR), line feed (LF), escape (ESC), etc.

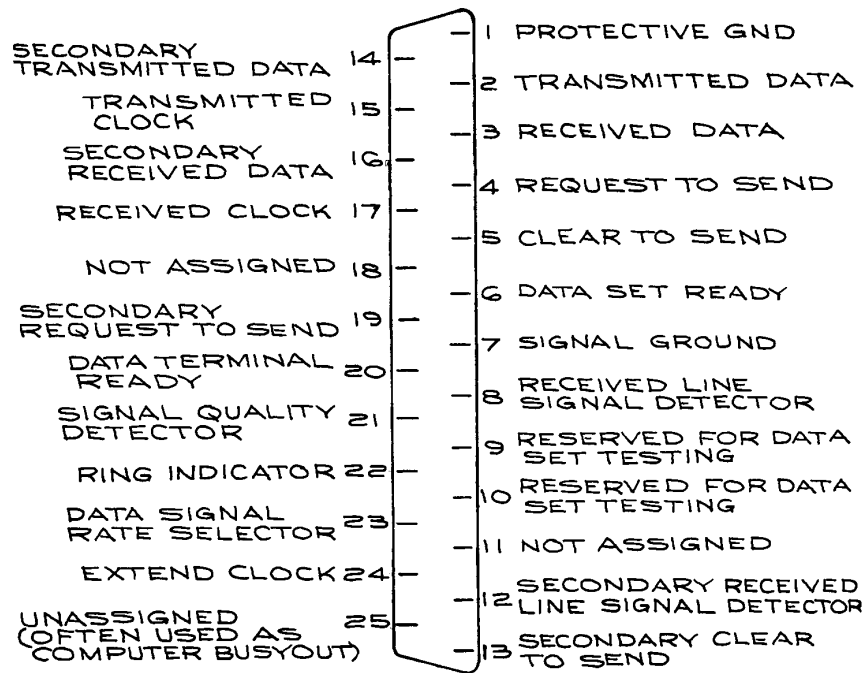


Figure 6.103 RS232C serial interface: 25-contact connector pin assignment.

assure synchronization between the two units. When data are transferred at a variable rate that depends on other independent parameters of the system, they are said to be *asynchronous*. In this case, the sending unit must alert the receiving unit of the coming of data, and the receiving unit then should give a *ready to receive* signal, after which the data are transferred. When data transfer is complete, an *end of data* signal is also usually sent. Without these signals, there can be no reliable data transfer.

To ensure the reliability of data transfer and detect errors in the transmission, an extra *parity bit* can be sent. One scheme requires the parity bit to be 1 when the sum of the data bits is even and 0 when the sum is odd. At the receiving unit, data bits are summed and compared with the parity bit – if there is consistency, the data are accepted; if not, they are rejected. A more elaborate way of ensuring integrity is to have the receiving unit echo the received data back to the sender, where they are compared with

the original data. Any differences cause the data to be rejected. This system minimizes transmission errors at the cost of additional complexity and lower speed.

Returning to the example of the spectroscopy experiment in Section 6.7.1, if the expected signal produces  $10^4$  counts per second and the background is 10 counts per second, the statistical error after one second is  $\sqrt{10^4 + 10}/10^4 \cong 1\%$ . If the counting interval is one second, at least 13 bits are needed for the data alone. To be safe, 16 bits should be used, with the extra bits reserved for housekeeping functions. This data transfer rate of 2 bytes per second is very modest. It is necessary, however, to consider the true rate of data transfer. If the data are held on the output lines of the counter for only 10 ms, the effective rate is 200 bytes per second.

In order to reduce the timing requirements of the counter, a *latch* can be used on the output lines to hold the data for a predetermined amount of time regardless of the subsequent status of the lines. The operation of a 4-bit latch is

illustrated in Figure 6.104. Data at the input lines to the latch ( $A_0$ ,  $A_1$ ,  $A_2$ , and  $A_3$ ) are transferred to the output lines ( $B_0$ ,  $B_1$ ,  $B_2$ , and  $B_3$ ) when the control line is activated. In the absence of a start signal on the control line, the output levels cannot change even though the input levels do. In this way, data can be held on the output lines for a time sufficient for them to be processed by the subsequent data-receiving unit. The timing relations among the various signals are important for proper operation of the latch. The control signal must arrive when there are valid data on the input lines.

A data *buffer* serves a purpose similar to a latch. It holds received data for processing by subsequent units or circuits. Printers, which are inherently slow because they are mechanical devices, often have data buffers that hold the input data until the printing mechanisms can transfer

them to paper. The buffer is then ready to accept another batch of data.

### 6.7.5 Analog Input Signals

The spectroscopy experiment discussed above could be modified by using a detector with an output current proportional to the incident light intensity, as in Figure 6.105. To convert the current to a digital signal requires an ADC. The resolution of the ADC is determined by the number of output bits, and the analog signal must be compatible, both in magnitude and in sign, with the input requirements of the ADC. If the ADC accepts only a voltage with a maximum full-scale value of 1.0 V and the analog signal has a current maximum of  $10^{-8}$  A, conversion from current to voltage can be made using an operational amplifier with a variable-gain element. Response and sampling times must be considered. The analog input to the ADC must remain constant long enough for the ADC to respond. This is called the *settling time*. For times less than the settling time, the ADC output will not be a valid representation of the analog input.

Voltage-to-frequency converters provide an easy way to convert analog signals into digital form. Consider the AD 537 converter manufactured by Analog Devices as an example. Through the selection of two external resistors and a capacitor, frequencies up to 100 kHz can be generated with a maximum input current of 1.00 mA. The circuit for a full-scale input of 1 V and a 10 KHz frequency maximum is given in Figure 6.106. The combination of the 8.2 k $\Omega$  resistor and 5 k $\Omega$  potentiometer sets the gain of the circuit. The potentiometer is adjusted so that the output frequency is 10 kHz for a 1 V input. The 20 k $\Omega$  potentiometer nulls the offset voltage of the input amplifier. The linearity is 0.05 % of full scale at 25  $^{\circ}$ C, with a temperature coefficient of +30 ppm of full scale per  $^{\circ}$ C. To minimize noise, a simple low-pass filter is used at the input. The 5 k $\Omega$  load resistor  $R_{out}$  allows the circuit to be reliably connected to TTL logic. The circuit can be used to convert an analog voltage to a digital word by connecting the output to a gated counter. A four-digit decade counter gated for 1 s will register the input voltage to four significant figures. A similar circuit can be used to convert the output of an analog electrometer to a digital format.

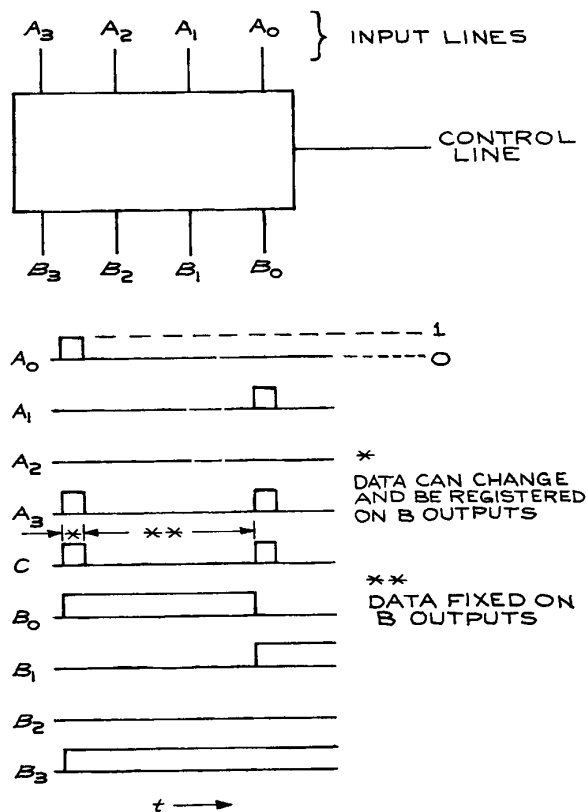
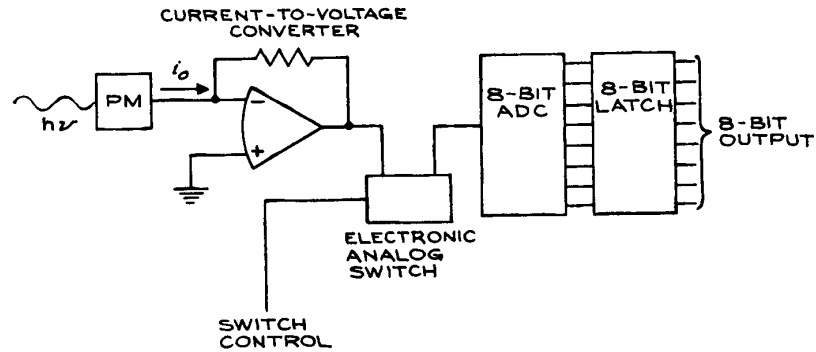
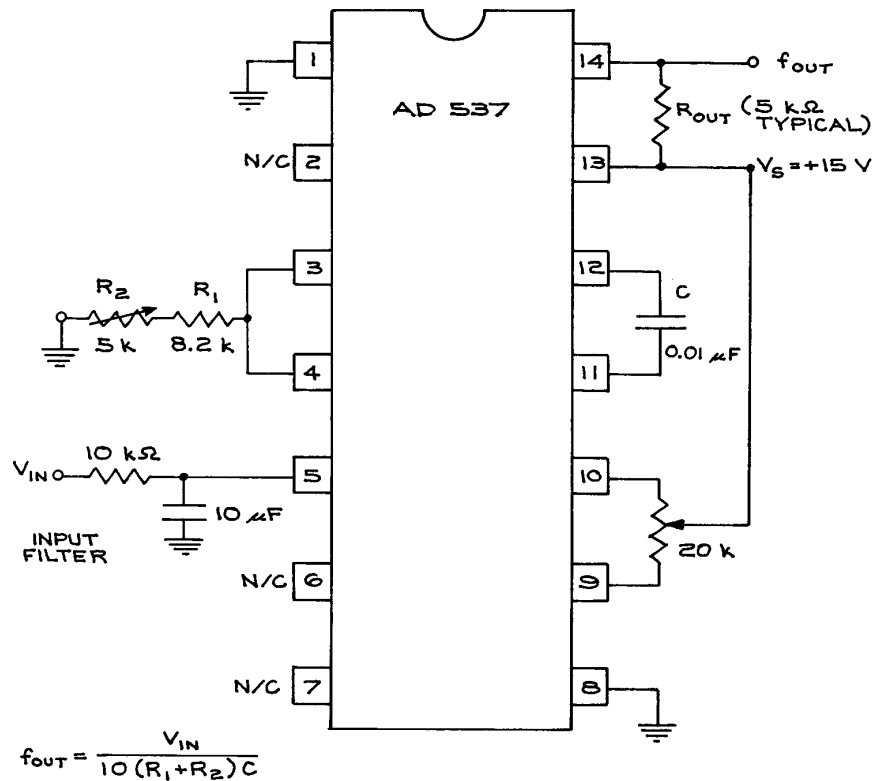


Figure 6.104 A 4-bit latch.





**Figure 6.105** Example of a light-measuring system using an ADC with a sample-and-hold circuit at the input.



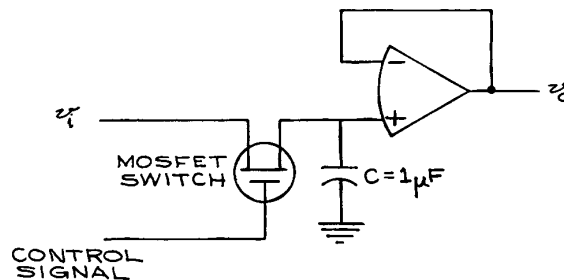
FOR TEMPERATURE STABILITY C TO BE NPO CERAMIC OR POLYSTYRENE

**Figure 6.106** Voltage-to-frequency conversion circuit. A 1.00 V input gives a 10 kHz output frequency.

The circuit can also be used to integrate a varying analog signal, as long as the signal variation occurs in a time that is much longer than the reciprocal of the maximum frequency. The area under the peaks in a gas chromatogram, for example, can be recorded by connecting the output from the chromatograph to a voltage-to-frequency converter and using a counter to total the number of output pulses. The scale factor depends on the maximum frequency and full-scale input voltage.

Voltage-to-frequency converters can also be used for frequency-to-voltage conversion. In this configuration, the input can be a TTL pulse train and the output is a voltage. The circuit can be used as an FM demodulator, frequency monitor, or motor speed controller. The AD537 voltage-to-frequency converter, made by Analog Devices, accommodates positive or negative input voltages and negative currents. It operates from a single 5 to 36 V supply and has a maximum full-scale frequency output of 100 kHz.

For the systems under consideration, the output current must not be changing faster than the ADC can respond. When the analog signal is changing too rapidly, a S/H circuit can be used ahead of the ADC. Important parameters of a S/H circuit are the *sampling time* and the *output decay time*. S/H circuits use a capacitor as a memory element, with the sampling time proportional to the capacitance. Because of leakage, capacitors cannot permanently maintain their voltage – fast sampling circuits with small capacitors maintain their output voltages for much less time than slower circuits. The *decay time* is the time for the output to fall to a fixed fraction of the original value. A figure of merit for S/H circuits is the ratio of sampling time to decay time. For the simple S/H circuit shown in Figure 6.107, a control signal closes the MOSFET switch, and the  $1\ \mu\text{F}$  capacitor is charged with a time constant  $R_{\text{on}}C$ , where  $R_{\text{on}}$  is the *on* resistance of the switch. Typical values of  $R_{\text{on}}$  are  $10^2$  to  $10^4\ \Omega$ , giving a time constant of  $10^{-4}$  to  $10^{-2}$  s for the sampling time. The decay time depends on capacitor leakage, the *off* resistance of the MOSFET, and the bias currents of the operational amplifier. When high-quality capacitors with polyethylene, polystyrene, polycarbonate, or Teflon dielectrics are used, most of the decay will be due to current through the switch ( $\sim 10^{-9}$  A) and bias currents ( $10^{-9}$  to  $10^{-8}$  A for an average, low bias-current operational amplifier). In this example,



**Figure 6.107** A sample-and-hold circuit using a capacitor as a memory element.

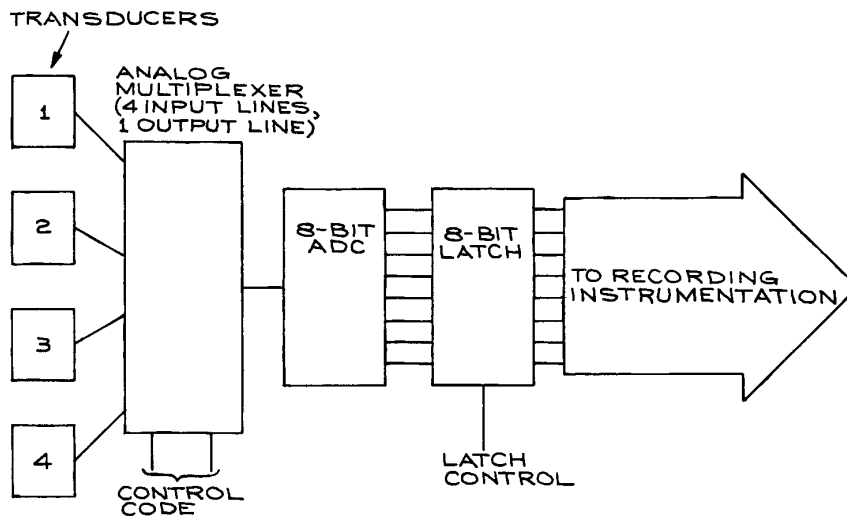
the decay will be of the order of 2 to 20 mV/s. Special MOSFET switches and operational amplifiers with picoampere bias currents can reduce these decay times even further.

## 6.7.6 Multiple Signal Sources: Data Loggers

A multiple-transducer, data-acquisition system is a straightforward extension of the single-transducer system already described. Figure 6.108 shows one possible arrangement. Such systems are useful when it is necessary to monitor and record several different signals at virtually the same time. If the transducers all have analog outputs, the most convenient arrangement is an *analog multiplexer*. The function of the multiplexer is to alternately connect each of the analog signal lines to an ADC circuit and the recording instrument. The switching between the various signal lines is accomplished by separate control lines. Multiplexing reduces the number of circuit components at the expense of more complex control and synchronization circuits. With the system in Figure 6.108, the data rate can be as much as four times that of a simple system.

## 6.7.7 Standardized Data-Acquisition Systems

A number of companies manufacture complete data-acquisition units specifically designed for the laboratory. As long as the input signals are compatible and the data



**Figure 6.108** A multiple-transducer data-acquisition system using a multiplexer.

rates are within the limits of the unit, they provide a fast and efficient means of constructing a data-acquisition system. Data acquisition from multiple sensors is a common industrial situation and there are several kinds of industrial data loggers. These, however, are designed for reliability rather than flexibility and are generally not suited to laboratory data acquisition, where many different kinds of transducers are used. There are two standardized laboratory data-acquisition systems, nuclear instrumentation module (NIM) and CAMAC.

The NIM system consists of individual electronic units that fit into a mounting frame or *bin*, which has a central power supply. The NIM bin is capable of powering several individual units, depending on their size. The bin power supply provides  $\pm 24$  V,  $\pm 12$  V, and usually  $\pm 6$  V. Each full-size bin has 12 slots in which instrument module units can be placed. Units can occupy anywhere from one to six slots. The bin power supplies and individual units have been designed so that a bin can accept the full complement of modules without becoming overloaded. A short list of standard NIM modules is given in Table 6.37. Interconnections between modules are made with coaxial cable. In older units, BNC connections are common newer units use higher-density, LEMO-type

**Table 6.37** Modules available in NIM

Power supplies
Preamplifiers, amplifiers
Discriminators
Coincidence units
Delays
Signal generators, pulse generators
Counters/timers
Gates
DACs and ADCs
Analog adders, subtractors, dividers, multipliers
Time to amplitude converters (TACs)
Signal averagers
Multichannel analyzers

connectors. Input and output specifications are standardized so that NIM modules can be connected together without compatibility concerns. The NIM logic level specifications are given in Table 6.38.

Originally the NIM standard voltages were  $\pm 12$  V and  $\pm 24$  V. As the use of integrated circuits increased, a  $\pm 6$  V

**Table 6.38 NIM Logic Levels**

	<i>Output (must deliver)</i>	<i>Input (must accept)</i>
<i>Digital Data</i>		
Logic 1	+ 4 to + 12 V	+ 3 to + 12 V
Logic 0	+ 1 to -2 V	+ 1.5 to -2 V
<i>Fast Logic (50 <math>\Omega</math> impedance)</i>		
Logic 1	-14 to -18 mA	-12 to -36 mA
Logic 0	-1 to + 1 mA	-4 to + 20 mA

voltage was included in the standard. Modules operating from  $\pm 12$  V and  $\pm 24$  V supplies, however, are still the most flexible, because all bins have these voltages. Where  $\pm 6$  V is required for a module in a  $\pm 12/\pm 24$  V bin, it is best to derive it from within the module using the available  $\pm 12$  V. Use of high-current, 6 V modules in a  $\pm 12/\pm 24$  V bin should be avoided because the regulation of the  $\pm 12$  V and  $\pm 24$  V supplies will be degraded by the high current in the common return line. Precise reference voltages should be produced in the modules themselves, since the bin voltages have relatively wide tolerances,  $\pm 24.0$  (+0.7%),  $\pm 12.0$  (+1.0%), and  $\pm 6.0$  (+3.0%). The NIM power-connector pin assignments and other details are given in Figure 6.109. The NIM standard logic levels are not compatible directly with common logic families such as TTL or ECL. The power-supply voltages are also different from common logic supplies (+5V) and operational-amplifier supplies (+15V). Some conversion circuits between NIM, ECL, and TTL levels are given in Figure 6.110.

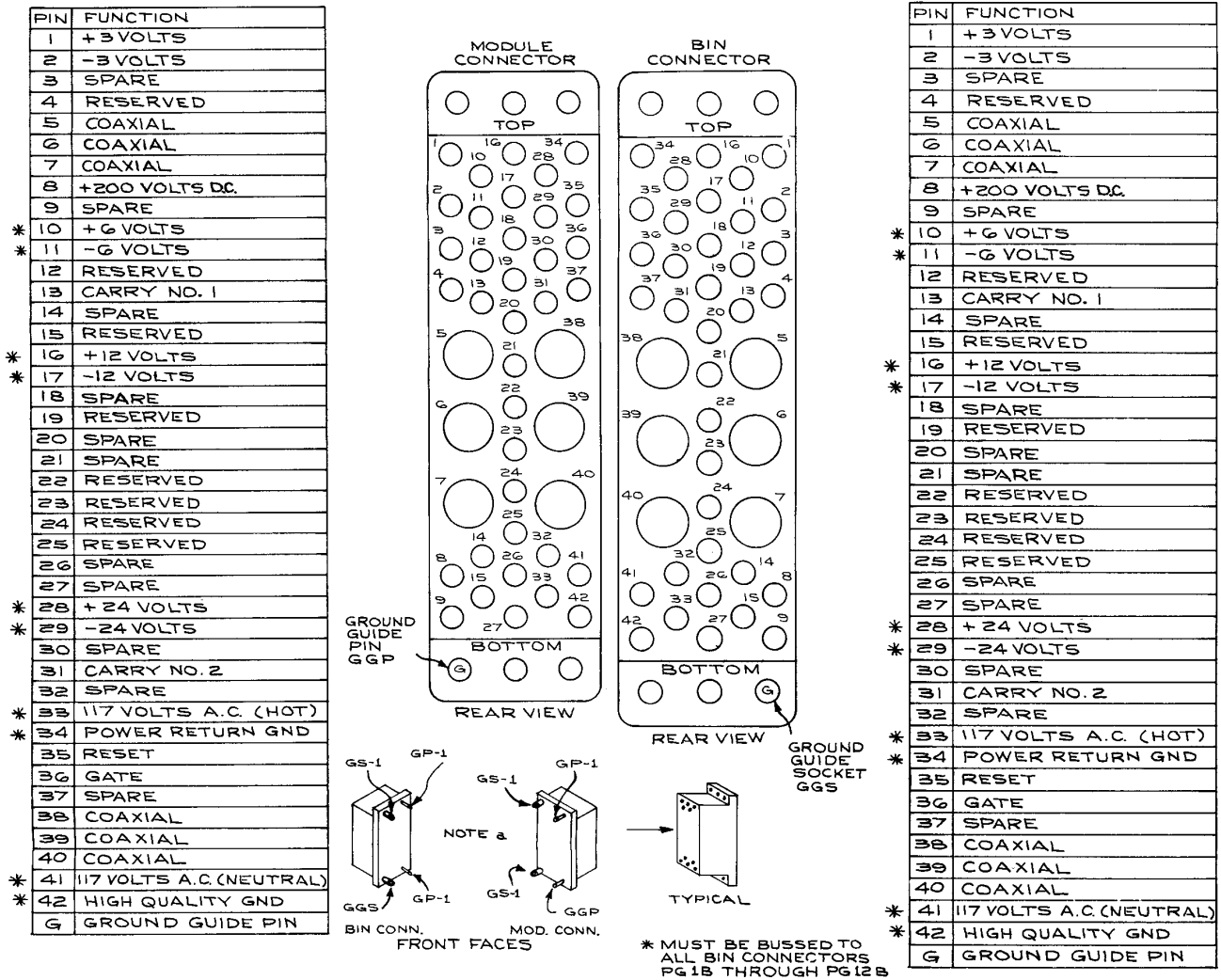
Nuclear instrumentation module systems are limited in complexity to hard-wired interconnections between modules. This becomes cumbersome when there are several modules that must be linked together. To circumvent the limitations of NIM systems, the CAMAC system was developed. This system retains the use of individual modules that can be plugged into a central power supply unit – in this case, a *crate*. The CAMAC crate, however, has a large number of additional lines besides those for the power.

The additional lines form the *data bus*. The bus has 24 *read* and 24 *write* lines, 4 station *subaddress* lines, and 5

station *function* lines – each station has its own identification line and *look-at-me* (LAM) line. Common to all stations are two *timing* lines, three *status* lines, and three *control* lines. These lines are connected to the control unit (crate controller). Also connected to the controller are the lines from the computer, generally divided into a set of data bits and address bits. The controller interprets the signals from the computer and sends appropriate signals to the individual CAMAC units or *stations*. This is shown schematically in Figure 6.111, with the line designations given in Table 6.39. With each of the units under digital control, it is possible, through the use of a computer connected to the crate controller, to have systems capable of extremely complex logic and control functions. Data can be collected by the computer and stored, displayed, and manipulated with a degree of complexity limited only by the programming language used by the computer.

The CAMAC system offers hardware and software compatibility among individual units. Programming can be a formidable task because the information transferred between the crate controller and the computer is a binary machine-language code that has no relation to the high-level languages used for scientific calculations. An alternative is to write the codes in a high-level language, such as FORTRAN, and at appropriate points in the program call a machine-language subroutine that converts the FORTRAN variables to the bit code necessary for control of the crate modules. When done in this way, machine-language programming can be relatively simple.

A CAMAC system can represent a considerable investment in money and in the time necessary to learn the required programming techniques. The need for a computer to run the system properly – and the computer peripherals such as terminals, printers, and mass storage units – are additional expenses. The advantage of the CAMAC system is the complete flexibility that it provides. As in NIM systems, the individual modules are not costly because there is no need for internal power supplies or complex logic circuits. There are now several manufacturers offering lines of CAMAC crates, controllers, and individual function modules. There are also NIM–CAMAC adapters that permit one to run certain NIM modules under CAMAC control.



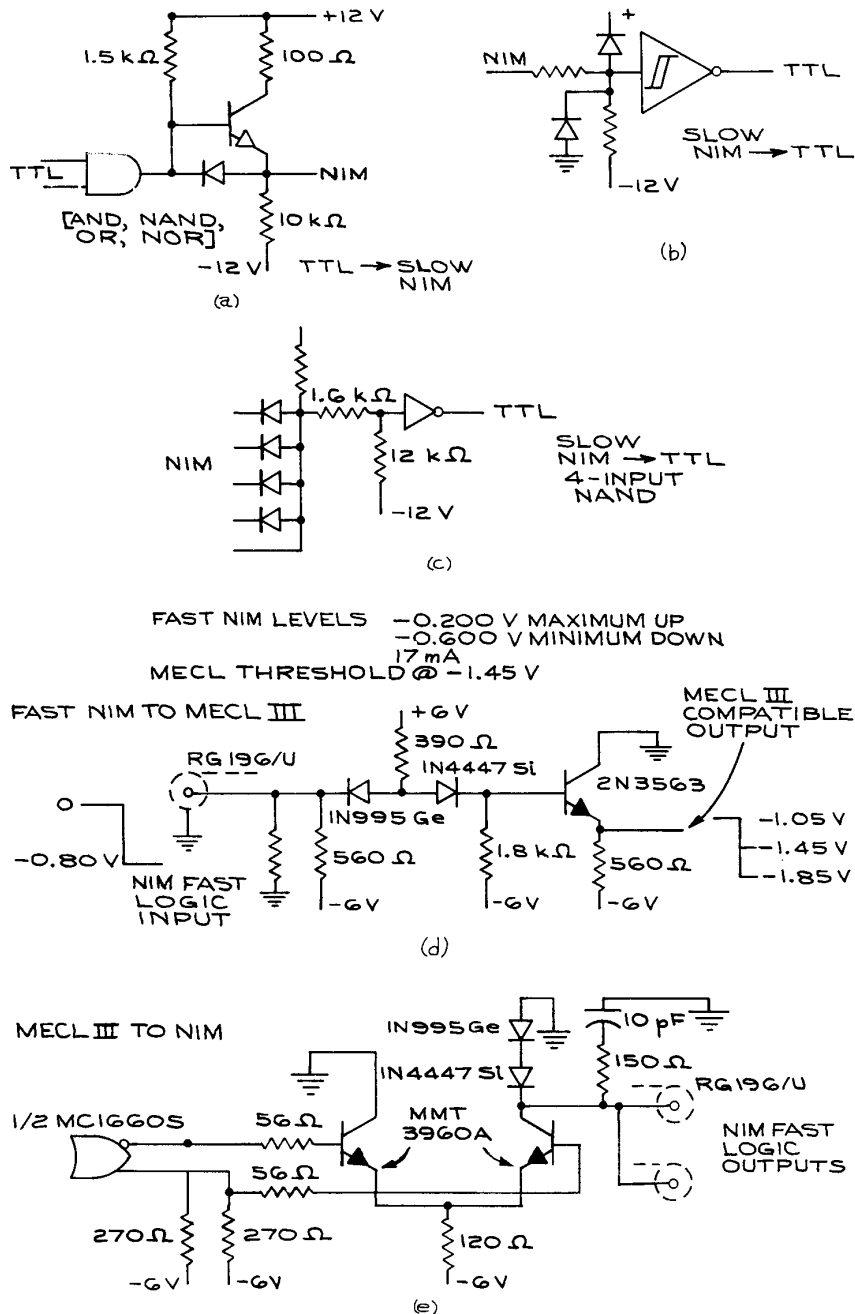
**Figure 6.109** NIM connector pin assignment (GP = guide pin, GGP = ground guide pin, GS-1 = guide socket, GGS = ground guide socket).

### 6.7.8 Control Systems

Control theory establishes a framework that enables one to optimize an active system for accuracy, stability, transient response, resistance to external perturbations, and internal component changes. The method of analysis uses the Laplace transform formalism in a way that is similar to that used in circuit theory (see Section 6.1.5). Control

theory is an established engineering subject and there are many excellent texts available. The purpose of this section is to give a brief introduction to the subject and provide general information on the analysis of systems and the proper configurations for accomplishing given tasks.

All control systems, whether electrical, mechanical, hydraulic, pneumatic, optical, thermal, acoustical, or



**Figure 6.110** (a) NIM-TTL and (b) NIM-ECL conversion circuits (fast NIM levels 17 mA,  $-0.200$  V maximum high,  $-0.600$  V minimum low; MECL threshold  $-1.45$  V. (From R. F. Nagel, NIM Fast Logic Modules Utilizing MECL III Integrated Circuits, *IEEE Trans. Nucl. Sci.*, **NS-19**, 520, 1972. Copyright 1972 IEEE.)

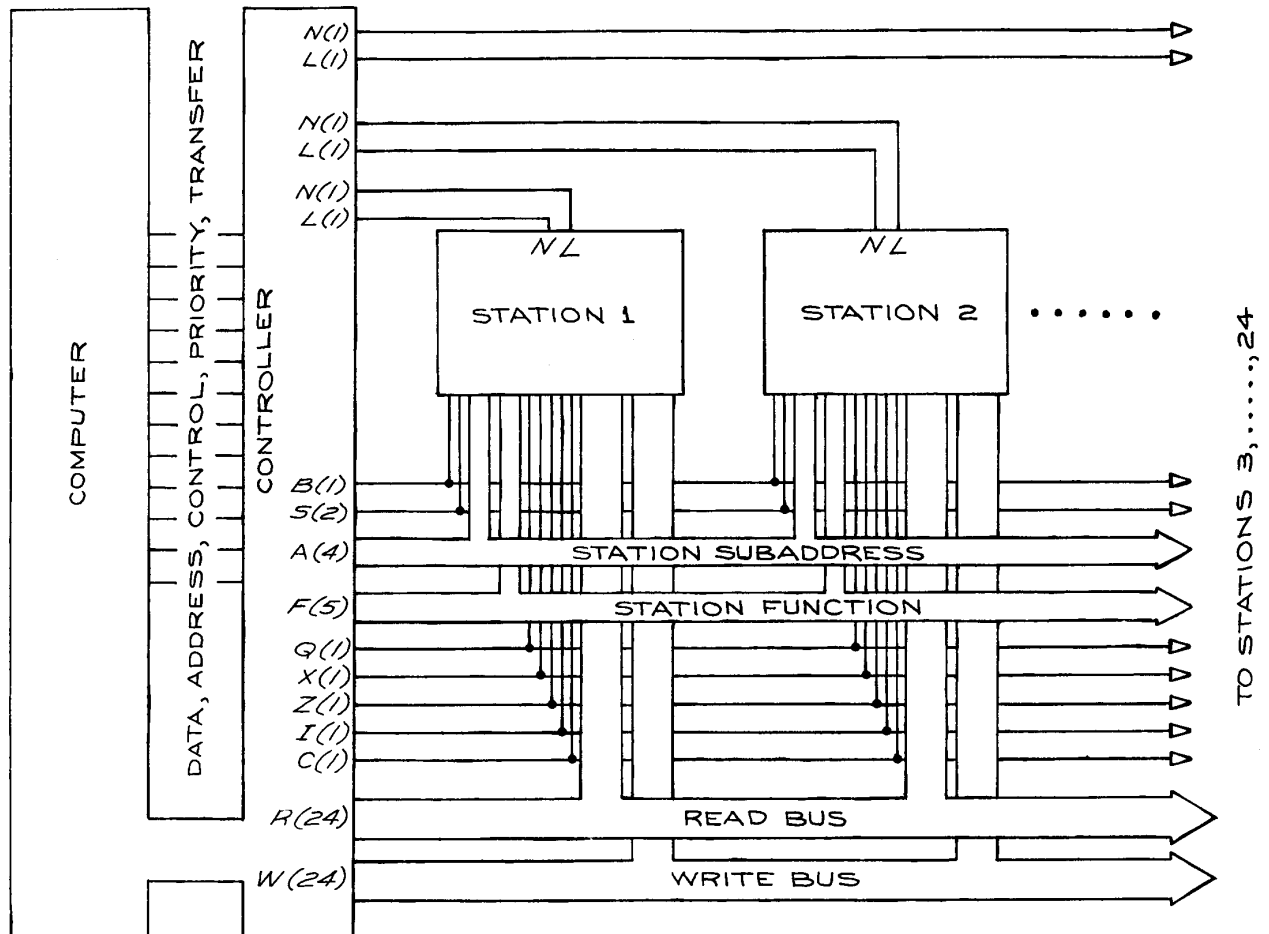


Figure 6.111 The CAMAC bus.

electromechanical can be represented in terms of one of the two block diagrams shown in Figure 6.112.

The first system is *open loop* where there is no signal path from the output back to the input. An example is the burner on the top of a stove. Once the position of the burner switch or valve is set, there is no further control over the burner output. The second system is *closed loop*, in which a portion of the sampled output is sent back to the input and compared with the input signal. An example of a closed-loop system is the oven of a stove. The temperature is set by an external input and this input is compared to a signal proportional to the temperature of the oven. If the

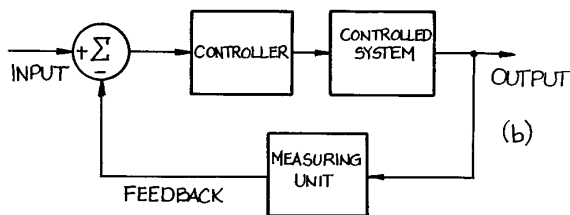
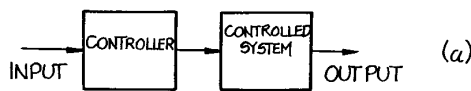
oven temperature exceeds the set temperature, the oven heat source is turned off; if the oven temperature is less than the set temperature, the source is turned on.

The standard method for testing an open-loop system is to apply signals at the inputs of each stage, starting with the last and proceeding to the first – verifying the input and output signal at each point. This sequential testing method cannot be used for a closed-loop system because the input to each stage is related to the output of the overall system.

All closed-loop systems have an input, an output, a summing or comparison unit, a controller, the working unit that is controlled, a measuring or sampling unit, and a feedback

**Table 6.39 CAMAC line designations**

Line Code (no. of lines used)	Command	Function
N(1)	Station number	Selects station
A(4)	Subaddress	Selects section of station
F(5)	Function	Defines function
S(2)	Strobe	Controls phase of operation
R(24)	Read	Parallel data to module
W(24)	Write	Parallel data from module
L(1)	Look-at-me	Request for service
B(1)	Busy	Operation in progress
X(1)	Command accepted	Module ready
Q(1)	Response	Status
Z(1)	Initialize	Sets module
I(1)	Inhibit	Disables module
C(1)	Clear	Clears registers
P(5)	Extra lines	
+24V } +6V }		Power



**Figure 6.112** Control system block diagram:  
(a) open loop, (b) closed loop.

path. Once mathematical models for each of the units in the system are developed and combined appropriately, the overall performance of the system can be evaluated and optimized by varying the parameters of the model.

Each unit of a control system can be characterized by a transfer or gain function. These functions describe how the

input to the unit is transformed by it and converted to the output. The individual transfer functions themselves are, in general, the solutions of ordinary, linear differential equations that are most easily expressed in terms of Laplace transforms (see Section 6.1.5). The full representation of the system in terms of Laplace transforms is particularly easy to assess with respect to stability, accuracy, and transient response.

Consider the general system shown in Figure 6.113 consisting of an input variable  $r(t)$ , output variable  $c(t)$ , and transfer functions  $g_1(t)$ ,  $g_2(t)$  and  $\beta(t)$ . The Laplace transforms of these functions are  $\bar{R}(s)$ ,  $\bar{G}_1(s)$ ,  $\bar{G}_2(s)$ ,  $\bar{B}(s)$ , and  $\bar{C}(s)$  respectively. The transfer function for the overall system is:

$$g(t) = \frac{g_1(t)g_2(t)}{1 + \beta g_1(t)g_2(t)} \quad (6.52)$$

with the equivalent Laplace transform:

$$\bar{G}(s) = \frac{\bar{G}_1(s)\bar{G}_2(s)}{1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)} \quad (6.53)$$

$\bar{G}(s)$  can be written as the ratio of two polynomials in  $s$ :

$$\bar{G}(s) = \frac{P(s)}{Q(s)} \quad (6.54)$$

The roots of  $P(s)$  give the zeros of the transfer function, and the roots of  $Q(s)$  give the poles of the transfer function. The system is stable if the poles of the transfer function lie in the left half of the complex  $s$ -plane.

## System Response

(i) **Sensitivity.** Sensitivity  $S$  is a measure of the fractional change in  $\bar{G}(s)$  with a fractional change in any part of the system represented by the variable  $b$ . The smaller the  $S$ , the less sensitive the system is to changes in any element of the system:

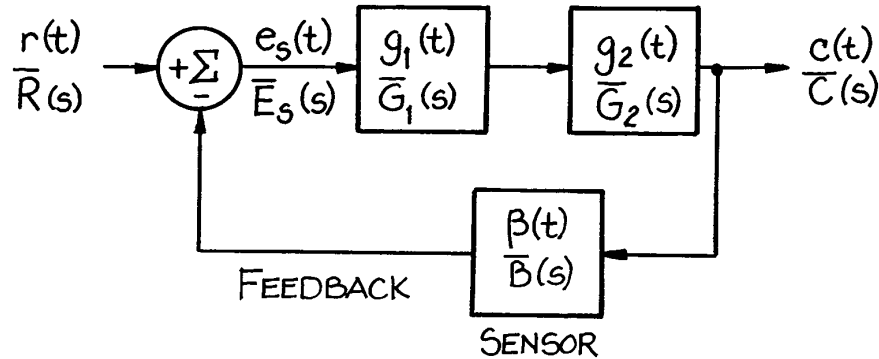
$$S = \frac{\Delta \bar{G}(s)/\bar{G}(s)}{\Delta b/b} \quad (6.55)$$

As examples,  $S$  will be calculated for variations in the controller transfer function  $\bar{G}_1(s)$  and for the feedback,  $\bar{B}(s)$ :

**Case 1:** Relative change in  $\bar{G}_1(s)$ ;

$$S = \frac{1}{1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)} \quad (6.56)$$





**Figure 6.113** Block diagram of a generalized closed-loop control system with schematic representations of the transfer functions and their Laplace transforms.

**Case 2:** Relative change in  $\bar{B}(s)$ ;

$$S = \frac{-\bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)}{1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)} \quad (6.57)$$

For Case 1,  $S$  is small when  $\bar{B}(s)$ ,  $\bar{G}_1(s)$ , and  $\bar{G}_2(s)$  are large. For Case 2,  $S$  is small only if  $\bar{B}(s)$ ,  $\bar{G}_1(s)$ , and  $\bar{G}_2(s)$  are small, so a system cannot simultaneously be made insensitive to changes in the transfer function of the controlled unit and transfer function of the feedback element. The solution is to use a high-quality feedback element that remains stable over a wide variety of conditions and to increase the transfer functions of the other system elements in order to stabilize the overall system with regard to their changes.

**(ii) Steady-state accuracy.** The steady-state accuracy of a system  $e_{ss}(t)$  is the difference between the input signal  $r(t)$  and the closed-loop signal from the output  $\beta c(t)$  taken in the limit as  $t \rightarrow \infty$ . In terms of Laplace transforms, the difference between the input signal and the closed loop signal from the output is:

$$\frac{\bar{R}(s)}{\bar{B}(s)[1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)]} \quad (6.58)$$

In  $s$ -space, the limit as  $t \rightarrow \infty$  is obtained by multiplying:

$$\frac{\bar{R}(s)}{\bar{B}(s)[1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)]} \quad (6.59)$$

by  $s$  and taking the limit as  $s \rightarrow 0$ :

$$\bar{E}_{ss} = \lim_{s \rightarrow 0} \frac{\bar{R}(s)}{\bar{B}(s)[1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)]} \quad (6.60)$$

To evaluate  $\bar{E}_{ss}$ , it is necessary to specify the signal  $\bar{R}(s)$ . For a unit step input  $\bar{R}(s) = 1/s$  (see Table 6.2) so that

$$\bar{E}_{ss} = \lim_{s \rightarrow 0} \frac{\bar{R}(s)}{\bar{B}(s)[1 + \bar{B}(s)\bar{G}_1(s)\bar{G}_2(s)]} \quad (6.61)$$

For simplification assume a unity feedback system so that  $\bar{B}(s) = 1$ . If then  $\bar{G}_1(s)\bar{G}_2(s)$  is of the form  $F(s)/s^N Q(s)$ , where  $F(s)$  and  $Q(s)$  are polynomials in  $s$  and  $N \geq 1$ ,  $\bar{E}_{ss}$  will approach zero and the output will approach the input.

**(iii) Proportional, Integral, and Differential (PID) Control.** The most general type of controller is one that provides an output that is the sum of three terms – one proportional to the difference between the reference signal and controlled signal, one proportional to the time integral of the difference, and one proportional to the time differential of the difference. These *PID* controllers use operational amplifiers configured as direct gain elements,

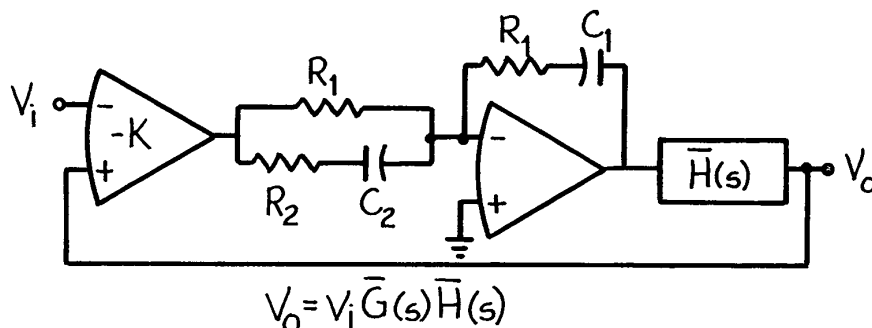


Figure 6.114 PID controller circuit.

integrators, and differentiators to attain these functions. The Laplace transforms of the three functions give a transfer function that has a form that depends on the details of the operational amplifier circuits. As an example, consider the circuit in Figure 6.114. It has the transfer function:

$$\bar{G}(s) = \frac{K(1 + sR_1C_1)^2}{sR_1C_1(1 + sR_2C_2)} \quad (6.62)$$

for  $R_1 + R_2/R_1 = C_1/C_2$ ,  $\bar{G}(s)$  can be rewritten as the sum of three terms:

$$\bar{G}(s) = K \left[ 1 + \frac{1}{sR_1C_1} + \left( \frac{R_1}{R_1 + R_2} \right) \frac{1 + sR_1C_1}{1 + sR_2C_2} \right] \quad (6.63)$$

where the first term is the *P*, the second the *I* and the last the *D*. The unit represented by the transfer function  $\bar{H}(s)$  has been added to limit the frequency response and increase the overall stability of the system. Commercial PID controllers allow one to adjust the absolute as well as the relative weights of each of the terms to attain the optimum overall system response. This generally means a response that has the maximum bandwidth consistent with a phase margin of approximately  $60^\circ$ .

### 6.7.9 Personal Computer (PC) Control of Experiments

Computer operation of experiments is best suited to routine measurements on systems with a high degree of reliability –

the latter condition is not always met for ordinary laboratory experiments. It is not worthwhile to add a computer to an experiment that is intended to make a limited number of specialized measurements. Though it may be tedious, it is often much more efficient to take measurements by hand rather than purchase specialized measuring instruments, a computer or a computer interface, and software, connect them together, write the necessary software, and place the system into operation. On the other hand, routine measurements in an experiment that continually produces large amounts of data are well suited to computer operation. Before considering computer control it is important to completely understand the experiment. Special attention should be given to the range of variables to be measured and controlled, the response time of the system, and the accuracy and precision of the quantities that are measured and controlled.

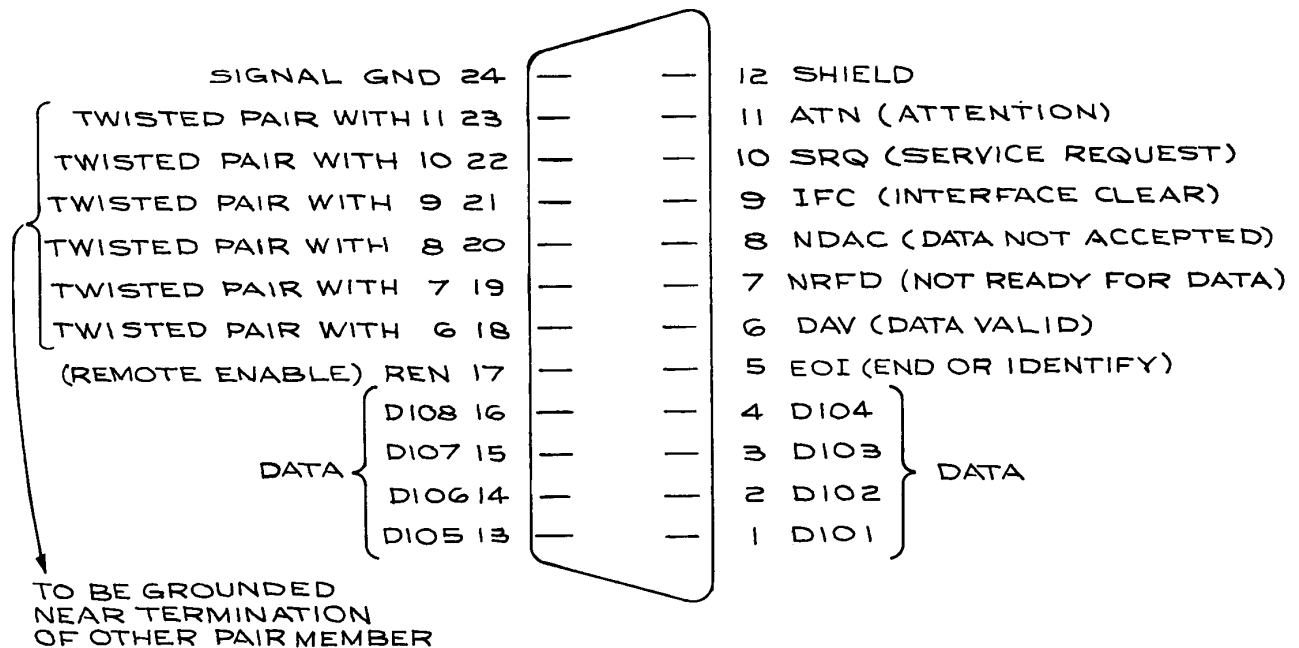
Two classes of computer-based measurement and control systems will be discussed. The first is based on the IEEE 488 or GPIB bus and requires measuring and control instruments compatible with the bus as well as the computer, interface circuit, and software. The second type of system is based on an interface circuit that either fits into one of the computer expansion slots, or resides in a separate chassis and is connected to the computer through a standard data transmission link like the RS232 serial interface. With the IEEE 488 system, sensors and transducers are connected to individual instruments, which, in turn, are connected to the IEEE 488 bus. In the second type of system, by contrast, sensors and transducers are directly connected to the interface board via screw terminals or coaxial connectors. In both systems, software is needed for the host computer.

It is possible to purchase an IEEE 488 interface card for a PC that will effectively control up to 12 IEEE 488-compatible instruments, provided that there are no restrictive timing or synchronization requirements. Within the IEEE 488 system, all signal conditioning and digital-to-analog conversion is done within the individual instruments attached to the bus. The IEEE 488 bus is separate from the bus structure of the control computer, and communication is entirely through the interface. Adapters for converting from the IEEE 488 bus to the RS 232C interface are readily available. IEEE 488 data paths are 8 bits (1 byte) wide, and a complete data transmission consists of two serial 8-bit bytes. Each unit connected to the bus is autonomous, with its own power supply.

Connections to the bus are through standard 24-pin connectors (see Figure 6.115) arranged so that they can be stacked one on top of the other and connected in a daisy-chain configuration. There can be up to 15 interconnected devices on a single bus, but the total transmission path length cannot exceed 20 m. Instruments connected to the

bus are designated as *talkers*, *listeners*, and *controllers*. A tape reader is a talker, for example, while a signal generator is a listener. Some instruments may be able to talk *and* listen, or even (like a variable-range frequency counter) talk, listen, *and* control. The bus itself is passive, and all the circuitry enabling an instrument to talk, listen, or control is contained within the instrument itself. Simple systems built around the bus need not even use a controller, but can have just one talker and one listener, as long as the instruments have the built-in interface circuitry that allows their functions to be assigned under some form of local control. The more usual configuration with a personal computer uses interface messages to assign the function of talker or listener to instruments on the bus.

For PC-based IEEE 488 systems, there is variation among the software offered by the different manufacturers. Three criteria for choosing software are ease of use, flexibility, and speed. Software that is easy to use is generally slow and difficult to adapt to special tasks. For specialized applications, software must have many options and a

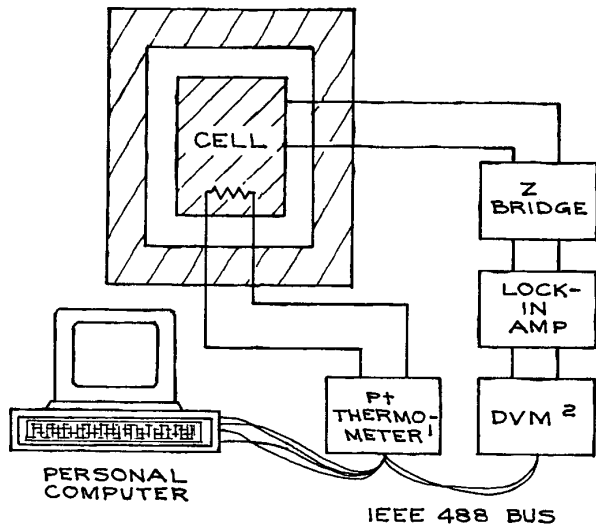


**Figure 6.115** IEEE 488 interface-bus connector pin assignment. Logic levels are TTL-compatible negative logic ("1", 0 to 0.4 V; "0", 2.5 to 5.0 V).

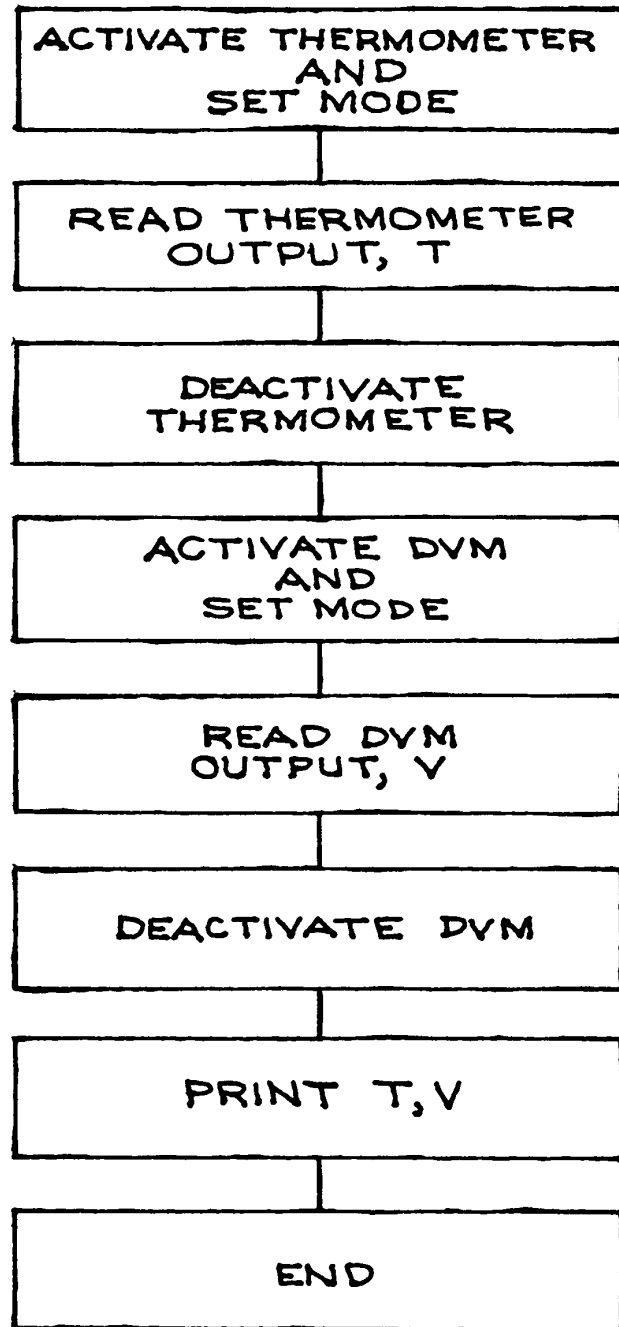
variety of functions. This requires that the experimenter spend additional time organizing and assembling the final program. If programs are to be fast they must be written in assembly language – a time-consuming task.

A simple PC-controlled experiment for measuring the dielectric constant of a liquid as a function of temperature is shown in Figure 6.116. In this experiment, a digital impedance bridge, lock-in amplifier, and digital voltmeter are connected to a computer via an IEEE 488 bus. The dielectric constant and conductivity of the liquid are calculated from the measurement of the capacitance and resistance of a liquid-filled cell. Measurements are made over a range of temperatures, and a platinum resistance thermometer is used to measure the thermostat temperature. Programming the IEEE 488 bus usually requires proficiency in a high-level language such as BASIC and complete knowledge of the operation of each instrument on the bus and its device-dependent commands. For the system in Figure 6.116, timing is not a consideration and the measurement sequence can follow the simple flow chart in Figure 6.117.

Implementing the flow chart with the IEEE 488 bus requires knowledge of the characteristics of the instruments on the bus. The thermometer is a *talker*, and the



**Figure 6.116** A personal-computer-controlled IEEE 488 temperature-measuring system.



**Figure 6.117** Flow chart for measurement sequence in system of Figure 6.116

digital voltmeter is a *listener* when its ranges are being set, and a *talker* when sending the voltage readings to the computer. To begin, the address of each instrument is set with switches on their back panels. The instruments are wired in daisy-chain fashion to the computer with ribbon cable and the IEEE 488 piggyback plugs. Sufficient time must be provided in the program for instrument readings to stabilize. This is done with *wait* loops inserted in the program (see Figure 6.118). The input and output data formats for each instrument must also be accommodated in the program so that the commands and data from the computer are correctly read by the instruments, and the data from the instruments correctly read and stored by the computer. Even for this relatively simple system, the BASIC program requires several lines of code. For more complex systems with special timing requirements, simple BASIC programs may be too slow or too inflexible. When this is the case, systems with separate software packages are required. Such systems are offered by Keithley Instruments and National Instruments, among others.

Data acquisition and control systems based on I/O modules and boards that connect directly to a personal computer are justified when there are a large number of standard sensors and the cost of separate IEEE 488 instruments is too great. For an I/O board system, it is necessary

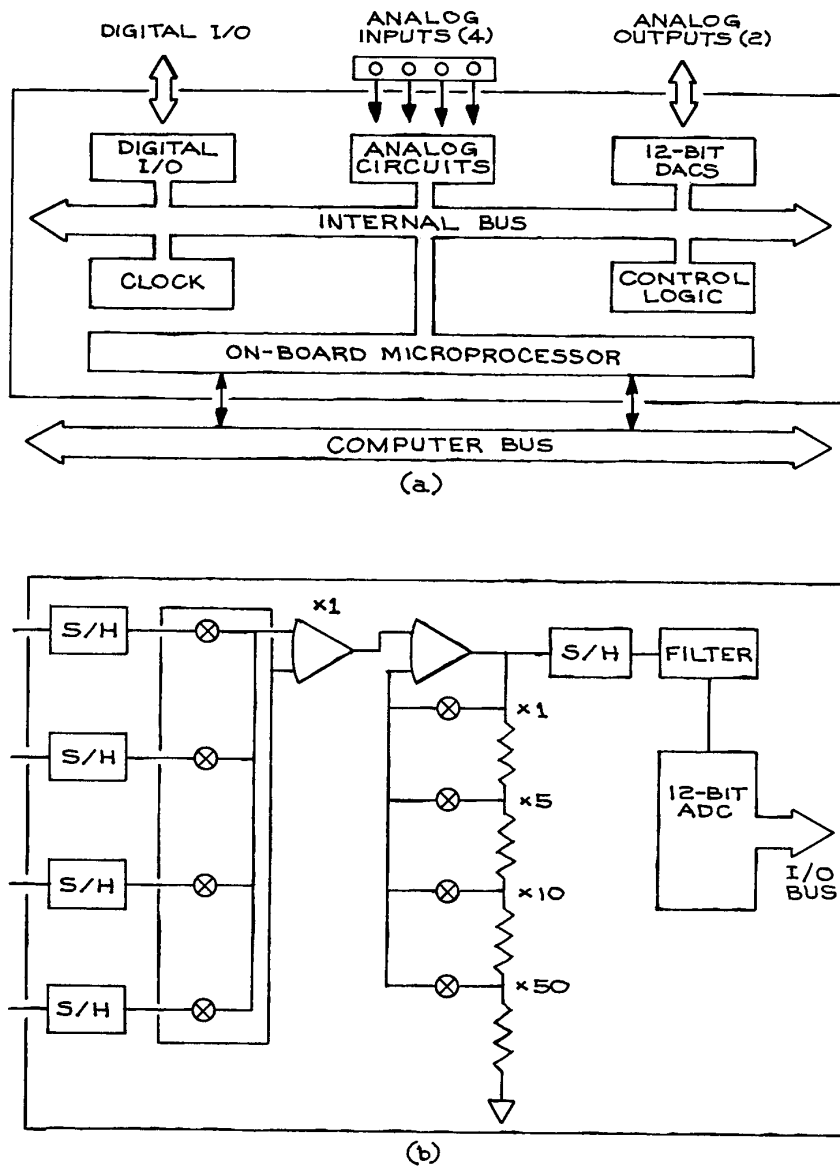
to consider the number of input channels, the number of output channels, input and common-mode signal levels, sampling rates, measurement precision, and computer and software requirements.<sup>9</sup>

Interface boards can be programmed directly in BASIC once the addresses of all the I/O ports are known. As with all dynamic systems, timing is crucial, and routines must be written to allow time for ADC conversions and data storage. Interpreted BASIC programs are slow, while compiled BASIC has a considerable speed advantage. Optimum speed and ease of programming are usually accomplished with software supplied by the manufacturers specifically for their interface boards. Such software has machine language subroutines for initializing the ports and reading and writing data. These routines are designed to be easy to use and can generally be called from within a BASIC program. The more sophisticated manufacturer-supplied software is menu driven, with true foreground – background computer operation and graphic data display with full interfacing to standard spreadsheet packages. Since much more time is spent with software than hardware, it is important that the programming materials supplied by the manufacturer be clear, easy to use, and well documented.

A typical small, commercial, data-acquisition interface circuit is shown in Figure 6.119(a). The details of analog

10 OPEN 7,10	Thermometer assigned device number 10
20 PRINT #7, "CR"	Set thermometer to Celsius, continuous reading
30 FOR I=1 TO 1000: NEXT I	Wait loop
40 INPUT #7, TC\$	Read thermometer output
50 FOR I=1 TO 2000: NEXT I	Wait loop
60 TC=VAL(MID\$(TC\$))*10E-4	Scale the output data
70 CLOSE 7	Deactivate thermometer
80 OPEN 8,22	DVM assigned device number 22
90 A\$="F1R7T1M3A1H1D0"	Set DVM to DC volts, auto range, internal
100 PRINT #8, A\$	trigger, math "off", autocal "off", high
	resolution "on", data ready request "off"
110 FOR I=1 TO 2000: NEXT I	Wait loop
120 INPUT #8, B\$	Read DVM
130 FOR I=1 TO 5000: NEXT I	Wait loop
140 CLOSE 8	Deactivate DVM
150 VC=VAL(B\$)	Convert B\$ to a number
160 PRINT "TC="; TC; " VC="; VC	Output data
170 END	

Figure 6.118 IEEE-488 BASIC program example.



**Figure 6.119** (a) Data-acquisition interface; (b) analog input circuits.

input circuits are shown in Figure 6.119(b). In this example, there are four single-ended or two differential analog inputs, each connected to a S/H circuit and analog multiplexer. The output of the multiplexer is connected to a unit-gain instrumentation amplifier, a programmable-gain

amplifier, and a second S/H circuit. The output of the second S/H is connected to a 12-bit ADC through a signal-conditioning network. The ADC output goes to the internal interface bus. Boards of this type plug directly into the expansion slot of a personal computer.

Connections to the input circuits are made through the screw terminals on the board.

Sample and hold circuits capture the analog input signals and hold for a time sufficient for the ADC to convert the analog signal to a binary output. The operation of the S/H has already been discussed. Of particular importance are the *acquisition time*, the *aperture time*, the *decay time*, and the *slew rate* (the maximum large-signal change that the S/H circuit can produce).

Input circuits can typically accommodate full-scale input voltages from 20 mV to 10 V. The number of samples per second is from 5000 to 10 000. With a 12-bit ADC, the system has a maximum resolution of 1 part in 4000, or 0.025%. In practice, this resolution cannot be realized for low-level input signals because of amplifier zero and gain drifts, noise, common-mode error signals, and ADC nonlinearities. These must all be taken into account and an *error budget* established when assessing the final accuracy of the system.

Important analog output specifications for instrument control are the resolution and the settling time required to achieve the quoted resolution. Since the analog output voltages are the input signals to other devices, output voltage and current ranges are important. The digital input and output ports are usually 8-bits wide and operate with TTL levels.

The primary concern when connecting sensors to a data-acquisition system is elimination of noise. Noise can be generated on the signal lines from magnetic fields, electric fields, thermal gradients, friction, and improper grounding. These noise sources can be controlled by proper choice of input and output impedances and of lead-wire composition, length, shielding, and placement.

High output-impedance, current-source transducers are less sensitive to magnetically induced noise than low output-impedance, voltage-source transducers. It is important to distinguish between ground and common (or return) lines in transducer circuits. Ground lines do not normally carry current, but return lines do. Return lines carry the signal currents that complete the transducer circuit. Operational amplifiers normally have a single output terminal and terminals for positive and negative power-supply voltages. There is no separate common terminal, and the return current must flow in the power supply lines. For this reason, it is important to

bypass the power-supply leads with large, high-frequency ceramic disk capacitors located as close to the amplifier as possible.

When a return line is connected to ground at more than one point in the circuit, it is possible to develop a voltage across the line due to small differences in ground potentials and the finite resistance of the return line. Grounding the signal return line at a single point in the circuit changes a single-ended transducer configuration to *differential*, and any difference in ground potentials appears as a common-mode voltage. When it is impossible to avoid multiple grounds in a system, an isolator can be used between the transducer and the input to the data-acquisition circuit. Because there is no galvanic connection across an isolator, potential differences between grounds cannot result in current flow.

Return lines should have the lowest possible inductance and should overlap the signal lines as much as possible. Shielded, twisted-pair cable is generally best for transducer connections. The shielding does not carry a current and serves to reduce electric and magnetic noise pickup. With twisted-pair cable, pickup is induced equally in both wires and can be canceled by an amplifier with a differential input. Twisted-pair cables in several sizes and configurations are available from Belden and Alpha. Cable-length guidelines are given in Table 6.40. In coaxial cable, the outer conductor is the return line.

A smaller source of noise is static discharges when insulators of dissimilar materials rub against each other. This is called *triboelectric induction* and can be eliminated with special low-noise cable that has graphite lubricant between dielectric surfaces. The ultimate obstacle to complete noise elimination comes from thermoelectric effects when junctions of dissimilar metals are at different temperatures.

**Table 6.40 Instrumentation cable length guidelines**

<i>Signal Source</i>	<i>Bandwidth (Hz)</i>	<i>Accuracy (%)</i>	<i>Maximum Cable Length<sup>a</sup></i>	
4–20 mA	<10	0.5	1000–5000	300–500
±1–±10 V	<10	0.5	50–300	15–90
±10 mV–±1 V	<10	0.5	5–100	1.5–30

<sup>a</sup> Shielded twisted pairs.

Typical values are  $0.1 \mu\text{V}/^\circ\text{C}$ . Thermoelectric solder (70% cadmium, 30% tin) can be used for low-level circuits, and precautions should be taken to keep all circuit elements and devices at the same temperature. Carbon and metal-film resistors produce larger thermoelectric voltages than precision wire-wound resistors.

Once the sources of noise in the system have been reduced to acceptable levels, input filtering can be used to improve the signal-to-noise ratio. Filtering cannot, however, eliminate the effects of improper wiring.

**Example of Sensor and Control Electronics.** In this example, a silicon substrate inside an ultrahigh vacuum chamber is heated to a predetermined temperature and then maintained at that temperature while an electron gun directs electrons at the surface and a mass spectrometer monitors the masses of the atoms and molecules emitted from the surface. An ion gauge monitors the pressure in the experimental chamber. The data from the experiment are count rates as a function of mass spectrometer setting. The rates are recorded and stored for subsequent analysis. A schematic representation of the system is given in Figure 6.120. The system is a mixed digital and analog system with a computer supplying control and data-acquisition functions and the heater and thermocouple circuits operating with analog signals. D/A and A/D circuits provide the translation between the digital and analog domains.

It is most convenient to divide the overall system into subsystems and determine the operating conditions for each of them. In the example, the subsystems are:

- Substrate/heater
- Thermocouple temperature monitor
- DMM
- Picoammeter current monitor
- Mass spectrometer
- Electron gun
- Ion gauge.

The substrate/heater and thermocouple temperature monitor form the temperature control system that both sends and receives signals to and from the computer. The electron gun receives signals from the computer while the DMM and ion gauge send signals. The mass spectrometer both receives and sends signals. The signals it receives set

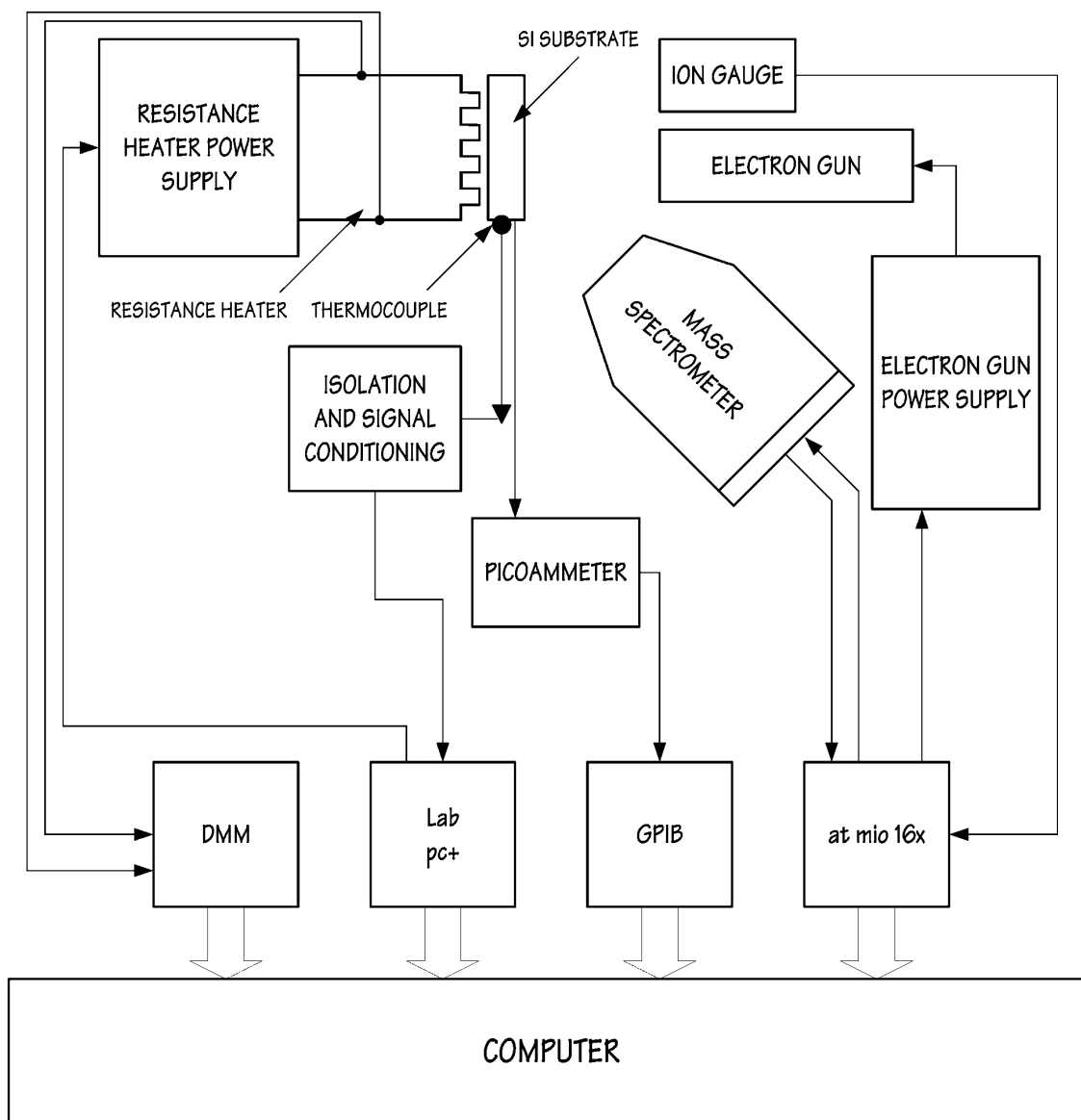
the mass range and scanning rate. The signals it sends are pulses from the detector.

There are a number of manufacturers of computer interface boards that are specifically designed for the control and monitoring of experiments – among them are National Instruments (manufacturer of LabVIEW), Hewlett-Packard (Agilent), Keithley Instruments, and Dattel. All have catalogs that describe their products in detail and provide very useful information on the general configuration of data-acquisition and control systems. The purpose here is to provide a plan of organization and guidelines.

The first task is to define the operating parameters for each of the subsystems. For the substrate/heater, this means defining the input power necessary to obtain the required temperature of the silicon substrate. Once the power is known, it is possible to define the power-supply requirements and its programming input requirements. For this example, the heater is a coil of tungsten wire that has a resistance of  $3 \Omega$  at room temperature. The required maximum power was determined to be 81 W (27 V at 3 A). This is supplied by a 300 W power supply with a range of 0 to 30 V at 0 to 10 A. The power supply is voltage programmable with variable gain, so that an input voltage from 0 to 10 V will vary the output from 0 to 30 V when the internal gain of the power supply is set to three. The outputs of D/A cards used for power supply control are generally 0 to 1 and 0 to 10 V. When this is not sufficient to drive the output of the power supply to the required maximum level, an intermediate gain stage will be required. The supply that is used is current limiting. For this application, the current-limiting control is set so that 3 A are available at all voltage settings. This prevents excessive current draw when the filament is cold – a necessary precaution for tungsten that has a temperature coefficient between  $4.5$  and  $8.9 \times 10^{-3}$  per  $^\circ\text{C}$ .

Because the programming voltage to the heater power supply is in the range 0 to 10 V, a computer interface card with a DAC is used that has a maximum analog output voltage of 10 V. The input to the card is a 12-bit code giving a maximum possible resolution of one part in 4096 or 0.025% at 30 V output. At 27 V, the resolution is 0.0225% slightly exceeding the maximum available resolution. The thermocouple is a type K NiCr/NiAl, with a working temperature range of  $-200$  to  $1300 \text{ }^\circ\text{C}$ . It is

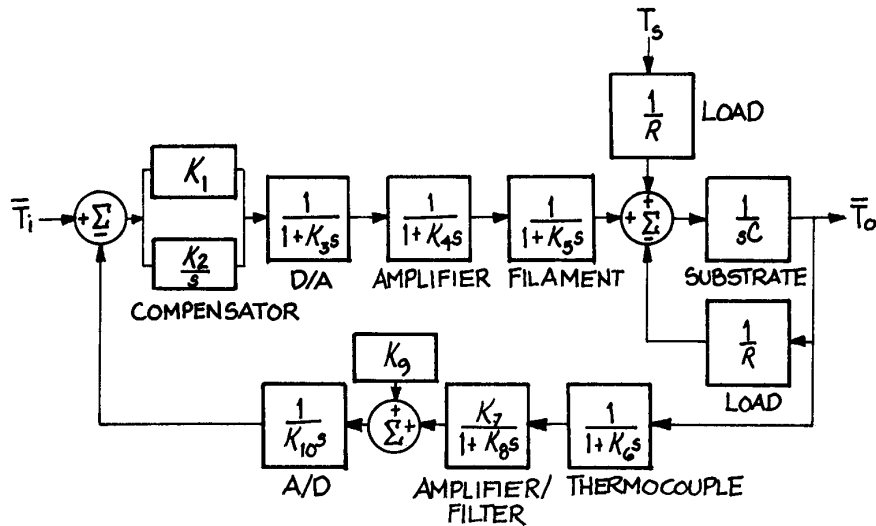




**Figure 6.120** System for controlling the temperature of a silicon substrate while monitoring the atoms and molecules emitted from the surface under electron bombardment.

non magnetic and compatible with the ultrahigh vacuum environment in which it is located. The thermocouple wires are crimped to a copper lug that is fixed to a stud on the stage holding the silicon substrate. The wires are

connected to the leads of a UHV thermocouple electrical feedthrough. The leads on the atmospheric pressure side of the feedthrough are connected to the signal conditioner/isolator unit that converts the output of the thermocouple to



**Figure 6.121** Block diagram of the silicon substrate-temperature-control system.

an analog voltage that is changed to a temperature by the software that interprets the output from the interface card. A block diagram of the substrate temperature control system is shown in Figure 6.121. In this Figure, the individual functions are represented by blocks. The Laplace transform of the transfer function for each block is indicated (see discussion of Laplace transforms in Section 6.1.5). The transforms are obtained from consideration of the differential equations that relate the output of each function to the input. The *compensator* is a proportional/integral type with constants  $K_1$  and  $K_2$  adjusted to attain the optimum response and stability of the circuit (see Section 6.7.8). One detail to be noted is that the output power of the filament amplifier must be linearly related to the difference between the set temperature  $T_i$  and the output temperature  $T_o$  for a Laplace transform analysis to be used.

**Determination of Block Parameters.** The time response of the D/A, amplifier, filament, thermocouple, amplifier/filter, and A/D are obtained from the specifications of the individual elements. The constants  $K_9$  and  $K_{10}$  were chosen so that the input to the A/D in the thermocouple circuit is 0 V at approximately  $-100^\circ\text{C}$  and 4 V at  $1000^\circ\text{C}$ . For the system under consideration:

$$K_3 = 1.2 \times 10^{-5} \text{ s}$$

$$\begin{aligned} K_4 &= 5 \times 10^{-5} \text{ s} \\ K_5 &= 1 \times 10^{-3} \text{ s} \\ K_6 &= 1 \times 10^{-3} \text{ s} \\ K_7 &= 0.3535 \text{ V} \\ K_8 &= 86.67 \text{ s} \\ K_9 &= 0.25 \text{ s} \\ K_{10} &= 1.2 \times 10^{-5} \text{ s} \end{aligned}$$

The remaining parameters – thermal resistance,  $R_T$ , and heat capacity  $C_T$ , are characteristic of the experimental arrangement and are determined experimentally from heating and cooling curves and a measurement of the maximum temperature at maximum power input. The assumption is made that the heat loss from the substrate to the surroundings is proportional to the temperature difference between the substrate and surroundings ( $T_o - T_s$ ). The initial slope of the heating curve  $dT_o/dt$  ( $T_o$  as a function of time for a constant heat input) when the substrate is at the temperature of the surroundings and the heat loss is zero is  $dT_o/dt = C_T dQ/dt$ , where  $C_T$  is the heat capacity of the substrate and  $dQ/dt$  is the rate at which heat is added to the substrate. Knowing  $dQ/dt$ , the input power, and  $dT_o/dt$ ,  $C_T$  can be calculated.  $R_T$  is correspondingly determined from the asymptotic temperature of the substrate for a constant power input. Assuming heat loss proportional to ( $T_o - T_s$ ), the thermal resistance of the system  $R_T$  is given

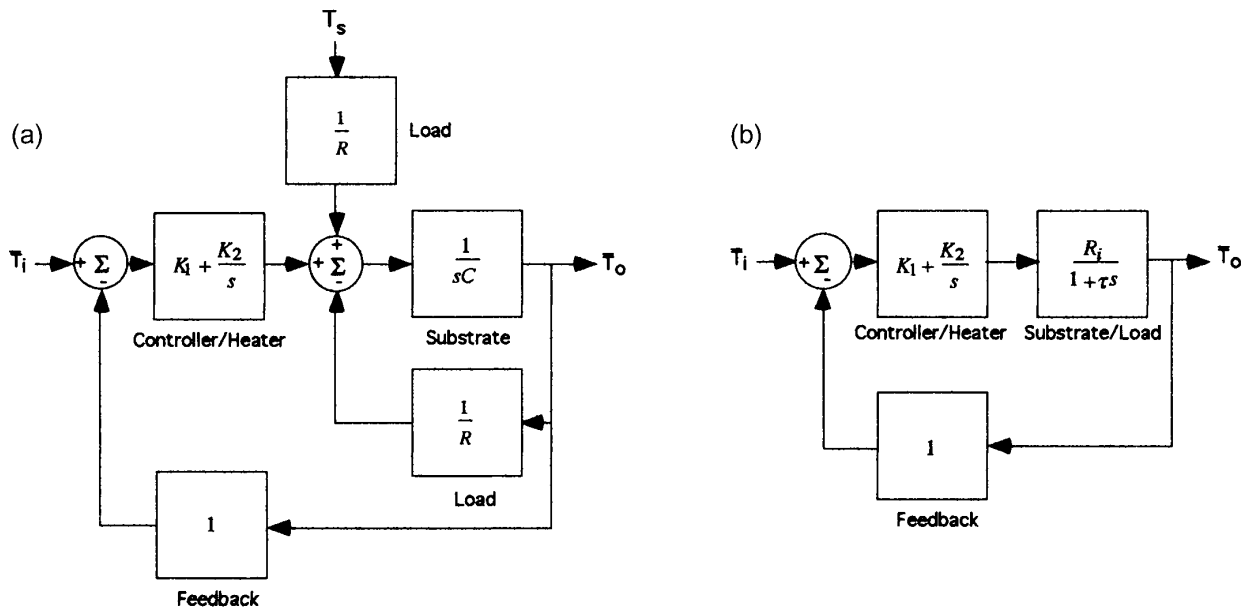
by  $(T_o - T_s)/\text{input power}$ . The values of  $C_T$  and  $R_T$  can be checked by a cooling-curve measurement. With the substrate initially at  $T_o > T_s$  and no heat input, the time for the temperature to fall to  $1/e$  of  $(T_o - T_s)$  is  $R_T C_T = 300$  s. This is consistent with the values for  $R_T$  and  $C_T$  of 5.6 K/W and 54 J/K.

Because the time constant for the substrate and load is at least two orders of magnitude larger than any other time constant in the system, it is possible to simplify the circuit to the one shown in Figure 6.122(a), where the overall transfer function of the D/A, amplifier, filament, and feedback transfer function are all set equal to one. In Figure 6.122(b), the substrate, load, and summing junction are replaced by a single transfer function  $R_i/(1 + \tau s)$ , where  $R_i$  is the effective thermal load equal to  $R_T + T_s/Q_i$ . In this case,  $\tau = R_T C_T = 300$  s,  $T_s$  is the temperature of the surroundings, and  $Q_i$  is the power input to the substrate at the set temperature  $T_o$ . It should be noted that the value of  $R_i$  will change with the temperature of the surroundings and the set temperature of the substrate. To

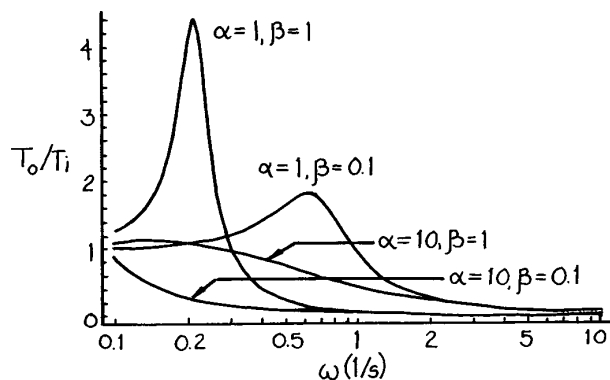
maintain the substrate at 500 K, the system requires a heat input of 36 W, giving a value of 13.9 K/W for  $R_i$ . The overall transfer function for Figure 6.122(b),  $\bar{T}_o(s)/\bar{T}_i(s)$ , is

$$\begin{aligned}\bar{T}_o(s)/\bar{T}_i(s) &= \frac{13.9(K_1 s + K_2)}{300s^2 + (13.9K_1 + 1)s + 13.9K_2} \\ &= \frac{\alpha s + 1}{21.6\beta s^2 + (0.072\beta + \alpha)s + 1}\end{aligned}\quad (6.65)$$

where  $\alpha = K_1/K_2$  and  $\beta = 1/K_2$ . The response and stability of the system are optimized by the appropriate choice of  $K_1$  and  $K_2$ . Figure 6.123 shows the frequency dependence of the system transfer function for different values of  $\alpha$  and  $\beta$ . This can be done empirically or by using standard control analysis methods – some of which are discussed in Section 6.7.8 and in more detail in the references at the end of the chapter. Programs such as Mathematica™ and Matlab™ have special packages for control analysis and design.



**Figure 6.122** (a) Simplified block diagram of the silicon-substrate-temperature control system; (b) block diagram with load and substrate transfer functions combined.



**Figure 6.123** Frequency dependence of steady-state transfer functions for a range of system parameter values. See text for definition of  $\alpha$  and  $\beta$ .

A DMM card is used to monitor the voltage across the leads of the heater and compare the results with the programmed voltage. The DMM is specified to have a resolution of 5 1/2 digits  $\pm 1$  in the next to least significant digit. When set on the 10 V range, the card can resolve 1 mV (0.01%) – more than adequate for verifying the output voltage of the power supply considering that the maximum resolution of the DAC is only 0.025 %.

The AT-MIO-16X interface board from National Instruments is supplied with software and connectors. It can be installed in any available 16-bit expansion slot on a PC. It has 16 single-ended and 8 differential analog input channels of 16 bits each, and a maximum sampling rate of 100 kilo-samples/s. Gains of 1, 2, 5, 10, 20, 50, and 100 are software-selectable. The output of the ion gauge is connected to one of the analog input channels. There are two analog output channels driven by 16-bit DACs. The two channels are used to control the electron-gun power supply and mass spectrometer. The maximum update rate is 100 kilo-samples/s and the output software-selectable voltage ranges are 0 to 10 V unipolar, and  $\pm 10$  V bipolar. The board also has eight digital TTL-compatible I/O lines and three independent 16-bit counter/timers. One of the counters is used to register the output counts from the mass spectrometer. A standard GPIB interface board is used to monitor the output of the picoammeter that registers the current from the electron gun that is collected by the silicon substrate.

## 6.8 EXTRACTION OF SIGNAL FROM NOISE

### 6.8.1 Signal-to-Noise Ratio

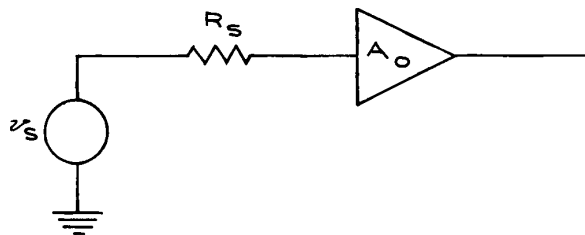
With many experiments, an important consideration is the minimum detectable signal that can be recovered. This depends on the *signal-to-noise ratio*<sup>10</sup> (SNR or S/N) rather than the absolute value of the signal. In counting experiments, where the minimum signal is one event, the uncertainty associated with  $C_s$  signal counts is  $\sqrt{C_s}$  so that the relative uncertainty is  $\sqrt{C_s}/C_s$  or  $1/\sqrt{C_s}$ . In principle, one could attain arbitrarily high precision by counting for long times, but this is often not practical because of instabilities in the experimental system.

With analog systems, where the signal is a voltage, current, or charge, the SNR depends on the source resistance and capacitance, shot noise,  $1/f$  or flicker noise, and amplifier noise. In this discussion, we assume that sources of noise such as r.f.i. do not exist. The noise will be assumed to be confined to the source and preamplifier and to represent the theoretical limit for such a combination.

Consider a detector that is a source of voltage with output impedance  $R_s$ , connected to an amplifier with a gain  $A_o$  as shown in Figure 6.124. *Johnson noise* is associated with the random motion of electrons in  $R_s$ . Its rms value over a frequency bandwidth  $\Delta f$  in Hz is:

$$v_{\text{Johnson}} = \sqrt{4kTR_s\Delta f} \quad (6.66)$$

where  $k$  is Boltzman's constant ( $1.38 \times 10^{-23}$  J/K), and  $T$  is the absolute temperature of the resistor. Johnson noise is frequency-independent – that is, the rms noise per unit bandwidth (noise density) is the same at all frequencies.



**Figure 6.124** Voltage source  $v_s$  with output impedance  $R_s$  connected to an amplifier of gain  $A_o$ .

For this reason, Johnson noise is often called *white noise*. For the purposes of circuit analysis, the noise from  $R_S$  can be represented by a voltage source giving a voltage  $v_{\text{Johnson}}$  in series with a noiseless resistor  $R$ . The noise current from this source is  $i_{\text{Johnson}} = v_{\text{Johnson}}/R$ .

*Shot noise* has its origin in the discrete nature of the electronic charge. The shot current noise accompanying a d.c. current  $I$  is

$$i_{\text{shot}} = \sqrt{2qi\Delta f} \quad (6.67)$$

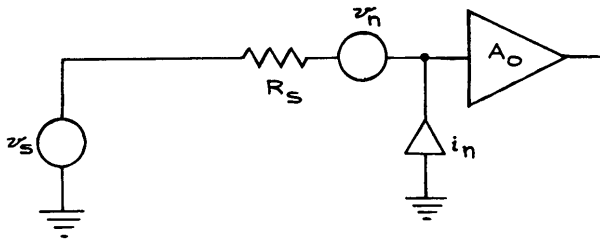
where  $q$  is the electronic charge in coulombs. Like Johnson noise, shot noise is frequency independent.

*Flicker* or *1/f noise* tends to dominate Johnson and shot noise below 100 Hz. The frequency dependence of this noise is of the form  $1/f^n$ , where  $n$  is usually between 0.9 and 1.35. Flicker noise imposes an important limitation on the SNR at low frequencies. For this reason, measurements should, if at all possible, be made at frequencies where white noise dominates.

## 6.8.2 Optimizing the Signal-to-Noise Ratio

An ideal amplifier of gain  $A_o$  can represent a real amplifier with current and voltage noise generators at the input terminals. It is usually assumed that the noise sources are frequency independent. Consider the system in Figure 6.125, where  $v_n$  and  $i_n$  represent the rms voltage and current noise per  $\text{Hz}^{1/2}$ . The SNR is then the output voltage due to the signal alone divided by the total output voltage, including noise:

$$\text{SNR} = \frac{v_s}{\sqrt{4kTR_s + v_n^2 + (i_n R_s)^2} \sqrt{\Delta f}} \quad (6.68)$$



**Figure 6.125** Real amplifier and signal source  $v_s$  with noise sources represented by voltage and current sources  $v_n$  and  $i_n$ .

From this formula it is clear that the SNR can be improved by reducing  $R_S$ ,  $\Delta f$ , and  $T$ . When characterizing noise in an amplifier, the *noise Figure* (NF) is often used:

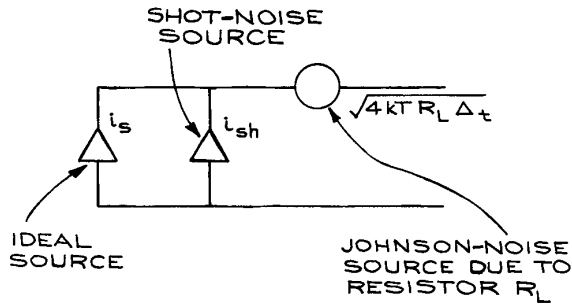
$$\text{NF(dB)} = 20 \log_{10} \frac{\text{input voltage SNR without amplifier}}{\text{output SNR from amplifier}} \quad (6.69)$$

An ideal amplifier will have the same SNR at both the input and output terminals. The smallest possible NF is 0, and an NF of 3 dB means that the amplifier has reduced the SNR by  $1/\sqrt{2}$  at its output terminals from what it was at the input terminals. Exceedingly good commercial amplifiers can have NFs as low as 0.05 dB over a limited range of frequency and source resistance. To evaluate the importance of the NF, it is necessary to know the input signal level, source resistance, and frequency bandwidth. If the SNR is  $10^3$  at the input terminals of an amplifier, for example, an amplifier with a NF of 20 dB will reduce it to  $10^2$  at the output terminals. This may still be entirely satisfactory for the intended application. Manufacturers of preamplifiers and amplifiers often supply contour plots of constant NF in frequency source-resistance space.

For optimum SNR, it is important to match the source resistance to the amplifier. The optimum value of the source resistor  $R_{SO}$  is  $v_n/i_n$ . This results in a minimum NF given by:

$$\text{NF}_{\text{min}} = 10 \log \left[ 1 + \frac{2v_n^2}{4kTR_{SO}} \right] \quad (6.70)$$

where  $R_{SO}$  is the optimum value of the source resistance. The source resistance is generally a property of the source. To transform the source's intrinsic resistance  $R_S$  to the optimum resistance  $R_{SO}$ , a transformer is used (never an additional series or shunt resistor). If the ratio of secondary to primary turns is  $\alpha$  then  $\alpha^2 R_S = R_{SO}$ . Although it was stated that the maximum SNR occurs when  $R_S = 0$ , in practice all sources have finite resistance and the SNR can be optimized by transformer matching. Photomultiplier tubes, photodiodes, and electron multipliers are best represented (see Figure 6.126) as noiseless constant-current sources in parallel with shot-noise and Johnson-noise sources resulting from the addition of a load resistor  $R_L$  to the circuit. In this case, the SNR is optimized by using as large a value of  $R_L$  as possible and an amplifier



**Figure 6.126** Representation of noise sources in a detector.

with high input impedance and low  $v_n$  and  $i_n$ . Amplifiers with FET input stages are the best choice. Maximum practical values of  $R_L$  are limited by the capacitance  $C$  of the input circuit (including cables) and the input impedance of the amplifier. There is no benefit in having  $R_L$  greater than the input impedance of the amplifier or so large that the time constant  $R_L C$  requires working at low frequencies where  $1/f$  noise dominates. The justification for using large values of  $R_L$  comes from the fact that the voltage signal from a current source is proportional to  $R_L$ , while the Johnson noise is proportional to  $\sqrt{R_L}$ .

An alternative way of specifying amplifier performance is with the *equivalent noise temperature*  $T_e$  defined as the necessary increase in temperature of the source resistor to produce the observed noise at the amplifier output—the amplifier being considered noiseless for this purpose:

$$T_e = T(10^{NF/10} - 1), \quad (6.71)$$

where  $T$  is the absolute temperature of the source resistor. Using the example of the amplifier with an NF of 3 dB and a source resistor at  $T = 300$  K:

$$T_e = 300 \text{ K}(10^{0.3} - 1) = 300 \text{ K} \quad (6.72)$$

In this case, the amplifier introduces an amount of noise equal to the noise from the source resistor.

### 6.8.3 The Lock-In Amplifier and Gated Integrator or Boxcar

The proper matching of a signal source to an amplifier is the first step to be taken in the recovery of a signal accom-

panied by noise. Once this has been done correctly, a number of signal-enhancing techniques can be used to extract the signal.

Because of the difficulty in making d.c. measurements due to zero drifts, amplifier instabilities, and flicker noise, the signal of interest should, if at all possible, be at a frequency sufficiently high that the dominant noise is white noise. A d.c. or low-frequency signal can be converted to a higher frequency by chopping the signal at the higher frequency. It is essential to choose a chopper frequency far from the power-line frequency and its harmonics. Frequencies near known noise frequencies should be avoided.

Signal-enhancing techniques are mainly based on bandwidth reduction. Because the white-noise power per unit bandwidth is constant, reducing the bandwidth will proportionally reduce the noise power. Of course, as the bandwidth goes to zero, the noise and signal power both go to zero as well. Common bandwidth-narrowing circuits are the resonant filter and low-pass filter. For resonant filters, the width of the resonance is given in terms of  $Q$ , which is approximately equal to  $f_0/\Delta f$ , where  $f_0$  is the frequency to which the filter is tuned and  $\Delta f$  is the bandwidth. Practical values of  $Q$  for electronic resonant filters vary from 10 to 100. A simple detection system might take the form shown in Figure 6.127. The rectifier converts the a.c. signal back to d.c., so it can be recorded on a chart recorder or read from a meter. With this arrangement, the noise passed by the filter is rectified along with the signal and adds to the recorded d.c. level.

The effective bandwidth can be substantially narrowed if, in the above arrangement, the ordinary rectifier is replaced by a synchronous rectifier. This kind of rectifier acts as a switch that is opened and closed in synchronization with the chopper. Because the phase relations for each of the noise components are random (a characteristic of white noise), they will tend to cancel each other when averaged—a low-pass filter at the output of the rectifier can be used to do the averaging. The  $RC$  time constant of the filter is related to the effective pass band of the system  $\Delta f$  by:

$$\Delta f = \frac{1}{4RC} \quad (6.73)$$

for a single-section, low-pass filter, and

$$\Delta f = \frac{1}{4nRC} \quad (6.74)$$

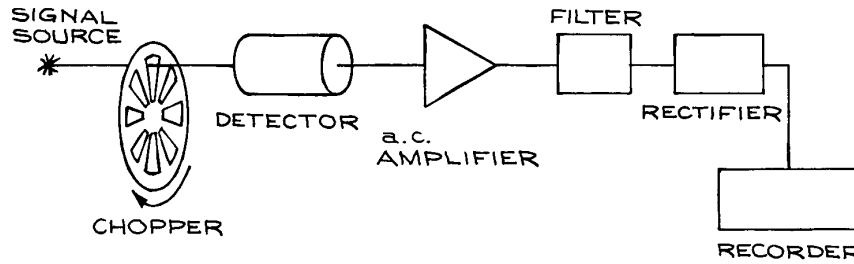


Figure 6.127 Simple detection and signal-enhancement system.

for  $n$  concatenated low-pass filters – each with time constant  $RC$ . Quite small values of  $\Delta f$  are possible with such a system. The only limitation is the length of time necessary for the measurement. If  $\Delta f$  is the pass band, the measurement time is approximately  $1/\Delta f$  so that it is necessary for the source signal to remain stable over times comparable to  $1/\Delta f$ . The chopping frequency is normally chosen to be 10 times the highest component frequency to be recovered in the source signal.

The system described above is often called a *lock-in amplifier* or *phase-sensitive detector*. The details of the operation of such devices are well documented.<sup>11</sup> As far as signal enhancement is concerned, however, the formulas given above are sufficient for estimating the benefits of such devices. It is well to remember that the SNR is proportional to  $1/\sqrt{\Delta f}$ , so that narrowing the pass band by a factor of four increases SNR by a factor of two, but increases the time of measurement by a factor of four.

When the signal to be detected has the form of a repetitive low-duty-cycle train of pulses, the lock-in amplifier may not be the best method for signal enhancement. The *duty cycle* is the fraction of time during which the signal of interest is present. With low duty cycles, signal information is available for only a fraction of the total time, while noise is always present. With timing and gating circuits, it is possible to connect the signal line to an  $RC$  integrating circuit only during those times when the signal is present. The time constant of the integrator is then chosen to be very much larger than the period of the pulse train. The time required for the capacitor to charge to 99% of the final voltage level is  $4.6RC$ , so that five time constants after the first gate opening the capacitor should be charged to within 1% of the final steady-state value – provided the signal is

continuously present. If the signal is present for a fraction  $\gamma$  of the time, the time constant of the integrator must be increased to  $5RC/\gamma$ . With this system, known as a *gated* or *boxcar integrator*, the SNR is increased only by increasing the time of the measurement. The effective bandwidth of the instrument is  $\gamma/4RC$ . Table 6.41 compares the important parameters associated with lock-in amplifiers and gated integrators.

#### 6.8.4 Signal Averaging

With a repetitive signal, merely averaging the signal over many cycles can improve the SNR. Let  $\text{SNR}_0$  represent the SNR in one cycle, and  $N$  the number of cycles over which one averages. The SNR improvement is proportional to  $\sqrt{N}$ . If each cycle lasts for a time  $\tau$ , the time  $T$  necessary to arrive at a specified SNR is given by:

$$T = \left[ \frac{\text{SNR}_0}{\text{SNR}} \right]^2 \tau \quad (6.75)$$

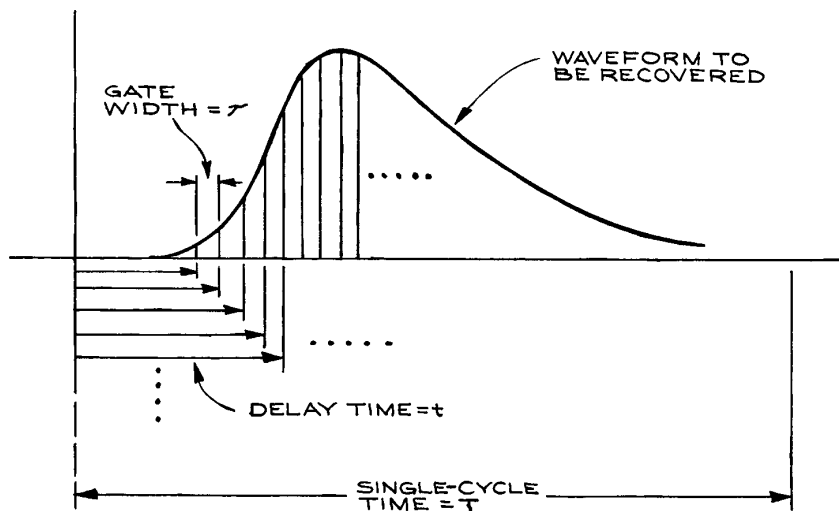
Lock-in amplifiers and gated integrators are inherently superior to simple signal averaging because of the band-narrowing functions they perform. The bandwidth improvement is only an advantage, however, when the highest frequency component in the signal to be recovered permits a long integrating time.

#### 6.8.5 Waveform Recovery

When it is necessary to know the variation of a signal with time, waveform recovery techniques are needed. The gated integrator can be converted to an instrument for waveform

**Table 6.41 Comparison of lock-in amplifier and gated integrator**

	<i>Lock-in Amplifier</i>	<i>Gated Integrator</i>
Duty cycle, $\gamma$	$> 0.50$	$< 0.50$
Bandwidth, $\Delta f$	$1/8RC$	$\gamma/4RC$
Minimum measurement time	$5RC$	$5RC/\gamma$
Highest recoverable signal frequency, $f_{\max}$	$1/20\pi RC$ for chopping frequency $\geq \pi RC/2$	$\gamma/20\pi RC$ for repetition frequency $\geq \pi RC/2$
Design notes	<ol style="list-style-type: none"> <li>1. Determine <math>f_{\max}</math>, the highest frequency component of the signal to be recovered.</li> <li>2. Choose <math>f_s</math>, the chopping frequency, where <math>f_s = 10f_{\max}</math>.</li> <li>3. Choose low-pass filter constants <math>1/RC = 2\pi f</math>, so as to pass frequency <math>f_{\max}</math>.</li> <li>4. The bandwidth <math>\Delta f</math> is <math>1/8RC</math>, and a tuned amplifier with a <math>Q</math> of 10 is sufficient to pass all frequency components of the signal to <math>f_{\max}</math>.</li> </ol>	<ol style="list-style-type: none"> <li>1. Required measurement time is <math>5RC/\gamma</math>.</li> <li>2. Bandwidth is <math>\gamma/4RC</math>.</li> <li>3. Preferred to a lock-in amplifier for signal repetition rate <math>\leq 10</math> Hz and low duty cycle.</li> </ol>



**Figure 6.128** Relation between gate width  $\tau$ , delay time  $t$ , and single-cycle time  $T$  in waveform recovery. The duty cycle  $\gamma$  is given by  $\tau/T$ .

recovery by the inclusion of variable delay and variable gate-width functions. A separate  $RC$  network is required for each delay time in such a scheme. A schematic representation of such a system is shown in Figure 6.128. Each

separate delay corresponds to a different part of the waveform. The duty cycle for each gate opening is  $\tau/T = \gamma$ , so that the effective bandwidth is  $\gamma/4RC$ . An integration time of  $5RC/\gamma$  is needed for the charge on each capacitor to



reach a steady-state value. If the waveform has been divided into  $n$  parts, the total time of measurement is  $5nRC/\gamma$ .

Digital schemes for recovering waveforms use ADCs to convert the analog signal level to a digital number, which is then recorded in the appropriate time channel with the aid of delay and gating signals. With such instruments, called *digital signal analyzers*, SNRs are only limited by the length of time of the measurement. The long-term stability of the various components, however, limits the use of very long measuring times. If the absolute value of the waveform is needed, one must record the number of cycles processed by the instrument. Data rates are limited by the conversion speed of the ADC that encodes the amplitude information. An advantage of digital signal analysis is the increased versatility in data manipulation and the production of permanent records on paper, magnetic, or semiconductor media. Such data are then available for computer analysis.

Instruments related to the digital waveform recorders are *transient recorders* and *memory oscilloscopes*. While these offer no gain in SNR over that in the original signal, they do provide outputs that are a record of a single waveform. An inefficient, but commonly used, method of SNR enhancement with these instruments is the separate recording of many waveforms and subsequent digital averaging by a computer.

### 6.8.6 Coincidence and Time-Correlation Techniques

There are a few general principles that govern all time-correlation measurements, which will be discussed in sufficient detail to be useful in assembling and optimizing the operating parameters of experiments where time is a parameter.

Correlated events are related in time, and this time relation can be established either with respect to an external clock or to the events themselves. Random or uncorrelated events bear no fixed time relation to each other but, on the other hand, their very randomness allows them to be quantified. Consider the passing of cars on a busy street. It is possible to calculate the probability  $P_{\bar{n}}(n)$  that  $n$  cars pass within a given time interval in terms of the average number

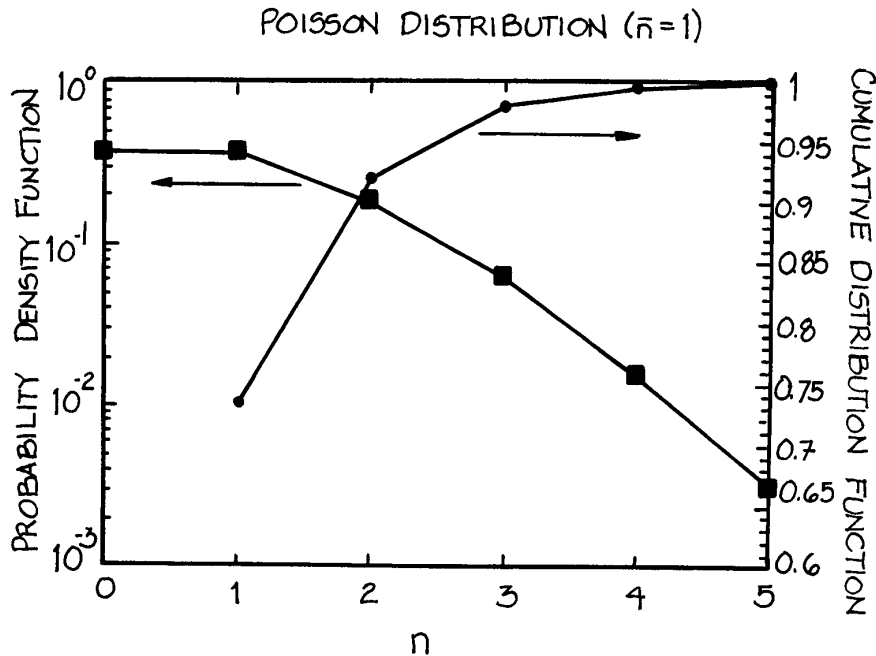
of cars  $\bar{n}$  in the interval where  $P_{\bar{n}}(n)$ , the Poisson distribution, is given by:

$$P_{\bar{n}}(n) = \frac{\bar{n}^n}{n!} e^{-\bar{n}}. \quad (6.76)$$

If 100 cars pass a fixed position on a highway in an hour, for example, then the average number of cars per minute is  $100/60$ . The probability that two cars pass in a minute is given by  $0.6^2/2e^{-0.6} = 0.10$ . The probability that three times the average number of cars pass per unit time is 0.02.  $P_{\bar{n}}(n)$  and the integral of  $P_{\bar{n}}(n)$  for  $\bar{n} = 1$  are shown in Figure 6.129. It is worthwhile to note that only a single parameter, the average value  $\bar{n}$ , is sufficient to define the function. The function is also not symmetric about  $\bar{n}$ .

This example of passing cars has implications for counting experiments. Consider again the particle-counting experiment shown in Figure 6.72(a). It consists of a detector, preamplifier, amplifier/shaper, discriminator, and counter. The detector converts the energy of an incoming particle to an electrical signal that is amplified by the preamplifier to a level sufficient to be amplified and shaped by the amplifier. The discriminator converts the signal from the amplifier to a standard electronic pulse of fixed amplitude and time duration, provided that the amplitude of the signal from the amplifier exceeds a set threshold. The counter records the number of pulses from the discriminator for a predetermined period of time. The factors that affect the measurement are counting rate, signal duration, and processing times.

In counting experiments, the instantaneous rate at which particles arrive at the detector can be significantly different from the average rate. In order to assess the rate at which the system can accept data, it is necessary to know how the signals from the detector, preamplifier, amplifier, and discriminator all vary with time. If signals are arriving randomly at an average rate of 1 kHz at the detector, for example, the average time between the start of each signal is  $10^{-3}$  s. If we consider a time interval of  $10^{-3}$  s, the average number of signals arriving in that interval is one. If as many as three pulses can be registered in the  $10^{-3}$  s, then 99% of all pulses will be counted (see Figure 6.129). To register the three randomly distributed pulses in the  $10^{-3}$  s interval, another factor of three in time resolution is required, with the result that the system must have the capability of registering events at a 10 kHz uniform rate, in



**Figure 6.129** Poisson distribution and cumulative Poisson distribution for  $\bar{n}=1$ .

order to be able to register randomly arriving events at a 1 kHz average rate with 99% efficiency. For 99.9% efficiency, the required system bandwidth increases to 100 kHz. Provided that the discriminator and counter are capable of handling the rates, it is then necessary to be sure that the duration of all of the electronic signals are consistent with the required time resolution. For 1 kHz, 10 kHz and 100 kHz bandwidths, this means that the signal durations must be no longer than 0.1, 0.01, and 0.001 ms, respectively. An excellent discussion of the application of statistics to physics experiments is given by Melissinos.<sup>12</sup>

The processing times of the electronic units must also be taken into consideration. There are propagation delays associated with the active devices in the electronics circuits and delays associated with the actual registering of events by the counter. The manufacturers of counters specify processing delays in terms of maximum count rate. A 10 MHz counter may only count at a 10 MHz rate if the input signals arrive at a uniform rate. For randomly arriving signals with a 10 MHz average rate, a system with

a 100 MHz bandwidth is required, to record 99% of the incoming events.

The result that is sought in counting experiments is a *rate* (number of events per unit time). For convenience, we divide the total time of the measurement  $T$  into  $n$  equal time intervals, each of duration  $T/n$ . If there are  $N_i$  counts registered in interval  $i$ , then the mean or average rate is given by:

$$\frac{1}{n} \sum_{i=1}^n \frac{N_i}{T/n} = \frac{N_N}{T} \quad (6.77)$$

where  $N_N$  is the total number of counts registered during the course of the experiment. To assess the uncertainty in the overall measured rate, we assume that the individual rate measurements are statistically distributed about the average value. In other words, they arise from statistical fluctuations in the arrival times of the events and not from uncertainties introduced by the measuring instruments. When the rate measurement is statistically distributed about the mean, the distribution of events can be described

by the Poisson distribution,  $P_{\bar{n}}(n)$ . The uncertainty in the rate is given by the standard deviation and is equal to the square root of the average rate  $\sqrt{\bar{n}}$ . Most significant is the relative uncertainty:

$$\frac{\sqrt{N/t_T}}{N/t_T}$$

The relative uncertainty decreases as the number of counts per counting interval increases. For a fixed experimental arrangement, increasing the time of the measurement increases the number of counts. As can be seen from the formula, the relative uncertainty can be made arbitrarily small by lengthening the measurement time. Improvement in relative uncertainty, however, is proportional to the reciprocal of the square root of the measurement time. To reduce the relative uncertainty by a factor of two, it is necessary to increase the measurement time by a factor of four. One soon reaches a point where the fractional improvement in relative uncertainty requires a prohibitively long measurement time.

Coincidence experiments require explicit knowledge of the time correlation between two events. Consider the example of electron impact ionization of an atom in which a single incident electron strikes a target atom or molecule and ejects an electron from it. Because the scattered and ejected electrons arise from the same event, there is a time correlation between their arrival times at the detectors.

In a two-detector coincidence experiment, of which Figure 6.72(b) is an example, pulses from the two detectors are amplified and then sent to discriminators – the outputs of which are standard rectangular pulses of constant amplitude and duration. The outputs from the two discriminators are then sent to the *start* and *stop* inputs of a time-to-amplitude converter or TAC. Even though a single event is responsible for ejected and scattered electrons, the two electrons will not arrive at the detectors simultaneously because of differences in path lengths and electron velocities. Electronic propagation delays and cable delays also contribute to the *start* and *stop* signals not arriving at the inputs of the TAC simultaneously – sometimes the *start* signal arrives first, sometimes the *stop* signal is first. If the signal to the *start* input arrives after the signal to the *stop*, that pair of events will not result in a TAC output. The result can be a 50% reduction in the number of recorded

coincidences. To overcome this limitation, a time delay is inserted in the *stop* line between the discriminator and the TAC. This delay can be a length of coaxial cable (typical delays are of the order of 1 ns/ft. – see Table 6.11), a lumped delay line, or an electronic delay circuit. The purpose of the delay is to insure that the stop signal always arrives at the *stop* input of the TAC after the *start* signal. A perfect coincidence will be recorded at a time difference approximately equal to the delay time. When a time window twice the length of the delay time is used, perfect coincidence will lie at the center of the time window, and it is possible to make an accurate assessment of the background by considering the regions to either side of the perfect coincidence region. An example of a time spectrum is shown in Figure 6.130.

An alternative to the TAC and PHA/ADC (see Section 6.4.7) is the time-to-digital converter (TDC) – a unit that combines the functions of the TAC and PHA/ADC. These units have *start* and *stop* inputs and an output that provides a binary number with a value that is directly proportional to the time difference between the *stop* and *start* signals. The TDC can be connected to an MCA or PC with the appropriate digital interface.

By storing the binary outputs of a PHA/ADC or a TDC in the form of a histogram in the memory of a MCA or PC,

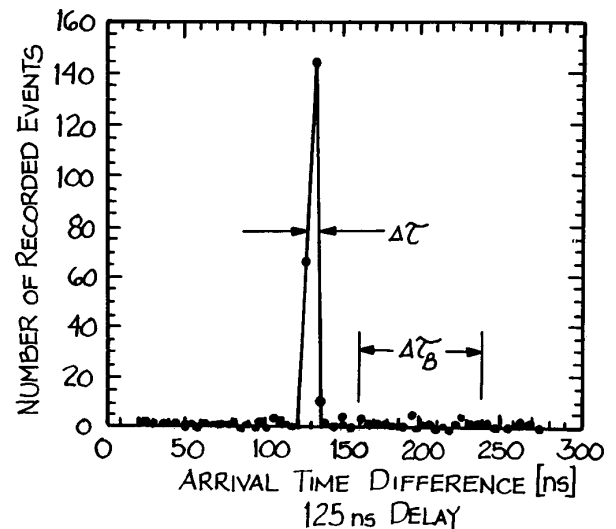


Figure 6.130 Coincidence time spectrum.

analyses can be performed on the data taking full account of the limitations of the individual components of the measuring system. For the preamplifiers and discriminators, time resolution is the principal consideration. For the TAC, resolution and processing time can be critical. The PHA/ADC or TDC provide the link between the analog circuits of the preamplifiers and discriminators and the digital input of the MCA or PC. The conversion from analog to digital by a PHA/ADC or TDC is not perfectly linear and the deviations from linearity need to be taken into consideration in precise work.

To assess the precision of a coincidence measurement, it is necessary to consider, not just the event under consideration, but also all other events arriving at the two detectors. While the events under study are correlated in time and result in a peak in the time spectrum centered approximately at the delay time, there are also background events that bear no fixed time relation to each other. If the average rate of the background events in each detector is  $R_1$  and  $R_2$ , then the rate that two such events will be recorded within time  $\Delta\tau$  is given by:  $R_B$ , where:

$$R_B = R_1 R_2 \Delta\tau \quad (6.79)$$

Let the rate of the event under study be  $R_A$ . It will be proportional to the probability of occurrence of the event  $p_A$ , the source function  $S$ , and the acceptances and efficiencies of the detectors  $d_1$  and  $d_2$ :

$$R_A = p_A S d_1 d_2 \quad (6.80)$$

For the background, each of the rates  $R_1$  and  $R_2$  will be proportional to the source function, the probability for single events  $p_1$  and  $p_2$ , and the properties of the individual detectors:

$$\begin{aligned} R_1 &= S p_1 d_1 \\ R_2 &= S p_2 d_2 \end{aligned} \quad (6.81)$$

Combining the two expressions for  $R_1$  and  $R_2$ :

$$R_B = S^2 p_1 p_2 d_1 d_2 \Delta\tau \quad (6.82)$$

Comparing the expressions for the background rate and the signal rate, one sees that the background increases as the square of the source function while the signal rate is

proportional to the source function. The signal-to-background rate  $R_{AB}$  is then:

$$R_{AB} = \frac{p_A}{p_1 p_2} \frac{1}{S \Delta\tau} \quad (6.83)$$

It is important to note that the signal is always accompanied by background. We now consider the signal and background after accumulating counts over a time  $T$ . For this, it is informative to refer to the time spectrum in Figure 6.130. The total number of counts within an arrival time difference  $\Delta\tau$  is  $N_T$  and this number is the sum of the signal counts  $N_A = R_A T$  and the background counts  $N_B = R_B T$ :

$$N_T = N_A + N_B \quad (6.84)$$

The determination of the background counts must come from an independent measurement, typically made in a region of the time spectrum outside of the signal region, yet representative of the background within the signal region. The uncertainty in the determination of the signal counts  $\delta N_A$  is given by the square root of the uncertainties in the total counts and the background counts:

$$\delta N_A = \sqrt{N_T + N_B} \quad (6.85)$$

The essential quantity is the relative uncertainty in the signal counts  $\delta N_A / N_A$ . This is given by:

$$\begin{aligned} \delta N_A / N_A &= \sqrt{N_T + N_B} / N_A \\ &= \sqrt{N_A + 2N_B} / N_A \\ &= \sqrt{R_A T + 2R_B T} / R_A T \\ &= \sqrt{\frac{1 + 2/R_{AB}}{R_A T}} \end{aligned} \quad (6.86)$$

Expressing  $R_A$  in terms of  $R_{AB}$  results in the following formula:

$$\delta N_A / N_A = \sqrt{\frac{(R_{AB} + 2)\Delta\tau}{\frac{p_A^2}{p_1 p_2} d_1 d_2 T}} \quad (6.87)$$

There are a number of conclusions to be drawn from the above formula. The relative uncertainty can be reduced to an arbitrarily small value by increasing  $T$ , but because the relative uncertainty is proportional to  $1/\sqrt{T}$ , a reduction in

relative uncertainty by a factor of two requires a factor of four increase in collection time. Reducing  $\Delta\tau$  can also reduce the relative uncertainty. It is understood that  $\Delta\tau$  is the smallest time window that includes the entire signal. Using the fastest possible detectors, preamplifiers, and discriminators and minimizing time dispersion in the section of the experiment ahead of the detectors all help to decrease  $\Delta\tau$ .

The signal and background rates are not independent, but are coupled through the source function  $S$ . As a consequence, the relative uncertainty in the signal decreases with the signal-to-background rate  $R_{AB}$ , a somewhat unanticipated result. Dividing  $\delta N_A/N_A$  by its value at  $R_{AB} = 1$  gives a reduced relative uncertainty  $[\delta N_A/N_{AB}]_R$  equal to  $\sqrt{(R_{AB} + 2)/3}$ . To review this result, consider a case where there is one signal count and one background count in the time window  $\Delta\tau$ . The signal-to-background ratio is 1, and the relative uncertainty in the signal is  $\sqrt{2 + 1/1} = 1.7$ . By increasing the source strength by a factor of ten, the signal will be increased by a factor of 10 and the background by a factor of 100. The signal-to-background ratio is now 0.1, but the relative uncertainty in the signal is  $\sqrt{110 + 100/10} = 1.45$  – a clear improvement over the larger signal-to-background case.

Another method for reducing the relative uncertainty is to increase the precision of the measurement of background. The above formulae are based on an independent measurement of background over a time window  $\Delta\tau$  that is equivalent to the time window within which the signal appears. If a larger time window for the background is used, the uncertainty in the background determination can be correspondingly reduced. Let a time window of width  $\Delta\tau_B$  be used for the determination of background, where  $\Delta\tau/\Delta\tau_B = \rho < 1$ . If the rate at which counts are accumulated in time window  $\Delta\tau_B$  is  $R_{BB}$ , the background counts to be subtracted from the total counts in time window  $\delta N_A/N_A$  become  $\rho R_{BB}T = \rho N_{BB}$ . The uncertainty in the number of background counts to be subtracted from the total of signal plus background is  $\sqrt{\rho^2 N_{BB}}$ . The relative uncertainty of the signal is then:

$$\sqrt{\frac{1 + 2\rho^2 N_{BB}/N_A}{N_A}} = \sqrt{\frac{1 + 2\rho N_B/N_A}{N_A}}$$

The expression for the reduced relative uncertainty is then:

$$[\delta N_A/N_{AB}]_R = \sqrt{\frac{R_{AB} + 2\rho}{3}} \quad (6.89)$$

The single coincidence measurement can be expanded to a multiple detector arrangement if it is necessary to measure coincidence rates as a function of more than one parameter. This is a much more efficient way of collecting data than having two detectors that are scanned over the range of the parameters. Depending on the signal and background rates, detector outputs can be multiplexed and a detector encoding circuit can be used to identify those detectors responsible for each coincidence signal. If detectors are multiplexed, it is essential to recognize that the overall count rate is the sum of the rates for all of the detectors. This can be an important consideration when rates become comparable to the reciprocal of system dead time.

It is often the case that signal rates are limited by the processing electronics. Consider a coincidence time spectrum of 100 channels covering 100 ns with a coincidence time window of 10 ns. Assume a signal rate of 1 Hz and a background rate of 1 Hz within the 10 ns window. This background rate implies an uncorrelated event rate of 10 kHz in each of the detectors. To register 99% of the incoming events, the dead time of the system can be no larger than 10  $\mu$ s. The dead time limitation can be substantially reduced with a preprocessing circuit that only accepts events falling within 100 ns of each other (the width of the time spectrum). One way to accomplish this is with a circuit that incorporates delays and gates to only pass signals that fall within a 100 ns time window. With this circuit the number of background events that need to be processed each second is reduced from 10 000 to 10, and dead times as long as 10 ms can be accommodated while maintaining a collection efficiency of 99%. This can be considered to be an extension of the multiparameter processing in which the parameter is the time difference between processed events. Goruganthu *et al.* discuss a preprocessing circuit.<sup>13</sup>

## 6.9 GROUNDS AND GROUNDING

### 6.9.1 Electrical Grounds and Safety

A battery or power supply produces a potential difference between its two output terminals. Only by defining the

potential of one of the terminals in an absolute way can the potential of the other terminal be given an absolute value. By convention, the electrical potential of the earth is defined to be zero. When one output terminal of a battery or power supply is connected to the earth, the absolute potential of the other terminal is then equal to the potential difference. Because the surface of the earth is a fairly good conductor of electricity, all points will be at nearly the same potential. A good connection to the earth via a copper rod buried in the ground is the most convenient way of establishing zero potential. Since such connections are not always readily available in the laboratory, one often uses connections to metal water pipes that go underground, to define zero potential.

Most connections to the a.c. line are now made with three-prong plugs and three-conductor wire. The wire is usually color coded in North America – black corresponding to the high-voltage or *hot* side of the line, white to the low-voltage or *neutral*, and green to the earth or *ground*. The terminals on a.c. plugs and receptacles are also coded – the hot terminal is brass, the neutral terminal is chrome-plated, and the ground terminal is dyed green. Switches and fuses for use with a.c. lines should always be placed in the hot line, otherwise the full line voltage will be present at the input even with an open switch or a blown fuse.

The three-wire system is designed for safety. Consider a piece of electrical equipment with a metal case operating from the a.c. line. Normal practice requires that the case be connected to the green ground wire of the line cord. Should the hot wire come in contact with the case, a very large current would flow in the hot line and blow a fuse or open a circuit breaker. If the case is not connected to ground, it will assume the potential of the hot line and could give an electric shock to anyone touching it. Of course, if the case were not a conductor, there would be no danger even if the hot line made a connection to it. There are a large number of appliances and hand power tools with so-called *double insulated* cases that do not require a separate ground connection for safety purposes. Ground-fault (circuit) interrupters (GFI or GFCI) are safety devices that are incorporated in a.c. power outlets where there is a danger of operators of electrical equipment attached to the outlet touching electrical ground while using the equipment. The GFI senses any current flowing directly to ground and switches off all power to the equipment to minimize elec-

trical shock. Ground-fault interrupter electrical outlets are standard in humid environments outdoors and in bathrooms and underground installations.

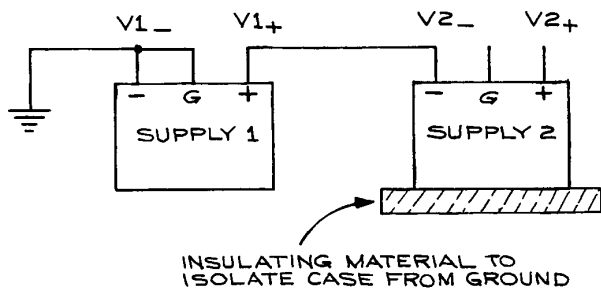
From this discussion it should be clear that fuses and circuit breakers are important safety elements in electrical circuits. Fuses are generally of either the standard quick-response type or the *slow-blow* type. The former will open whenever its maximum rated current is exceeded, even momentarily, while the *slow-blow* fuse is designed to sustain momentary current surges in excess of the maximum rated continuous value. The two types are not interchangeable, even though they may have identical maximum continuous current ratings. Circuit breakers are generally of the electromagnetic-relay or the bimetallic-strip type. They can be reset after opening from an overload.

It is sometimes necessary to *float* or remove all ground connections from a piece of electrical equipment. This is necessary when a signal between two points in a circuit, neither of which is at ground, is to be measured, amplified, or processed in some way. Power supplies are floated when it is necessary to produce a well-defined potential difference with reference to a nonzero potential. An example is the filament heater supply of a high-voltage electron gun with the filament at a high negative potential. The heater supply is connected across the filament with one lead at the high negative potential. In this configuration, neither heater supply lead is at ground potential. Floating introduces the possibility of additional noise and electrical breakdown within the circuit with a resulting destruction of components. There are safety hazards associated with floated circuits.

At this point it is worthwhile to make the distinction between d.c. and a.c. grounds. A *d.c. ground* connection is a direct low-resistance connection to ground, while an *a.c. ground* connection is one that does not permit d.c. currents to flow while offering low impedance to a.c. currents. In Figure 6.131, power supply 1 has the negative terminal and case (*G terminal*) at d.c. ground. Power supply 2 is floating on power supply 1 and, in the absence of any external connections, no current will flow through the terminals of the supplies. Since high-quality power supplies generally have very low output impedances, the a.c. impedances of the terminals marked  $V_{1+}$ ,  $V_{2-}$ , and  $V_{2+}$  with respect to ground are low, although the potentials at the terminals are not zero. It is good to keep in mind that the impedances to ground are frequency-dependent

because of the characteristics of the output circuits. At high frequencies, the impedances may become large due to the frequency characteristics of the output circuit. As a result, transient voltage spikes may not be effectively shunted to ground and may adversely affect the circuits to which the power supplies are connected. A remedy is the use of good high-frequency capacitors of an adequate voltage rating between the terminals  $V1_+$  and  $V2_+$  and ground.

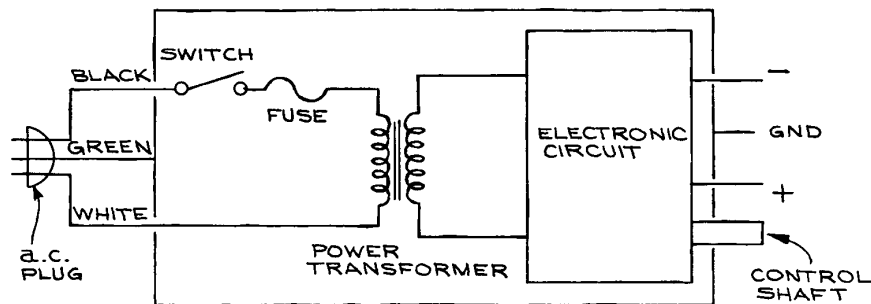
Figure 6.132 illustrates a typical line-operated power supply connected to an a.c. outlet via a three-wire line cord. Under normal operation, either the negative or the positive output terminal is connected to the ground terminal, depending on whether a positive or a negative voltage with respect to ground is desired. When floating the power supply, the ground connection is removed and the potential at which the supply is to be floated is applied to either the positive or negative output terminal. This may be satisfac-



**Figure 6.131** Power supply 2 floating on power supply 1. The  $V2_+$ ,  $V2_-$ , and G terminals of supply 2 are assumed to have no direct d.c. connection to ground.

tory for levels of as much as a few hundred volts, but has its hazards because the internal circuitry is floating, while the case is at ground through the green wire in the three-wire line plug. Potentials at least equal to the floating potential plus the potential difference at the output terminals now exist at the terminals of several components in the internal circuit. Components are often mechanically fixed to the grounded chassis. The electrical insulation between the components and the chassis, as well as that between the primary and secondary of the power transformer, must be sufficient to sustain the d.c. voltage difference. In the event of electrical breakdown, components will be destroyed and the shafts of the controls may attain a potential that is equal to the floating potential and thus present a shock hazard. The maximum potential for floating line-powered equipment in this way is generally 600 V.

When floating line-operated equipment at voltages above the rated maximum voltages, it is necessary to completely remove the ground connection to the case. The easiest method is to use a two-terminal to three-terminal a.c. line-plug converter and not connect the third terminal to ground at the a.c. outlet. If the case of the instrument is then electrically isolated from all conducting surfaces by placing it in a Plexiglas™ enclosure, it can be electrically floated at the desired potential without risk of electrical breakdown between the components of the circuit and the case. With the case at a nonzero potential, there is a shock hazard associated with contacting it and the controls so that *access must be provided via auxiliary insulated knobs and shafts*. A further potential problem is the possibility of electrical breakdown



**Figure 6.132** Mounting a line-operated power supply so that both output terminals are independent of the a.c. line ground.

between the primary and secondary windings of the power transformer. At the primary, the peak voltage values are +155 V, corresponding to the normal peak values of the a.c. line voltage. The secondary voltage, however, will have a d.c. component equal to the floating voltage superimposed on the normal a.c. value. If the resulting d.c. plus peak a.c. secondary voltage exceeds the voltage rating of the transformer insulation, electrical breakdown can occur, destroying the transformer and possibly the rest of the circuit. *Isolation transformers* are commonly used to solve this problem. These transformers have a 1:1 turns ratio and high-voltage insulation between the primary and secondary windings. The primary is connected to the a.c. line, while the secondary is connected to the primary of the input transformer of the instrument to be floated. In this way, the floating voltage appears across the isolation-transformer windings. The case and controls are still at the floating potential, and precautions similar to those in the former situation must be observed. When specifying an isolation transformer, the power-handling capacity in watts or kilovolt-amperes (KVA) as well as the maximum isolation voltage must be considered.

A substitute for an isolation transformer is two high-voltage filament transformers of the same turns ratio, connected back to back. Such transformers are often used in oscilloscope power supplies. This will produce the necessary 1:1 overall voltage ratio with isolation. Often such transformers are more readily available and less expensive than high-voltage, high-power, single-isolation transformers.

When floating instruments that are not directly powered from the a.c. line, the external power supply (whether it is a battery or line-operated supply) must be floated, too. The precautions that apply to the floating of line-operated instruments apply here also. Table 6.42 summarizes the important considerations involved when floating power supplies.

### 6.9.2 Electrical Pickup: Capacitive Effects

It is often necessary to detect and measure electrical signals in the millivolt and even microvolt range. If the signals are accompanied by noise that is generated by external sources, the signals can be so degraded as to render the measurement useless. The elimination of external noise sources – or, if this is not possible, the elimination of their effects – is an important element in the design and con-

---

**Table 6.42 Floating power supplies**

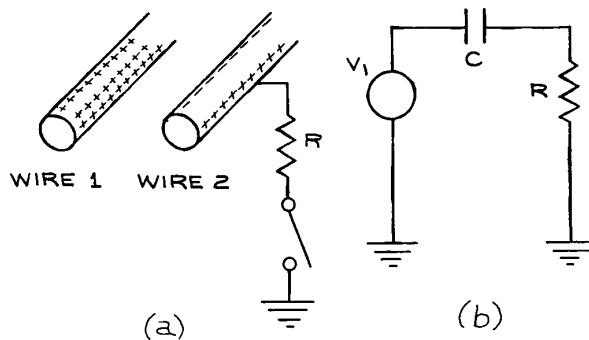
---

- (A) Normal Operation
    - (1) Input-line cord connected to three-terminal a.c. outlet
    - (2) Chassis at ground potential
    - (3) – or + output terminal to grounded center terminal for a + or – output voltage
  - (B) Low-voltage floating operation
    - (1) Input-line cord to three-terminal a.c. outlet
    - (2) Chassis at ground potential
    - (3) + or – terminal to floating potential
    - (4) Grounded output terminal unconnected
  - (C) Medium-voltage floating operation
    - (1) Input-line cord attached to 3-2 plug adaptor with ground terminal unconnected
    - (2) Chassis at floating potential
    - (3) + or – output terminal connected to chassis output terminal (at floating potential)
    - (4) Case and chassis isolated from all conductors
    - (5) Insulating shafts and knobs on all switches and dials
  - (D) High-voltage floating operation
    - (1) a.c. input from isolation transformer
    - (2) Instrument case at floating potential
    - (3) Case and chassis electrically isolated from all conducting surfaces
    - (4) Insulated shafts and knobs on all switches and controls
- 

struction of electronic measuring, control, and detection systems.

Consider a wire at a potential  $V_1$  in the vicinity of a second wire that is insulated from all conducting surfaces [see Figure 6.133(a)]. Coulomb forces between the charges on the surfaces of the wires will cause polarization of the free charges in wire 2. Because wire 2 is electrically isolated, it will remain electrically neutral. If wire 2 is now connected to ground through a resistance  $R$ , a current will momentarily flow due to the presence of the positive charges on wire 1 attracting negative charge to wire 2 from the ground. The equivalent circuit is shown in Figure 6.133(b). Typical values of the coupling capacitance are 1 to 100 pF. Two twisted #22 polyethylene-insulated wires, for example, will have a capacitance of 20 pF/ft. While the center-wire-to-shield capacitance of RG/58 coaxial cable





**Figure 6.133** (a) Capacitive coupling between two wires; (b) the equivalent electrical circuit with the effect of wire 1 on wire 2 represented by  $C$ .

is 33 pF/ft. If we assume that  $V_1$  is varying sinusoidally at frequency  $\omega$ , the rms potential across  $R$  is given by:

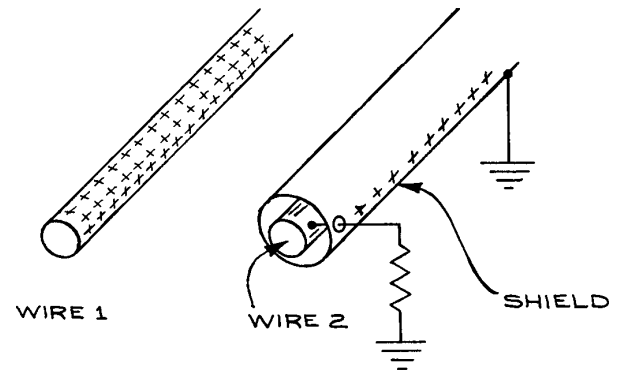
$$v_{1\text{rms}} \frac{R}{R + 1/\omega C} = v_{1\text{rms}} \frac{\omega RC}{1 + \omega RC} \quad (6.90)$$

(the voltage-divider formula). Thus the rms potential across  $R$  increases with  $R$ ,  $\omega$ , and  $C$ .

In the average laboratory, there are many sources of a.c. voltage that can couple capacitively to signal lines – open-line sockets, lighting fixtures, and line cords are a few examples. An estimate of the capacitance  $C_{\text{min}}$  necessary to induce a 1 mV rms voltage across 1 M $\Omega$  from a power-line source can easily be made by setting  $V_{1\text{rms}}$  equal to 120 V,  $\omega = 2\pi f = 188 \text{ rad/s}$ , and neglecting  $\omega RC$  with respect to 1:

$$C_{\text{min}} = \frac{1 \times 10^{-3} \text{ V}}{1.2 \times 10^2 \text{ V} \times 1.88 \times 10^2 / \text{s} \times 10^6 \Omega} \\ = 0.044 \text{ pF}$$

a value easily attained. Such induced voltages can be substantially reduced by the use of shielding. If a grounded shield is placed around wire 2, the situation illustrated in Figure 6.134 will occur. The conducting shield around wire 2 will have a current induced in it when connected to ground, due to the potential on wire 1. This current, however, flows through a low-resistance connection. In addition, and most important, any charge on the shield will



**Figure 6.134** Shielding to prevent capacitive coupling.

reside on the surface, and the interior will be field-free. In practice, common coaxial braided shield is about 90% effective against capacitatively coupled voltages from external sources. Foil shields are even more effective.

### 6.9.3 Electrical Pickup: Inductive Effects

A changing magnetic field can induce a current in any loop it cuts. The magnitude of the induced current depends on the area of the loop and the time rate of change of the magnetic field. High-resistance circuits are not greatly affected by such inductive effects, but they are important in low-resistance circuits, such as the input circuits of current and pulse amplifiers. Common sources of magnetic-field pickup are transformers, inductors, and wires carrying large a.c. currents. Effective shielding against low-frequency magnetic fields is accomplished with ferromagnetic enclosures, that concentrate the stray magnetic fields within the shield material. The usual practice is to enclose the source of the magnetic fields within such shields. Higher-frequency magnetic fields are best shielded with copper enclosures. Eddy currents induced in the copper produce counter magnetic fields. A rapidly changing current is a source of time-varying magnetic fields that can themselves induce currents in other circuits. Fast rise time pulses in the low-impedance output circuits of pulse amplifiers can generate rapidly changing magnetic fields, that can in turn induce unwanted currents in nearby low-impedance input circuits. Once again, the most effective shielding is high conductivity copper sheet or foil.

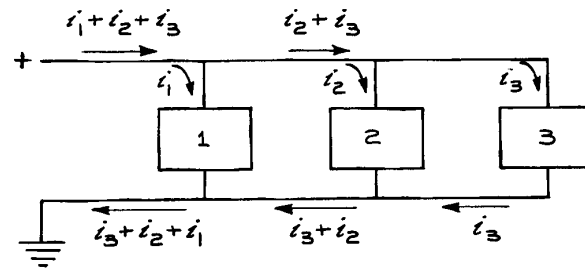
### 6.9.4 Electromagnetic Interference and r.f.i.

The high-frequency fluctuation of current or charge in a conductor results in the radiation of part of the energy in the conductor in the form of an electromagnetic wave. Such a wave propagates through space at the speed of light and, when not carrying useful information is called *radio-frequency interference* (r.f.i.). Sources of such radiation are automobile ignition systems, microwave ovens, electrical discharges, electric motors, electromechanical switches and relays, and electronic switches such as thyratrons, rectifiers, SCRs, and triacs. Power supplies with switching regulators are often sources of r.f.i.

This interference can be reduced by the use of zero-crossing switching regulators in which the switching only occurs when the voltage across the switch is zero. The most effective shield against r.f.i. is a grounded enclosure of a conducting material, such as copper or aluminum. On the surfaces of such a shield, the electric component of any incident electromagnetic wave is zero and further propagation is not possible. Conducting screens, rather than solid sheet, are often used for r.f.i. shielding because of the savings in weight. When using such a screen, it is important that the mesh size be small compared to the wavelength of the highest-frequency component of the r.f.i. Standards have now been established for r.f.i. emission from certain classes of electrical and electronic equipment. Sensitive, high-frequency-measuring equipment must have good immunity to r.f.i. and additional r.f.i. shielding can be often purchased as an option for equipment that must operate in especially noisy environments.

### 6.9.5 Power-Line-Coupled Noise

Consider the case illustrated in Figure 6.135, where a single supply line powers three circuits. If circuit 1 suddenly requires a large amount of current, circuits 2 and 3 may be affected because of the resistance and inductance of the power-supply line. A large current drawn by circuit 1 can cause a momentary voltage drop at circuits 2 and 3 that can be propagated as noise throughout the circuit. The greater the time rate of change of the current pulse through circuit 1, the larger the effect on the other



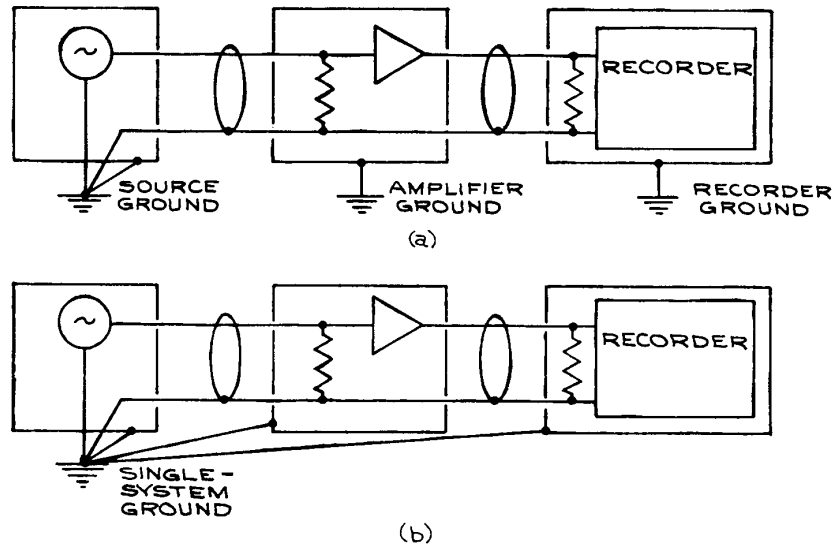
**Figure 6.135** Current flow in three circuits on a common power line.

circuits. This problem is very common with logic circuits, particularly TTL circuits, that have saturating output stages. If a number of gates switch simultaneously, very large transient current spikes can appear on the power line. These transients are then interpreted as a change in logic level by other gates, and transitions occur that are entirely spurious. Standard practice is to use low-resistance, low-inductance power supply lines to the logic gates and decouple the gates from the power supply with high-frequency capacitors from the power-supply terminal of the gate to ground. For TTL, every group of 20 gates must be decoupled from the power supply with a 0.1  $\mu\text{F}$  capacitor, and every group of 100 gates must be decoupled with an additional 0.1  $\mu\text{F}$  tantalum capacitor with good high-frequency characteristics.

Similar situations arise with switched, high-current devices such as temperature-controlled furnaces operated directly from the a.c. line. Large current transients may affect all equipment plugged into the same line. The only remedy here is to operate the high-current device from a completely separate line circuit.

As an example, consider a series of logic gates to be connected to a power supply by 1 ft. of #22 wire, which has a resistance of  $16 \Omega/1000 \text{ ft.}$  and an inductance of  $640 \mu\text{H}/1000 \text{ ft.}$  If a few gates switch simultaneously, resulting in a current spike of 10 mA with a rise time  $t_r$  of 10 ns, the momentary voltage drop along the line will be approximately  $iZ$ , where  $Z$  is  $\sqrt{R^2 + \omega^2 L^2}$ . In this case:

$$Z = \sqrt{(16 \times 10^{-3} \Omega)^2 + (2\pi \times 3.3 \times 10^7 / \text{s} \times 6.4 \times 10^{-7} \text{ H})^2} \quad (6.92)$$



**Figure 6.136** System grounding: (a) multiple grounds; (b) single ground to eliminate ground loops.

where  $\omega = (2\pi/3)t_r$  has been used in the estimate of the inductive reactance of the wire. The voltage drop is then:

$$10^{-3}\text{A} \times 133\ \Omega = 1.3\ \text{V} \quad (6.93)$$

For the TTL logic, this drop could be enough to cause some marginally substandard gates to change state momentarily.

### 6.9.6 Ground Loops

Instrument terminals are connected to ground in order to have the potentials of those terminals be at 0 V for reference purposes, assuming that every ground connection is made directly to the zero potential of the earth. This is often not the case, since most ground connections are through the third wire of the a.c. line. Since this line has finite resistance, currents flowing in it will cause potential drops and, depending on the point on the line where a ground connection is made, the potential can be substantially different from 0 V. When the 0 V references of two instruments that are connected together differ, there is danger of introducing noise into the system. The problem is made all the worse by the fact that the reference points can change their potentials with respect to each other in a way independent of other

system parameters. The solution to this problem is to have only a single ground point in the system and connect all the zero reference points and shields to it. This means removing the ground connections in the power-line cords of all line-operated instruments, and isolating all shields and cases from conducting surfaces that may be connected to ground at points other than the single system ground. All connections to this single system ground should be made as short as possible, with the lowest-resistance conductors available. By having the 0 V reference potential connection and shield connection at the same point in the system, no currents can be induced in the reference line.

To illustrate these ideas, two different ways of connecting a system composed of a signal source, amplifier, and recorder are shown in Figure 6.136. In the top arrangement each case is separately grounded and also connected to the other cases through the outer conductor of the coaxial cable between them. If the three grounds are at different potentials, currents will flow through the coaxial outer shields connecting the devices together. Since the reference-potential line is the shield, it will take on different potentials at different points in the circuit. As these potentials change, the signal levels will change – the result being seen as noise on the signal line.

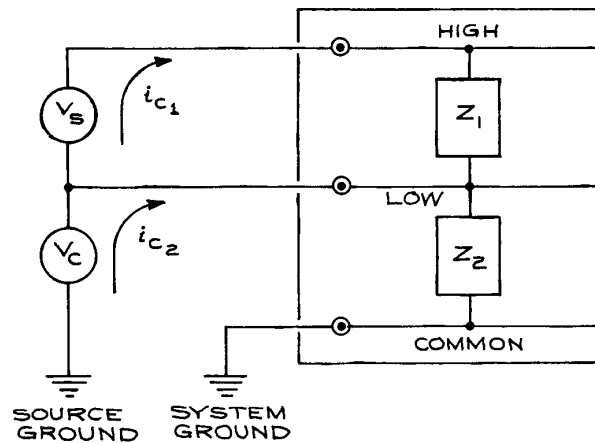
In the second arrangement, the common reference line and the instrument shields are all attached to ground at a single point – the reference ground. No ground loops can exist, and the reference line remains at the same potential everywhere in the system. Clearly this is a superior arrangement. The implementation may be difficult, however, since it is necessary to connect all the cases and shields separately to a single ground point. If all the instruments are mounted in a single rack, each must be electrically isolated from the conducting rack structure. The coaxial-shield connections must also be isolated from the cases, and the ground connection in the a.c. line cord must be removed from each of the instruments. It is not always necessary to take such precautions with all stages of a system – it may be sufficient to eliminate ground loops from the most sensitive element to which the signal source is connected.

It is worthwhile to understand the origins of ground loop and shielding problems in order to be able to solve them when they arise. An extensive treatment is given by Morrison.<sup>14</sup>

The problem of ground loops also occurs when measurements must be made across two electrical terminals, neither of which is at ground potential.<sup>15</sup> Bridge measurements are a common example of this. For floating measurements, instruments that have the reference input terminal electrically separate from the ground or common terminal must be used, otherwise the floating potential (common-mode potential) will be short-circuited.

In Figure 6.137,  $v_S$  is the source voltage to be measured,  $v_C$  is the common-mode voltage,  $Z_1$  is the input impedance of the measuring system, and  $Z_2$  is the impedance from the common of the measuring circuit to the low terminal – normally  $10^8$  to  $10^{10} \Omega$  in parallel with  $10^3$  to  $10^5$  pF. The presence of  $v_C$  results in currents  $i_{C1}$ , and  $i_{C2}$  flowing through  $Z_1$  and  $Z_2$ . This presents no difficulties so long as the resistance of the lines is zero; if the resistances are finite, the common-mode currents will result in potential differences at the low input terminal and add to the error of the measurement. Generally,  $Z_1 \ll Z_2$  and the critical parameter is the ratio of the resistance of the low signal line to  $Z_2$ . This ratio is frequency-dependent, because  $Z_2$  is complex. If the source and instrument grounds are not at the same potential, the difference will add to  $v_C$ .

The common mode rejection ratio (CMRR) is a quantity for specifying how well an instrument rejects a com-



**Figure 6.137** Measurement of  $v_S$  across two terminals, one of which is not at ground. Current flow through  $Z_2$ , the impedance from the low terminal of the measuring instrument to ground, will result in an error voltage.

mon-mode voltage superimposed on the voltage to be measured (the normal mode voltage):

$$\text{CMRR}(\text{dB}) = -20 \log \frac{v_{\text{NM}}}{v_{\text{CM}}}, \quad (6.94)$$

where  $v_{\text{NM}}$  is the portion of the common-mode voltage that appears as a normal-mode voltage, and therefore a source of error. Meters that can be floated have CMRRs from 80 to 120 dB at d.c. and 60 to 100 dB at line frequency. When higher values are required, special voltmeters having an additional guard terminal must be used. This terminal, when connected to the low terminal of the source by a low-resistance connection, shunts the common-mode current from the measuring terminals. Such meters offer CMRR values of 160 dB at d.c. and 140 dB at line frequency. For a meter with a CMRR of 160 dB, a 10 V common-mode signal will result in a  $0.1 \mu\text{V}$  normal-mode voltage.

## 6.10 HARDWARE AND CONSTRUCTION

It is almost always better to purchase a piece of electronic equipment than to design and construct it. This is certainly the case for power supplies, amplifiers, signal generators, and similar equipment that is mass-produced by a large number of companies in a wide variety of models.

When one does, however, decide to construct a piece of electronic equipment because of cost or the unavailability of commercial units, there are a number of well-defined steps to follow:

- (1) Design of the circuit
- (2) Selection of components
- (3) Construction and testing of a breadboard model
- (4) Construction of the final circuit
- (5) Mounting
- (6) Final testing.

Except for the simplest circuits, one is well advised to use proven designs that can be found in manufacturers' application books and in books and articles on circuit design. A few such publications are given in the list of references.<sup>16</sup> There are, in addition, well-established techniques that are used by electronics engineers (but not often included in textbook examples of circuit design) that can make the difference between a circuit that works and one that does not. These include bandwidth limiting, power-supply decoupling, and signal conditioning.

### 6.10.1 Circuit Diagrams

Two kinds of circuit diagrams are the block diagram and the schematic wiring diagram. The *block diagram* shows the logical layout of the circuit by grouping all elements necessary for a single function into a single rectangle. Although simple, such diagrams are very useful in understanding the general operation of a circuit. The block diagram is much like the flowchart of a computer program.

The *schematic wiring diagram* shows all the components of the circuit and the connections between them. Figure 6.138 lists symbols used in schematic diagrams. The more complete the schematic, the more useful it is for troubleshooting. A schematic will have the values and ratings of all the components, as well as numbers and color codes when necessary. Good schematics include the values of voltages at critical points in the circuit and drawings of waveforms where appropriate. One generally reads schematic diagrams starting at the upper left with the input and proceeding to the lower right to the output.

A useful addition to the schematic diagram is the *chassis layout diagram*, which shows the physical location of all parts. For ease of drawing, components that appear next to

each other on the schematic may be far away from each other on the actual chassis or circuit board.

With increasing use of integrated circuits, it is more and more difficult to identify the functions of different elements in a schematic. This is because the integrated circuits are often only represented by rectangles with pin numbers. To identify their function, it is necessary to consult a manufacturer's catalog. Abbreviations common with ICs are given in Table 6.43 along with representative package outlines. Schematic diagrams also help identify specific components. If there is an operational amplifier on a circuit board with a balance control near it for nulling purposes, it is useful to know where the amplifier is in the circuit and the nature of the input and output before proceeding with adjustments. The location of damaged components can be useful information when deciding on the probable cause of the damage and other components that may also have been affected.

### 6.10.2 Component Selection and Construction Techniques

When selecting components, the considerations that an electronic engineer finds important may be different from those of a laboratory scientist building only a single example of a circuit. It is always wise to overspecify components – that is, to use components of higher ratings and better quality than a critical cost-effective analysis of the circuit would show to be necessary. Usually the cost of components is a small fraction of total project cost when time is considered.

Once the circuit design has been decided upon and the components assembled, a *breadboard* model can be constructed. This may seem an unnecessary and time-consuming step, but care at this stage will save a great deal of time later. The breadboard circuit should be constructed from a complete, detailed schematic circuit diagram.

There are a number of prototype circuit boards and aids for making breadboard models. Prototype boards use sockets to which components and wires can be connected in a temporary way. The spacing of the sockets is usually on .1 in. centers to accommodate the usual 14- and 16-pin dual in-line integrated circuits. The sockets accommodate only a limited range of solid-wire sizes. Outside of this range, the wire is either not held securely or the socket is bent out

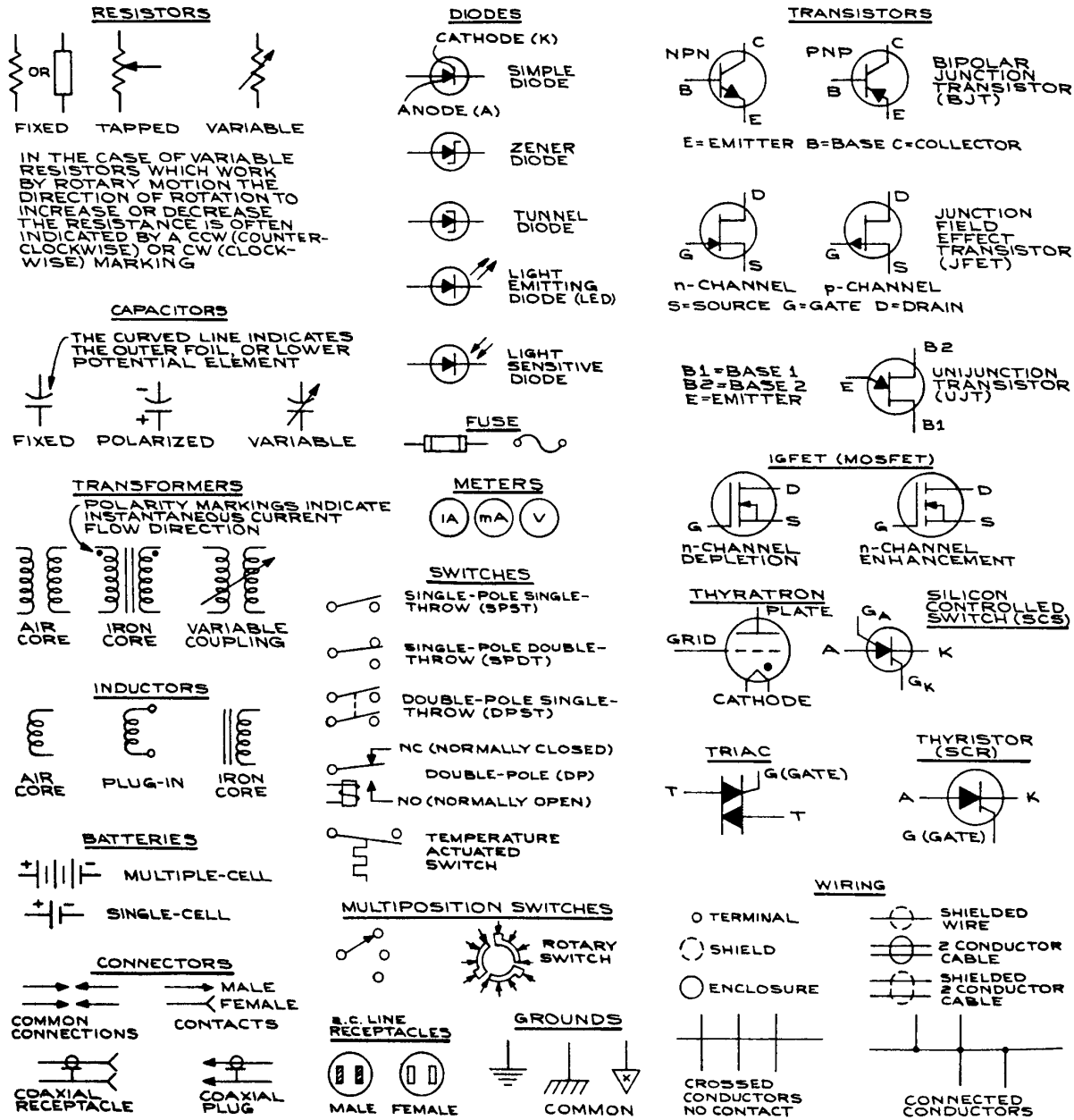


Figure 6.138 Symbols for schematic circuit diagrams.

of shape. Prototype boards of this type are satisfactory for low-frequency circuits and logic circuits using TTL and CMOS. For sensitive high-frequency circuits and ECL logic, stray capacitance between sockets is a problem, and such circuits must be built on a circuit board with a ground plane if they are to work properly.<sup>17</sup> At this stage it is necessary that the breadboard circuit follow the schematic in every detail. The mechanical layout should be neat, with interconnections between components as short as possible. This simplifies troubleshooting and avoids pickup problems.

Interfaces to external power, and input and output circuits present some difficulty when working with breadboards. Dangling wires and alligator-clip connections invite short or open circuits and damaged components. Whenever possible, the breadboard should be mounted on a larger structure with terminals and connectors to which semi-permanent power-supply and signal connections can be made (see Figure 6.139). Connections from these terminals to the breadboard can then be made with short pieces of hookup wire.

The breadboard circuit should be fully tested before proceeding to the final construction stage. During testing, the circuit should be operated with the same input and output connections the final circuit will have.

Once the breadboard circuit is functioning properly, the final circuit can be constructed. Depending on the complexity of the circuit, there are a number of construction techniques available. In general, all components are mounted on circuit boards with the boards fitted into a case or attached to a panel. The power supply can be incorporated into the circuit, or power-supply voltages can be brought in from the outside through connectors. For simple low frequency circuits involving only a few components and a moderate number of connections, perforated circuit board and push-in terminals can be used [see Figure 6.140(a)]. To avoid bending the terminals when they are inserted in the board, an insertion tool matched to the terminal should be used. All connections are individually made and soldered for such circuits. Whenever possible, active components such as ICs and transistors should be used with sockets, and all soldering done with the components removed from the sockets. Perforated circuit board with printed circuit wiring is also a convenient medium for making final circuits [see Figure 6.140(b)]. With this type

of circuit board, all soldering is done on the board and there is no need for separate terminals. To facilitate bringing power to the components, power-supply and ground-bus structures are often part of such boards. For more complex circuits, custom-printed circuit boards (see

---

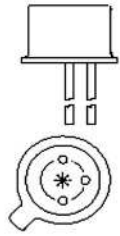
**Table 6.43 Transistor and integrated circuit packages**

---

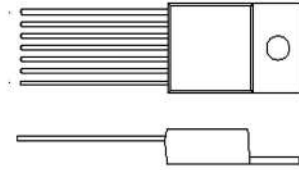
<b>Transistor</b>	
TO-3	Round Can on Diamond Base
TO-5	Round Can
TO-220	Rectangle with Heat Sink
<b>Through Hole IC</b>	
DIP	Dual In-Line Package
PDIP	Plastic In-Line Package
SIL	Single In-Line
SIM	Single In-Line Module
SIP	Single In-Line Package
VIL	Vertical In-Line
ZIP	Zig-Zag In-Line Package
<b>Surface Mount IC</b>	
PSMC	Plastic Surface mount Component
QSOP	Quarter Size Small Outline Package
SOIC	Small Outline IC
SOL	Small Outline Large
SOP	Small Outline Package
SOT	Small Outline Transistor
SSOP	Small Shrink Outline Package
TQFP	Thin Quad Flat-Pack
TSOP	Thin Small Outline Package
TSSOP	Thin Shrink Small Outline Package
<b>IC Modules</b>	
BGA	Ball Grid Array
DIMM	Dual In-Line Memory Module
JLCC	J-Lead Chip Carrier
LCCC	Leadless Ceramic Chip Carrier
LDCC	Leaded Ceramic Chip Carrier
PGA	Pin Grid Array
PLCC	Plastic J-Lead Chip Carrier
SIMM	Single In-Line Memory Module

---

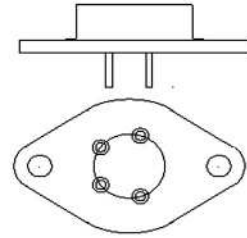
Table 6.43 (contd.)



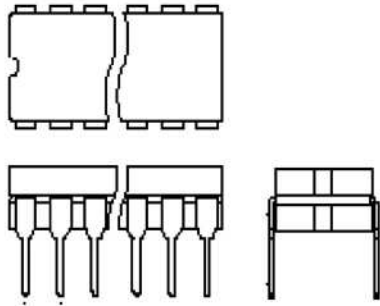
TO-5



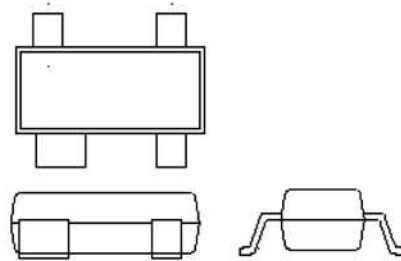
TO-220



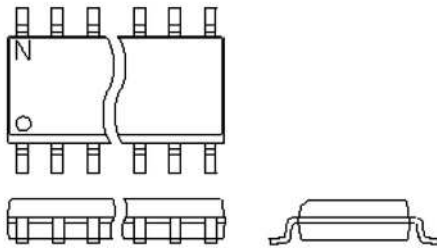
TO-3



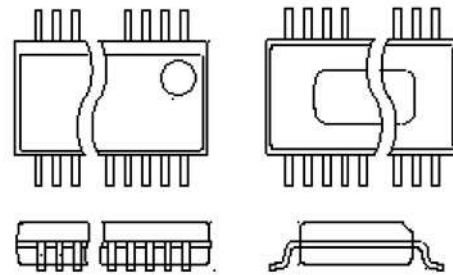
PDIP



SOT



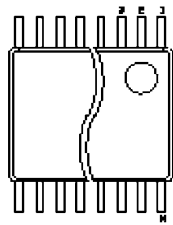
SOIC



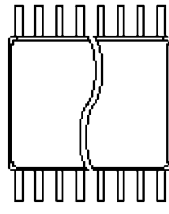
QSOP



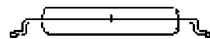
Table 6.43 (contd.)



Top View

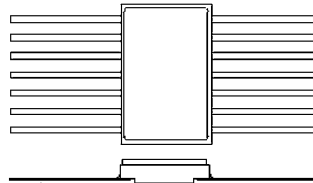


Bottom View

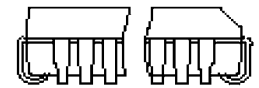
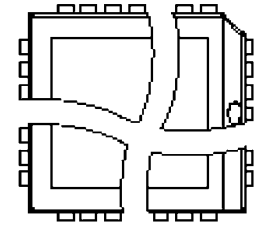


End View

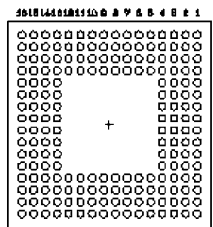
TSSOP



14 L Flatpack



PLCC (J Lead)

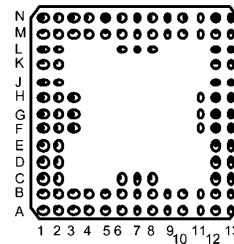


Bottom View

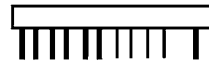


Side View

BGA (Ball Grid Array)

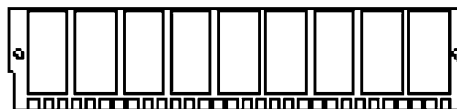


Bottom View

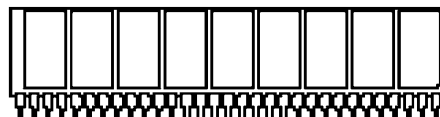


Side View

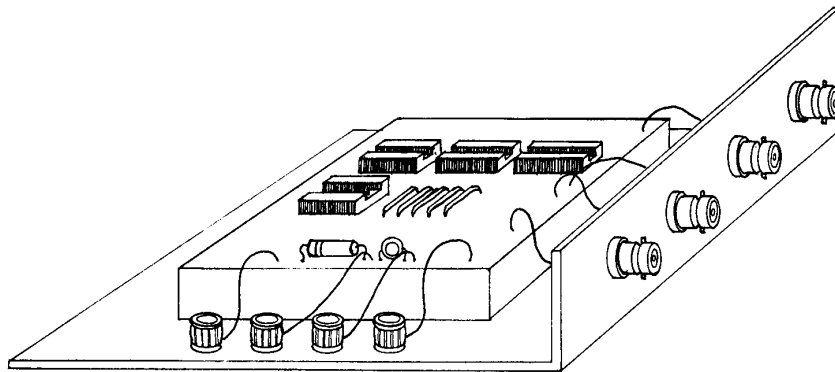
Pin Grid Array (PGAs)



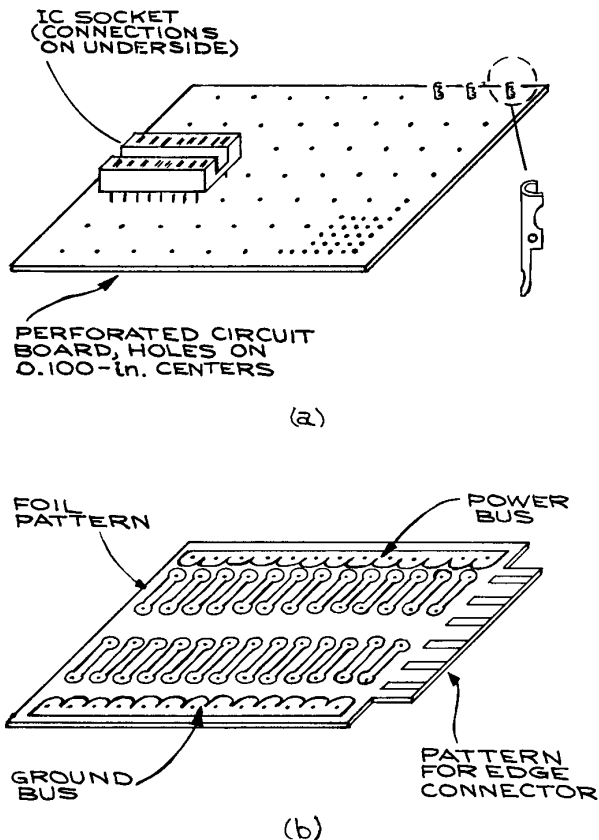
SIMM (Single In-Line Leadless Memory Module)



SIP (Single In-Line Leaded Memory Module)



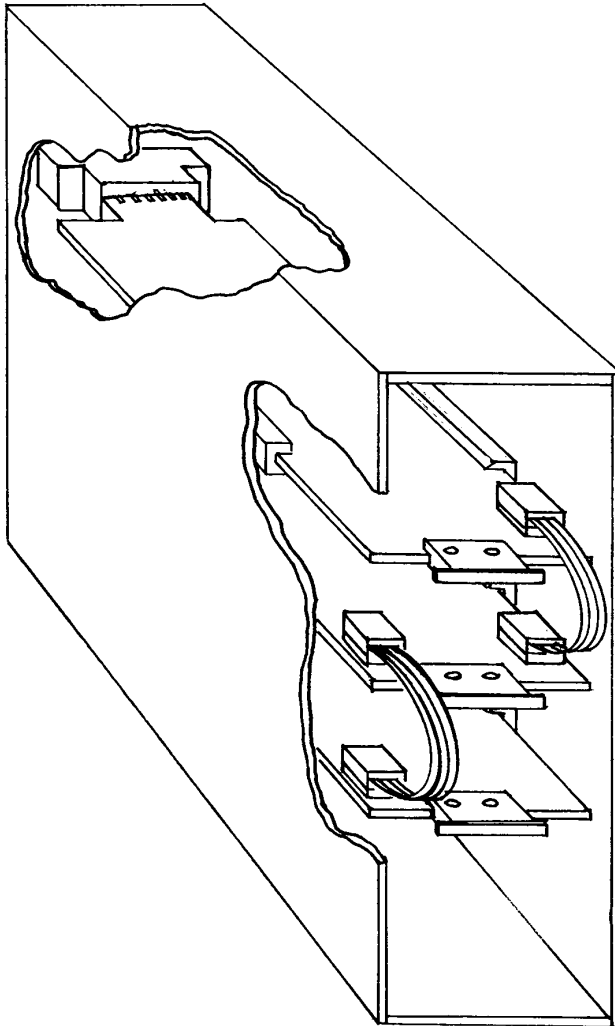
**Figure 6.139** Circuit breadboard mounted to chassis on standoffs to allow proper connection of power supply, input and output leads.



**Figure 6.140** Perforated circuit boards for use with (a) pin connections and (b) direct solder connections.

Section 6.10.3) or Wire Wrap™ boards (see Section 6.10.4) can be used.

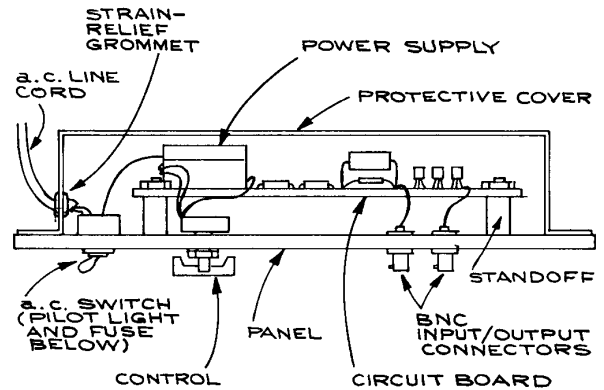
The circuit board containing all the electrical components and a power supply, if required, is typically mounted on a standard 19 in. aluminum rack panel on standoffs. When there is a need for more than two or three circuit boards, the means of mounting and of making the interconnections must be considered more carefully. For multiple circuit boards, *card-cage* mounting – using guides and edge connectors, as shown in Figure 6.141, is a common technique. Interconnections between boards are made between the edge connectors on the rear panel of the chassis. A variation of this method uses a *motherboard*, to which all the connections from the other boards are brought. In this method, the edge connectors provide power to the separate boards and help to mechanically retain the boards in the chassis. Signal leads are brought to the motherboard via multiple pin connectors and flat multiple-conductor cable, using mass termination hardware for increased reliability. A variation of the motherboard configuration includes the edge connectors on the motherboard. An on-off switch, pilot light, and fuse are put on the front panel, as well as input and output connectors and controls. When labeling the panel, it is wise to indicate schematically the electronic functions performed by each section of the circuit with standard symbols. Labels can be applied with the silk-screen technique or with dry transfer letters, India ink, or embossed



**Figure 6.141** Circuit boards mounted in a card cage showing interconnections and guide slots.

pressure-sensitive tape. A protective cover should be placed over the circuit (see Figure 6.142).

Construction can be simple. A portable electric drill or drill press is necessary for drilling proper-size holes in panels and covers. A list of electronic-circuit construction tools is given in Table 6.44 and a list of useful hardware is given in Table 6.45. Flathead screws should be used on exterior surfaces, and lock washers should be used under



**Figure 6.142** Mounting of a circuit board on a rack panel.

**Table 6.44 Electronic-circuit construction tools**

Nutdrivers
Sheet-metal nibblers
Universal electrician's tool
Needle-nose pliers
Slip-joint pliers
1/2-in tapered reamer
Wire cutters
Wire strippers
Soldering gun, pencil
Knife
Single-edge razor blades
Solder sucker
Solder wick
Heat gun
Clip-on heat sinks
Chassis punches (5/8, 3/4, 1(1/8), 1(1/4)in.)
Circuit board holder

*Note:* These tools supplement those listed in Table 1.1.

nuts. Standard .062 in. circuit board can be cut easily with a hacksaw and drilled with standard high-speed drills. A nibbling tool is useful for making irregular-shaped holes in sheet metal, and a set of chassis punches from 1/2 to 1 1/2 in. speeds up work and makes accurate holes in sheet metal. A tapered hand reamer is very useful for enlarging holes in sheet metal and circuit-board material.

**Table 6.45 Hardware and electronic equipment***Hardware*

Screws (flathead and roundhead), nuts, flat washers, lock washers (4–40, 6–32, 10–24)  
 Solder lugs and terminals  
 Binding posts  
 Grommets  
 Standoffs  
 Cable clamps  
 Line cord  
 Hookup wire  
 Spaghetti (flexible thin insulating tubing) in assorted sizes  
 Shrink tubing in assorted sizes  
 Eyelets  
 Fuse holders  
 Indicator lamps  
 Switches

*Electronic Equipment*

Signal sources:

- Signal generator
- Pulse generator
- Logic pulse source

VOM/DVM

Oscilloscope

Logic probe

Logic clips

Test leads for VOM and DVM

Oscilloscope probes:

- × 1 probe for low frequencies
- × 10 compensated probe for frequencies above 10 MHz
- Radio-frequency probe for demodulating r.f. signals

insulated substrate covered with a metal (usually copper) foil. Some of the common rigid substrates used for copper-clad boards are listed in Table 6.46 along with their codes. Nonrigid boards or *flex-boards* are used to fit into irregular confined spaces.

Commercial production of PCBs begins with copper-clad laminated boards, usually epoxy-impregnated fiberglass .062 in. thick, from which copper is removed by chemical etching to form the pattern of traces, pads, and busses that are the electrical connections and mounting areas for the individual electronic components and connectors. To increase circuit density, patterns can be formed on both sides of the board with electrical connections between the layers made by plated holes or *vias*. Sandwiches of up to four boards (eight layers) are common, and current manufacturing capabilities permit as many as 40 layers. After fabrication of the board pattern, holes are drilled for component leads, vias, and fasteners – a thin layer of solder (solder plate) is flowed over the copper and through the vias to protect from oxidation and increase the ease of component soldering. Solder mask – a thin, tough, insulating film – is placed over the board, leaving only the solder terminals and pads uncovered. The mask protects the traces, acts as an insulator to prevent short circuits, and prevents solder bridges from occurring when components are soldered to the board. Component outlines and descriptive information are printed on the board using a silkscreen technique.

The circuit patterns are photographically transferred to the boards using photo-plots and photo-resists. The photo-plots are the equivalent of a photographic negative and contain all of the detail at the required precision of the final circuit board. In the past, photo-plots were made by hand with opaque tape and precut pads and pad arrays, but this has been commercially superseded by CAD/CAM (computer aided design/computer aided manufacture) programs that are now widely available for PCs. These programs have two parts – schematic capture and board layout. Simulation software is often available within the layout programs to verify the functioning of the circuits for a variety of input signals. Printed circuit boards produced with CAD/CAM software have a number of advantages, even at the single board prototype level:

- Uniform electrical and mechanical properties as a result of using standard materials and automated fabrication

### 6.10.3 Printed Circuit Boards

The use of PCBs in laboratory electronics is justified when several identical boards are needed or when the proper functioning of the circuit requires the kind of controlled geometry that printed circuit boards offer. Emitter-coupled logic (ECL) circuits require microstrip line geometries for interconnections and precise placement of components to achieve fast rise times while at the same time avoiding crosstalk between circuit elements. Sensitive low-level amplifier circuits require precise lead placement to eliminate noise pickup. Printed circuit boards consist of an

**Table 6.46 Circuit-board substrates**

<i>Type of board material</i>	<i>Designation</i>	<i>Application</i>
Paper-base phenolic	XXXXP	Hot punching
	XXXPC	Room-temperature punching
Paper-base phenolic	FR-2	Flame-resistant
Paper-base epoxy resin	FR-3	Flame-resistant
Glass-fabric-base epoxy resin	FR-4	General-purpose flame-resistant
	FR-5	Temperature and flame-resistant
	G-10	General purpose
	G-11	Temperature-resistant
	GT and GX	Controlled dielectric constant
Glass-fabric-base polytetrafluoroethylene resin		

*Note:* Standard thicknesses, including the copper cladding, range from 1/32 (0.031) to 1/4 (0.250) in., with 1/16 (0.062) in. the most common.

methods based on standard electronic and mechanical engineering conventions that are built into the CAD/CAM programs

- Error checking in the schematic-capture routines
- Linking the output of the schematic-capture program to circuit-analysis programs for evaluation of the circuit
- Electronic recording and transmission of the photo-plot (Gerber), drill, solder mask, silkscreen, and outline files
- Competitive pricing and fast turnaround.

These advantages have to be weighed against the inconvenience of modifying printed circuit boards that are only meant to be prototypes, and having to acquire and learn the CAD/CAM software. Manufacturers of schematic-capture and PCB-layout programs are listed at the end of this chapter. Trial versions of their software products are generally available and many offer student and educational institution discounts. A number of PCB fabricators make available layout programs that are easy to use for relatively simple circuits. Procedures for producing commercial PCBs using schematic-capture and board-layout programs are detailed in the following two sections.

**Schematic Capture.** A schematic diagram of the circuit is laid out on the computer using components from the program library. Components not in the library can sometimes be obtained from the manufacturer of the component. The library entry contains mechanical and electrical information about the component, as well as the numbering and designation of the pins and/or terminals. Layout is accomplished by arranging the components on the com-

puter screen with the mouse or by specifying coordinates and then interconnecting the pins of the components. Each connection is a node. From the schematic arrangement, a *netlist* is made. This is a list of all the components along with their electrical parameters and the nodes to which they are connected. The netlist is the basis of various error-checking routines to verify that inputs and outputs are not grounded and that no component is shorted.

**Board Layout.** The netlist from the schematic capture program is the input to the board layout program. Layout starts by defining the size and shape of the board and the number and designation of layers. For a two-sided board, one side (the top) is the component side and the other (the bottom) is the wiring side – even though there will generally be wiring on both sides.

Components are placed on the board from a stack of outline drawings (*rat nest*) with connections specified by the netlist. The components with their accompanying wires are initially arranged within the board area according to a few simple rules: separate input and outputs by as large a distance as possible; make power-supply and ground connections as short as possible; minimize the number of crossed traces; and horizontal and vertical routing of traces. Final routing of the connections is done manually for small, uncomplicated boards with only a few components. Autorouting routines are useful for complex boards, but hand routing may still be necessary in certain circumstances. For small boards, it is possible to skip the schematic capture part of the procedure and go directly to the board-layout routine. In this case, parts are placed manually and interconnections made individually.

**Table 6.47 PCB construction considerations**

- (1) Board size, shape material [ 1/16 in. G10 is most common with 1 oz./ft.<sup>2</sup> (35 μm) copper cladding]
- (2) External connections
  - (a) Solder or wrapped to terminals
  - (b) Plugs
  - (c) Edge connectors
- (3) Mounting
  - (a) Screws
  - (b) Guide slots and edge connector
- (4) Component layout
  - (a) Mechanical stress
  - (b) Thermal stress
  - (c) Inputs and outputs separated to avoid positive feedback
- (5) Lead layout
  - (a) 1/16 or 0.050 in. preferred width; no sharp corners
  - (b) 0.015 in. minimum lead clearance, larger in the vicinity of solder connections
  - (c) Crossovers
    - (i) Insulated wire bridges when few crossovers remain
    - (ii) Double-sided board and plated-through holes when many crossovers remain

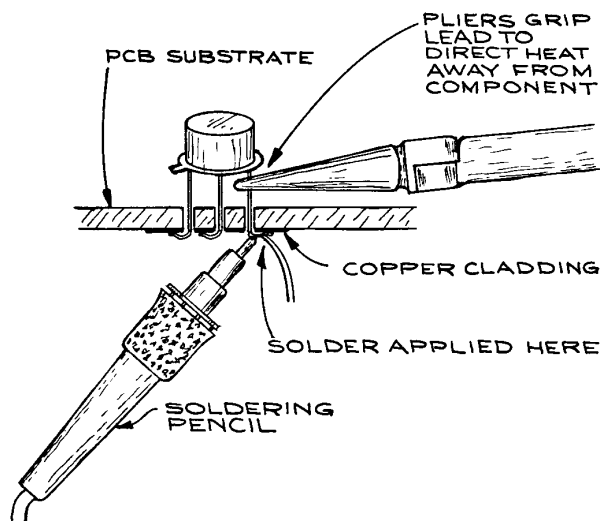
When parts are available in the program library, this can be an efficient method for board layout, but all automatic error checking is lost with manual parts placement. Table 6.47 lists important PCB layout considerations.

The PCB manufacturer requires computer files to fabricate the boards. These files are routinely transferred electronically over the Internet. The files specify the board layout, board size and shape, hole pattern, solder-mask pattern, and silkscreen labels and outlines. It is essential that the files follow precisely the format required by the manufacturer. There is one Gerber file and one aperture file per layer. The Gerber file specifies the interconnections and the aperture file gives the widths of the traces. If solder mask and silkscreen printing are used, each requires a Gerber file and aperture file per layer. Holes are specified by a drill file and a drill list file. For laboratory purposes, the FR-4 grade board is the current standard. The copper-foil cladding is usually cathode-quality, electrolytic copper. The most common thickness is 1.0 oz./ft.<sup>2</sup> (35 μm). Turn-around time for double-sided boards is generally three to five days. A rectan-

gular, double-sided board, 3 in. by 5 in., costs less than \$100. Per board prices fall substantially for two or more boards.

Components are fixed to PCBs by soldering. For large, complex, multilayer boards, the PCB manufacturer will often have the capacity to automatically locate parts on the board and solder them in place using wave soldering for through-hole components and reflow methods for surface-mount components. For surface-mount parts a solder-paste mask is used. This mask has openings corresponding to the solder pads. It is placed on the board, properly aligned in relation to the pads, and solder paste is squeezed over it. When the mask is removed, solder paste covers only the pads, the components can then be placed in position ready for soldering. When hand soldering, through-hole wire lead components are held in place on the board by inserting the leads through the appropriate holes in the board and bending them outward at 45°. The board can then be turned over for soldering the leads to the pads. After soldering, the leads are clipped short with flush cutting pliers, see Figure 6.143. Soldering surface mount components is more difficult and takes some practice.

When the pads on the board have a solder reflow finish, soldering is easier than if solder has to be added. With reflow finish pads, flux is first applied to the pads from a



**Figure 6.143** Pliers used as a heat sink. A heat sink is required when soldering to the leads of heat sensitive components.

flux pen or syringe and the component placed in position with leads aligned with the pads. The component is then tacked in position by heating opposite corner pins with a fine-tip, well-tinned soldering iron from which all excess solder has been removed. Each pin is then heated in sequence to melt the solder and form a high conductivity joint. It is wise to repeat this procedure a second time to insure the secure soldering of all the pins. The board can be examined under a strong light source to check for solder bridges. Bridges can be removed with solder braid.

For boards without a solder plate finish, surface mount components can be soldered in place using the same general technique, but with a thin coating of solder on the iron tip. As the tip is wiped over the component leads solder is transferred from the tip to the leads and pads making effective solder joints. With this technique, solder bridges are more likely and careful inspection is required.

To solder surface-mount resistors and capacitors solder must flow over the circuit board pad and component contact. This is illustrated in Figure 6.144. To solder surface-mount transistors and ICs with coarse-pitch leads to

circuit-board pads it is necessary that solder flow between the component contacts and pads. Fine solder, and fine-tipped soldering iron, a low-power binocular microscope, and steady hands are required for these components. Solder flux dispensed from a pen or syringe with a fine tip can be applied to the pads and the component contacts set down on top of the flux. The contact is then heated with a fine-tip soldering iron and fine wire solder applied until the solder melts and wets the component contact and circuit-board pad.

Although commercial production of PCBs from master artwork has been replaced by fabrication directly from computer files, it is still possible to make a PCB from hand-drawn artwork with kits that are available from electronics suppliers. The kits consist of circuit boards and sensitizers or presensitized boards, developer, ferric chloride etchant, and trays and brushes. The first step is creation of the artwork master, a precise representation of the final traces that will appear on the board. The master is produced using drafting techniques. To increase speed and accuracy of manual drafting, PCB dry-transfer drafting tapes, socket-hole patterns, and terminal pads are available. The tools needed are a

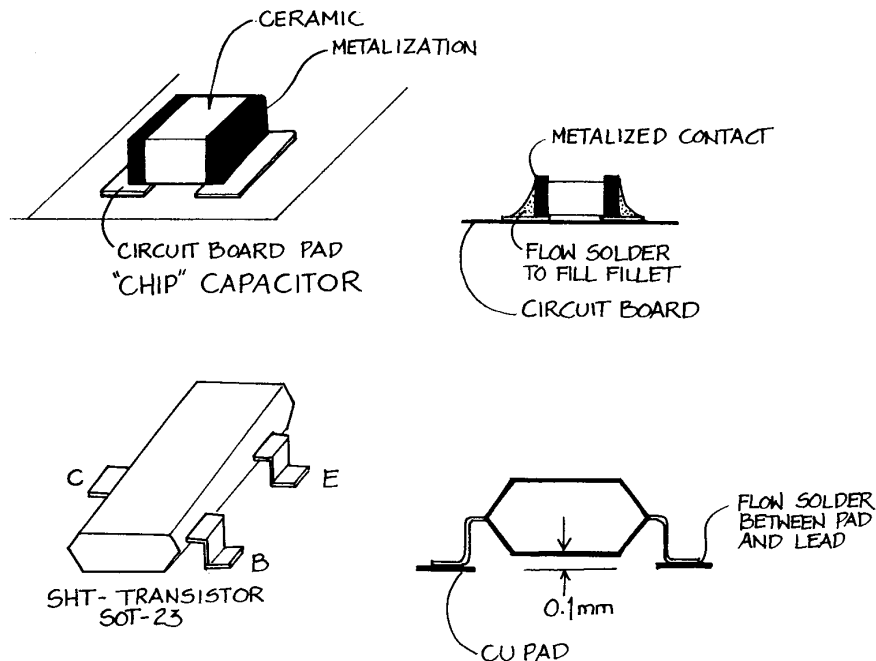


Figure 6.144 Soldering surface mount (SMT) components.

T-square, triangles, an X-Acto™ knife to cut the tape to length, and tweezers for removing the dry transfers from the backing sheet. To avoid dimensional changes with temperature and humidity, the artwork can be prepared on Mylar™ drafting film rather than ordinary paper. The pattern on the master photographic negative is transferred to the circuit board by the use of a *photoresist* sensitizer. Boards can be purchased presensitized, or unsensitized blanks can be coated with a photoresist that is in aerosol or liquid form. A *negative* resist, after exposure to UV light, is unaffected by a subsequent developer solution, whereas UV-exposed, *positive* resist is dissolved by developer. If the master artwork is a positive (that is, the desired foil pattern is represented by opaque areas on a transparent background), a positive resist is required. The sensitized board is exposed to strong light through the full-size master. The board is then placed in a glass tray containing developer. Areas on the board exposed to the light will be dissolved, while those areas protected by the artwork will remain. After full development, the board is rinsed, dried, and placed in an etching solution. It is also possible to laser print or photocopy circuit designs directly onto specially coated transfer paper – available from JDR Microdevices – that is then ironed on to a blank PC board, which is then ready for etching. Several electronics suppliers sell resist pens that are used to draw PCB artwork directly on bare circuit board. Ferric chloride is the most common etchant. It works best when gently heated and agitated. Full etching can take up to one hour per board and should always be done in a glass tray. When etching is complete, the board is rinsed and all etchant residues removed with steel wool or a solvent. A more complete description of the procedure is given in the 1999 edition of the ARRL Handbook for Radio Amateurs. The photoresists have limited shelf life, and the etchant is corrosive.

Transistor and integrated circuit packages for PCB mounting come in a wide variety of forms. Table 6.43 gives the designation of some common packages with representative diagrams.

A technique related to PCBs and based on thick-film technology is used to produce high-density circuit patterns on alumina ( $\text{Al}_2\text{O}_3$ ) substrates. A screen with 200 to 300 lines/in. transfers the desired circuit pattern with a special ink containing suspended metal particles and a binder. When the inked substrate is fired, the metal particles fuse to each other and the substrate, producing a tightly bonded conduct-

ing pattern. Components are then soldered to the appropriate pads on the substrate with microsoldering irons or solder paste and carefully controlled hot air. Thick-film, rather than discrete, resistors are often used with this technique because they can be produced by the same printing technique by using a different ink. Small, high-capacitance chip capacitors replace the more conventional disc and tubular capacitors. With the thick-film technique, it is often quite easy to incorporate ICs and discrete components on the same substrate to produce a self-contained hybrid circuit. Such a circuit can be hermetically sealed and has the appearance of a large IC. The advantages of such circuits are small size, excellent electrical properties, and good mechanical strength. There are a number of companies that produce hybrid circuits starting from schematic diagrams. Because of its specialized nature, thick-film work is best left to such firms.

Solder for electrical and electronic work is an alloy of tin and lead and is specified by the ratio in weight % tin to lead. Thus 40/60 solder is 40% tin and 60% lead by weight. An eutectic alloy with the sharp melting point of 183 °C (361 °F) is formed when the mixture is 63/37. As one departs from this ratio, the melting point becomes less sharp, extending over a larger temperature range. Solder wire can be purchased in diameters from .015 in. (28 gauge) to .062 in. (16 gauge).

Fluxes, whose purpose it is to dissolve the oxides on the surfaces of the metals to be joined, are almost always used in soldering. Rosin fluxes are commonly used for electrical and electronic work. Under no circumstances should acid fluxes be used. These are extremely corrosive and can severely damage components, insulation, and the circuit board itself. Fluxes come in paste and liquid form and are conveniently packaged in pens and syringes for surface-mount component work. There are three types of rosin fluxes: R (low activity), RMA (mild activity), and RA (active). The R type is normally used. There are now fluxes that can be removed by rinsing with water and also *no-clean* fluxes that do not have to be removed after soldering. Solder wire for electronic use generally has flux incorporated in it, though pure flux is also available.

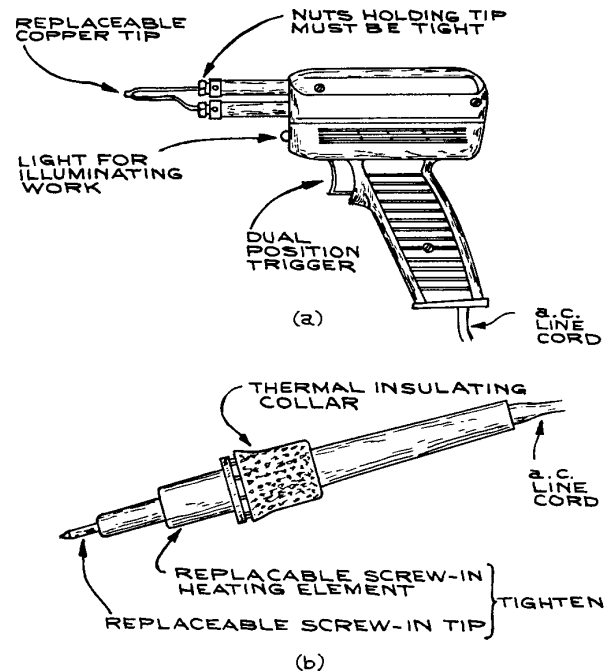
The solder should be maintained at 35 to 65 °C (60 to 120 °F) above its melting point for a time sufficient to completely wet the surfaces to be joined. Too little heat will result in insufficient wetting of the surfaces and a so-called *cold-solder joint*, which is mechanically and



electrically unsatisfactory. Too much heat can destroy components, melt plastic insulation, and burn the circuit board. When soldering heat-sensitive components, a heat sink can be used near the component to protect it. This can be in the form of a pair of long-nose pliers or special tweezers, which are clamped on the lead between the source of heat and component to be protected (see Figure 6.143). To ensure a good joint, the tip of the soldering iron should have a thin bright coat of molten solder covering it, for thermal contact. With time, this coating becomes contaminated with oxides and should be renewed by wiping the surface and reapplying fresh solder. This is called *tinning* the iron. The properly tinned hot tip is then brought in contact with the surfaces to be soldered, and only when they have reached the temperature necessary to melt the solder should solder be applied to them. A common mistake is to melt the solder on the surface of the iron and hope it will flow on to the surfaces to be soldered. Since the molten solder flows toward the hottest point, this method leads to wasted solder and cold-solder joints. Poor solder joints also result from insufficient heat and from contaminated surfaces that resist being wetted by the solder even in the presence of flux. When soldering is complete, the flux should be removed with a commercial flux solvent or isopropyl alcohol.

Increasing recognition of the toxic properties of lead and organic fluxes and solvents has resulted in the availability of lead-free solders combined with noncontaminating fluxes or water soluble fluxes in a variety of formulations. The lead-free solders are mostly tin with copper, silver, bismuth, and antimony added, to enhance to varying degrees mechanical strength, electrical conductivity, thermal properties, and wetting and flow. Many of these solders are compatible with lead-tin solder. Noncontaminating fluxes can remain on the circuit after soldering while water-soluble fluxes can be removed with a deionized water rinse.

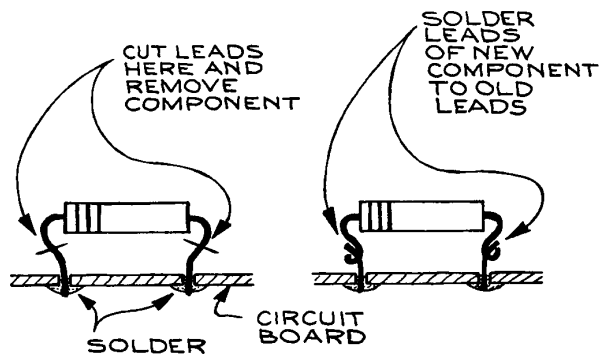
Soldering irons are generally of two types, the soldering gun [Figure 6.145(a)] and the single-element soldering pencil [Figure 6.145(b)]. Guns come in single- and dual-watt models from 100 to 325 W. They have the advantage of being able to be turned on and off quickly, but are generally too powerful, heavy, and bulky for PCB work. Soldering pencils come in wattages from 12 to 75 W with a large variety of tip shapes. Thermostatic control and stands with tip-cleaning sponges are often used with them. For best results, soldering-iron tips should be cleaned and tinned



**Figure 6.145** (a) Soldering gun; (b) single-element soldering pencil.

regularly, and the heating elements should be checked for good electrical, mechanical, and thermal contact to the tips.

It is often necessary to remove a component from a PCB and replace it. One method that does not require desoldering the old component is pictured in Figure 6.146. This technique does not work with ICs and multilead components, where desoldering from the board is necessary. Three common desoldering methods are solder wick, solder aspiration, and heat and pull. *Solder wick* is copper braid coated with rosin flux that withdraws solder from a joint by capillary action when placed on top of the molten solder to be removed. Generally, more heat is required to remove a component than to solder it in place initially. Braid is sold in widths from .030 in. to .210 in. The fine braid is most useful for removing solder bridges between the leads of fine-pitched surface-mount components. Solder can also be removed by using a sucking tool to aspirate the solder after melting it with a soldering iron. The tool can be a rubber bulb with a heat-resistant nonmetallic tip, or a triggered spring-loaded syringe.



**Figure 6.146** Replacing a component on a printed circuit board.

Special heat-and-pull soldering irons are needed to remove multilead components, such as ICs. These irons have tips shaped to heat all leads simultaneously. These tips often do not heat up evenly, resulting in nonuniform heating of the component leads. If the component is lifted before all leads are free, the circuit board traces can be damaged. After removal of the component, the holes must be cleaned of residual solder before proceeding with the insertion of another component. Solder removal by heating followed by shaking or blowing should be avoided. It results in solder splashes that can cause short circuits. Surface-mount components can be removed in a variety of ways. The simplest is to use solder braid to remove as much solder from each of the leads as possible and then heat each lead until the solder is melted and pry it away from its pad with a fine dental pick. When this is done to all the pads, the component can be removed from the board. This method results in the bending of all the component leads and may lift the pads from the board. Printed circuit board tracks that have been damaged by overheating or mechanical stresses can be removed with a knife and replaced with new sections soldered directly to undamaged foil. When pads become detached from the substrate, they can be replaced with new ones anchored to the board with swaged eyelets.

An alternative removal method uses a length of music wire that is inserted behind the leads of the surface mount component. As each pin is heated the wire is slid under it to free it from the pad. This is done for each of the leads until the component is freed from the board. Because the force

on the wire is parallel to the surface of the circuit board, the chance of lifting pads and traces is reduced.

When a large number of surface-mount components are to be soldered, a hot-air rework station gives faster and more uniform and reliable results than hand soldering with a fine tip iron. Such rework stations consist of a source of hot air, a hand piece or holder and a set of nozzles. The openings in the nozzles match the position and size of the contacts of the surface-mount component to be soldered. With these units, the temperature, air-flow rates, and duration of air flow can be individually controlled. The surface-mount components are placed on the circuit board pads, which have been coated with solder paste. The appropriate nozzle is inserted into the hand piece and the air-flow rate, temperature and flow duration selected. The nozzle is placed over the component and the heated air turned on. All contacts are heated simultaneously to the temperature at which the solder flows and the component contacts are bonded to the circuit-board pads. In a few seconds all connections are made. The procedure is repeated for all of the components. The more advanced units allow for programming of air flow, air temperature and heating cycle. Control over lead temperature and heating time means that there is no damage to the component nor to the circuit board. The rework station can also be used to remove surface-mount components from circuit boards by heating the contacts with the correct size nozzle and removing the component once the solder holding the contacts becomes molten. Some rework stations incorporate vacuum tweezers into the hot air nozzles to assist in the removal of the components. Because of the small size of surface-mount resistors and capacitors, and the fine pitch of surface-mount IC contacts, a low-power binocular microscope is very useful for assessing the quality of the solder joints and inspecting for solder bridges between contacts.

In the preceding discussion of PCBs, low-frequency applications have been assumed. If the board is made for high-frequency applications, a great deal more care is required in component placement and lead geometry. For fast logic circuits using ECL ICs, microstrip line geometries are recommended.<sup>17</sup>

#### 6.10.4 Wire Wrap™ Boards

Boards, made by the Gardner–Denver Company under the trade name Wire Wrap™, have sockets spaced at intervals

corresponding to the spacing of the leads on integrated circuits. The sockets have long, square cross-section posts, so that when the socket is swaged into a substrate material the post extends on the back-side. The length of the post determines how many separate connections are to be made to it. Three- and four-layer length posts are the most common (see Figure 6.147). A special wrapping tool is used to make connections to the pins. The tool can be manual, line-operated, battery-operated, or air-operated. The hand tool is entirely adequate for circuits with up to 10 ICs. As the wire is wrapped around the post, the high pressures generated between the wire and the sharp corners of the post form a cold weld, which has good electrical and mechanical properties. The advantages of such a system are:

- (1) No solder connections to cause heat damage to components and circuit board
- (2) Electrical components that plug into sockets and are therefore easily removed for replacement or testing
- (3) Higher component density than obtainable with double-sided circuit boards, because wire crossings are possible without short circuits
- (4) Ease of removing and remaking connections (an unwrapping tool that is a companion to the wrapping tool allows one to remove connections – however, if the connection that one wants to remove is not the highest of a stack, one must remove the upper ones to get to it).
- (5) No artwork, photographic reproduction, masking, or etching required

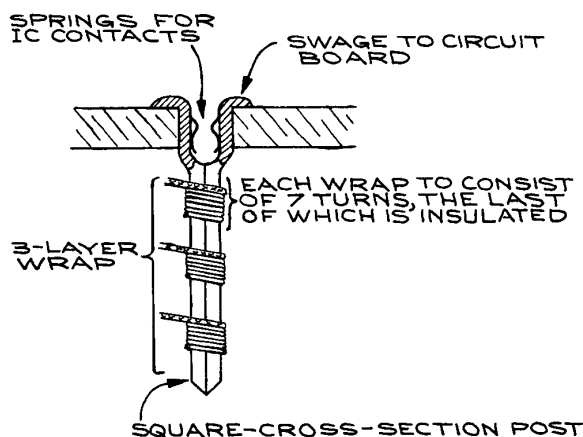


Figure 6.147 Wire Wrap™ post.

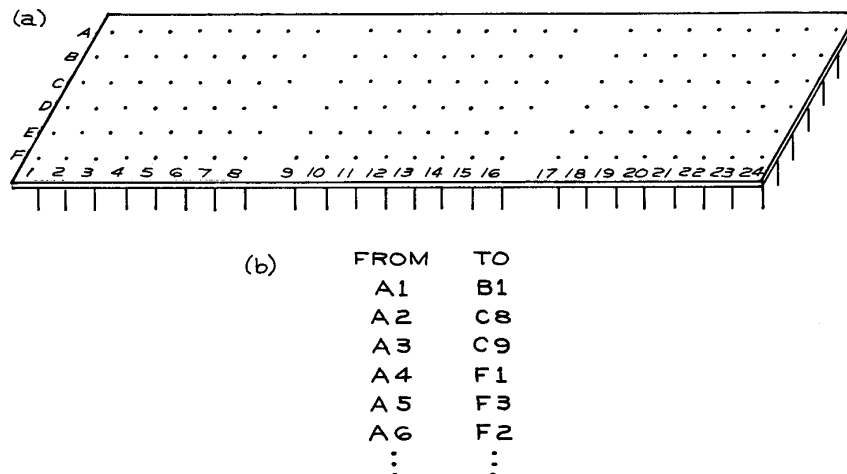
- (6) Reliability
- (7) Speed of fabrication.

In practice, one uses single-conductor wire specially made for wrapping. The three most common gauges are AWG 30 (0.25 mm), 28 (0.32 mm), and 26 (0.40 mm). Different wrap, unwrap, and insulation-stripping tools are used for each gauge. One can obtain the wire in rolls or in precut and prestripped lengths. If one uses rolls, it is necessary to have a wire stripper to remove 1 in. of insulation at either end of the length of wire to be wrapped. Thermal strippers ensure that the conductor will not be nicked when stripped, but special mechanical strippers that do an adequate job can also be obtained inexpensively.

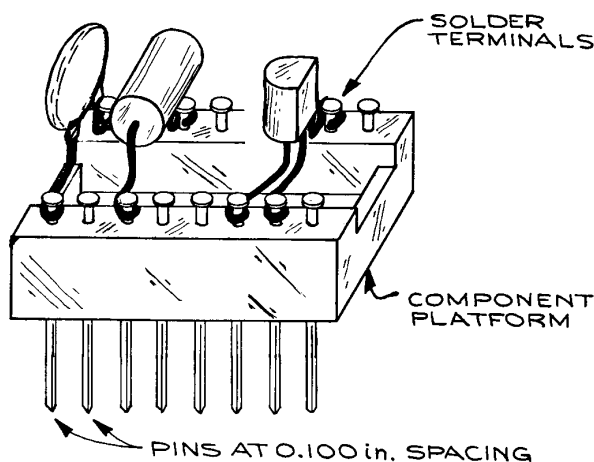
The usual Wire Wrap™ board consists of a matrix of sockets with spacing corresponding to the spacing of the terminals on the normal 14- or 16-pin DIP IC. After the placement of ICs on the board has been decided upon, a Wire Wrap™ list is made, which indicates the pins between which connections are to be made. Each pin is identified by a letter plus number code identifying the column and row of the pin [see Figure 6.148(b)]. Once the list is complete, all the wrapping can be done at one time. It is common for Wire Wrap™ boards to have a ground plane and a power-supply voltage plane or bus. The Wire Wrap™ posts can be connected to these planes at intervals of 14 to 16 pins. A solder bridge can sometimes be made between the appropriate pin and the exposed ground plane. Connections to the plane from outside the board can be made with a screw-and-lug connection to tie points on the board. Signal connections are usually via ribbon connectors to DIP sockets.

Discrete components are used with mounting platforms or, if sufficient space is available, they can be wrapped directly to the posts. The platforms have pins that fit the socket holes in the board, and the discrete components are soldered to the posts or forked terminals on the top of the platform (see Figure 6.149). When a circuit with only a few ICs must be constructed, an inexpensive method is to use individual Wire Wrap™ sockets and glue them to perforated circuit board with 0.100 in. hole-spacing. This is much less expensive than commercial boards and entirely adequate for low-frequency applications. A disadvantage of Wire Wrap™ boards is their high initial cost, but they can be reused.

Initial verification of the connections on a board is done by removing all components, attaching the ground



**Figure 6.148** (a) A Wire Wrap™ board; (b) example of a wrap list.



**Figure 6.149** Platform for mounting discrete components. The platform plugs into the Wire Wrap™ board.

and power-supply leads, and measuring the voltage at each pin. They should all be consistent with the Wire Wrap™ list.

### 6.10.5 Wires and Cables

There are many types of electrical wires used in the laboratory. Selection of the correct type is important for correctly functioning equipment. Wire should be selected according to its voltage and current rating for routine low-

frequency operation. For applications involving high frequencies and low signal levels, more care must be taken. The use of multiple-conductor cables can simplify wiring, and some knowledge of the kinds of multiple conductor configurations is useful. Table 6.48 lists the diameter, allowable current, and resistance per 1000 feet of B&S-gauge

**Table 6.48** Current-carrying capacities of copper-insulated wire

B&S Gauge	Diameter (in.)	Allowable Current <sup>a</sup> (A)	Resistance per 100 ft. <sup>b</sup> (Ω)
8	0.128	50	0.628
10	0.102	30	0.999
12	0.081	25	1.588
14	0.064	20	2.525
16	0.051	10	4.016
18	0.040	5	6.385
20	0.032	3.2	10.15
22	0.025	2.0	16.14
24	0.020	1.25	25.67
26	0.016	0.80	40.81
28	0.013	0.53	64.90
30	0.010	0.31	103.2

<sup>a</sup> For rubber-insulated wires the allowable current should be reduced by 30%.

<sup>b</sup> At 20 °C (68 °F).

**Table 6.49 Electrical properties of thermoplastics**

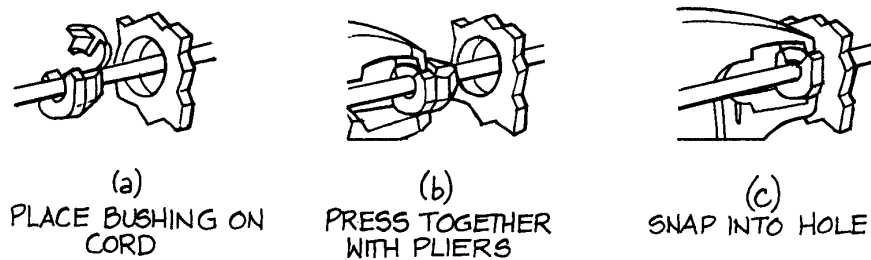
<i>Material</i>	<i>Trade Name</i>	<i>Volume Resistivity (<math>\Omega</math> cm)</i>	<i>Dielectric Strength (V/mil)<sup>a</sup></i>	<i>Power Factor at 60 Hz</i>	<i>Characteristics</i>
ABS	Lustran	$10^{15}$ – $10^{17}$	300–450	.003–.007	Tough, with average overall electrical properties
Acetals	Delrin	$10^{14}$	500	.004–.005	Strong with good electrical properties to 125 °C
Acrylics	Lucite, Plexiglas, Perspex	$>10^{14}$	450–480	.04–.05	Resistant to arcing
Fluorocarbons:					
CTFE	KEL-F	$10^{18}$	450	.015	Excellent electrical properties, some cold flow
FEP	Teflon FEP	$>10^{18}$	500	.0002	Properties similar to TFE, good to 400 °F
TFE	Teflon TFE	$>10^{18}$	400	<.0001	One of the best electrical materials to 300 to 500 °F; cold flow
Polyamides	Nylon	$10^{14}$ – $10^{15}$	300–400	.04–.6	Good general electrical properties; absorbs water
Polyamideimides and polyimides	Vespel, Kapton	$10^{16}$ – $10^{17}$	400	.002–.003	Useful operating temperatures from 400 to 700 °F; excellent electrical properties
polycarbonates	Lexan	$10^{16}$	410	.0001–.0005	Good electrical, excellent mechanical properties; low water absorption
Polyethylene and Polypropylenes	—	$10^{15}$ – $10^{18}$	450–1000	.0001–.006	Good electrical, weak mechanical and thermal properties
Polyethylene terephthalates	Mylar	$>10^{16}$	500–710	.0003	Tough, excellent dielectric properties
PVC	Saran	$10^{11}$ – $10^{16}$	300–1100	.01–.15	Low cost; general purpose; average electrical properties

<sup>a</sup> mil = .001 in.

insulated copper wire. Table 6.49 lists the electrical properties of common thermoplastics used for insulation.

The most common wire is a.c. line cord, the type used for connection to the a.c. outlet. Twin-conductor *zip cord* should be avoided in laboratory applications because of its limited resistance to mechanical stresses. Three-conductor

color-coded line cord is best suited for the laboratory. The double insulation of the wires provides extra mechanical and electrical protection. The standard code is black for hot, white for neutral, and green for ground, and should be observed at both the plug and the chassis end of the cord. To minimize damage to the line cord at the chassis end,



**Figure 6.150** Mounting line cord at the chassis to relieve strain on the electrical connections.

strain relief and mechanical protection of the insulation should be provided. This is often accomplished with a single strain-relief grommet or combination plastic grommet and clamp, as in Figure 6.150. An alternative to direct connection of an a.c. line with plug to a chassis is an *a.c. receptacle* or *power entry* module. These accept standard a.c. line cords with male connectors on the a.c. outlet end and female connectors that mate to the receptacle or module. The receptacles and modules may contain on-off switches, fuses, pilot lights, and rfi filters. Some examples are shown in Figure 6.151.

Line cord is almost always stranded to enhance flexibility. If the wire is to be used in high-current-carrying applications, care should be taken not to cut the strands when stripping, since that will limit the capacity of the wire and cause heating at the stripped end. Sometimes the individual strands are insulated with varnish. When making electrical connections with wire of this kind, the insulation must be removed with fine emery paper or steel wool.

Another type of wire used for a.c. lines is the solid-conductor #12 or #14 gauge wire for power distribution. The most common are two- and three-conductor Romex, which is PVC-insulated, and two-conductor BX, which has a metal-armored outer covering. Both types can be routed within walls or outside, in either metal or PVC conduit. Because of the heavy gauge of the conductors, connections are made to screw terminals or with *wire nuts* when splices are involved. The installation of a.c. lines requires a professional electrician who knows the conventions, regulations (codes), and procedures for fusing and connection to existing power lines.

For electronic circuits where the voltages are less than a few hundred volts and the currents are a few amperes,

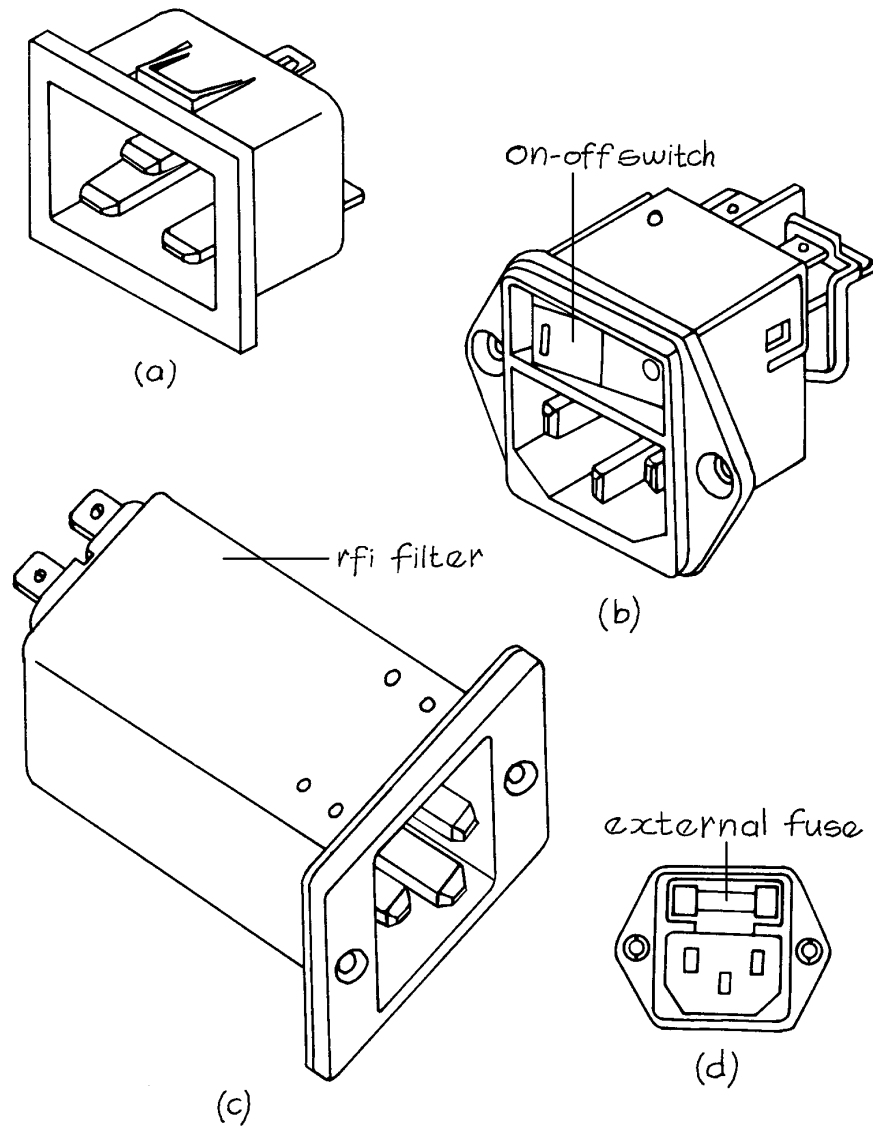
hookup wire is used. This can be solid conductor or multi-stranded. The solid-conductor wire is easier to use when making connections because it does not fray, but it lacks the flexibility of multistranded wire. Hookup wire is generally *tinned* – that is, coated with a tin-lead alloy to enhance solderability. The insulation is usually PVC, polyethylene, or Teflon™.

When stripping wire, care should be taken to avoid nicking the conductor of the solid wire and cutting strands of the multistranded wire. There are a large variety of wire strippers on the market (see Figure 6.152), but none is foolproof and all require a certain amount of skill if the conductor is to remain undamaged. Thermal strippers that melt the insulation locally before it is withdrawn from the wire work very well, but are not portable. For Teflon™-insulated wire, thermal strippers are very useful because the slippery material is difficult to grip.

When working with stranded wire, it is recommended that the ends be tinned. It should be remembered, however, that solder will render multistrand wire rigid, so tinning should be confined to a short length at the end. For coding purposes, the insulation of hookup wire comes in many colors and combinations of colors. Color coding is a very useful way to avoid wiring errors and makes subsequent troubleshooting easier.

*Test-prod wire* is highly flexible multistrand wire with rubber insulation rated at a few thousand volts. Such wire comes with red or black insulation and is used with multi-meters and in high-voltage circuitry.

For voltages from a few kilovolts to several tens of kilovolts, special high-voltage cable must be used. The most common jacket materials are silicone rubber, Teflon™, and Kapton™. Silicone-rubber insulation is



**Figure 6.151** A.c. receptacle/power line modules: (a) simple; (b) onoff switch incorporated; (c) with internal r.f.i. protection; (d) external fuse.

very flexible, but has only modest dielectric breakdown strength and volume resistivity, so that high-voltage cables made with it have large diameters. Kapton™, a polyimide, has high breakdown strength and volume resistivity, and cables made with it are relatively

small in diameter. Stiffness is the principal disadvantage of Kapton™. The properties of Teflon™ are intermediate between those of silicone rubber and Kapton™, and it is usually the best compromise. Typical properties of high-voltage Teflon™ cable are given in

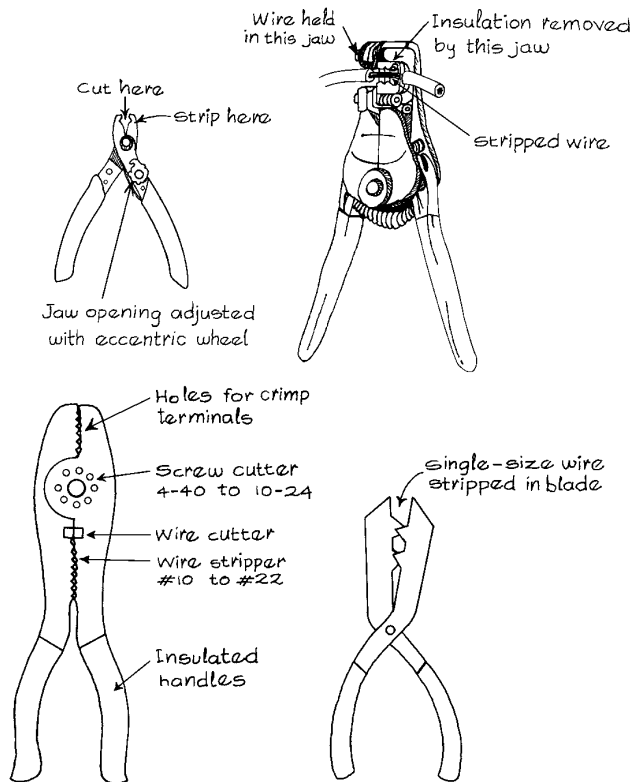


Figure 6.152 Wire strippers.

Table 6.50. One source of high-voltage cable is the ignition cables used in automobiles. The silicone-rubber-insulated stainless-steel or copper stranded-conductor type found in high-performance cars can be used to 50 kV. The more common carbon-filament conductor cable used to minimize r.f.i. is not suitable.

*Magnet wire* is varnish-insulated, solid-conductor wire used for winding electromagnets, transformers, and inductors. The thin insulation means that a large number of turns can be packed into a given volume. To make electrical connections with such wire, the varnish must be removed with emery paper, steel wool, or special solvents. Magnet wire is not a suitable substitute for hookup wire. When using it for electromagnets, one should be aware of the possible buildup of heat around the inner windings, which can degrade the insulation and create

Table 6.50 Typical high-voltage Teflon-insulated cable

d.c. Voltage Rating <sup>a</sup> (kV)	Conductor	Cable Diameter (in.)
10	19/36 <sup>b</sup>	0.077
15		0.097
20		0.117
25		0.137
40		0.197
50		0.237

<sup>a</sup> a.c. voltage rating is typically 20–25% of d.c.

<sup>b</sup> 19 strands of #36 wire.

short circuits. The varnish insulation is rated for voltage break-down and heat resistance.

For high-current electromagnets, rectangular cross-section wire is used, because the high surface-area-to-volume ratio facilitates convection cooling. In applications that require conduction cooling, hollow-core wire is used, through which cooling water can be circulated.

In addition to two- and three-conductor line cord, there are many other multiconductor, round-cable configurations. The conductors can be solid or stranded. The wires are color-coded for identification. For low-level, low-frequency signals, the wires can be enclosed in a single shield, as is done for microphone cable. Configurations where there is individual shielding of single wires also are available. Shielding can be in the form of a braid or a foil – if it is foil, a *drain wire* is also included, which is electrically connected to the shield and to which connections can be made. For high-frequency applications there is coaxial cable and 300  $\Omega$ , parallel-conductor cable with transmission-line properties. These are discussed in the sections on coaxial cable and connectors (Sections 6.2.3 and 6.2.4).

Flat multiconductor cable is very useful when a large number of wires are needed, as with computer and data interfaces. There are many different arrangements of the wires within such cable. The most common consists of round parallel conductors. There are also flat conductors and alternating round and flat conductors. Flat conductors minimize interconductor capacitance and interference among signals. Twisted pairs of wires are used for transmission lines because constant impedance



can be maintained and the wires are flexible and easy to route. An alternating twisted pair–straight geometry is often used for data transmission. The great advantage of flat cable is that mass termination insulation displacement connectors can be used. These connectors are designed so that the cable is clamped and electrical contact is made to each of the wires in a single operation with a special tool. No stripping, soldering, or crimping is required. The reliability of such connections is excellent.

### 6.10.6 Connectors

Probably the best-known connector combination is the *binding post* and *banana plug* or *tip plug* (see Figure 6.153). A wide variety of conductors and conductor

terminations can be made to binding posts in a rapid, reliable semipermanent manner. Binding posts have the disadvantage of being bulky and highly susceptible to the pickup and radiation of electromagnetic energy. When mounting binding posts on a chassis, the mounting hole should be made large enough to accommodate the shoulders on the insulating sleeves. In this way, the central conductor is kept well away from the chassis. The standard spacing between posts is 0.75 in. This corresponds to the spacing of twin-conductor banana plugs (see Figure 6.154), which are quite common and useful, especially when it is necessary to connect coaxial cable to binding posts. The twin plug usually has one terminal marked *ground*, to which the shield of coaxial cable is attached.

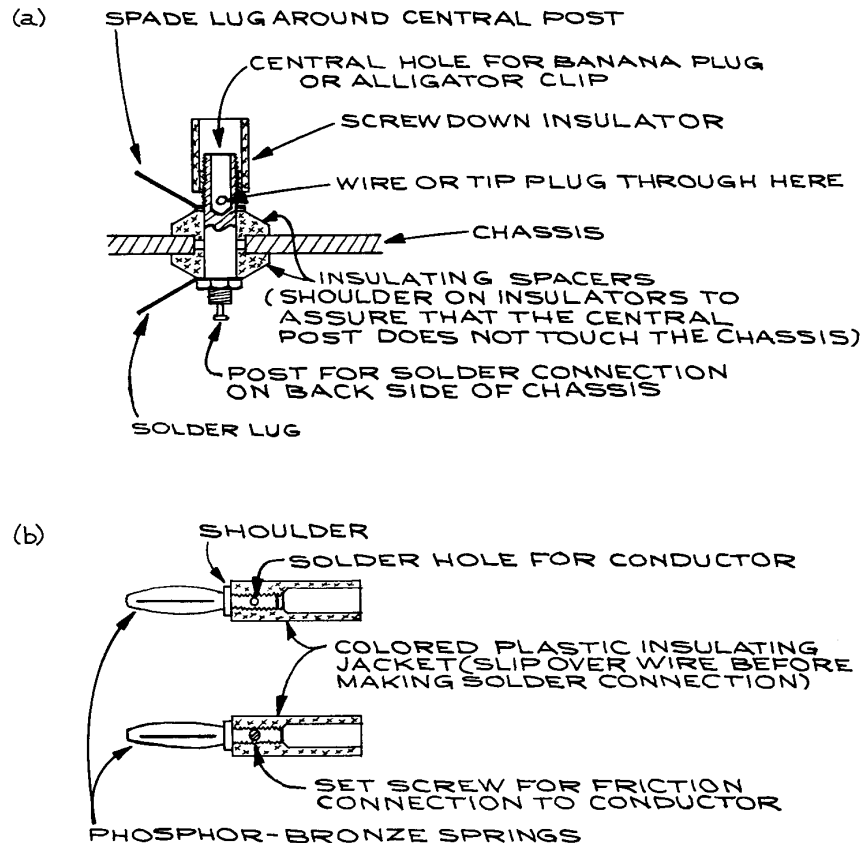
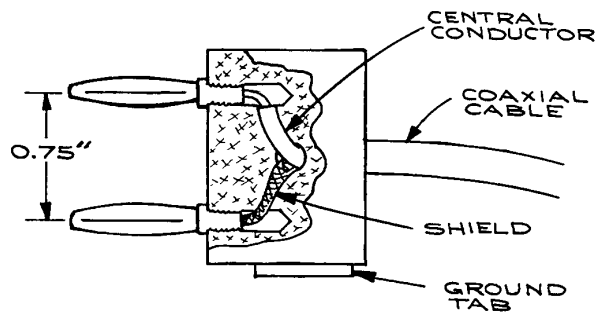
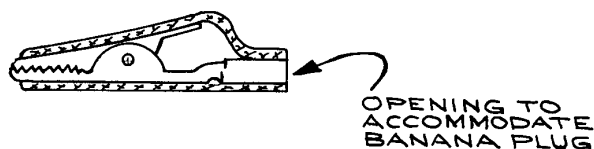


Figure 6.153 (a) Binding post; (b) banana plug.



**Figure 6.154** Twin-conductor banana plug.



**Figure 6.155** Alligator clip with insulated plastic cover.

*Alligator clips* (see Figure 6.155), although useful with test leads, should not be used in any permanent or even semipermanent installation. The type of clip into which one inserts a banana plug is more useful than a clip requiring solder or screw connections. Flexible insulating sleeves that fit over the clips prevent short circuits and are always used when the clip is at the end of a test lead.

*Phone plugs* (see Figure 6.156) come in several different sizes and provide one means for the termination and connection of coaxial cable carrying low-frequency signals. Because this type of plug is polarized (that is, the connections can only be made in one way), it is sometimes used for low-voltage, low-current d.c. power-supply connections. There are many variations on the basic plug-jack design. There are plugs that can accommodate three or more separate connections, and there are jacks that remain shorted until the insertion of the plug.

Insulated *barrier strips* (see Figure 6.157) with screw terminals are a good way of making semipermanent connections for d.c. applications, and are often found on the rear panels of power supplies.

For terminating the ends of wires, there are a variety of terminal connectors that can be attached to the wire by soldering or crimping (see Figure 6.158). The terminal used should be of the correct size with respect to both the wire used and the opening in the terminal. Quick-connect, friction-type, push-on terminals are very useful in d.c. applications, when the connection must be repeatedly made and broken. Such terminals are often also found on mechanical relays, circuit breakers, and mechanical switches. Crimp connectors are color-coded according to the wire-size range they can accommodate. It is important to use the proper crimp-tool opening – color-coded on many tools – to avoid too loose a crimp (too large a hole) or a severed wire (too small a hole). Three different types of crimping tool are shown in Figure 6.159.

When soldering connectors, only enough heat should be used to make a good joint, since too much heat will melt the insulation. Some type of *third hand* is helpful in this regard – a small vice or an alligator clip at the end of a heavy piece of solid copper wire anchored to a metal base.

A very large part of the electronics-hardware industry is devoted to the manufacture of connectors. For laboratory applications, only a few of the most common ones will be described, and some guidelines for connector selection will be given. Electrical contacts between the wires and connector pins can be made by soldering or crimping. For some connectors, the pins must be removed to make the electrical connection and then inserted into the connector block. Removal of the pin then requires a special tool. A common power connector using crimped sockets and pins is shown in Figure 6.160. High-density connectors using crimped terminals and integral holding latches are shown in Figure 6.161.

D-connectors derive their name from the cross-sectional shape of the connector body (see Figure 6.162). The usual RS232C interface uses a 25-pin, subminiature D-connector. The connector can have up to 50 contacts and connection is made in a variety of ways: by soldering the conductors into a hollow recess at the back of the contact (solder pot); with crimp pins and sockets; and with ribbon-cable mass termination contacts. Such connectors can be attached directly to a panel with the correct cutout. When fitted with a shell (either plastic or metal) and a cable clamp, they become plugs. Locking accessories are

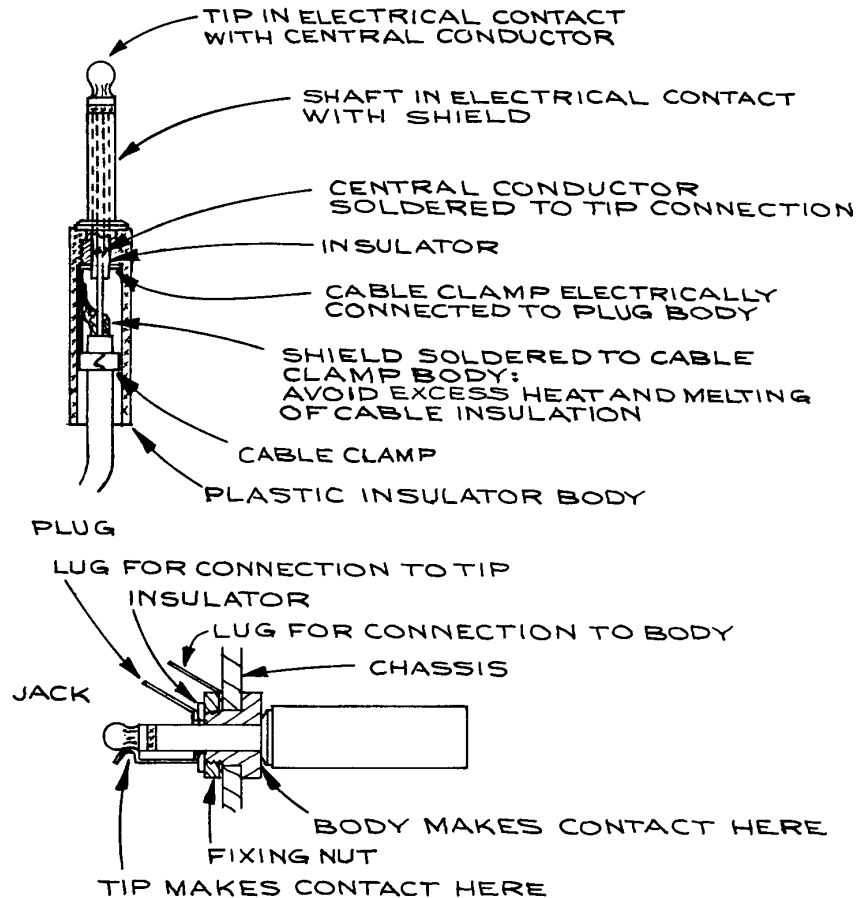


Figure 6.156 Phone plug.

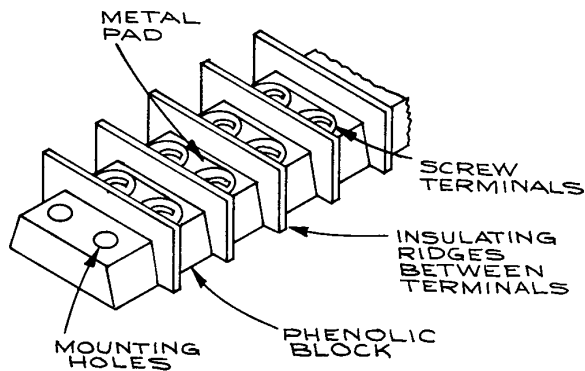


Figure 6.157 Barrier strip.

also available for securing the plug to the mating jack, and the D-shape assures correct mating. Double-density connectors of this type are also made, and there exist a wide variety of configurations with high-voltage, high-current, and coaxial contacts, in addition to the standard single pins. Because pins are very closely spaced, *shrink tubing* is routinely used over the solder connections. This tubing comes in a variety of sizes, materials, and shrink ratios. When used to insulate solder connections, the tubing serves a strain-relief as well as an insulating function. It is important to choose the correct diameter tubing for the application. If the tubing is too small in diameter, it will split upon being shrunk over the enclosed wire or

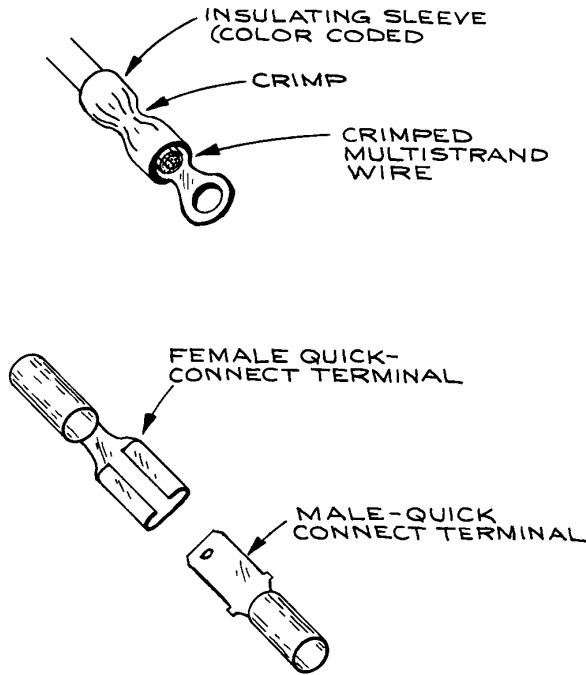


Figure 6.158 Wire terminals.

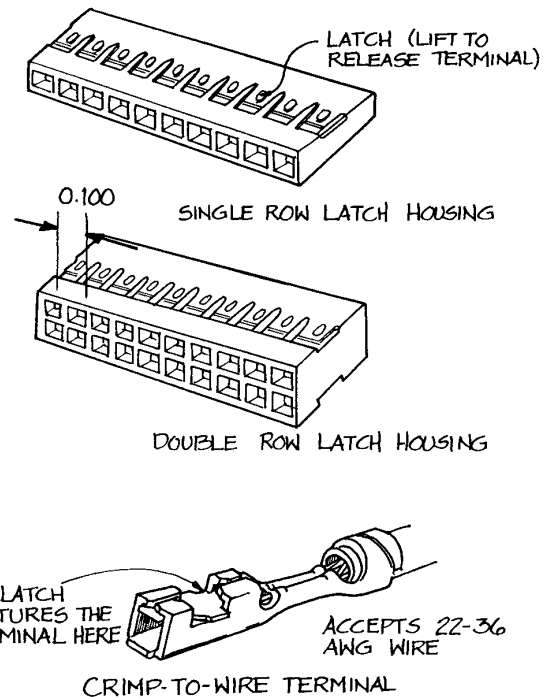


Figure 6.160 Multiple contact connector, crimp terminals, insertion/extraction tool.

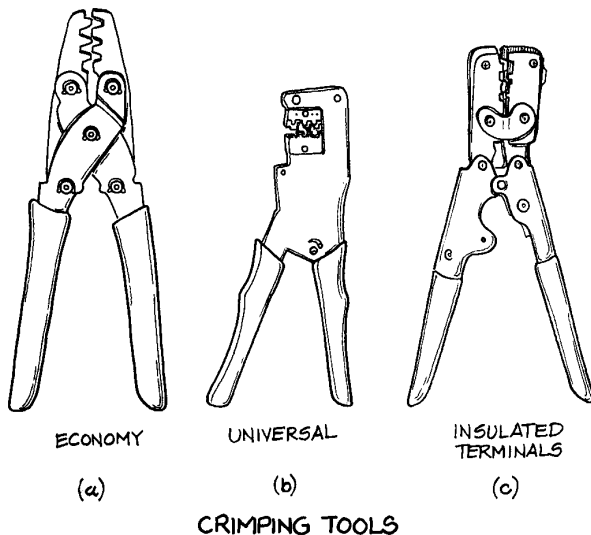


Figure 6.159 Crimping tools.

connector – if the diameter is too large, it will not fit tightly. Heating should be done with a heat gun rather than a soldering iron.

Threaded circular connectors (see Figure 6.163) are based on designs originally used by the military for aircraft. They employ separate removable pins and a threaded mating collar to make a mechanically secure union between male and female connectors. There is a broad range of sizes, styles, and pin arrangements, and for this reason these connectors are used on a variety of laboratory electronic equipment. There are various ways of retaining the pins in the insulator block. Some of them require special insertion and removal tools, while others use a second backup block held in place inside the connector.

Nuclear instrumentation module (NIM) and CAMAC equipment are examples of the connection of a modular electronic unit to a main chassis. The basic connector is a rectangular insulator block, which holds contact pins

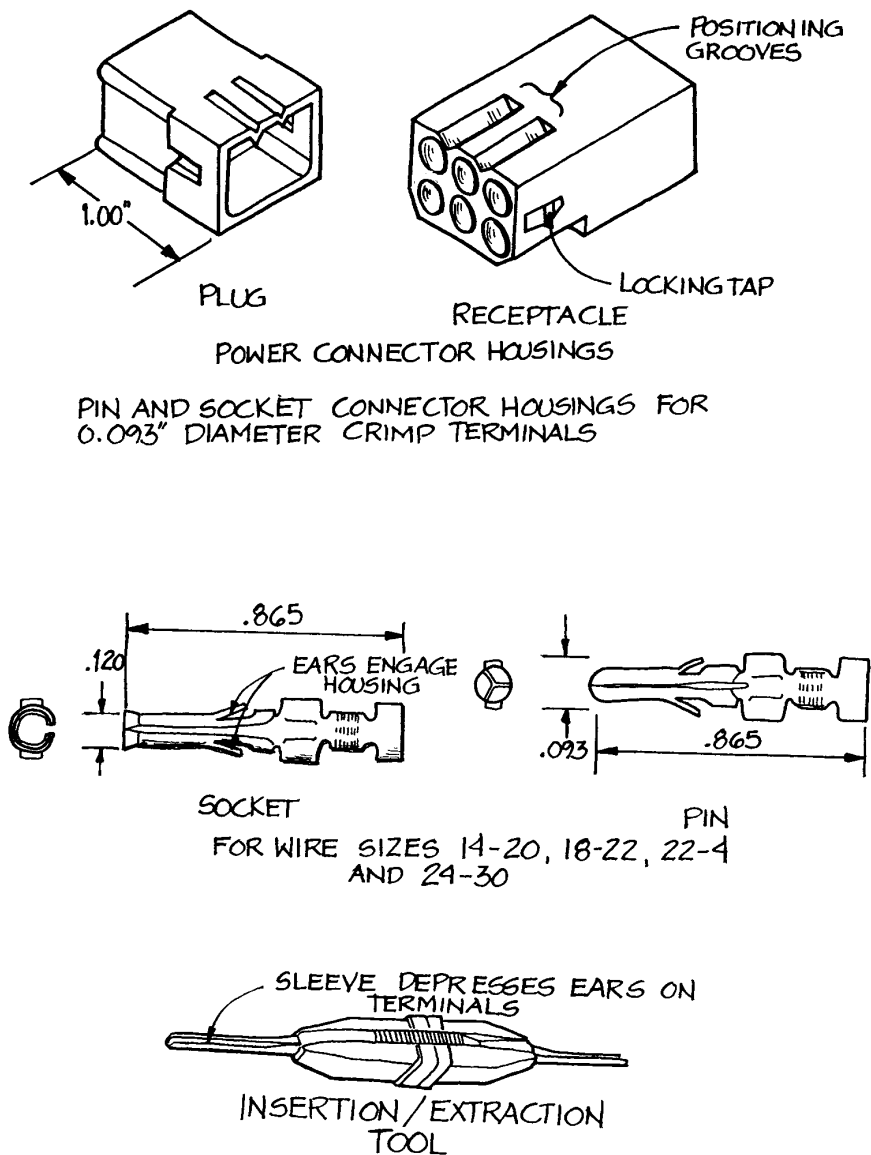
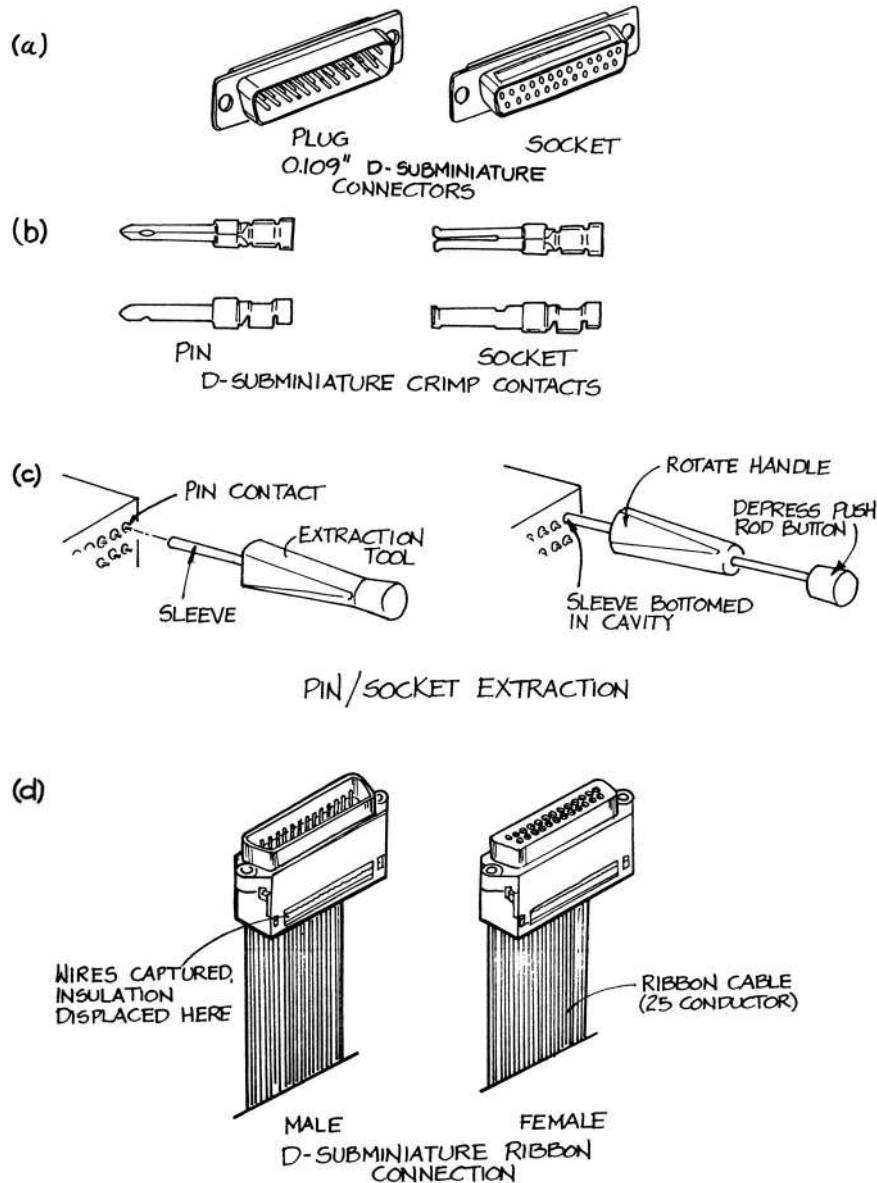


Figure 6.161 Multiple conductor connector and crimped terminal.

maintained in place by one-way spring action. Wires are crimped to the pins that are then inserted into the block. An extraction tool is required for pin removal. When the connector has a large number of pins, considerable force is required for mating and threaded screw jacks are often used.

Extra pins can be purchased separately and shrouds are available to convert the connector to a plug.

No high-voltage connector is entirely satisfactory, and for this reason they should be avoided and permanent connections with ceramic feedthroughs or standoffs made



**Figure 6.162** (a) D-connector; (b) pin and socket crimp contacts; (c) pin and socket extraction with tool; (d) ribbon cable connections.

whenever possible. Reynolds/Teledyne manufactures a wide range of single-conductor, multiple-conductor and co-axial high-voltage wire and connectors with voltage ratings to 70 kV. European-type phenolic sparkplug

connectors make acceptable connectors when used with ignition cable; a drilled-out Teflon™-insulated r.f. coaxial connector can also be used. These are illustrated in Figure 6.164.

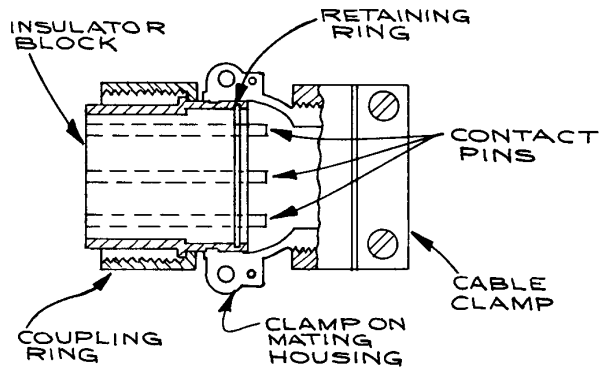


Figure 6.163 Threaded military-type connector.

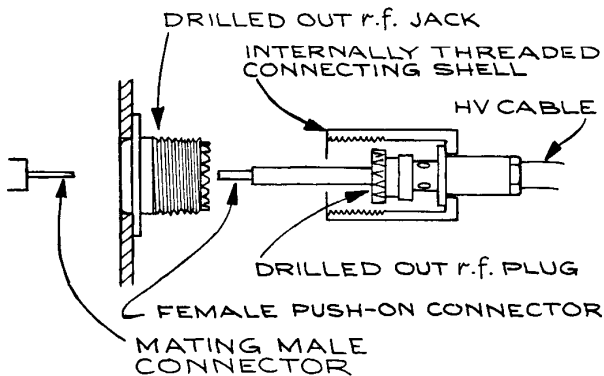
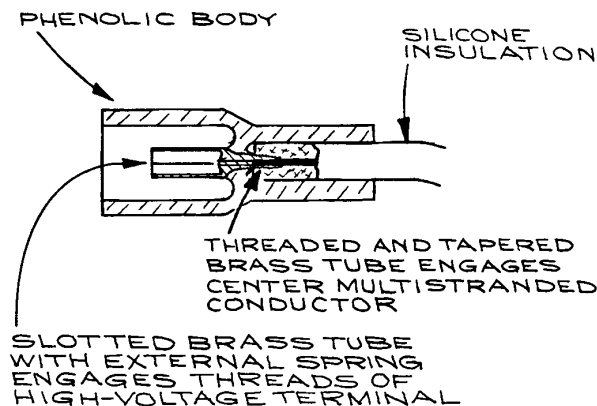


Figure 6.164 High-voltage connectors made from a spark-plug and r.f. connector.

## 6.11 TROUBLESHOOTING

### 6.11.1 General Procedures

There are a number of steps that can be taken when a piece of electronic equipment fails to operate properly. These depend on the complexity of the equipment and the availability of test equipment and diagrams.

The worst symptoms are usually the easiest to treat. An experienced TV repairman once noted that 95% of all malfunctions in TV sets could be found and corrected in less than one half hour. For the other 5% the best solution was to sell the customer a new set. This also applies to laboratory equipment. With increasing use of highly reliable integrated circuits and resistance and capacitance modules, the most unreliable elements of a circuit are the mechanical parts (such as switches, dials, and fans) and especially the connectors. When an amplifier or power supply no longer works, one's first impulse should not be to take out screwdriver and wrench and begin disassembling the chassis, but to be sure that it is plugged in, that the on-off switch is functioning, that all fuses are intact, that the cooling fans are operating, and that the air filters are clean and not blocked. Switch contacts can be cleaned with spray cleaners made for the purpose. If these measures produce no results, the various dial and switch settings of the instrument should be checked. Often what is viewed as a malfunction is merely an incorrect setting, causing the instrument to operate in an unexpected mode.

If these simple actions are not effective, the next course of action depends on the complexity of the equipment and the availability of circuit diagrams. To repair a complex circuit without circuit diagrams is almost impossible. If diagrams are not at hand, the manufacturer should be contacted for the necessary diagrams and troubleshooting procedures. Often these are available online, so a web search is a good way to begin.

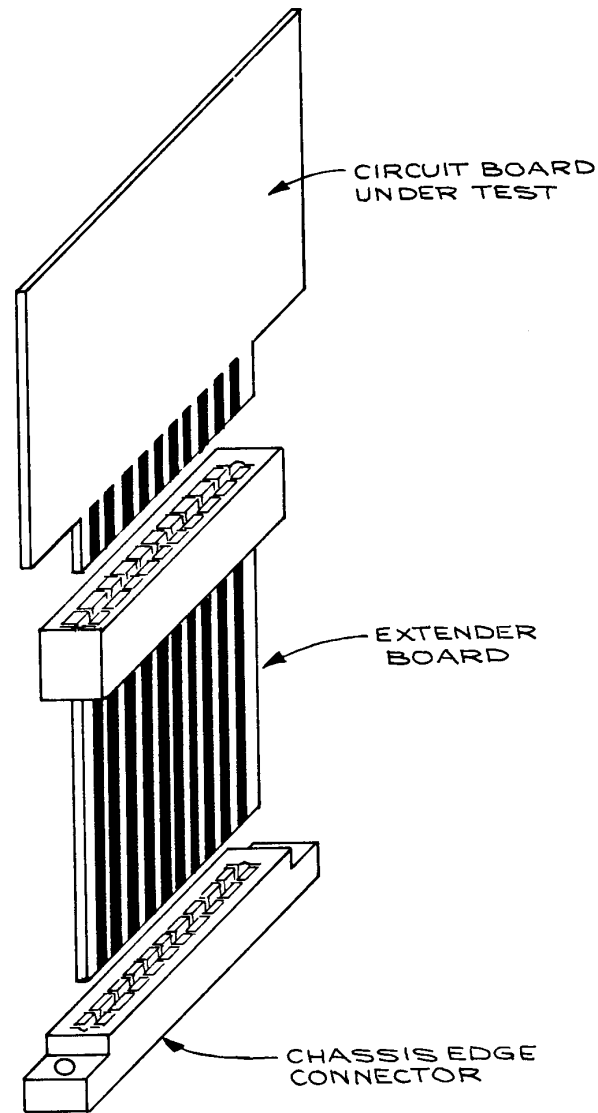
There are, of course, certain visual checks that one can perform without knowledge of the circuit. Charred resistors and an overheated circuit board are readily apparent, as are loose and dangling wires. One difficulty with merely replacing a damaged component, however, is that only the symptom may be treated. The charred resistor may be the result of a failure in another component.

Troubleshooting is most effective and efficient when a complete set of circuit diagrams, schematic drawings, chassis

diagrams, and functional block diagrams are available. These are often accompanied by a troubleshooting guide, which lists the most frequently encountered malfunctions, the symptoms, and the remedies. The area in which the malfunction is occurring can often be localized by studying the functional block diagram of the unit. As an example, consider a counter-timer that responds normally to input signals but gives erratic readings when counting for preset time intervals, and erratic times when in the preset-counts mode. In this case, the gating circuit and the time base are obviously suspect. One should begin by looking at that part of the circuit.

Good schematic diagrams include waveforms and voltage levels at all critical points in the circuit. These should be checked first. Well-designed circuit boards often have the d.c. voltage levels printed directly at the appropriate test points and components identified by a printed letter-and-number code that corresponds to an identical code on the schematic diagram.

A common difficulty is gaining access to components and connections in the restricted space between boards in a multiple circuit-board assembly. *Extender boards* (see Figure 6.165) are circuit boards arranged in such a way that they mate to the circuit-board connector at one end and the circuit board itself at the other end. The extender board is of sufficient length to place the board in a position where the components and circuit interconnections are easily accessible. Signal injection is a common method for isolating a circuit fault. With this technique, an appropriate signal is injected at the input of the suspected circuit element, and the output is monitored. This works well for linear circuits, such as amplifiers, but becomes quite complex for digital circuits, where it may be necessary to stimulate several input terminals and monitor several output terminals simultaneously – and often in synchronization with a clock signal. Logic pulsers, probes, and clips, are useful for small-scale testing. They are shown in Figure 6.166. For more complex circuits, it is necessary to capture the signals and store them for subsequent analysis. There are logic analyzers made expressly for this purpose. They are expensive, however, but worthwhile in situations where a large amount of digital circuit troubleshooting is done – more the domain of the electronic technician than the experimental scientist. Often it is much more economical to replace all the suspected ICs in a given section (espe-



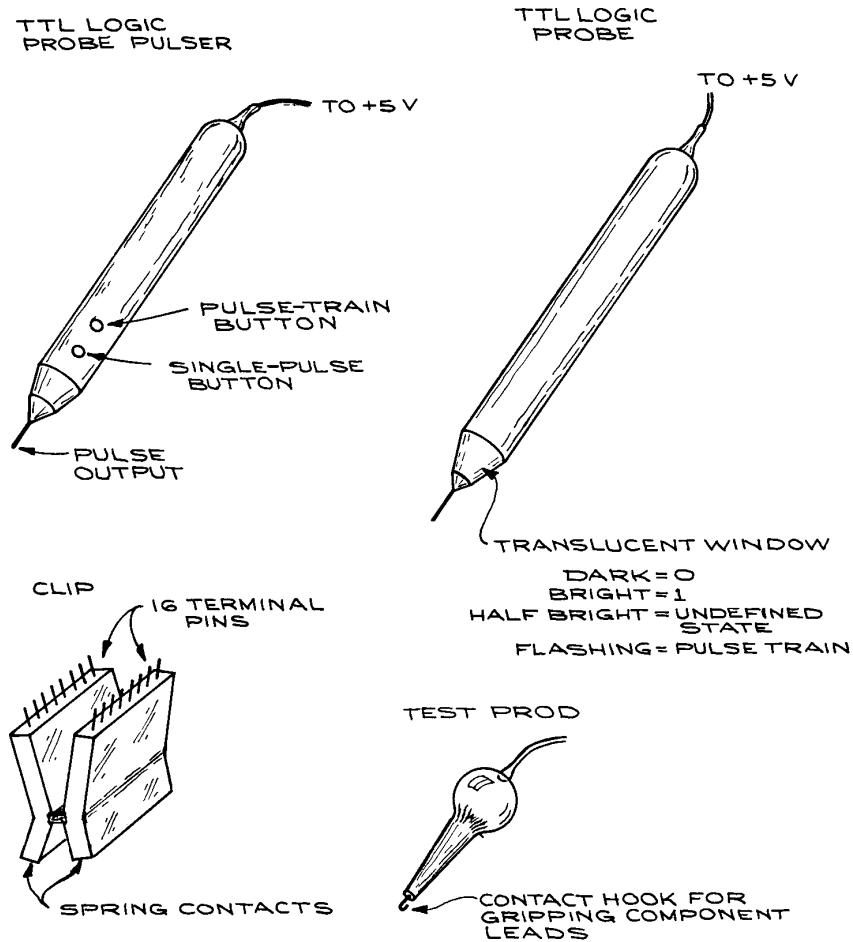
**Figure 6.165** Use of the extender board for troubleshooting.

cially if they are easily removable and in sockets) than to test each one individually.

Some common sources of faults in electronic equipment are:

- (1) Electrolytic capacitors in general
- (2) Pass transistors in the output circuit of power supplies

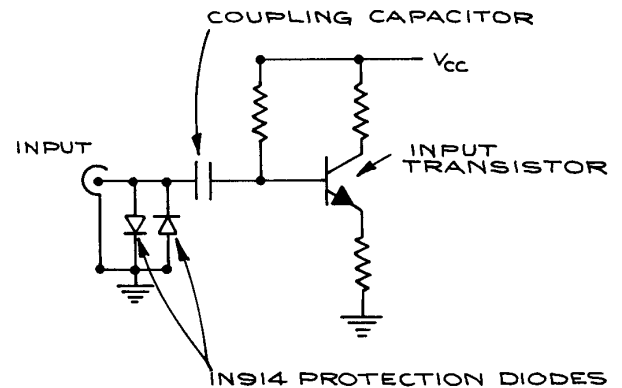




**Figure 6.166** Logic pulser, probe, clip, and test prod.

- (3) Input transistors in the input circuits of amplifiers and preamplifiers
- (4) Mechanical switches and potentiometers.

Repeated failure of input transistors in low-level preamplifier circuits can often be cured by placing two diodes, such as 1N914s, across the input, as shown in Figure 6.167. So long as the input signal does not exceed a few tenths of a volt, both diodes remain nonconducting, allowing the unattenuated signal to pass. Should the signal exceed 0.6 V – positive or negative – one or the other diode will



**Figure 6.167** Use of protection diodes at a low-level input.

**Table 6.51 Semiconductor integrated-circuit code prefixes**

Company	Prefix <sup>a</sup>
Analog Devices	AD
Advanced Micro Devices	Am
General Instrument	AY, GIC, GP
Intel	C, I
RCA	CA, CD, CDP
TRW	CA, TDC, MPY, CMP, DAC, MAT, OP
Precision Monolithics	PM, REF, SSS
National Semiconductor	DM, LF, LFT, LH, LM, NH
Fairchild	F, A, $\mu$ L, <i>Unx</i>
Ferranti	FSS, ZLD
GE	GEL
Harris	HA
Motorola	HEP, MC, MCC, MCM, MFC, MM, MWM
Intersil	ICH, ICL, ICM, IM
ITT	ITT, MIC
Siliconix	L, LD
Fugitsu	MB
Mostek	MK
Plessey	MN, SL, SP
Signetics	N, NE, S, SE, SP
Raytheon	R, RAY, RC, RM
Texas Instruments	SN, TMS
Sprague	ULN, ULS
Westinghouse	WC, WM
Hewlett-Packard	5082- <i>nmn</i>
Vishay	Si

<sup>a</sup> *x* = number; *n* = letter.

conduct, shunting the signal to ground and protecting the input from excessive voltage levels.

### 6.11.2 Identifying Parts

One can often localize the malfunction to a circuit component, but may not know enough about the component to be able to replace it. The parts list may give the equipment manufacturer's component code rather than the standard code of the component manufacturer. When this is the case, it is necessary to identify the component from the code printed on it. Most ICs, transistors, and other compo-

nents have a four-digit date code giving the date of manufacture. The first two digits are the year, and the next two are the week of the year. Thus, 9812 is the 12th week (last week in April) in 1998. The other codes are the manufacturer's component designation. To help identify the manufacturer, a list of logos can be found at the following web sites: <http://www.icmaster.com/LogosA-M.asp> and <http://www.icmaster.com/LogosN-Z.asp>. The site <http://www.elektronikforum.de/ic-id> has links to icmaster and others. Semiconductor manufacturer homepages and data sheets are given at <http://www.bgs.nu/sdw/p.html>. A list of the prefix codes of semiconductor manufacturers is given in Table 6.51. Once the manufacturer is known, the appropriate data book or web site can be consulted. One should replace components with caution, paying attention to the package type and temperature range. The three standard ranges are commercial (0 to 70 °C), industrial (−25 to +85 °C), and military (−55 to +125 °C). Sometimes specially selected or matched components are used. Replacement with off-the-shelf units may not work in this case.

### Cited References

1. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, Princeton, NJ, 1945.
2. *Electr. Design*, **24**, 63, 1976.
3. J. Millman and C. C. Halkias, *Integrated Electronics: Analog and Digital Circuits and Systems*, McGraw-Hill, New York, 1972, pp. 244–245.
4. J. G. Tobey, G. E., Huelsman, and L. P., Graeme, *Operational Amplifiers, Design and Application*, McGraw-Hill, New York, 1971; J. Graeme, *Designing with Operational Amplifiers*, McGraw-Hill, New York, 1977.
5. E. Fairstein and J. Hahn, Nuclear pulse amplifiers – fundamentals and design practice, *Nucleonics*, **23**(7), 56, 1965; *Nucleonics*, **23**(9), 81, 1965; *Nucleonics*, **23**(11), 50, 1965; *Nucleonics* **24**(1), 54, 1966; *Nucleonics*, **24**(3), 68, 1966.
6. *Voltage Regulator Handbook*, National Semiconductor Corporation, Santa Clara, CA, 1975.
7. *FAST Data Manual*, Signetics Corp., Sunnyvale, CA, 1984.
8. *Standard Nuclear Instrument Modules*, adopted by AEC Committee on Nuclear Instrument Modules, US Government Publication TID-20893 (Rev. 3).

9. P. H. Garrett, *Analog I/O Design*, Reston Publishing Co., Reston, VA, 1981.
10. S. Letzter and N. Webster, Noise in amplifiers, *IEEE Spectrum*, August, 67–75, 1970.
11. John C. Fisher, Lock in the Devil, educe him or take him for the last ride in a boxcar?, *Tek Talk*, Princeton Applied Research, 6, No. 1.
12. A. C. Melissinos, *Experiments in Modern Physics*, Academic Press, New York, 1966, Chapter 10.
13. R. R. Goruganthu, M. A. Coplan, J. H. Moore, J. A. Tossell, (e, 2e) momentum spectroscopic study of the interaction of  $-CH_3$  and  $-CF_3$  groups with the carbon-carbon triple bond, *J. Chem. Phys.*, **89**, 25, 1988.
14. R. Morrison, *Grounding and Shielding Techniques in Instrumentation*, John Wiley & Sons, Inc., New York, 1967.
15. *Floating Measurements and Guarding*, Application Note 123, Hewlett-Packard, Palo Alto, CA, 1970.
16. *Circuits for Electronics Engineers*, S. Weber (Ed.), Electronics Book Series, McGraw-Hill, New York, 1977; *Circuit Design Idea Handbook*, W. Furlow (Ed.), Cahner's Books, Boston, 1974; *Electronics Circuit Designer's Casebook*, Electronics, New York; *Signetics Analog Manual, Applications, Specifications*, Signetics Corporation, Sunnyvale, CA; *Linear Applications Handbook, Vols. 1 and 2*, National Semiconductor Corporation, Santa Clara, CA.
17. W. R. Blood, Jr., *MECL Applications Handbook*, 2nd edn., Motorola Semiconductor Products, Pheonix, AZ, 1972.
18. D. Lindsey, *The Design and Drafting of Printed Circuits*, 2nd edn., Bishop Graphics, Westlake Village, CA, 1984.
19. *Transducer Interfacing Handbook*, D. H. Sheingold (Ed.), Analog Devices, Norwood, MA, 1980.

## General References

### CAMAC and IEEE-488 (GPIB or HP-IB)

- CAMAC: A Modular Instrumentation System for Data Handling*, ESONE Committee, Report EUR 4100, 1972, Chapters 4–6.
- CAMAC Tutorial Issue, *IEEE Trans. Nucl. Sci.*, **NS-20**(2), April 1973.
- D. Horelick and R. S. Larsen, CAMAC: "A Modular Standard," *IEEE Spectrum*, April 50, 1976.
- HP-IB General Information*, Hewlett-Packard Publication No. 5952–0058.
- IEEE Standard 488*, available from the IEEE Standards Office, 345 E. 47th St., New York, NY 10017.

*Specifications for the CAMAC Serial Highway and Serial Crate Controller Type 62*, Report EUR 6100e, Commission of the European Communities, Greel, Belgium, 1976, Chapter 14.

## Circuit Theory

- P. Grivet, *The Physics of Transmission Lines at High and Very High Frequencies*, Academic Press, New York, 1970.
- H. V. Malmstadt, C. G. Enke, and S. R. Crouch with G. Horlick, *Electronic Measurements for Scientists*, Benjamin, Menlo Park, CA, 1974.
- J. Millman and A. Grabel, *Microelectronics*, 2nd edn., McGraw-Hill, New York, 1987.
- C. J. Savant, Jr., *Fundamentals of the Laplace Transform*, McGraw-Hill, New York, 1962.
- A. I. Zverev, *Handbook of Filter Synthesis*, John Wiley & Sons, Inc., New York, 1967.

## Printed Circuit Board Software

### Schematic Capture and Board Layout

Electronics Workbench  
 908 Niagra Falls Boulevard, Suite 068  
 North Tonawanda, New York  
 14120-2060  
 800-263-5552  
 FAX: 416-977-1818  
<http://www.electronicworkbench.com>

PADS Software Inc.  
 165 Forest Street  
 Marlborough MA 01752  
 508-485-4300  
 FAX: 508-485-7171  
<http://www.pads.com>

Ivex Design International  
 15232 NW Greenbrier Parkway  
 Beaverton, OR 97006  
 503-531-3555  
 FAX: 503-629-4907  
 WinDraft, WinBoard, IvexView, IvexSpice

Cadence Design Systems  
 2655 Seely Road  
 San Jose, CA 95134  
 1-800-746-6223

FAX: 408-943-0513  
<http://www.cadence.com>

Mentor Graphics Corporation  
 8005 S.W. Boeckman Road  
 Wilsonville, OR 97070  
 1-800-547-3000  
<http://www.mentor.com/pcb/>

P-CAD  
 12348 High Bluff Drive  
 San Diego, CA 92130  
 858-350-3000  
 FAX: 858-350-3001  
<http://www.pcad.com>

Protel Technologies, Inc.  
 5252 North Edgewood Drive  
 Suite 175  
 Provo, UT 84604  
 1-800-544-4186  
 FAX: 801-224-0558  
<http://www.protel.com>

## Printed Circuit Board Fabricators

Golden Gate Graphics  
 South Bay Circuits  
 Advanced Circuits  
 21100 East 33rd drive  
 Aurora CO 80011  
<http://www.advancedcircuits.com>

## Control Analysis and Design

- R. C. Dorf, *Modern Control Systems*, 6th edn., Addison-Wesley, New York, 1992.
- T. E. Fortmann and K. L. Hitz, *An Introduction to Linear Control Systems*, Marcel Dekker, New York, 1977.
- G. F. Franklin, D. J. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, 2nd edn., Addison-Wesley, New York, 1991.
- B. C. Kuo, *Automatic Control Systems*, 6th edn., Prentice-Hall, New Jersey, 1990.
- K. Ogata, *Modern Control Engineering*, 2nd edn., Prentice-Hall, New Jersey, 1991.
- C. L. Phillips and R. D. Harbor, *Feedback Control Systems*, 2nd edn., Prentice-Hall, New Jersey, 1991.

## Components

- M. Grossman, Focus on r.f. Connectors, *Electr. Design*, **24**(11), 60, 1976.
- C. A. Harper, To compare electrical insulators, *Electr. Design*, **24**(11), 72, 1976.
- Handbook of Components for Electronics*, C. A. Harper (Ed.), McGraw-Hill, New York, 1977.
- T. H. Jones, *Electronic Components Handbook*, Reston Publishing, Reston, VA, 1978; Selecting capacitors properly, *electr. design*, **13**, 66, 1977.

## Data Books

- D.A.T.A. Books*, D.A.T.A., San Diego, CA, Volumes include *Optoelectronics, Digital IC, Linear IC, Transistor, Diode, Thyristor, Power Semiconductor, Interface IC*.
- MECL System Design Handbook*, W. R. Book, Jr., and E. C. Tynan, Jr. (Eds.), 2nd edn., Motorola Semiconductor Products, Phoenix, AZ, 1972.
- Motorola Semiconductor Data Library*: Vol. 1, *EIA Type Numbers to 1N5000 and 2N5000*; Vol. 2, *Discrete Products, EIA Type Numbers 1N5000 and Up, 2N5000 and Up, 3N ... and 4N ...*; Vol. 3, *Discrete Products, Motorola Non-Registered Type Numbers*; Vol. 4, *MECL Integrated Circuits*.
- National Semiconductor Data Books: Linear, TTL, CMOS, MOSFET, Memory, Discrete*, National Semiconductor Corporation, Santa Clara, CA.
- Optoelectronics Designer's Catalog*, Hewlett-Packard, 1984.
- Signetics Data Manual; Logic, Memories, Interface, Analog, Microprocessor, Military*, Signetics, Sunnyvale, CA.
- The TTL Data Book for Design Engineers*, 1st edn., Texas Instruments, Dallas, TX, 1973.
- Voltage Regulator Handbook*, National Semiconductor Corporation, Santa Clara, CA.

## Handbooks

- The Electronics Handbook* (The Electrical Engineering Handbook Series), J. C. Whitaker (Ed.), CRC Press/IEEE Press, Boca Raton, FL, 1996.
- ARRL Handbook*, 83rd edn., American Radio Relay League, Newington, CT, 2006.
- Electronics Designers' Handbook*, L. J. Giacoletto (Ed.), 2nd edn., McGraw-Hill, New York, 1977.
- T. D. S. Hamilton, *Handbook of Linear Integrated Electronics for Research*, McGraw-Hill, New York, 1977.

ITT Engineering Staff, *Reference Data for Radio Engineers*, 6th edn., Sams, Indianapolis, 1979.

*Radio Engineering Handbook*, K. Henney (Ed.), McGraw-Hill, New York, 1959.

## Master Catalogs

*Electronic Design's Gold Book: Master Catalog and Directory of Suppliers to Electronics Manufacturer.*

*Electronic Engineers Master*, United Technical Publications, Garden City, NJ.

## Noise

S. Letzter and N. Webster, Noise in amplifiers, *IEEE Spectrum*, August, 67–75, 1970.

A. V. D. Ziel, Noise in solid-state devices and lasers, *Proc. IEEE*, **58**(8), 1178, 1970.

## Particle and Radiation Detection

*Electronics for Nuclear Particle Analysis*, L. J. Herbst (Ed.), Oxford University Press, London, 1970.

*Electro-Optics Handbook: A Compendium of Useful Information and Technical Data*, RCA Defense Electronics Products, Aerospace Systems Division, Burlington, MA, 1968.

Glenn F. Knoll, *Radiation Detection and Measurement*, John Wiley & Sons, Inc., New York, 1979.

*Optoelectronics Designer's Catalog*, Hewlett-Packard, Palo Alto, CA, 1980.

*Modular Pulse-Processing Electronics*, ORTEC, 801 South Illinois Avenue, Oak Ridge, TN 37831–0895, 2001.

## Practical Electronics

*Design Techniques for Electronics Engineers*, Electronics Book Series, McGraw-Hill, New York, 1977.

P. Horowitz and W. Hill, *The Art of Electronics*, 2nd edn., Cambridge University Press, Cambridge, 1989.

D. Lancaster, *CMOS Cookbook*, Sams, Indianapolis, 1977.

D. Lancaster, *TTL Cookbook*, Sams, Indianapolis, 1974.

## Trade Publications

*Digital Design*, Benwill, Boston: computers, peripherals, systems; monthly.

*Electronic Component News*, Chilton Col, P.O. Box 2010, Radnor, PA 19089; monthly.

*Electronic Design*, Hyden Publishing Co., Inc., a subsidiary of VNU, 10 Mulholland Dr., Hasbrouck Heights, NJ 07604.

*Electronic Engineering Times*, CMP Publications, Manhasset, N.Y.; biweekly tabloid.

*Instrument and Apparatus News (IAN)*, Chilton, Radnor, Pa.: instruments, industrial controls, digital systems; monthly.

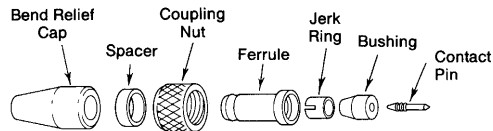
## Troubleshooting

R. E. Gasperini, *Digital Trouble Shooting*, Hayden, Rochelle Part, NJ, 1976. *Techniques of Digital Trouble Shooting*, Hewlett-Packard Application Note 163–1.

## Electronic Measurements

*Low Level Measurements: Precision D.C. Current, Voltage, and Resistance Measurements*, 5th edn., Keithley Instruments, Cleveland, OH, 1998.

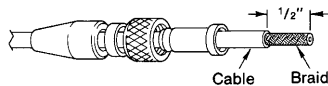
# MICRODOT®



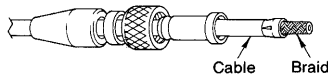
## Coaxial Cable RG-196/U



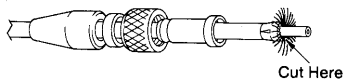
STEP 1 . . . Slide bend relief cap, spacer, coupling nut and ferrule over cable.



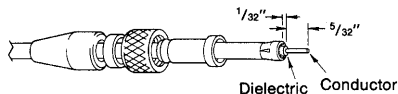
STEP 2 . . . Strip cable to the dimension shown.



STEP 3 . . . Place jerk ring over shield, slotted end of ring should butt against cable jacket. Squeeze the ring snugly around the shield with a slight taper at the slotted end. Be careful not to squeeze the ring out of round.



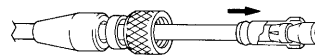
STEP 4 . . . Comb out braid and cut it close to the jerk ring as shown.



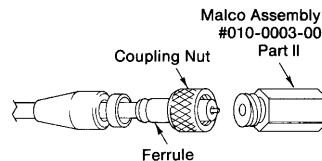
STEP 5 . . . Strip dielectric from end of conductor  $5/32$ " and to  $1/32$ " from the jerk ring (be careful not to nick conductor). If conductor is stranded it must be tinned.



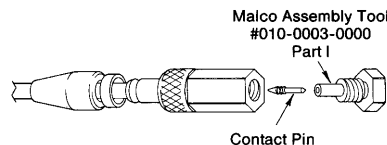
STEP 6 . . . Slip bushing over conductor and dielectric until bushing is against braid.



STEP 7 . . . Hold ferrule in one hand, with other hand push on the bushing until jerk ring and part of bushing start inside the ferrule.



STEP 8 . . . Slide coupling nut over ferrule. Then screw coupling nut to Malco Tool, Part II.



STEP 9 . . . Insert contact pin with barbed end out, in small hole at the end of Malco Tool, Part I, and screw into Part II as far as possible, thus forcing contact pin completely into bushing.



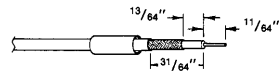
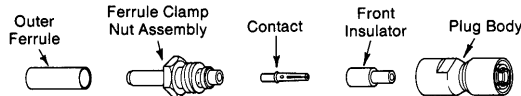
STEP 10 . . . Slide the bend relief cap up and over the end of the ferrule until it snaps into the groove. Remove Malco Tools, Part I and Part II.

All dimensions are in inches  
Microdot is a registered trademark of Microdot Corp.

To obtain the Malco Assembly Tools contact  
Malco, 306 Pasadena Ave., South Pasadena, CA 91030,  
Phone (818) 799-9171

\*Reproduced by permission of Ceramaseal.

## SMB

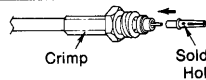


### Coaxial Cable RG-174/U

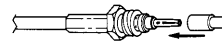
STEP 1 . . . Slide outer ferrule over cable. Strip cable jacket, braid, and dielectric to the dimensions shown. Make all cuts sharp and square. **IMPORTANT:** Do not nick braid, dielectric, or center conductor. Tin center conductor, avoid excessive heat.



STEP 2 . . . Slightly flare out end of cable braid as shown. **DO NOT** comb out braid. Install ferrule clamp nut assembly onto cable so that the ferrule portion slides under braid, and insulator inside the assembly butts against cable dielectric.



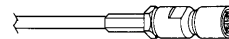
STEP 3 . . . Slide outer ferrule over and up against nut. Make sure no slack exists in braid. Crimp outer ferrule with Ceramaseal's Crimp Tool #880A2840-04 or Thomas & Betts tool #WT-400, keeping cable dielectric bottomed out inside. Install center contact and solder in place. Use only rosin or noncorrosive flux. Do not get solder on the outside of the contact.



STEP 4 . . . Insert front insulator over contact.

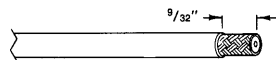
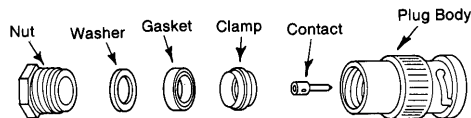


STEP 5 . . . Screw plug body onto prepared cable termination and wrench-tighten by holding the ferrule clamp nut stationary while rotating the plug body.



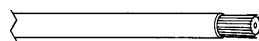
Finished assembly.

## BNC

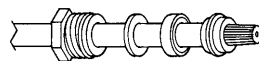


### Coaxial Cables RG-58/U RG-140/U RG-141/U

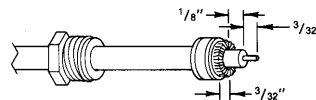
STEP 1 . . . Cut jacket and strip to dimension shown.



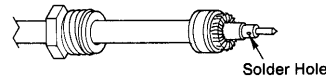
STEP 2 . . . Comb out braid and taper toward the center.



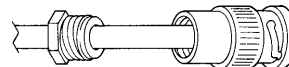
STEP 3 . . . Place the nut, washer, and gasket over the cable, then slide the clamp on over the braid so that the inner shoulder fits against the end of the cable jacket.



STEP 4 . . . With the clamp in place, fold back braid as shown and trim  $3/32$ " from the end. Trim dielectric to the dimensions shown.



STEP 5 . . . Slip contact in place so it butts against the dielectric and solder in place. Remove excess solder from outside of contact. Be sure cable dielectric is not heated excessively and swollen so as to prevent dielectric from entering into connector body.

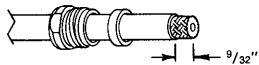
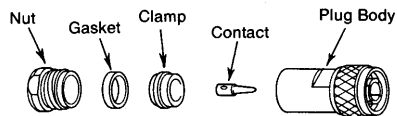


STEP 6 . . . Push assembly into body as far as it will go. Slide nut into body and screw in place with wrench until tight. For this operation, hold cable and shell rigid and rotate nut.

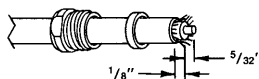
All dimensions are in inches

To obtain the Thomas & Betts Crimp Tool contact:  
Thomas & Betts Corp., 920-T Route 202, Raritan, NJ 08869,  
Phone (201) 685-1600, Telex 833190

## N



STEP 1... Place nut and gasket with "V" groove toward clamp, over cable and cut jacket to the dimension shown.



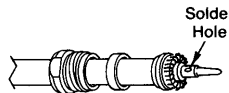
STEP 2... Comb out braid and fold out. Cut off cable dielectric to the dimensions shown.



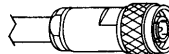
STEP 3... Pull braid wires forward and taper toward conductor. Place clamp over braid so that the inner shoulder fits against the end of the cable jacket.

## Coaxial Cables

RG-8/U  
RG-9/U  
RG-213/U  
RG-214/U  
RG-225/U

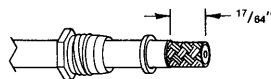
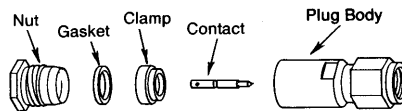


STEP 4... Fold back braid as shown, trim to proper length and form over clamp as shown. Solder contact to center conductor.

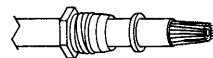


STEP 5... Insert cable and parts into connector body. Make sure sharp edge of clamp seats properly in gasket. Tighten nut.

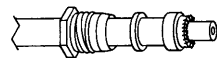
## SMA



STEP 1... Place nut and gasket over cable and cut jacket to dimension shown.



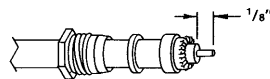
STEP 2... Comb out braid and taper forward toward the conductor.



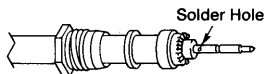
STEP 3... Place clamp over braid and push back against jacket. Fold braid back against clamp and trim as necessary so that wires do not touch shoulder of clamp.

## Coaxial Cables

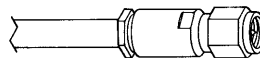
RG-55/U  
RG-58/U  
RG-141/U  
RG-142/U  
RG-223/U



STEP 4... Trim dielectric 1/8" from the end of the cable. Do not nick center conductor.



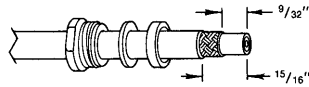
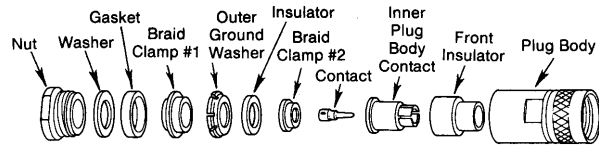
STEP 5... Solder contact in place so as to be seated squarely against dielectric. Clean all surfaces thoroughly.



STEP 6... Thread connector assembly onto prepared cable assembly. Tighten to 20-25 in./lbs. torque.

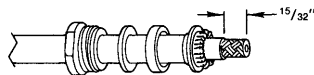


## TRUE TRIAX

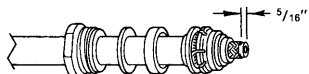


**Coaxial Cables**  
**RG-8/U**  
**RG-9/U**  
**RG-10/U**

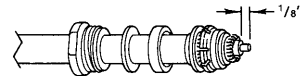
STEP 1 . . . Slide nut, washer and gasket over cable. Cut off outside jacket to  $15/16$ " dimension shown. Make a clean cut, being very careful not to nick braid. Cut outer braid to  $9/32$ " dimension as shown.



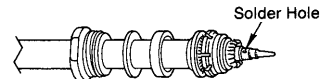
STEP 2 . . . Slide first braid clamp over braid up to outer jacket. Fold first braid back over clamp, making sure braid is evenly distributed over the surface of clamp. Trim second jacket to  $15/32$ " dimension as shown.



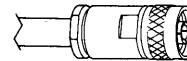
STEP 3 . . . Trim second braid to dimension  $5/16$ " as shown. Slide on outer ground washer, insulator and second braid clamp over inner braid.



STEP 4 . . . Fold second braid back over braid clamp, again making sure that braid is evenly distributed over the surface of the clamp. Trim inner dielectric to the  $1/8$ " dimension as shown.

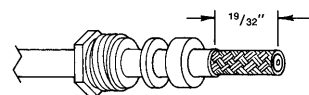
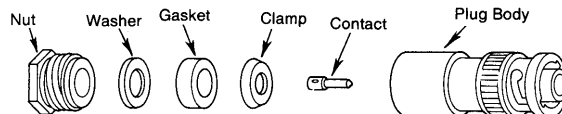


STEP 5 . . . Tin the inside hole of the contact. Tin wire and insert into contact and solder. Remove any excess solder. Be sure cable dielectric is not heated excessively and swollen so as to prevent dielectric from entering the plug body.



STEP 5 . . . Place front insulator and inner plug body contact into back of plug body and push into proper place. Insert cable-contact assembly into plug body. Screw nut into body with wrench until moderately tight.

## MHV

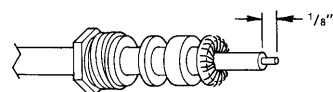


**Coaxial Cable**  
**RG-59/U**

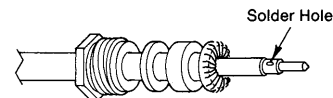
STEP 1 . . . Slide nut, washer and gasket (in improved version, with "V" groove toward clamp) over jacket, and cut off jacket  $19/32$ " from end of cable as shown.



STEP 2 . . . Comb out braid and pull braid wires forward and taper toward conductor.



STEP 3 . . . Place clamp over braid and push back against cable jacket. Fold back braid wires as shown, trim to proper length and evenly form over clamp as shown. Cut dielectric to  $1/8$ " dimension shown. Tin exposed conductor using minimum amount of heat.

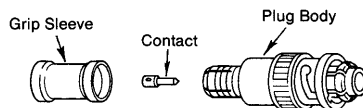


STEP 4 . . . Solder contact to conductor. Remove excess flux and solder from outside of contact.

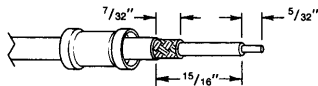


STEP 5 . . . Insert prepared cable termination into plug body. (In improved version make sure sharp edge of clamp seats properly in gasket.) Tighten nut moderately, holding plug body stationary.

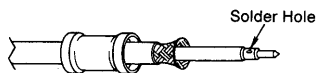
## SHV, 5 kV



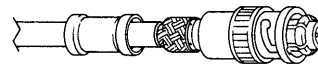
### Coaxial Cable RG-59/U



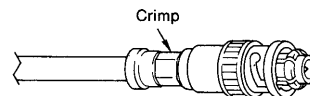
STEP 1 . . . Slide the grip sleeve over the cable. Trim the end of the cable to the dimensions shown.



STEP 2 . . . Tin end of the cable and inside of contact. Solder the contact to the center conductor. Flare out the braid without fraying.

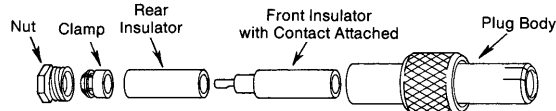


STEP 3 . . . Push the assembled contact into the plug body until the dielectric bottoms against the shoulder within the plug body. At this point, the body grip fingers should be under the flared braid as shown.

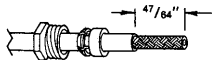


STEP 4 . . . Slide the grip sleeve forward against the plug body, over the braid and crimp as shown. Use Kings Hand Crimp Tool #KTH-1000 and Crimp Die #KTH-2062.

## BSHV, 7.5 kV



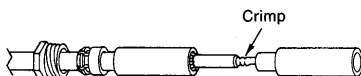
### Coaxial Cable RG-59/U



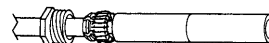
STEP 1 . . . Slide nut and clamp over cable. Cut the jacket to the dimension shown. Do not nick braid.



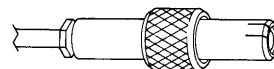
STEP 2 . . . Comb out braid and fold back against jacket. Cut dielectric to the dimension shown, do not nick conductor.



STEP 3 . . . Carefully slide rear insulator over cable just so it covers the braid. Insert the conductor into the crimp end of the front insulator and crimp as shown.



STEP 4 . . . Carefully slide the rear insulator up against the front insulator as shown. Fold braid out and slide clamp up to meet the braid. Fold the braid back over the clamp and trim so as to fit around the clamp as shown.

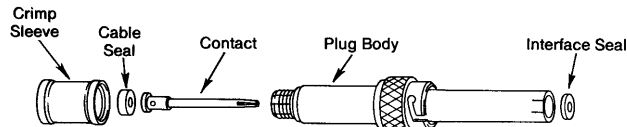


STEP 5 . . . Slide prepared cable end into plug body and tighten nut moderately.

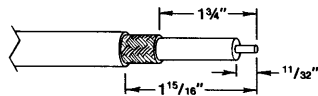
All dimensions are in inches

To obtain the Kings Hand Crimp Tool contact  
Kings Electronics Co. Inc., 40 Marblehead Road, Tuckahoe, N.Y. 10707,  
Phone (914) 793-5000, Fax (914) 793-5092

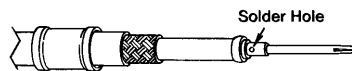
## SHV, 15 kV



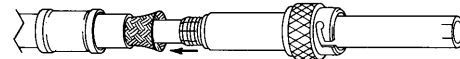
NOTE: Before assembly be sure you can distinguish between the cable seal and the interface seal. The cable seal has a .09" hole and is .10" thick, however the interface seal has a .13" hole and is .07" thick. Do not interchange.

Coaxial Cable  
RG-8/U

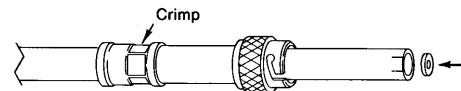
STEP 1 . . . Slide the crimp sleeve over the cable. Trim the end of the cable to the dimensions shown and tin the exposed conductor.



STEP 2 . . . Slip cable seal over center conductor then slide contact over conductor. Push contact against cable seal and maintain this slight pressure while soldering contact in place through solder hole.

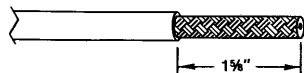
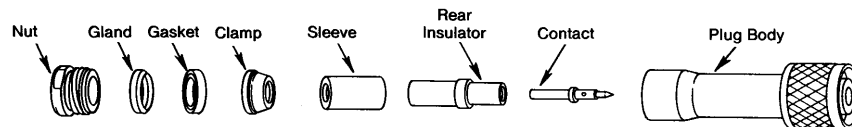


STEP 3 . . . Flare out braid without fraying the ends. Insert prepared cable assembly into the plug body carefully. DO NOT pinch or otherwise damage cable seal. Guide braid smoothly over the spined crimp area of the body until the contact shoulder butts against the inner insulator.

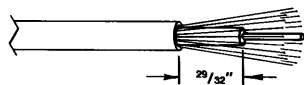


STEP 4 . . . Slide the grip sleeve into position and crimp as shown. Use Ceramaseal's Crimping Tool #880A2840-02 or Thomas & Betts Crimping Tool #WT-540 with Crimping Die #5452. Make sure braid does not extend beyond the end of the grip sleeve. Insert interface seal into the open end of the plug. Carefully slide the interface seal over contact until it bottoms evenly around the contact.

## HN



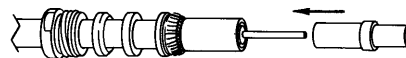
STEP 1 . . . Cut end of cable even. Strip off vinyl jacket to the dimension shown.



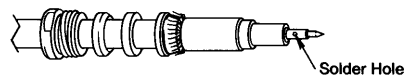
STEP 2 . . . Comb out braid and cut dielectric to the dimension shown.



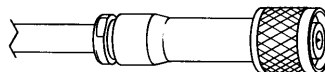
STEP 3 . . . Taper braid wires forward and slide nut and gland onto jacket. Make certain knife edge of gland is toward end of cable. Then slide gasket onto jacket with "V" groove toward gland. Clamp is now pushed over braid so that the internal shoulder butts flush against the cable jacket.



STEP 4 . . . Fold braid back over clamp and trim. Tin exposed center conductor using minimum amount of heat. Slide sleeve and rear insulator over cable dielectric.



STEP 5 . . . Solder contact to center conductor. Rear insulator must seat against cable dielectric and contact shoulder must be flush with insulator face as shown. Coat cable dielectric and insulator mating surfaces with Amphenol #53-307 Silicone Compound or equal to achieve 5 kV peak rating under operating conditions.



STEP 6 . . . Slide prepared cable termination carefully into body. Be sure knife edge of gland remains in groove of gasket. Tighten nut with wrench, holding body stationary. Gasket should be cut in half during tightening.

All dimensions are in inches

## DETECTORS

Phenomena that are studied experimentally often manifest themselves as sources of electromagnetic radiation or particles. To be useful, the radiation or particles that are involved in the experiment must be detected. In this chapter we will consider the operating characteristics, selection criteria, and performance of various types of radiation and particle detectors. We will focus primarily on detectors of optical radiation, from X-rays to far infrared, and on charged-particle detectors. For a discussion of detectors of high-energy photons, such as  $\gamma$  rays, and elementary particles, such as neutrons, neutrinos, and the many particles involved in high-energy nuclear physics, we refer the reader to the more specialized treatises on these subjects.<sup>1-4</sup>

### 7.1 OPTICAL DETECTORS

Optical detectors of various kinds detect electromagnetic radiation from the X-ray to the far infrared region of the electromagnetic spectrum. In common with other branches of electronics, the development of optical detectors has occurred by a series of advances through the use of gas-filled tubes and vacuum tubes to various semiconductor devices. Gas-filled and vacuum-tube devices still retain some niche electronics applications, such as high-voltage switching, microwave power amplification, and specialized audio. Gas-filled photodetectors, which offer some degree of signal amplification in low-cost consumer applications, have essentially disappeared. Vacuum-tube photodetectors, especially photomultiplier tubes, remain in widespread use for detection of radiation below about 1  $\mu\text{m}$ . This is especially true for low-light-level signal detection

(photon counting), and in applications where a large detector area is required (as in scintillation counters). The designer of an optical system is confronted with a broad range of detector types: vacuum-tube, semiconductor, and thermal. The last category includes thermopiles, pyroelectric detectors, Golay cells, and bolometers, each with its own advantages and disadvantages.

The choice of detector for a given application will almost invariably involve a choice from among the numerous commercially available devices in each of these categories. The construction of optical detectors is rarely undertaken in the laboratory except by the specialist. The choice of detector will be governed by factors such as:

- (1) The wavelength region of the radiation to be detected
- (2) The intensity of the radiation to be detected
- (3) The time response required to resolve high-speed events
- (4) The detector surface area required
- (5) The environmental conditions in which the detector is to be used
- (6) The cost
- (7) The electronics to be used in conjunction with the detector.

The choice of a detector for a particular experiment should not be an exercise performed in isolation. The other components of the experimental system may influence it, and vice versa. This is particularly true if very weak radiation is to be detected. In such a situation, action should be taken to maximize the light signal actually reaching the detector; this will involve a careful choice of optical components to be placed between source and detector – for

example, suitable light collection and focusing optics. If the light from the source must pass through a dispersing element (monochromator, etalon, etc.) before reaching the detector, then the choices of dispersing element and detector are likely to be interrelated. Some dispersing elements have high resolution (see Section 4.7), but low light throughput; others the reverse. Still others have high throughput and resolution, but can only be scanned in a restricted way, if at all. In a given experimental situation, the various requirements of resolution, light throughput, scanability, and available detector sensitivities must be offset one against another in arriving at a sensible compromise.

Detectors can be classified as photon detectors or thermal detectors. In *photon detectors*, individual incident photons interact with electrons within the detector material. This leads to detectors based on *photoemission*, where the absorption of a photon frees an electron from a material; *photoconductivity*, where absorption of photons increases the number of charge carriers in a material or changes their mobility; the *photovoltaic effect*, where absorption of a photon leads to the generation of a potential difference across a junction between two materials; the *photon-drag effect*, where absorbed photons transfer their momentum to free carriers in a heavily doped semiconductor; and *photochemical detectors*, such as the photographic plate.

In *thermal detectors*, the absorption of photons leads to a temperature change of the detector material, which may be manifest as a change in resistance of the material, as in the *bolometer*; the development of a potential difference across a junction between two different conductors, as in the *thermopile*; or a change of internal dipole moment *polarizability* in a temperature-sensitive ferroelectric crystal, as in a *pyroelectric* detector. With the exception of the last, thermal detectors have slow time response when compared to photon detectors. They do, however, respond uniformly at all wavelengths (in principle).

Photon detectors can be further classified by the spectral region to which they are sensitive. Most photoemissive detectors operate from the soft X-ray region through to about 1.2  $\mu\text{m}$  and rapidly become insensitive beyond 1  $\mu\text{m}$ . New semiconductor photocathodes, however, are becoming available that are sensitive out to 1.7  $\mu\text{m}$ . Photodiodes operated in a “Geiger” mode provide single photon

detection capability from about 250 nm to 1.8  $\mu\text{m}$ . Photoconductive and photovoltaic detectors based on different materials are used right from the visible into the very far infrared, at least to 1000  $\mu\text{m}$ . Photographic films, which are also photon detectors, generally respond best in the visible, although films marginally sensitive out to about 1.2  $\mu\text{m}$  are available. Films with appropriate fluorescent coatings respond, to ultraviolet light.

This brief survey is not intended to be exhaustive. Many other detection mechanisms have been described, but not all are commonly used or are commercially available. For more information and details about detectors in general than will be found here, the reader should consult References 5–16, and the catalogs of manufacturers.

## 7.2 NOISE IN THE OPTICAL DETECTION PROCESS

The issue of *noise* becomes of great importance when an optical detector is used for detection of a weak signal. Noise is the randomly fluctuating voltage or current produced by the detector and its associated electronics, which interferes with the ability of the system to see the signal being looked for. There are several sources of noise, and an understanding of their origin and magnitude is very important in designing a detection system that has the ability to see the desired signal. A noise voltage or current fluctuates both positively and negatively about zero on a timescale that is determined by the frequency response of the system generating the noise. Consequently, if the noise is characterized by a current  $i_N$ , the average noise current is  $\langle i_N \rangle = 0$ . The noise must therefore be characterized by its mean square value, averaged over a sufficiently long time  $\langle i_N^2 \rangle$ . This is often called the *noise power*, although strictly speaking, the noise power is  $\langle i_N^2 \rangle R$ , where  $R$  is an effective resistance through which the noise current flows. There are several important sources of noise, which deserve separate consideration.

### 7.2.1 Shot Noise

*Shot noise*, also called *photon noise*, or *quantum noise*, is inescapable. This is noise associated with the signal being

detected. If a light signal being detected gives an average detected current of  $\langle i_s \rangle$ , then the shot noise is:

$$\langle i_N^2 \rangle_{SN} = 2e\langle i_s \rangle \Delta f \quad (7.1)$$

where  $e$  is the magnitude of the electronic charge, and  $\Delta f$  is the frequency bandwidth being detected. Filtering the output signal and reducing the operating bandwidth will reduce this noise. Shot noise arises because of the discrete nature of the charge carriers produced in the detection process. A given detected current examined microscopically reveals itself to be a sequence of current impulses, which, although arriving at an average rate over time, fluctuate in a random way. The individual current impulses have a characteristic time duration, determined by the time response of the detector. If each current impulse was infinitely narrow – as  $\delta$  function – then the frequency spectrum of the noise would be flat or *white* over all frequencies. In reality, the shot noise spectrum is flat over frequencies up to a maximum value determined by the time response of the detector. If the detector has a time constant  $\tau$ , determined by resistance and capacitance, then its 10% to 90% *rise time* is  $t_r = 2.197\tau$ . The bandwidth  $\Delta f$  is related to  $\tau$  by:

$$\Delta f = \frac{1}{2\pi\tau} \quad (7.2)$$

## 7.2.2 Johnson Noise

*Johnson noise*, also called *thermal noise* or *Nyquist Noise*, is associated with the effective resistance of the detector circuit. For a resistance  $R$  it is described by a noise power:

$$\langle i_N^2 \rangle_{JN} = \frac{4kT\Delta f}{R} \quad (7.3)$$

where  $k$  is Boltzmann's constant, and  $T$  is the absolute temperature. Johnson noise arises because of the blackbody radiation associated with an enclosure at absolute temperature  $T$ . A resistor at temperature  $T$  acts as an antenna for the electric fields of the local blackbody radiation. If the resistor did not reradiate its detected fields, then it would not remain in thermal equilibrium. Consequently, it generates a noise current.

## 7.2.3 Generation-Recombination (gr) Noise

*Generation-recombination* (gr) noise is most important in lightly doped semiconductors. It results from the random recombination of electrons and holes in the material. The noise power for gr noise can be written as:

$$\langle i_N^2 \rangle_{gr} = \frac{4\langle i^2 \rangle \tau}{\bar{N}(1 + 4\pi^2 \Delta f^2)} \quad (7.4)$$

where  $\tau$  is the characteristic time constant for recombination in the semiconductor, and  $\bar{N}$  is the average number of carriers.

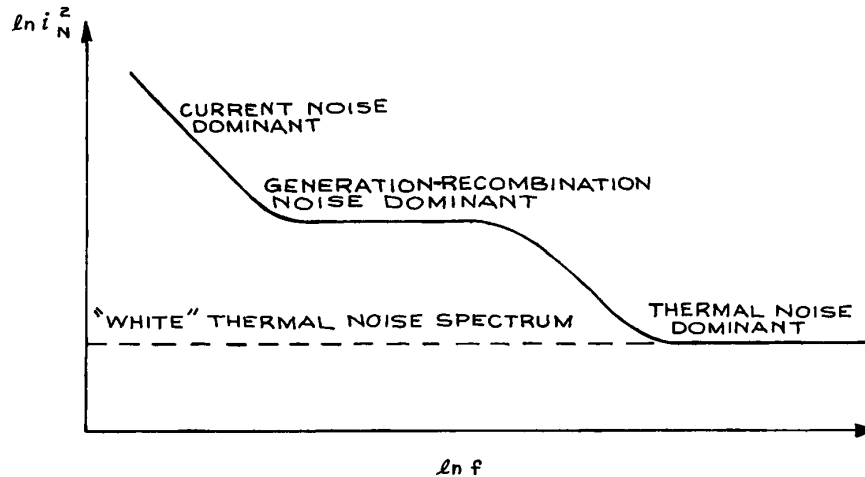
## 7.2.4 1/f Noise

*1/f noise*, also called *flicker noise*, *current noise* or *excess noise*, is still not well understood. It contributes at low frequencies, and has been shown to continue to increase down to frequencies as low as  $10^{-4}$  Hz. 1/f noise power can be written as:

$$\langle i_N^2 \rangle_{1/f} = \frac{K\langle i \rangle^\alpha \Delta f}{f^n} \quad (7.5)$$

where  $K$  is a constant for a particular device,  $\langle i \rangle$  is the average (dc) current through the device,  $\alpha$  lies between 0.5 and 2, and  $n \sim 1$ . Because 1/f noise dominates at low frequencies, measurements at very low frequencies should be avoided whenever possible. In experiments where light emission results from some form of stimulation, for example in fluorescence or light scattering, it is generally possible to periodically excite the phenomenon and move the detection to a frequency synchronous with the excitation. Good examples involve lock-in detection, boxcar averaging, or digital signal averaging (see Chapter 6 for more details).

Thermal noise and shot noise have so-called white spectra – they contribute equal noise signals in a given detection bandwidth at any frequency. Current noise generally shows a 1/f frequency dependence. As shown in Figure 7.1, gr noise contributes at frequencies up to a value of the order of the reciprocal of the carrier lifetime, and then falls off rapidly with increasing frequency.



**Figure 7.1** Schematic frequency dependence of noise in semiconductor detectors;  $i_N$  is the noise current.

## 7.3 FIGURES OF MERIT FOR DETECTORS

Before discussing individually the characteristics of commonly used and commercially available detectors, the various *figures of merit* used by manufacturers to specify their products will be described. These are the noise-equivalent power (NEP), detectivity ( $D^*$ ), responsivity ( $\mathfrak{R}$ ), quantum efficiency ( $\eta$ ), and time constant ( $\tau$ ), which allow such questions to be answered as: What is the minimum light intensity falling on the detector that will give rise to a signal voltage equal to the noise voltage from the detector? What signal will be obtained for unit irradiance? How does the electrical signal from the detector vary with the wavelength of the light falling on it? What is the modulation frequency response of the detector or its ability to respond to short light pulses?

### 7.3.1 Noise-Equivalent Power

The *noise-equivalent power* (NEP) is the rms value of the sinusoidally modulated radiant power falling upon a detector, that gives an rms signal voltage equal to the rms noise voltage from the detector. The NEP is usually specified in

terms of a blackbody source at 500 K, the reference bandwidth for the detection of signal and noise (usually 1 or 5 Hz), and the modulation frequency of the radiation (usually 90, 400, 800 or 900 Hz). For example, a noise-equivalent power written NEP (500 K, 900, 1) implies a blackbody source at 500 K, a 900Hz modulation frequency, and a 1 Hz detection bandwidth. Thus, if a radiant intensity  $I$  ( $\text{W}/\text{m}^2$ ) falls on a detector of sensitive area  $A$  ( $\text{m}^2$ ), and if signal and noise voltages  $V_S$  and  $V_N$  are measured with bandwidth  $\Delta f$  (Hz) (small enough so that the noise voltage frequency spectrum is flat within it), then the NEP measured is:

$$\text{NEP} = \frac{IA}{\sqrt{\Delta f}} \frac{V_N}{V_S} \text{ W Hz}^{-1/2} \quad (7.6)$$

### 7.3.2 Detectivity

At one time, detectivity was defined as the reciprocal of the NEP. Most detectors exhibit an NEP that is proportional to the square root of the detector area; so a detector-area-independent *detectivity*  $D^*$  is now used, specified by:

$$D^* = \frac{\sqrt{A}}{\text{NEP}} \quad (7.7)$$

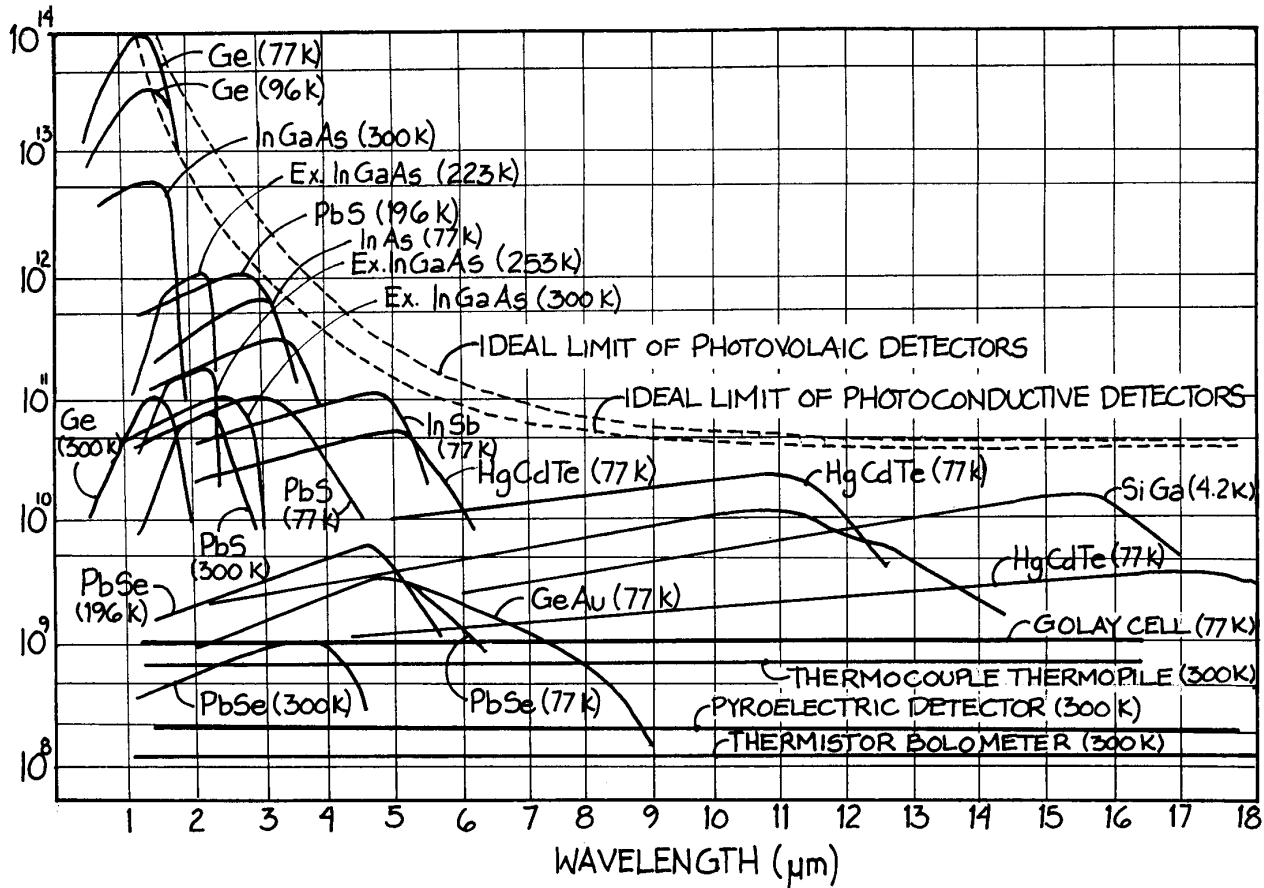


Figure 7.2 Variation of the Detectivity ( $D^*$ ) as a function of wavelength for a comprehensive set of different detectors.

$D^*$  is specified in the same way as NEP: for example,  $D^*(500 \text{ K}, 900, 1)$ . To specify the variations in response of a detector with wavelength, the spectral detectivity is used. Thus, the symbol  $D^*(\lambda, 900, 1)$  would specify the response of the detector to radiation of wavelength  $\lambda$ , modulated at 900 Hz and detected with a 1 Hz bandwidth. The units of  $D^*$  are  $\text{cm Hz}^{1/2} \text{ W}^{-1}$ . Curves of  $D^*(\lambda)$  for various detectors are shown in Figure 7.2.

For a thermal detector of absorptivity and emissivity  $\epsilon$ , at absolute temperature  $T_1$ , completely enclosed by an environment at temperature  $T_2$ , the photon-noise-limited detectivity can be shown to be:<sup>5</sup>

$$D^* = \frac{4.0 \times 10^{16} \epsilon^{1/2}}{(T_1^5 + T_2^5)^{1/2}} \text{ cm Hz}^{1/2} \text{ W}^{-1} \quad (7.8)$$

For a photon detector, the limiting spectral detectivity is:<sup>5</sup>

$$D^*(\lambda) = \frac{c_0 \eta(\nu)}{2h\nu\pi^{1/2} \left[ \int_{\nu_0}^{\infty} \frac{\eta(\nu') \nu'^2 h^{\nu'/kT_2} d\nu'}{(e^{h\nu'/kT_2} - 1)^2} \right]^{1/2}} \quad (7.9)$$

where  $\eta(\nu)$  is the *quantum efficiency* of the photoemissive surface (the average number of free electrons produced for



each incident photon absorbed),  $\lambda = c_0/\nu$ , and  $\nu_0$  is the lowest frequency to which the detector is sensitive. Spectral detectivities predicted by Equation (7.9) are described by Kruse, McGlauchlin, and McQuistan.<sup>5</sup>

### 7.3.3 Responsivity

The responsivity  $\mathfrak{R}$  of a detector specifies its response to unit irradiance:

$$\mathfrak{R} = V_S/IA \quad (7.10)$$

A similar figure of merit, usually used to characterize photoemissive detectors, is the *radiant sensitivity*  $S$ , which is the current per unit area of the photoemissive surface produced by unit irradiance:

$$S = \frac{i_S}{P} \quad (7.11)$$

where  $i_S$  is the total current from the detector and  $P$  the total radiant power falling on it. Some curves showing the spectral variation of radiant sensitivity for different photoemissive surfaces are given in Figure 7.3.

### 7.3.4 Quantum Efficiency

For a photon detector, the *quantum efficiency* is defined as:

$$\eta = \frac{\text{number of electrons or electron - hole pairs produced}}{\text{number of photons absorbed}} \quad (7.12)$$

It is easy to show that the quantum efficiency of a photon detector and its responsivity are related according to:

$$\mathfrak{R} = \frac{\eta e}{h\nu} \quad (7.13)$$

where  $e$  is the magnitude of the electronic charge.

### 7.3.5 Frequency Response and Time Constant

The frequency response of a detector is the variation of responsivity or radiant sensitivity as a function of the modulation frequency of the incident radiation. The signal from the detector should be a.c.-coupled; otherwise the generated d.c. signal that is produced as the detector begins

to fail to respond to the modulation will also be detected. The frequency variation of  $\mathfrak{R}$  and the time constant  $\tau$  of the detector are generally related according to:

$$\mathfrak{R}(f) = \frac{\mathfrak{R}(0)}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}} \quad (7.14)$$

like a low-pass filter, as shown in Figure 7.4.

Thus the responsivity has decayed to a value of  $\mathfrak{R}(0)/\sqrt{2}$  at a frequency  $1/2\pi\tau$ . The time constant  $\tau$  is a simple measure of the ability of the detector to respond to a sharply rising or falling optical signal.

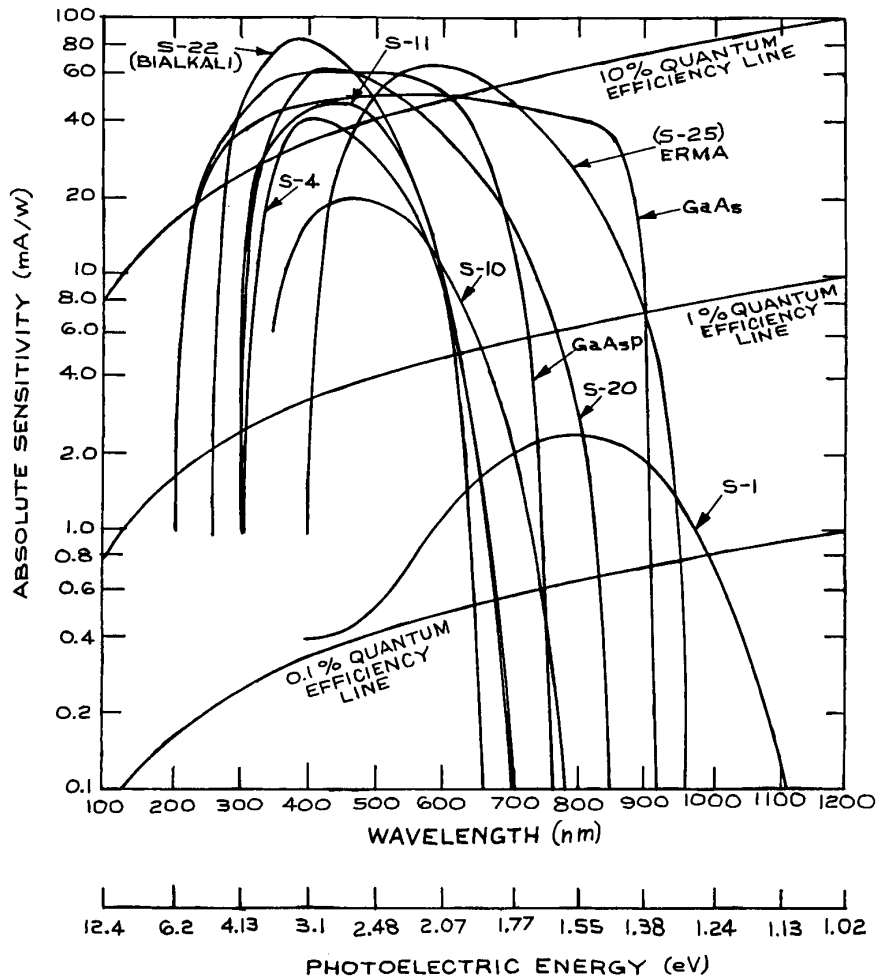
A very important parameter that determines the frequency response of a detector is its capacitance. For driving a 50 ohm load, for example, a capacitance of 1 pF will limit time constant to 50 ps. With the exception of photomultiplier tubes (PMTs) there is a direct relationship between detector size and time constant. Photomultiplier tubes with a sensitive area of 2000 mm<sup>2</sup> can respond as quickly as 1–2 ns. However, for semiconductor detectors there is no escaping the size/time-constant relationship. For example, a semiconductor with a 1 mm sensitive area will have a limiting time response of 1 ns and a bandwidth of 160 MHz. On the other hand, a small size photodiode with a diameter of 25  $\mu\text{m}$ , and designed for direct coupling to optical fiber will have a bandwidth out to 20 GHz.

### 7.3.6 Signal-to-Noise Ratio

The random fluctuations in the output voltage or current from a detector set a lower limit to the radiant power that can be detected, given the detector temperature and operating conditions, source modulation frequency, and electronic-detection-system bandwidth. The signal-to-noise ratio (S/N ratio) is the ratio of electrical power produced by the detected light to the noise power, which can be written in a simple form as:

$$\frac{S}{N} = \frac{\langle i_S^2 \rangle}{\langle i_N^2 \rangle} \quad (7.15)$$

where  $i_S$  is the current produced by the detected light, and  $i_N$  is the noise current. Angular brackets  $\langle \dots \rangle$  indicate time averaging. Noise can often be reduced by operating the detector in an appropriate way. In a practical detection system, the detector must be coupled to various forms of electronic processing systems, pre-amplifiers, amplifiers,



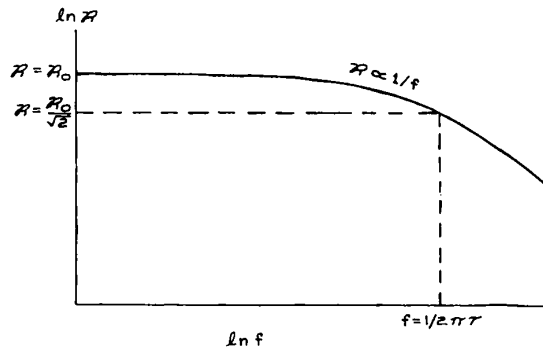
**Figure 7.3** Wavelength dependence of the radiant sensitivity of several commercially available photocathode materials. (From *Handbook of Lasers*, R. J. Pressley (Ed.), CRC Press, Cleveland, 1971; by permission of the CRC Press.)

filters, lock-in detectors, and so on. This electronic system necessarily creates additional noise; in a well-designed system this is kept to a minimum.

It will be seen in the following sections that there are various ways to reduce, and even essentially eliminate, detector noise. Photon noise, however, which arises from the signal itself cannot be eliminated. Thus, the ultimate performance of any detector is *photon-limited*. The photon-noise limit is different for thermal detectors, which

are sensitive to total absorbed radiation, and photon detectors, which respond, at least in a microscopic way, to the absorption of individual quanta.

The signal-to-noise (S/N) performance of a detector in a given situation can usually be calculated from the detector characteristics specified by the manufacture and the characteristics of the electronic system that the detector will drive. Specifically important are the dark-noise level, the noise figure [especially for devices with internal gain such



**Figure 7.4** Typical frequency dependence of detector responsivity  $\mathcal{R}$ .

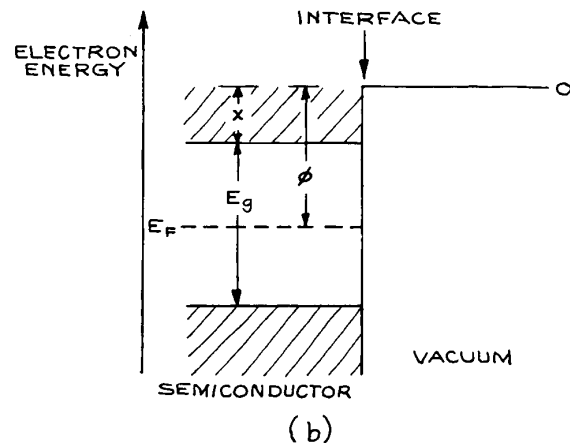
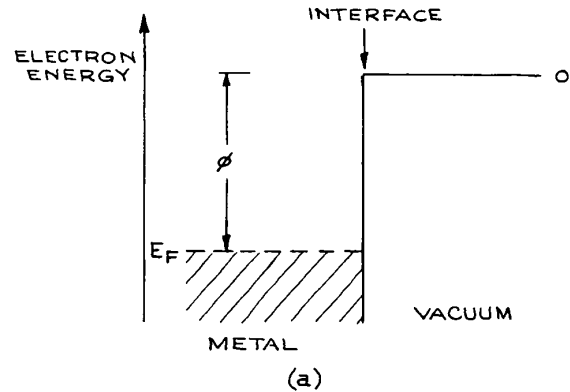
as avalanche photodiodes (APDs)], the quantum efficiency (or responsivity) of the detector, and the input impedance of the following electronics and its noise characteristics.

Davis<sup>6</sup> has provided several examples of how to calculate detector noise performance in various situations.

If the detector manufacturer specifies  $D^*$  or NEP, and not all do, signal-to-noise calculations can be simplified. Frequently the dark-noise current will be specified, usually in units of nA or pA. Noise generally increases with the square root of the bandwidth used. For example, a detector with a bandwidth of 100 MHz and a dark current of 5 nA has a dark noise of  $0.5 \text{ pA}/\sqrt{\text{Hz}}$ .

## 7.4 PHOTOEMISSIVE DETECTORS

Photoemissive detectors include vacuum photodiodes, photomultiplier tubes, microchannel plates and photo-channeltrons. These are all photon detectors that utilize the photoelectric effect. When radiation of frequency  $\nu$  falls upon a metal surface, electrons are emitted if the photon energy  $h\nu$  is greater than the *work function*  $\phi$  characteristic of the material being irradiated. A simplified energy-level diagram illustrating this effect for a metal–vacuum interface is shown in Figure 7.5(a). For most metals,  $\phi$ ; is in a range from 4–5 eV (1 eV equiv.  $1.24 \mu\text{m}$  although for the alkali metals it is lower, for example, 2.4 eV for sodium and 1.8 eV for cesium. Pure metals, alloys, (particularly beryllium–copper,) and cesium iodide, are used as photoemissive surfaces in ultraviolet and vacuum-



**Figure 7.5** Band structure of: (a) a metal vacuum interface and (b) a pure semiconductor–vacuum interface ( $\phi$  = work function;  $\chi$  = electron affinity;  $E_g$  = band-gap energy;  $E_F$  = Fermi level).

ultraviolet detectors. Lower work functions, and consequently sensitivities that can be extended into the infrared, can be obtained with special semiconductor materials. Figure 7.5(b) shows a schematic energy-level diagram of a semiconductor–vacuum interface. In this case the work function is defined as  $\phi = E_{\text{vac}} - E_F$ , where  $E_F$  is the energy of the Fermi level. In a pure semiconductor, the Fermi level is in the middle of the band gap, as illustrated in Figure 7.5(b). In a *p*-type doped semiconductor,  $E_F$  moves down toward the top of the valence band, while in *n*-type material it moves up toward the bottom of the conduction band. Consequently,  $\phi$ ; is not so useful a measure of the

minimum photon energy for photoemission as it is for a metal. The electron affinity  $\chi$  is a more useful measure of this minimum energy for a semiconductor. Except at absolute zero, photons with energy  $h\nu > \chi$  cause photoemission. Photons with energy  $h\nu > E_g$  lead to the production of carriers in the conduction band: this leads to intrinsic photoconductivity, which is the operative detection mechanism in various infrared detectors, such as InSb.

### 7.4.1 Vacuum Photodiodes

Once electrons are liberated from a photoemissive surface (a photocathode), they can be accelerated to an electrode (the anode), positively biased with respect to the cathode, and generate a signal current. If the acceleration of photoelectrons is directly from cathode to anode through a vacuum, the device is a vacuum photodiode. Because the electrons in such a device take a very direct path from anode to cathode and can be accelerated by high voltages – up to several kilovolts in a small device – vacuum photodiodes have the fastest response of all photoemissive detectors. Risetimes of 100 ps or less can be achieved. External connections and electronics are generally the limiting factors in obtaining short risetimes from such devices. Vacuum photodiodes, however, are not very sensitive, since, at most, one electron can be obtained for each photon absorbed at the photocathode.

If the space between photocathode and anode is filled with a noble gas, photoelectrons will collide with gas atoms and ionize them, yielding secondary electrons. Thus, an electron multiplication effect occurs. However, because the mobility of the electrons moving from cathode to anode through the gas is slow, these devices have a long response time, typically about 1 ms. Gas-filled photocells are no longer competitive with solid-state detectors and have disappeared from use.

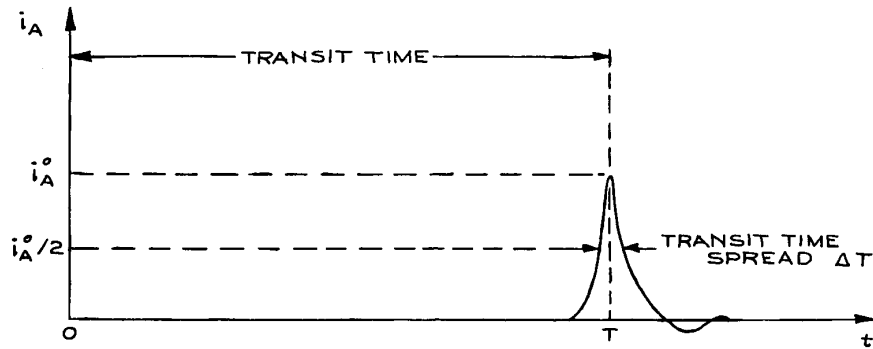
### 7.4.2 Photomultipliers

If photoelectrons are accelerated *in vacuo* from the photocathode and allowed to strike a series of secondary electron emitting surfaces, called *dynodes*, held at progressively more positive voltages, a considerable electron multiplication can be achieved and a substantial current can be collected at the anode. Such devices are called *photomul-*

*tipliers*. Practical gains of  $10^9$  (anode electrons per photoelectron) can be achieved from these devices for short light pulses. Continuous gains must be held lower, to perhaps  $10^7$ . This limit is imposed primarily by the ability of the final dynodes in the chain to withstand the thermal loading produced by continuous electron impact, and by the time needed for the dynodes to be recharged when large currents are drawn from the anode. Because of their very high gain, photomultipliers can generate substantial signals when only a single photon is detected: for example, an anode pulse of 2 ns duration containing  $10^9$  secondary photoelectrons produced from a single photoelectron will generate a voltage pulse of 4 V across 50 ohms. This, coupled with their low noise, makes photomultipliers very effective single-photon detectors. Photomultiplier  $D^*$  values can range up to  $10^{16}$  cm Hz<sup>1/2</sup> W<sup>-1</sup>. Only the dark-adapted human eye, which can detect bursts of about 10 photons in the blue, comes close to this sensitivity.

The time-response characteristics of photomultiplier tubes depend to a considerable degree on their internal dynode arrangement. The response of a given device can be specified in terms of the output signal at the anode that results from a single photoelectron emission at the photocathode. This is illustrated in Figure 7.6. Because electrons passing through the dynode structure can generally take slightly different paths, secondary electrons arrive at the anode at different times. The anode pulse has a characteristic width called the *transit-time spread*, which usually ranges from 0.1 to 20 ns. The time interval between photoemission at the cathode and the appearance of an anode pulse is called the *transit time*, and is usually a few tens of nanoseconds. The transit time and transit-time spread also fluctuate slightly from one single-photoelectron-produced anode pulse to the next. Although this is a fine point, it may be something to worry about in precision-timing experiments. The principle suppliers of photomultiplier tubes are Hamamatsu, Burle Industries<sup>17</sup> (now merged with Photonis), Electron Tubes, Phillips (Amperex) Photek, and Sens-Tek.

There are seven main types of dynode structure in common usage in photomultiplier tubes; these are illustrated in Figure 7.7. The circular cage structure [Figure 7.7(a)] (used, for example, in the Hamamatsu 1P28 photomultiplier tube) is compact and can be designed for good electron-collection efficiency and small transit-time spread.



**Figure 7.6** Typical anode pulse produced by a single photoelectron emission at the cathode of a photomultiplier tube.  $t$  is the time following photoemission. The transit-time  $T$ , the transit-time speed  $\Delta T$ , and the peak anode current  $i_A^0$  all fluctuate from pulse to pulse.

This dynode structure works well with opaque photocathodes, but is not very suitable for high-amplification requirements, where a larger number of dynodes is required. The box-and-grid [Figure 7.7(b)] and Venetian-blind structures [Figure 7.7(c)] (used, for example, in the Hamamatsu Models R464 and R1513, respectively) offer very good electron-collection efficiency. Because they collect multiplied electrons independent of their path through the dynode structure, a wide range of secondary-electron trajectories is possible, leading to a large transit-time spread and slow response. Typical response times of these types of tube are 10–20 ns. Venetian-blind tubes can easily be extended to many dynode stages and have a very stable gain in the presence of small power supply fluctuations. These tubes also have an optically opaque dynode structure providing very low dark-current noise when operated under appropriate conditions.

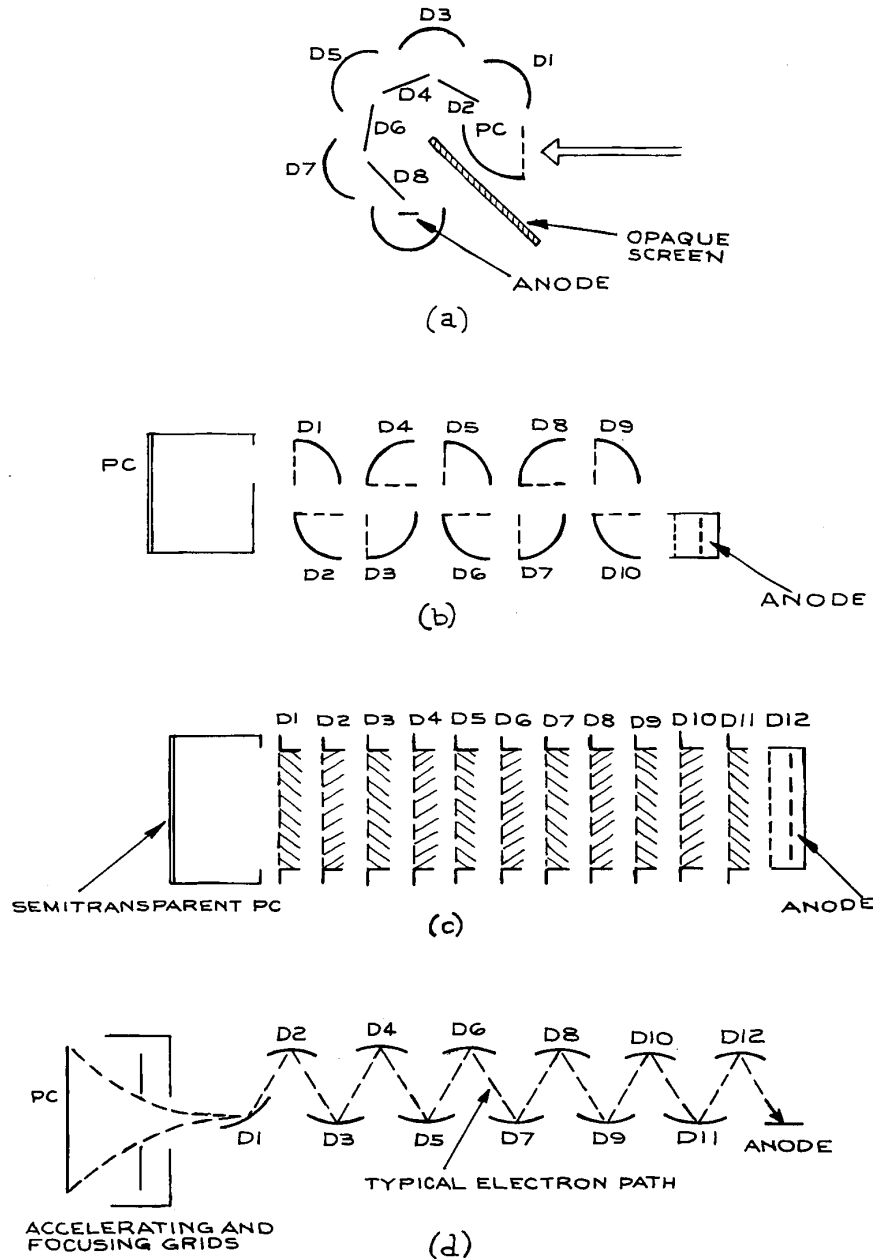
The focused dynode structure [Figure 7.7(d)] (used, for example, in the Hamamatsu R331, R2059, and the old Phillips 56 AVP and other 56-series tubes<sup>18</sup>) is designed so that electrons follow paths of similar length through the dynode structure. To accomplish this, electrons that deviate too much from a specified range of trajectories are not collected at the next dynode. These types of tube offer short response times, typically 1–2 ns; some recently developed tubes are even faster.

The mesh type of dynode structure [Figure 7.7(e)] has a series of crossed fine-mesh dynodes arranged in close

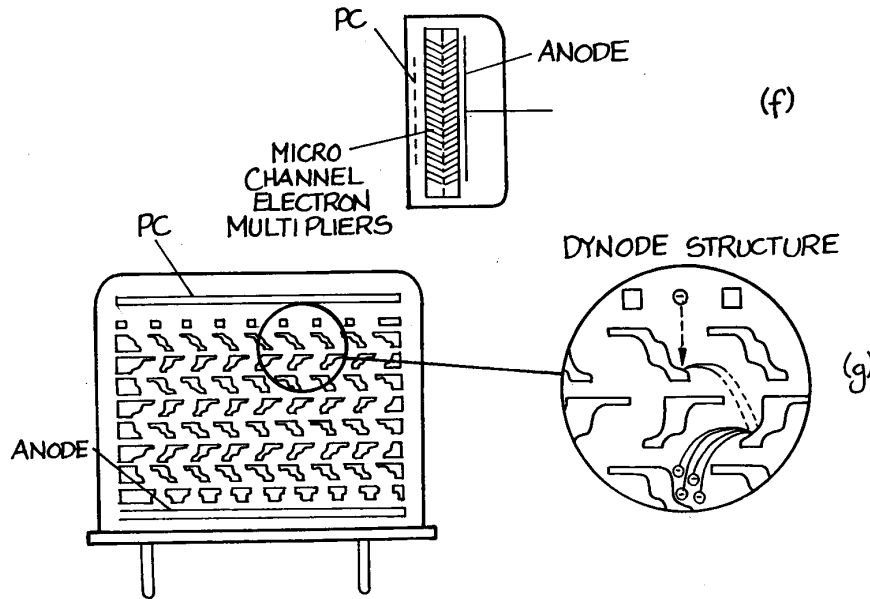
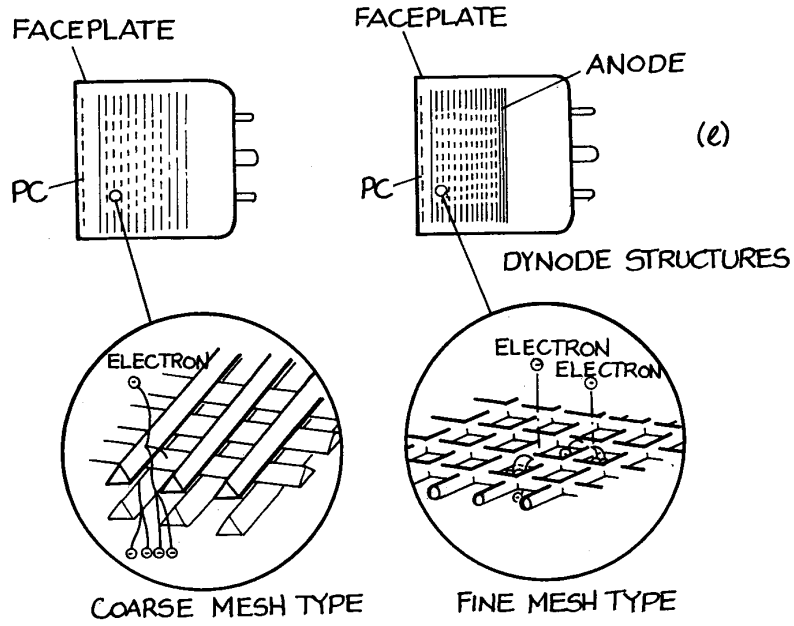
proximity. The structure allows compact construction, high immunity to magnetic fields, as well as uniformity of response and good pulse linearity. Versions of this type with special anodes provide position-sensitive performance. In this type of device the output signal from a series of anodes provides direct information on the location of illumination on the photocathode surface.

The metal-channel dynode structure [Figure 7.7(g)] provides a very compact dynode structure and fast response becomes of the closeness between each stage of dynodes. This structure also allows position-sensitive detection. Good examples of this type of tube are in the Hamamatsu R5900 series. These types of tube can tolerate quite large magnetic fields. The Hamamatsu R5946 tube has been shown to operate satisfactorily in fields of 4kG.<sup>19</sup>

Closely related to these fundamental photomultiplier dynode structures is the *microchannel plate* (MCP) [Figure 7.7(f)]. Instead of traditional dynodes, photoelectrons enter a bundle of capillary tubes fused together in an array. Each tube has a resistive coating on its inner surface that acts as a continuous electron multiplier. Each capillary individually can be referred to as a channel electron multiplier (CEM) [or *channeltron* (a trademark of Burle Industries)]. In these capillaries electrons make multiple reflections from the tube walls generating secondary electrons. MCPs have fast time response, good immunity from external magnetic fields, and position-sensitive capability when provided with multiple anodes. These devices are available



**Figure 7.7** Schematic diagram of the internal structure of various types of photomultiplier tube; (a) squirrel cage; (b) box and grid; (c) venetian blind; (d) focused dynode; (e) mesh type; (f) microchannel electron multiplier; (g) metal-channel dynode structure. D = dynode; PC = photocathode.



commercially from Burle (Photonis), DelMar Ventures, El-Mul, Hamamatsu, Photek, TecTra, and Topag. CEMs are primarily used for particle detection (see later).

Other types of multiplier structures, not shown in Figure 7.7, are also available. Traveling-wave phototubes and various types of crossed-field photomultiplier have very short response times, down to 0.1 ns.<sup>20</sup>

### 7.4.3 Photocathode and Dynode Materials

The performance of a photomultiplier depends not only on its internal structure, but also on the photoemissive material of its photocathode and the secondary-electron-emitting material of its dynodes. The selection of a photodiode or photomultiplier for a particular spectral response depends on an appropriate selection of photocathode material. The wavelength dependence of various commercially available photocathode materials is shown in Figure 7.3; some radiant sensitivities and quantum efficiencies are given in Table 7.1. Figure 7.8 shows the radiant sensitivities and quantum efficiencies of some new negative electron affinity (NEA) photocathode materials. Practical quantum efficiencies for photoemissive materials range up to about 0.4. The short-wavelength cutoff of a given material depends on its work function. This cutoff is not sharp because, except at absolute zero, there are always a few electrons high up in the conduction band available for photoexcitation by low-energy photons. Some few of these electrons, because of their thermal excitation, will be emitted without any photostimulation. This contributes the major part of the dark current observed from the photocathode. Materials that have low work functions, and are consequently more red- and infrared-sensitive, have higher – often much higher – dark currents than materials that are optimized for visible and ultraviolet sensitivity. Such photocathodes are best operated cooled. Unless a phototube is to be used primarily for detecting long-wavelength radiation, we recommend selecting a photocathode with the shortest possible wavelength response.

Photocathodes are available in both opaque and semitransparent forms, depending on the model of phototube. In the semitransparent form the photoelectrons are ejected from the thin photoemissive layer on the side opposite to the incident light. In both types the photocathode has to perform two important functions: it must absorb incident

photons and allow the emitted photoelectrons to escape. The latter event is inhibited if photons are absorbed too deep in a thick photoemissive layer, or if the photoelectrons suffer energy loss from scattering in the layer. If the emitted photoelectron has too much energy, it can excite a further electron across the band gap. This pair-production phenomenon inhibits the release of photoelectrons from the photoemissive layer, and accounts for the ultraviolet cutoff characteristics of the different materials shown in Figure 7.2. With reference to Figure 7.5(b), it can be shown that for pair production to occur, the incident photon energy must be greater than  $2E_g$ . For materials where  $\chi < E_g$ , photoelectrons have their best chance of escaping, and the photocathode its highest quantum efficiency.

Because semitransparent photocathodes allow photons to penetrate sufficiently for photoelectrons to escape on the far side of the layer, they are deposited as semiconductor layers on the insulating glass substrate of the entrance window. These photocathodes are less tolerant of relatively high light levels than are opaque photocathodes, which can be deposited on a conductive substrate. A conductive substrate preserves field distribution within the PMT even at relatively large cathode currents. When an opaque photocathode is used, photons pass through the entrance window of the PMT and strike the photocathode, which is mounted separately, as shown schematically in Figure 7.7(a). Photoelectrons are emitted on the same side of the photocathode as photons are absorbed.

Practical photoemissive materials fall into two main categories: classical photoemitters and NEA materials. Classical photoemitters generally involve an alkali metal or metals, a group V element, such as phosphorus, arsenic, antimony, or bismuth, and sometimes silver and/or oxygen. Examples are the Ag-O-Cs (S1) photoemitter, which has the highest quantum efficiency beyond about 800 nm of any classical photoemitter, and Na<sub>2</sub>KSbCs, the so-called tri-alkali (S-20) cathode.

Negative electron affinity photoemitters have only been developed within the last 15 years. These materials generally utilize a photoconductive single-crystal semiconductor substrate, with a very thin surface coating of cesium and usually a small amount of oxygen. The cesium (oxide) layer lowers the electron affinity below the value it would have in the pure semiconductor, achieving an effectively



Table 7.1 Characteristics of photoemissive surfaces

Cathode	Radiant Sensitivity (mA/W)		1.06 $\mu\text{m}$	Peak Quantum Efficiency (%)	$\lambda_{\text{peak}}$ ( $\mu\text{m}$ )	Cathode Dark Current <sup>a</sup> (A/cm <sup>2</sup> )
	515 nm	694 nm				
S-1 Cs-O-Ag	0.6	2	0.4	0.08	800	$9 \times 10^{-12}$
S-10 Cs-O-Ag-Bi <sup>b</sup>	20	2.7	0	5	470	$4 \times 10^{-16}$
Cs <sub>3</sub> Sb on MnO <sup>c</sup>	39	0.2	0	13	440	$10^{-16}$
(tri-alkali)	53	20	0	18	470	$9 \times 10^{-16}$
S-22 (bi-alkali)	42	0	0	26	390	$2 \times 10^{-15}$
GaAs <sup>d-g</sup>	48	28	0	14	560 <sup>h</sup>	$10^{-16}$
GaAsP <sup>d</sup>	60	30	0	19	400	$3 \times 10^{-15}$
InGaAs(Cs) <sup>d,f</sup>	7	2.6	0.1	15	380	$2 \times 10^{-14}$
InP/InGaAs <sup>d,e,f</sup>	3	3		11(1.5 $\mu\text{m}$ )	1.55 $\mu\text{m}^h$	$4 \times 10^{-14}$
InP/InGaAsP <sup>d,f</sup>	3	3		10 (1.3 $\mu\text{m}$ )	1.3 $\mu\text{m}^i$	$4 \times 10^{-15}$
S-25 (ERMA) <sup>j</sup>	53	26	0	25	430	$1 \times 10^{-15}$
Cs-Te (solar blind)	—	—	—	15	254	$2.5 \times 10^{-17}$

Note: Table shows typical values, but these can vary greatly from one manufacturer to another. We recommend that manufacturers' data sheets be consulted for more details.

<sup>a</sup> At room temperature.

<sup>b</sup> Cathode designated S-3 is similar.

<sup>c</sup> Several types of CsSb photocathodes exist where the CsSb is deposited on different opaque and semitransparent substrates and various window materials are used. These photocathodes have the designations S-4, 5, 13, 17, and 19 as well as S-11.

<sup>d</sup> NEA photoemitters.

<sup>e</sup> Available from Burle Industries.

<sup>f</sup> Available from Hamamatsu Corporation.

<sup>g</sup> Available from Hamamatsu Corporation.

<sup>h</sup> Sensitive out to 1.7  $\mu\text{m}$ . Exact spectral characteristics will depend on thickness of photoemitter and whether it is used in transmission or not.

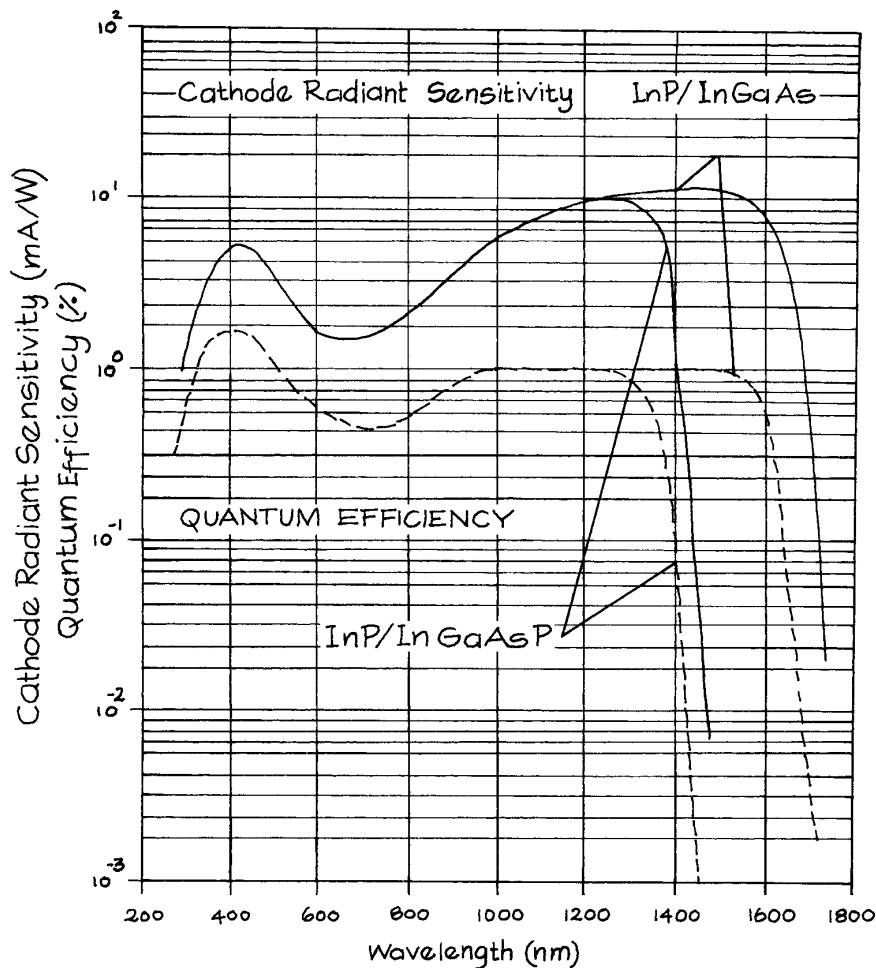
<sup>i</sup> Sensitive out to 1.4  $\mu\text{m}$ . Exact spectral characteristics will depend on thickness of photoemitter and whether it is used in transmission or not.

<sup>j</sup> Extended-red S-20.

negative value. Examples of such NEA photoemitters are GaAs (CsO) and InP (CsO). Negative electron affinity emitters can offer very high quantum efficiency and extended infrared response. For example, GaAs (CsO) has higher quantum efficiency in the near infrared than an S-1 photocathode. Other new NEA photocathodes include InP/InGaAsP (sensitive to 1.4  $\mu\text{m}$  and InP/InGaAs (sensitive to 1.7  $\mu\text{m}$ ). Commercial photomultipliers with NEA photocathodes are available from Burle Industries and Hamamatsu, that offer response out to 1.7  $\mu\text{m}$  in a tube cooled to  $-100^\circ\text{C}$ . Development of NEA photoemitters is continuing<sup>21</sup>, and the experimentalist seeking long-wavelength response would be well advised to follow

the literature and check for the introduction of new commercial tubes. For further details about NEA materials, the interested reader should consult the article by Zwicker.<sup>22</sup>

The performance of the dynode material in photomultiplier tubes is specified in terms of the secondary-emission ratio  $\delta$ , as a function of energy. For a phototube with  $n$  dynodes the gain is  $\delta^n$ . In the past, the most common dynode materials were CsSb, AgMgO, and BeCuO. The last is also used as the primary photoemitter in windowless photomultipliers that are operated *in vacuo* for the detection of vacuum-ultraviolet radiation. BeCuO can be reactivated after exposure to air. The above materials have



**Figure 7.8** Quantum efficiencies and radiant sensitivities of near infrared semiconductor photocathode materials.

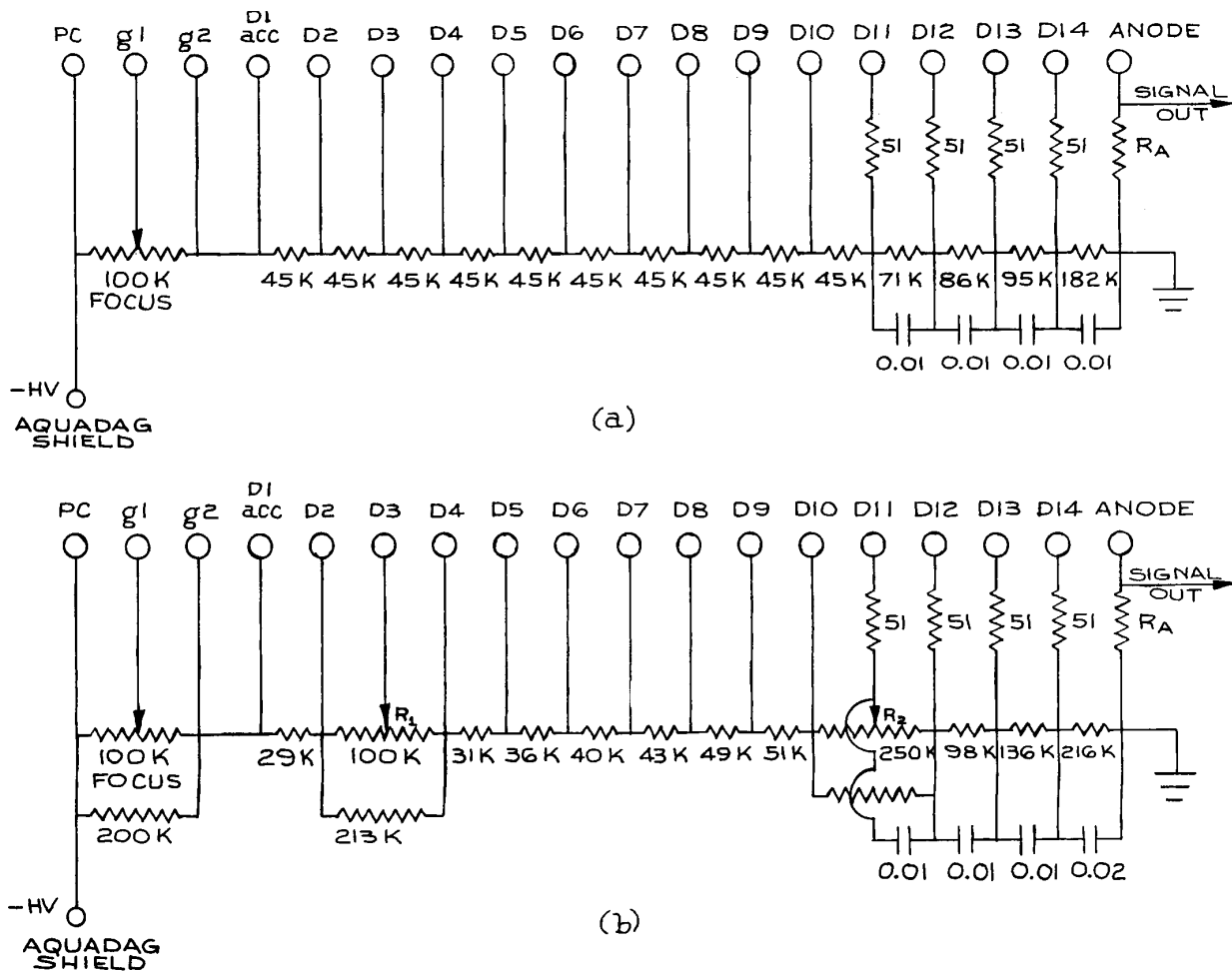
$\delta$ -values of 3–4. Newer NEA dynode materials have much higher  $\delta$ -values; in particular, that of GaP can range up to 40 for an incident-electron energy of 800 eV. With such high  $\delta$ -values, a photomultiplier tube needs fewer dynodes for a given gain, but with larger interdynode accelerating voltages. Such tubes are compact and provide fast time response. In very many commercial photomultipliers, at least the first dynode is now often made of GaP. This offers improved characterization of the single-photo-electron response of the tube, which is important in designing a system for optimum signal-to-noise ratio.

#### 7.4.4 Practical Operating Considerations for Photomultiplier Tubes

Complete photomultiplier assemblies are available, in which the tube, mounting base, dynode chain and output connector are pre-packaged. Suppliers include C&L Instruments, Hamamatsu, OE Technologies, and Products for Research. For the experimentalist who desires more control over tube operation, the following sections should be useful. (Note also that Hamamatsu have published a comprehensive guide to photomultiplier use.<sup>23</sup>)

**Dynode Chains.** The accelerating voltages are supplied to the dynodes of a photomultiplier by a resistive voltage divider called a *dynode chain*. The relative resistance values in the chain determine the distribution of voltages applied to the dynodes. The total chain resistance  $R$  determines the chain current at a given total photocathode-anode applied voltage. Some phototubes are supplied with an integral dynode chain, but most are not. Adequate dynode-chain designs are

generally supplied by the manufacturer with each tube. Alternative designs intended for certain types of response, such as maximum gain for short pulses or highest linearity, can frequently be found by consulting the literature. Two examples of such dynode-chain designs for use with Phillips 5-series tubes are given in Figure 7.9. The total dynode-chain resistance is chosen so that the chain current is at least 100 times greater than the average d.c. anode current to be drawn from the

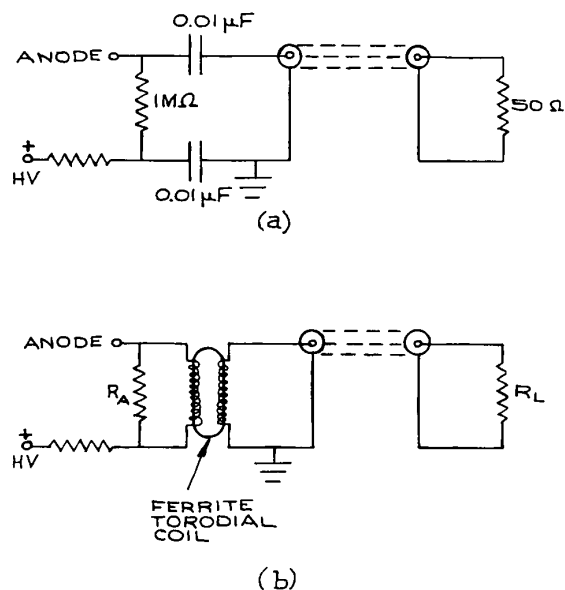


**Figure 7.9** Dynode chains for Phillips 56-series photomultiplier tubes: (a) chain for high gain and fast time response; (b) chain for high linearity ( $R_1$  and  $R_2$  adjusted for optimum performance). PC = photocathode; g1, g2 = grids; acc = accelerator grid; D1–D14 = dynodes. Resistance values are in ohms, capacitance values in  $\mu$  F. For very low-level light detection, all resistances can be scaled upward to reduce overall chain current.

tube. This current is specified for each tube and may range as high as 1 mA for high-gain tubes. The operation of photomultiplier tubes at such high average currents for very long is not recommended. Long-term d.c. anode currents should be kept between 1 and 10  $\mu\text{A}$  for most tubes.

There are several interesting features of the dynode chains shown in Figure 7.9. The photocathode is generally operated at negative potential with the anode near ground. This makes it easy to couple the output signal to other electronics. The tube can be operated with the cathode grounded, but in this case the signal from the anode, which is now at high positive potential, must be coupled through high-voltage capacitors or an insulated transformer, as shown in Figure 7.10. The photocathode is also often connected to a conducting shield, which is either a painted coat of colloidal graphite (Aquadag) on the outside of the tube envelope, or a metal foil wrapped around the tube envelope. If the photocathode is at negative potential, appropriate insulation must be used around the tube if the tube is in close proximity to any grounded metal parts inside its housing.

Many photomultiplier tubes have one or more focusing electrodes between the photocathode and the first dynode. The voltage on these electrodes can be adjusted to optimize the collection of photoelectrons from the photocathode. The last three or four dynodes are decoupled with high-voltage, high-frequency capacitors (disc ceramics work well in this application). These capacitors prevent depression of the dynode-chain voltage on the last few dynodes when a large pulse of electrons passes through the tube (the secondary-electron current drawn from each dynode must be supplied by the dynode chain). For high-frequency applications the inclusion of small damping resistors (typically between 50  $\Omega$  and 1 k $\Omega$ ) between the decoupling capacitors and the dynodes is recommended. The resistor  $R_A$ , the anode resistor, determines the actual voltage that will be produced by a given anode current, since the photomultiplier serves as a current source.  $R_A$  should not be so large that the voltage developed across it is significant compared with the voltage between the anode and the last dynode. A value of  $R_A$  between 1 and 10 M $\Omega$  is usual. In high-frequency applications, such as single-photon counting, the signal from  $R_A$  should be coupled out through a 50  $\Omega$  coaxial cable terminated in 50  $\Omega$ . In many cases, the



**Figure 7.10** Anode signal coupling methods for a photomultiplier operated with the anode at high positive potential: (a) capacitor coupling for fast-risetime, short-pulse operation (the component sizes are typical; the capacitors should be high-frequency, high-voltage types); (b) transformer coupling for lower-frequency, modulated operation (the high-voltage winding on the transformer should be sufficiently well insulated for isolation from the core; the size of  $R_A$  will depend on various factors, such as transformer primary impedance and operating frequency).

effective value of the anode load is set by the input impedance of the electronics to which the tube is connected. In such applications the cable connection to the anode should be shielded as close to the anode as possible, even going so far as to insert  $R_A$  under the grounded screen of the coaxial cable. Some high-frequency tubes are supplied with an integral coaxial connector on the anode. In any case, the aim is to reduce parasitic inductance and capacitance associated with the anode connection. For example, even with a 50  $\Omega$  load, a parasitic capacitance greater than 20 pF will limit the response time of the tube to 1 ns.

The actual response-time behavior of the photomultiplier can be determined by observing its single-photo-electron response. This is done by looking at the anode pulses with a fast oscilloscope, which generally requires that the 50  $\Omega$

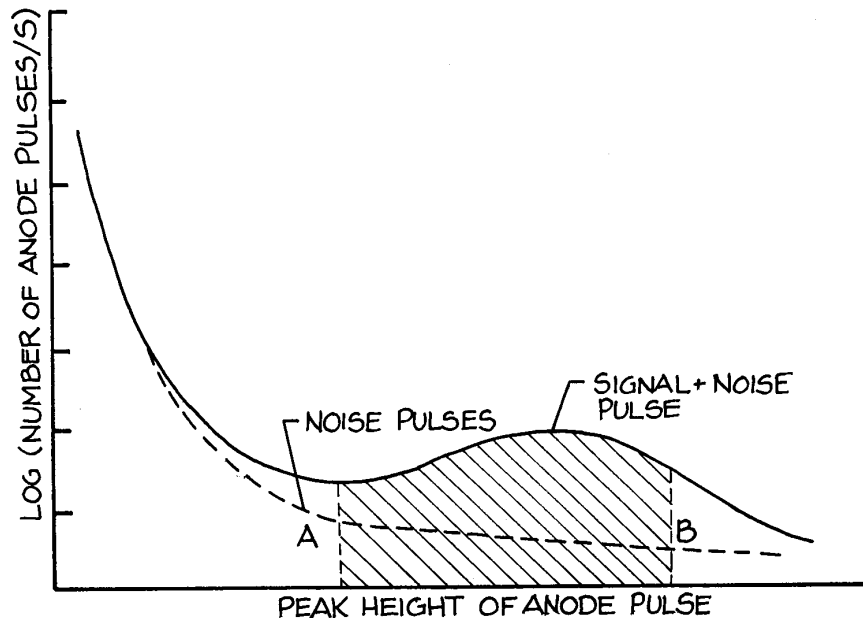
coaxial cable from the anode be terminated in  $50\ \Omega$ . The photocathode should not need to be illuminated for this to be done; sufficient noise pulses will usually be observed. The pulses should appear as in Figure 7.6. If they have too much of an exponential tail they are probably limited by the anode resistor and parasitic capacitance. These anode pulses reflect the time distribution and number of secondary electrons reaching the anode following single (or multiple) photoelectron emissions from the photocathode. If the height distribution of anode pulses is measured, a distribution such as is shown in Figure 7.11 will probably be seen.

**Mounting Photomultiplier Tubes.** Photomultiplier tubes are generally mounted by clamping their plastic socket inside a cylindrical tube housing. The housing should be light-tight: a photomultiplier should never be exposed to ambient lighting when its high-voltage dynode chain is on, for the resultant large secondary-electron current will disintegrate the last dynodes and may strip

the photocathode layer itself. If the total power dissipation of the dynode chain is low, the whole dynode chain may be mounted directly on the tube socket. If space is not at a premium, the dynode chain can be remote from the tube base and connected to it with high-voltage-insulated wire. The damping resistors, decoupling capacitors, and anode resistor should, however, be kept in close proximity to the tube. The photomultiplier should be mounted so there is no optical path from its base to its photocathode.

Photomultipliers are fragile and should be mounted so they are not subject to stress. This is particularly important if a tube is to be cooled. In practice, this means that the tube should be grasped at only one point in its mount. Rugged (and expensive) photomultipliers that can withstand severe vibrations and accelerations are available from ADIT and Hamamatsu.

**Noise in Photomultiplier Tubes.** Noise in photomultipliers comes from several sources:



**Figure 7.11** Schematic photomultiplier anode pulse-height distribution likely to be observed in practice. The best signal to noise ratio would be obtained in a photon-counting experiment by collecting only anode pulses in a height range roughly indicated by the shaded region *AB*.

- (1) Thermionic emission from the photocathode
- (2) Thermionic emission from dynodes
- (3) Field emission from dynodes (and photocathode) at high interdynode voltages
- (4) Radioactive materials in the tube envelope (for example,  $^{40}\text{K}$  in glass)
- (5) Electrons striking the tube envelope and causing fluorescence
- (6) Electrons striking the dynodes and causing fluorescence
- (7) Electrons colliding with residual atoms of vapor in the tube (cesium for example) and causing fluorescence
- (8) Cosmic rays.

Noise from photomultipliers is always greater after they have been exposed to light (without high voltage applied). They gradually become quieter after operation under dark conditions for an extended period.

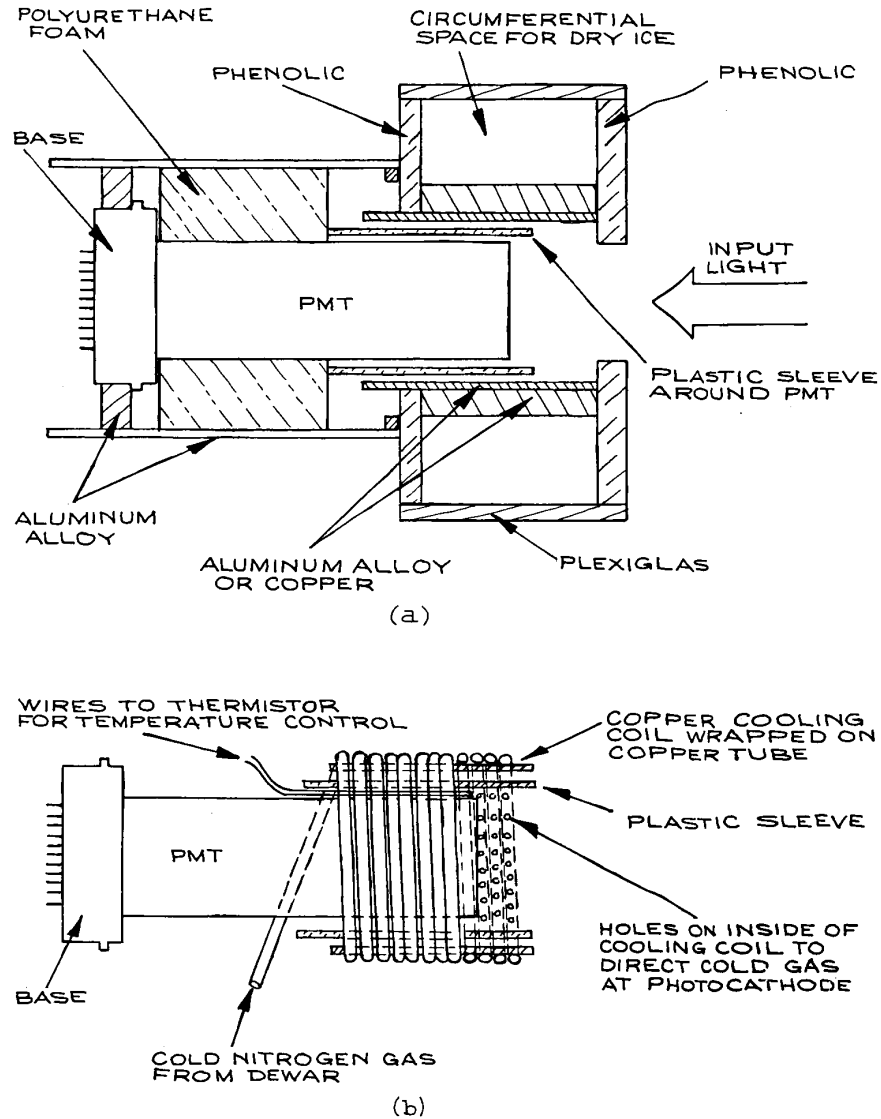
Noise from thermionic emission can be reduced considerably by cooling the tube (particularly for S-1 tubes). A factor of 10–1000 reduction in noise can be obtained by cooling to about  $-20^\circ\text{C}$ . S-1 tubes need to be cooled to low temperatures, but temperatures below about  $-70^\circ\text{C}$  are not recommended. Too much cooling has a marginal effect on reducing thermionic emission – which will already be negligible for most tubes at  $-20^\circ\text{C}$  – and may have deleterious effects. At low temperatures, the conductivity of the photocathode layer falls, which causes it to no longer be an equipotential. The resultant distorted field distribution between cathode and dynodes will increase the possibility of noise from sources such as item (5) above. Two convenient designs of photomultiplier tube-coolers are shown in Figure 7.12. In the design shown in Figure 7.12(a), a copper or aluminum cylinder encloses the tube (without touching it). This tube is in thermal contact with a chamber containing Dry Ice (or liquid nitrogen). Condensation on the face of the photomultiplier tube is minimized, since moisture preferentially condenses and freezes on the much colder metal tube. In the second design, shown in Figure 7.12(b), cold nitrogen gas boiled off from a liquid-nitrogen Dewar flows through a copper tube surrounding the photomultiplier, and finally sprays over its photocathode surface to minimize water-vapor condensation. The rate of supply of cold nitrogen gas can be adjusted to provide a range of temperatures.

Photomultiplier tube coolers are available from Amherst Scientific, Electron Tubes, Hamamatsu, and Products for Research.

When noise from field emission is a problem with certain high-voltage, high-gain tubes, one solution is to reduce the voltage. In single-photon-counting experiments, larger than normal anode pulses may result from field emission and can be rejected with an appropriate window discriminator.<sup>24</sup>

Noise from radioactive envelope materials should generally be negligible. Noise from electrons striking the tube envelope can be reduced by a shield around the tube held at photocathode potential. Noise from electrons striking the dynodes and causing fluorescence causes the fewest problems in Venetian-blind tubes, which have little or no direct optical path from dynodes to photocathode. Noise from gas in the tube can be a problem in old tubes, which frequently become gassy; solution: buy a new tube. Noise from cosmic rays is usually unimportant; it can be reduced by shielding.

Of all these sources of noise, thermionic emission from the photocathode is generally by far the most important. For a given photocathode material the total noise will depend on the photocathode area. Most manufacturers will list the dark current observed at the anode. The values in Table 7.1 have been derived from manufacturers' anode dark currents corrected for tube gain and photocathode area. Under comparable conditions, tubes with a small photocathode area, like the old S-20 ITT FW-130 with a photocathode diameter of 0.25 cm, are less noisy than tubes with large photocathodes, like the 46 mm photocathode of the S-20 Hamamatsu R550. In many experiments a large photocathode area is unnecessary – light coming through a monochromator slit illuminates a very small area. The effective photocathode area, and consequently its thermionic noise, can be reduced by wrapping a magnetic coil around the photocathode, which prevents electrons other than those from the center of the photocathode from reaching the first dynode. Under circumstances where photomultipliers must be operated in close proximity to magnetic fields, magnetic shields to enclose the tube are available from Ad-Vance Magnetics, Amuneal, Magnetic Shield Corp. (Perfection Mica), Mu Shield, and Products for Research.



**Figure 7.12** Photomultiplier-tube coolers: (a) using Dry Ice reservoir in thermal contact with a metal tube close to, but electrically insulated from, the photocathode; (b) using cold nitrogen gas circulated around tube and sprayed onto photocathode surface (a thermistor monitors temperature and allows control of the rate of supply of cold gas).

In the anode pulse-height distributions shown in Figure 7.11, small anode pulses are much more likely to correspond to noise than are pulses in the middle of the distribution. In photon-counting experiments these

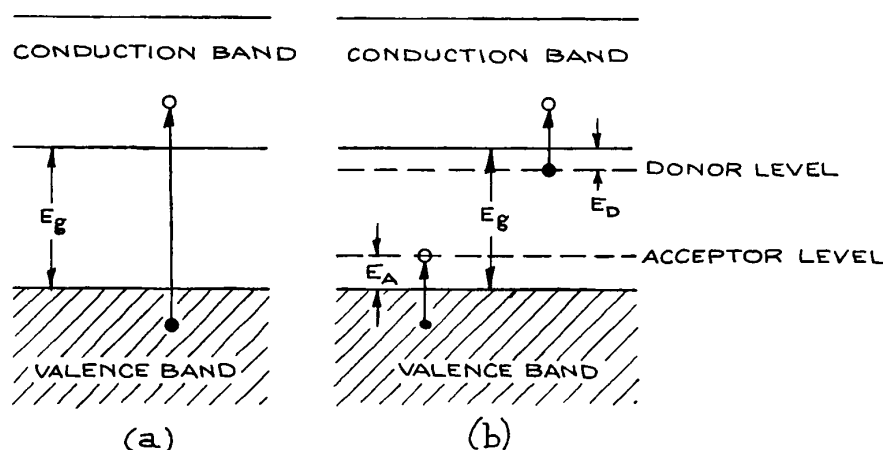
small pulses should be rejected with a discriminator. A window discriminator provides simultaneous rejection of these pulses and those which are much larger than average.<sup>25</sup>

**Ultraviolet and Vacuum-Ultraviolet Detection with Photodiodes and Photomultipliers.** For detection of radiation down to about 105 nm, phototubes with windows of LiF are available, as are tubes with windows of MgF<sub>2</sub> or sapphire. Windowless phototubes or CEMs are used for detection of even shorter wavelengths. For detection of radiation down to 58 nm or beyond, conventional photomultiplier tubes sensitive to visible light (S-11) can be used if the outer surface is coated with a layer of sodium salicylate – or if a sodium-salicylate-coated disc is placed close to the photocathode. Sodium salicylate fluoresces in the visible with almost unity quantum efficiency and converts incident vacuum-ultraviolet radiation to a wavelength where a conventional tube can detect it. The response of the sodium salicylate is also fast, so that 1 ns ultraviolet detection is possible. To prepare a sodium-salicylate-coated window, a solution of sodium salicylate in reagent-grade methanol (80 g/L) is atomized and sprayed onto the window, at a distance of about 10 cm from the atomizer. Commercial nasal sprays and air brushes pressurized with dry nitrogen are suitable for this purpose. The atomizer should be far enough from the window that no large drops can be deposited. The method works better if the window is kept warm (50–70 °C). Spraying should be continued until the density of the sodium salicylate layer reaches about 20 g/m<sup>2</sup>, which can be determined by weighing.

## 7.5 PHOTOCONDUCTIVE DETECTORS

Photoconductive detectors can operate through either intrinsic or extrinsic photoconductivity. The physics of *intrinsic photoconductivity* is illustrated in Figure 7.13(a). Photons with energy  $h\nu > E_g$  excite electrons across the band gap. The electron–hole pair that is thereby created for each absorbed photon leads to an increase in conductivity (mostly from the electrons.) Semiconductors with small band gaps respond to long-wavelength infrared radiation but must be cooled accordingly; otherwise thermally excited electrons swamp any small photoconductivity effects. Table 7.2 lists commonly available intrinsic photoconductive detectors together with their usual operating temperature and the limit of their long-wavelength response  $\lambda_0$ , together with some representative figures for detectivities and time constants. Note that silicon and germanium are also operated in both photovoltaic and avalanche modes (see Section 7.6).

If a semiconductor is doped with an appropriate material, impurity levels are produced between the valence and conduction bands, as shown in Figure 7.13(b). Impurity levels that are able to accept an electron excited from the conduction band are called *acceptor levels*, whereas impurity levels that can have an electron excited from them into the conduction band are called *donor levels*. Thus, in



**Figure 7.13** Mechanism for: (a) intrinsic photoconductivity; (b) extrinsic photoconductivity.



Table 7.2 Intrinsic photoconductive detectors

Semiconductor	$T$ (K)	$E_g$ (eV)	$\lambda_0$ ( $\mu\text{m}$ )	$D^*$ (max) ( $\text{cm Hz}^{1/2} \text{W}^{-1}$ )	$\tau$
CdS	295	2.4	0.52	$3.5 \times 10^{14}$	$\approx 50$ ms
CdSe	295	1.8	0.69	$2.1 \times 10^{11}$	$\approx 10$ ms
Si <sup>e</sup>	295 <sup>a</sup>	1.12	1.1	$\leq 2 \times 10^{12}$	— <sup>b</sup>
Ge <sup>e</sup>	295 <sup>a</sup>	0.67	1.8	$10^{11}$	10 ns <sup>c</sup>
PbS	295	0.42	2.5	$\leq 2 \times 10^{11}$	— <sup>d</sup>
	195	0.35	3.0	$\leq 5 \times 10^{11}$	— <sup>d</sup>
	77	0.32	3.3	$\leq 8 \times 10^{11}$	— <sup>d</sup>
PbSe	295	0.25	4.2	$1 \leq 10^9 - 5 \times 10^9$	1 $\mu\text{s}$
	195	0.23	5.4	$1.5-4 \times 10^{10}$	30–50 $\mu\text{s}$
	77	0.21	5.8	$\leq 3 \times 10^{10}$	50 $\mu\text{s}$
InSb <sup>e</sup>	77	$\approx 0.23$	5.5–7.0	$\leq 3 \times 10^{10}$	0.1–1 $\mu\text{s}$
Hg <sub>x</sub> Cd <sub>y</sub> Te <sup>f</sup>	77	$\leq 0.1$	12–25	$10^9-10^{11}$	>1 ns

<sup>a</sup> Increased sensitivity can be obtained by cooling.

<sup>b</sup> Detectors with time constants from 50 ps upward and various detectivities are available.

<sup>c</sup> Detectors operated in a photovoltaic mode with time constants from 120 ps upward and various detectivities are available.

<sup>d</sup> Detectors with time constants ranging from about 100  $\mu\text{s}$  to 10 ms and varying detectivities are available.

<sup>e</sup> More commonly operated in a photovoltaic mode.

<sup>f</sup> Wavelength range depends on stoichiometry

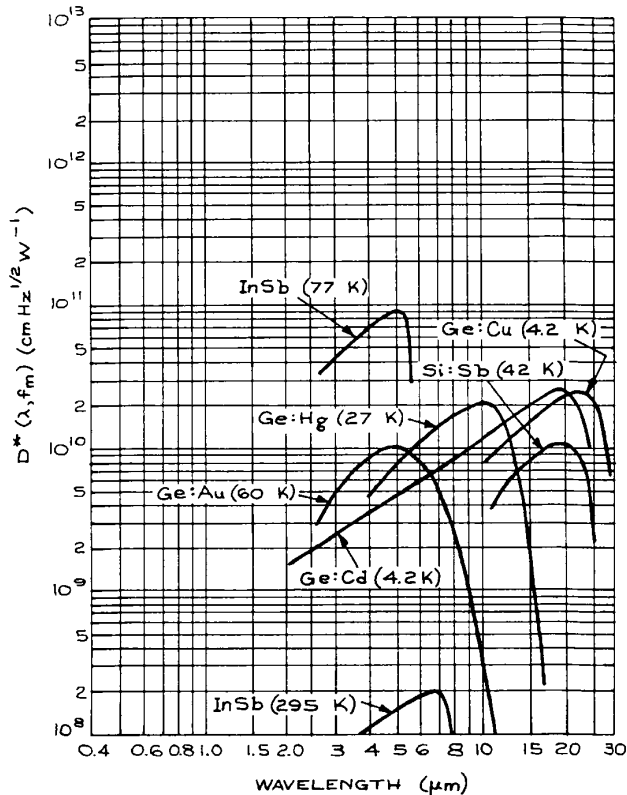
Silicon detectors are available from many suppliers.<sup>26</sup> Ge detectors are available from Jobin Yvon and Judson. PbS and PbSe detectors are available from Cal Sensors, Hamamatsu, Judson and New England Photoconductor, InSb detectors are available from Hamamatsu, InfraRed Associates, Judson, and Kolmar Technologies. MCT detectors are available from Boston Electronics, Hamamatsu, InfraRed Associates, Judson, and Kolmar Technologies.

Figure 7.13(b) photons with energy  $h\nu > E_A$  excite an electron to the impurity level, leaving a hole in the valence band and thereby giving rise to *p*-type *extrinsic photoconductivity*. Photons with energy  $h\nu > E_D$  will excite an electron into the conduction band, giving *n*-type extrinsic photoconductivity. For example, gold-doped germanium has an acceptor level 0.15 eV above the valence band and is an extrinsic *p*-type photoconductor, as is copper-doped germanium, which has an acceptor level 0.041 eV above the valence band. These used to be the two most commonly used extrinsic photoconductive detectors, responding out to about 9  $\mu\text{m}$  and 30  $\mu\text{m}$ , respectively, but they have been largely superseded by mercury cadmium telluride (MCT) detectors. Curves showing the variation of their  $D^*$  with wavelength are given in Figure 7.14. Table 7.3 lists the operating characteristics of some extrinsic photoconductive detectors. Detectors listed with operating temperatures below 28 K will in practice frequently be operated at 4 K, since liquid helium is a safe, available coolant, although its

use involves more complicated technology than the use of liquid nitrogen. Only long wavelength response gallium-doped germanium detectors are currently commercially available, although other extrinsic detectors are still custom-made in specialized research laboratories.

All photoconductive detectors, whether intrinsic or extrinsic, are operated in essentially the same way, although there are wide differences in packaging geometry. These differences arise from differing operating temperatures and speed-of-response considerations. A schematic diagram showing the main construction features of a liquid-nitrogen-cooled photoconductive or photovoltaic detector is given in Figure 7.15. Uncooled detectors can be of much simpler construction – for example, in a transistor or flat solar-cell package.

One feature of the cooled detector design shown in Figure 7.15 is worthy of note. The field of view of the detector is generally restricted by an aperture, which is kept at the temperature of the detection element. This shields the detector from ambient infrared radiation, which peaks at



**Figure 7.14**  $D^*(\lambda)$  as a function of wavelength for various photoconductive detectors. (Courtesy of Hughes Aircraft Company.)

9.6  $\mu\text{m}$ . For detection of low-level narrow-band infrared radiation the influence of background radiation can be further reduced by incorporating a cooled narrow-band filter in front of the detector element. The filter will only radiate beyond the cutoff wavelength of the detector, and thus restricts transmitted ambient radiation to a narrow band. Liquid-helium-cooled detectors generally have a double Dewar; an outer one for liquid nitrogen surrounds the inner, for liquid helium. Liquid-nitrogen-cooled detectors, such as  $\text{HgCdTe}$ , are recommended in preference to liquid-helium-cooled detectors, such as  $\text{Ge:Cu}$ , whenever there is a choice. Some miniature-package commercial detectors can be cooled with Joule–Thompson refrigeration units driven with compressed gas, which eliminates the need for externally supplied liquified-gas coolant. Thermo-

electrically cooled (Peltier effect) infrared detector packages are available from Cal Sensors, Hamamatsu, Marlow, and New England Photoconductor, for the operation of MCT, PbS, or PbSe detectors down to temperatures of 193 K. Further details of the operating principles behind these and other refrigeration techniques used with infrared detectors are given by Hudson.<sup>7</sup>

Figure 7.16 shows a basic biasing circuit commonly used for operating photoconductive detectors.  $R_d$  is the detector dark resistance. It is easy to see that the change in voltage,  $\delta V$ , that appears across the load resistor  $R_L$  for a small change  $\delta R$  in the resistance of the detector is:

$$\Delta V = \frac{-V_0 R_L \Delta R}{(R_d + R_L)^2} \quad (7.16)$$

This is at a maximum when  $R_L = R_d$ . Thus it is common practice to bias the detector with a load resistance equal to the detector's dark resistance. The bias voltage is selected to give a bias current through the detector that gives optimum detectivity. This bias current will generally be specified by the manufacturer. Figure 7.17 shows a simple op-amp circuit for use with a photoconductive detector.

To obtain a fast response from a photoconductive detector, great care must be taken to minimize the stray capacitance  $C_s$  in the input circuit to the preamplifier. Otherwise, the time constant of the detector will be limited by  $R_L C_s$ . The ultimate limits to the speed of an actual detector are set by its internal capacitance  $C_d$  and the majority-carrier lifetime. Many commercial detectors are manufactured so that the stray capacitance of the detector and its connection leads is very small. This is generally true of high-speed commercial detectors packaged in all-metal Dewars with integral coaxial bias connections. Metal Dewar packages are preferable to glass ones: although the latter are cheaper, they are very fragile.

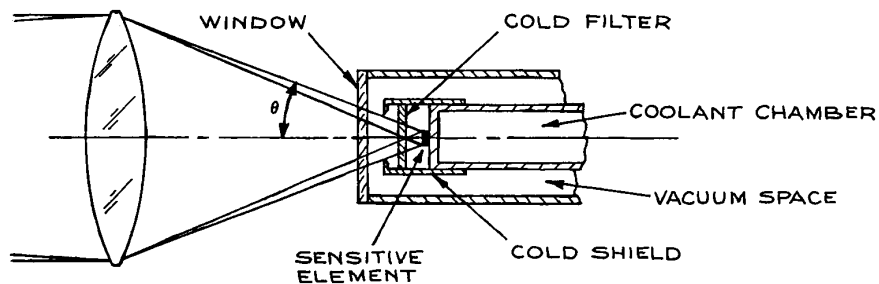
If a detector is not specifically packaged to minimize stray inductance, it is possible to reduce the stray capacitance of the detector package and improve its speed of response by the technique of “bootstrapping,” which is illustrated in Figure 7.18. All bias leads to the detector and the leads to the preamplifier are double-shielded. The preamplifier should ideally have a high input impedance, a low input capacitance, a wide bandwidth, and 50 ohm output impedance. Such preamplifiers can be constructed, or bought from detector suppliers or Perry

**Table 7.3 Extrinsic photoconductive detectors**

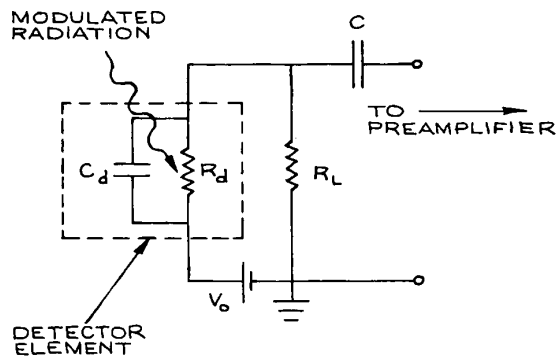
Semiconductor	Impurity	$T$ (K)	$\lambda$ ( $\mu\text{m}$ )	$D^*$ (or NEP)	$\tau$
Ge: Au <sup>a</sup>	p-type	77	8.3	$3 \times 10^9 - 10^{10}$	30 ns
Ge: Hg <sup>a</sup>	p-type	<28	14	$1 - 2 \times 10^{10}$	$4 > 0.3$ nm
Ge: Cd	p-type	<21	21	$2 - 3 \times 10^{10}$	10 ns
Ge: Cu <sup>a</sup>	p-type	<15	30	$1 - 3 \times 10^{10}$	>0.4 ns
Ge: Zn <sup>a,b</sup>	p-type	<12	38	$1 - 2 \times 10^{10}$	10 ns
Ge: Be	p-type	<3	115	$2 \times 10^{10}$	>1 $\mu\text{s}$
Ge: In	p-type	4	111	—	<1 $\mu\text{s}$
Ge: Ga	p-type	4	>200	$8 \times 10^{-13}$ (NEP)	<1 $\mu\text{s}$
Si: Ga	p-type	4	17	$10^9 - 10^{10}$	>1 $\mu\text{s}$
Si: As	n-type	<20	23	$1 - 3.5 \times 10^{10}$	0.1 $\mu\text{s}$

<sup>a</sup> Sometimes also contain silicon.

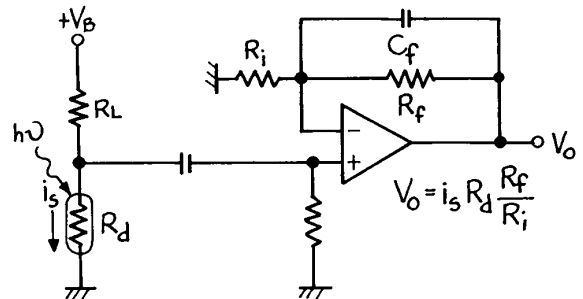
<sup>b</sup> Sometimes also contain antimony.



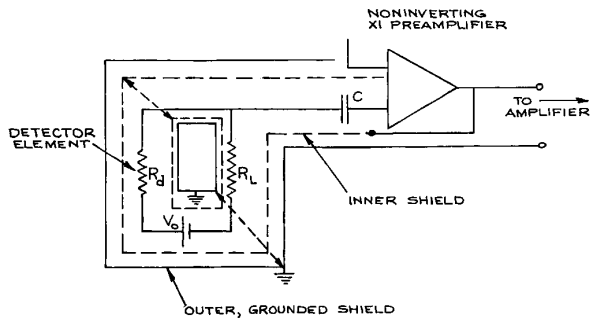
**Figure 7.15** Radiation-shielded liquid-nitrogen-cooled photoconductive or photovoltaic infrared detector assembly.



**Figure 7.16** Simple biasing circuit for operating a photoconductive detector with modulated radiation.



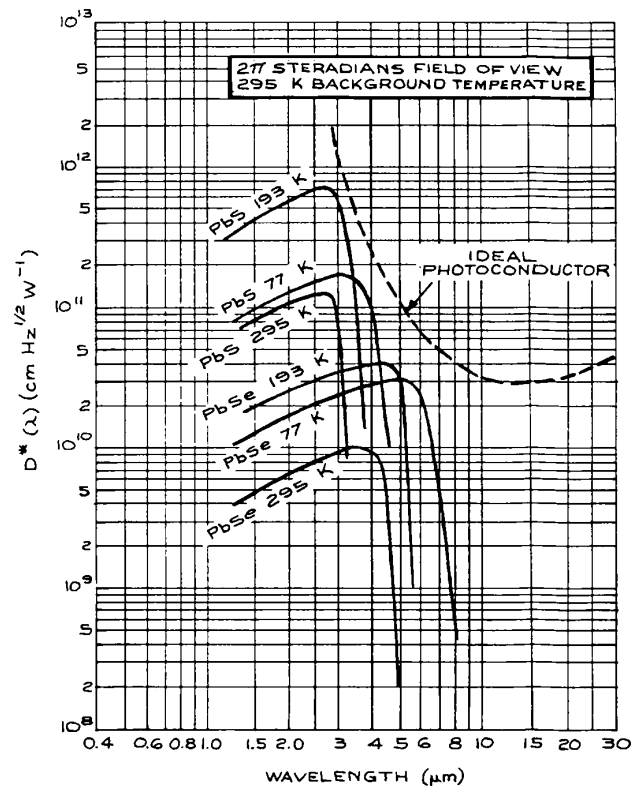
**Figure 7.17** Simple op-amp circuit for operating a photoconductive detector.



**Figure 7.18** Arrangement for “bootstrapping” an infrared detector to minimize effects of parasitic capacitance.

Amplifier. The inner shield is not grounded, but is connected directly to the low-impedance output of the unit-gain preamplifier. Thus the inner shield is “bootstrapped” to the signal voltage and stray capacitance is effectively eliminated. By this means the speed of response of a detector can be improved substantially, say from 1  $\mu$ s down to 50 ns. An alternative way to improve the speed of response, at the expense of signal (but not detectivity), is to reduce the value of the load resistor.

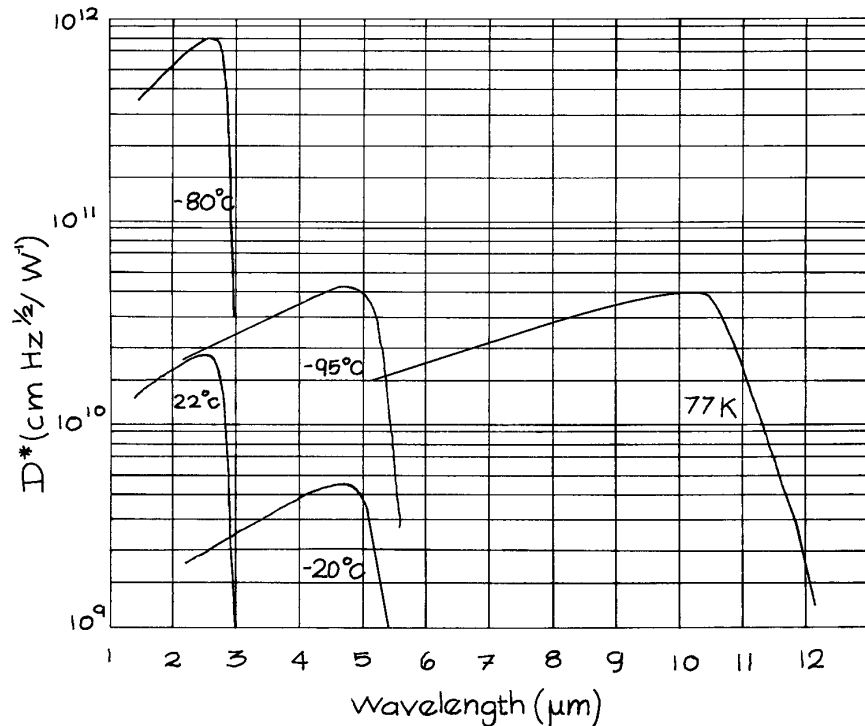
A few photoconductive detectors are worthy of brief extra comment. Silicon and germanium are much more commonly used for photodiodes, frequently in an avalanche mode. These devices are discussed in Section 7.6. Lead sulfide detectors have high impedance, 0.5 to 100 M $\Omega$ , and slow response, but are the most sensitive detectors in the spectral region between 2 and 3  $\mu$ m and can be used uncooled.  $D^*(\lambda)$  curves for these detectors are shown in Figure 7.19. They are available from Cal Sensors, Hamamatsu, Judson, and New England Photoconductor. Lead selenide is sensitive to longer wavelengths than lead sulfide, but InAs operated in a photovoltaic mode is to be preferred in this spectral region (3–4  $\mu$ m). InSb operated in a photovoltaic mode is probably the detector of choice between 4 and 5.3  $\mu$ m. Beyond about 5  $\mu$ m, HgCdTe (MCT) is the detector of choice and is to be preferred to detectors such as Ge:Hg and Ge:Cu, which must be operated below liquid-nitrogen temperature. MCT detectors can be operated in either a photoconductive or photovoltaic mode. HgCdTe comes in different stoichiometries, generally represented as  $\text{Hg}_x\text{Cd}_{1-x}\text{Te}$ . Different stoichiometries have different spectral responses. By varying its



**Figure 7.19**  $D^*$  as a function of wavelength for various lead-salt photoconductive detectors. (Courtesy of Hughes Aircraft Company.)

composition, HgCdTe exceeds in performance all other 77 K detectors between 6 and 12  $\mu$ m. Sometimes these detectors also contain zinc and are designated HgCdZnTe. Thermoelectrically cooled MCT detectors have higher detectivity at shorter wavelengths than MCT detectors operated at 77 K, as shown in Figure 7.20. HgCdTe detectors are available from Boston Electronics, Hamamatsu, InfraRed Associates, Judson, and Kolmar Technologies. For further details of the above detectors, the reader should consult references 5, 12, 19, and 26. Suppliers of other detectors are listed in reference 25.

The use of extrinsic photoconductivity for the detection of far infrared radiation requires the introduction of appropriate doping material into a semiconductor in order to generate an acceptor or donor impurity level extremely close to the valence or conduction bands, respectively.



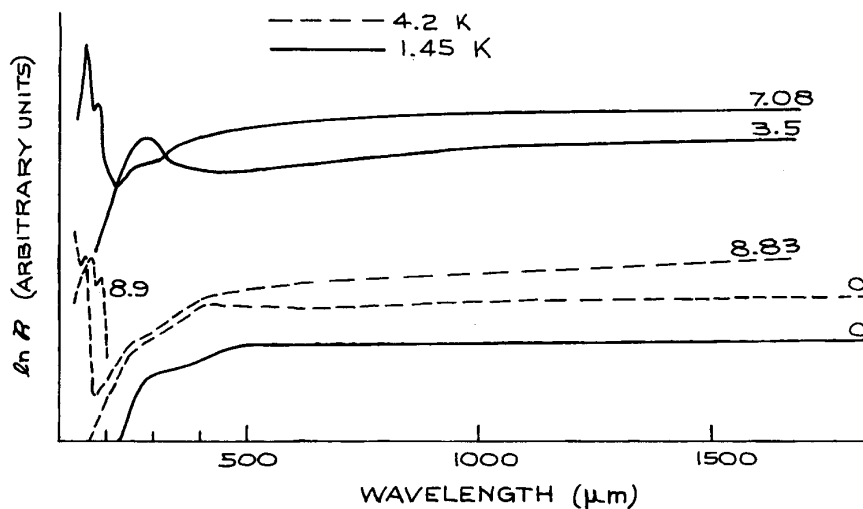
**Figure 7.20**  $D^*$  as a function of wavelength for room temperature, thermoelectrically cooled, and liquid nitrogen cooled (77 K) mercury cadmium telluride photovoltaic detectors manufactured by Judson.

Well-characterized impurity levels have been generated in germanium in this way using gallium,<sup>27</sup> indium,<sup>28</sup> boron,<sup>29</sup> or beryllium doping. The long-wavelength sensitivity limit is restricted to about 120  $\mu\text{m}$ , but can be extended to 200  $\mu\text{m}$  in Ge:Ga by stressing the crystal. Ge:Ga detectors are available from QMC Instruments. Longer-wavelength-sensitive, extrinsic photoconductivity can be observed in appropriately doped InSb in a magnetic field. Bulk InSb can be used more efficiently for infrared detection in a different photoconductive mode entirely.<sup>30</sup> Even at the low temperature at which far-infrared photoconductive detectors operate ( $\leq 4$  K), there are carriers in the conduction band. These free electrons can absorb far infrared radiation efficiently and move into higher-energy states within the conduction band. This change in energy results in a change of mobility of these free electrons, which can be detected as a change in conductivity, using a circuit such as the one shown previously in Figure 7.17. Some typical

wavelength response curves are shown in Figure 7.21. These *hot-carrier-effect* photoconductive detectors or hot-electron bolometers can be used successfully over a wavelength range extending from 50 to  $10^4$   $\mu\text{m}$ ; they have detectivities up to about  $1 \times 10^{11}$   $\text{cm Hz}^{1/2} \text{W}^{-1}$  and response times down to 1  $\mu\text{s}$ . They are frequently operated in a large magnetic field (several hundred kA/m or more). These detectors are available from QMC Instruments. Magnetically tuned detectors are also available from QMC Instruments. These detectors are tuned in wavelength response by operating them in a magnetic field.<sup>31,32</sup>

## 7.6 PHOTOVOLTAIC DETECTORS (PHOTODIODES)

In a *photovoltaic detector* photoexcitation of electron-hole pairs occurs near a junction, when radiation of energy greater than the band gap is incident on the junction region.



**Figure 7.21** Responsivity as a function of wavelength of a hot-carrier InSb photoconductive detector operated at two different temperatures and various magnetic field strengths. The number on each curve is the magnetic field strength in units of  $10^5$  A/m.

Extrinsic photoexcitation is rarely used in photovoltaic photodetectors. The internal energy barrier of the junction causes the electron and hole to separate, creating a potential difference across the junction. This effect is illustrated for a  $p$ - $n$  junction in Figure 7.22.

Other types of structure are also used, such as  $p$ - $i$ - $n$ , Schottky-barrier (a metal deposited onto a semiconductor surface), and heterojunction (a junction between two different semiconductors). The  $p$ - $n$  and  $p$ - $i$ - $n$  structures are the most commonly used. All these devices are commonly called *photodiodes*. The characteristics of some important photodiodes are listed in Table 7.4. Important photodiodes include silicon for detection of radiation between 0.1 and 1.1  $\mu\text{m}$ , germanium for use between 0.4 and 1.8  $\mu\text{m}$  and indium (gallium) arsenide between 1 and 3.8  $\mu\text{m}$ . Responsivity variations with wavelength for photodiodes that are most important for near infrared (communications) applications are shown in Figure 7.23. Other important photodiodes include indium antimonide between 1 and 7  $\mu\text{m}$ , and mercury-cadmium telluride between 1 and 20  $\mu\text{m}$ . Some typical curves of  $D^*(\lambda)$  are shown in Figure 7.24. These spectral response regions are not all necessarily covered by a detector operating for example, InSb

responds to 7  $\mu\text{m}$  at 300 K, but to wavelengths no longer than 5.6  $\mu\text{m}$  at 77 K. The wavelength response of HgCdTe detectors depends also on the stoichiometric composition of the crystal.<sup>32</sup> The maximum wavelength of response of  $\text{Hg}_x\text{Cd}_{1-x}\text{Te}$  can be estimated from its bandgap  $E_g$  and the operating temperature  $T$ .  $E_g$  is approximated by the formula:<sup>33</sup>

$$E_g(\text{eV}) = -0.302 + 1.93x - 0.81x^2 + 0.832x^3 + 5.35(1 - 2x)10^{-4}T \quad (7.17)$$

All these photodiodes have very high quantum efficiency, defined in this case as the ratio of photons absorbed to mobile electron-hole pairs produced in the junction region. Values in excess of 90% have been observed in the case of silicon.

When a photodiode detector is illuminated with radiation of energy greater than the band gap, it will generate a voltage and can be operated in the very simple circuit illustrated in Figure 7.25(a). It is much better, however, to generate a photodiode detector in a reverse-biased mode, as shown in Figure 7.25(b), where positive voltage is applied to the  $n$ -type side of the junction and negative to the  $p$ -type. In this case, the observed photosignal is seen as

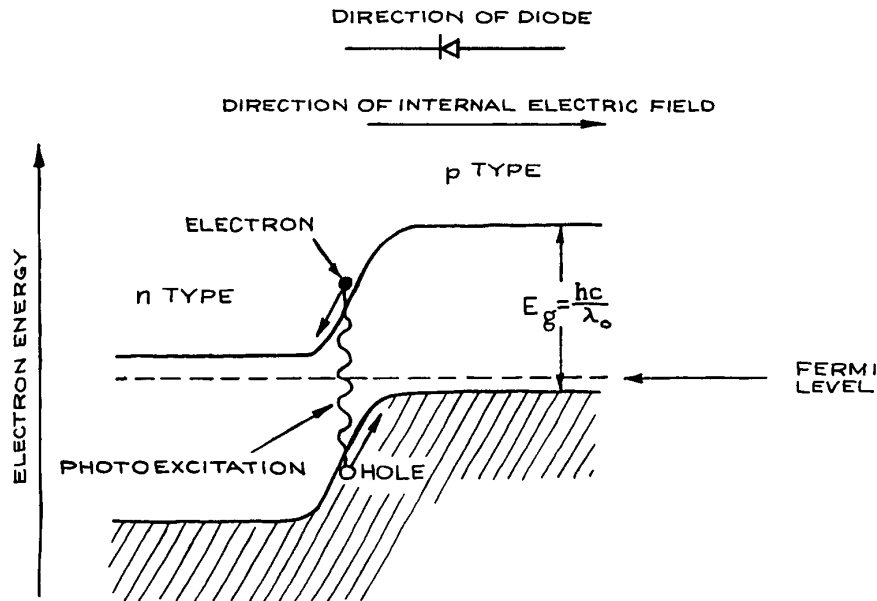


Figure 7.22 Photoexcitation at a  $p$ - $n$  junction.

Table 7.4 Photovoltaic detectors (photodiodes)

Semiconductor	$T$ (K)	Wavelength Range ( $\mu\text{m}$ )	$D^*$ (max)	$\tau$	Suppliers <sup>a</sup>
Si	300	0.2–1.1	$\leq 2 \times 10^{13}$	—	— <sup>b</sup>
GaAlAs	300	0.8–0.93	$10^{-14}$ (NEP)	1 $\mu\text{s}$	(1)
Ge	300	0.4–1.8	$10^{11}$	0.3 ns	(2)
InGaAs	300	1–2.5	$8^{11}$	4 ns	(2)
InAs	300	1–3.8	$< 4 \times 10^9$	5 ns–1 $\mu\text{s}$	(3)
InAs	77	1–3.2	$4 \times 10^{11}$	0.7 $\mu\text{s}$	(3)
InSb	300	1–7	$1.5 \times 10^8$	0.1 $\mu\text{s}$	(4)
InSb	77	1–5.6	$< 2 \times 10^{11}$	>25 ns	(5)
HgCdTe	77	1–25 <sup>c</sup>	$10^9$ – $10^{11}$	>1.6 ns <sup>c</sup>	(6)

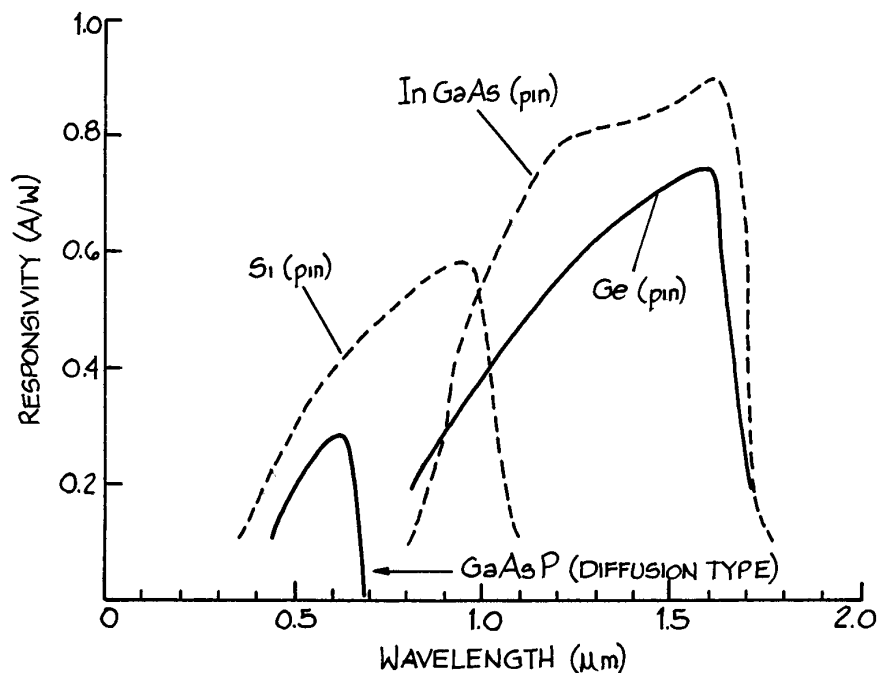
<sup>a</sup> (1) Very sensitive, narrow spectral response detectors. Available from Advanced Photonix, Anadigics, and Opto-Diode;

(2) Available from Electro-Optical Systems, Fermionics, Hamamatsu, Judson, and Perkin-Elmer Optoelectronics (formerly EG&G Optoelectronics) Optoelectronics, and Newport; (3) Electro-Optical Systems, Judson, and Melles Griot and Judson;

(4) Electro-Optical Systems, Hamamatsu, InfraRed Associates, and Kolmar Technologies; (5) Electro-Optical Systems, Kolmar Technologies; Technologies; (6) Boston Electronics, Infrared Associates, Judson, Hamamatsu, Kolmar Technologies, Newport/Oriel.

<sup>b</sup> A very wide range of silicon photoconductive and photovoltaic detectors are available with NEP figures down to  $10^{-16}$  W Hz<sup>1/2</sup> and time constants down to 6 ps. Reference 25 contains a list of suppliers.

<sup>c</sup> Range varies from one supplier to another.



**Figure 7.23** Responsivity of important near-infrared photodiodes.

a change in current through the load resistor. The difference between the two modes of operation can be easily seen from Figure 7.26, which shows the  $I$ - $V$  characteristic of a photodiode in the dark and in the presence of increasing levels of illumination. At a given level of illumination the photodiode can generate either an open-circuit voltage  $V_{OC}$  or a short-circuit current  $I_{SC}$ . A photodiode responds much more linearly to changes in light intensity and has greater detectivity when operated in the reverse-biased mode. Ideal operation is obtained when the diode is operated in the current mode with an operational amplifier that effectively holds the photodiode voltage at zero—its optimum bias point. Two simple practical circuits which can be used to operate a photodiode in this way are shown in Figures 7.27 and 7.28.

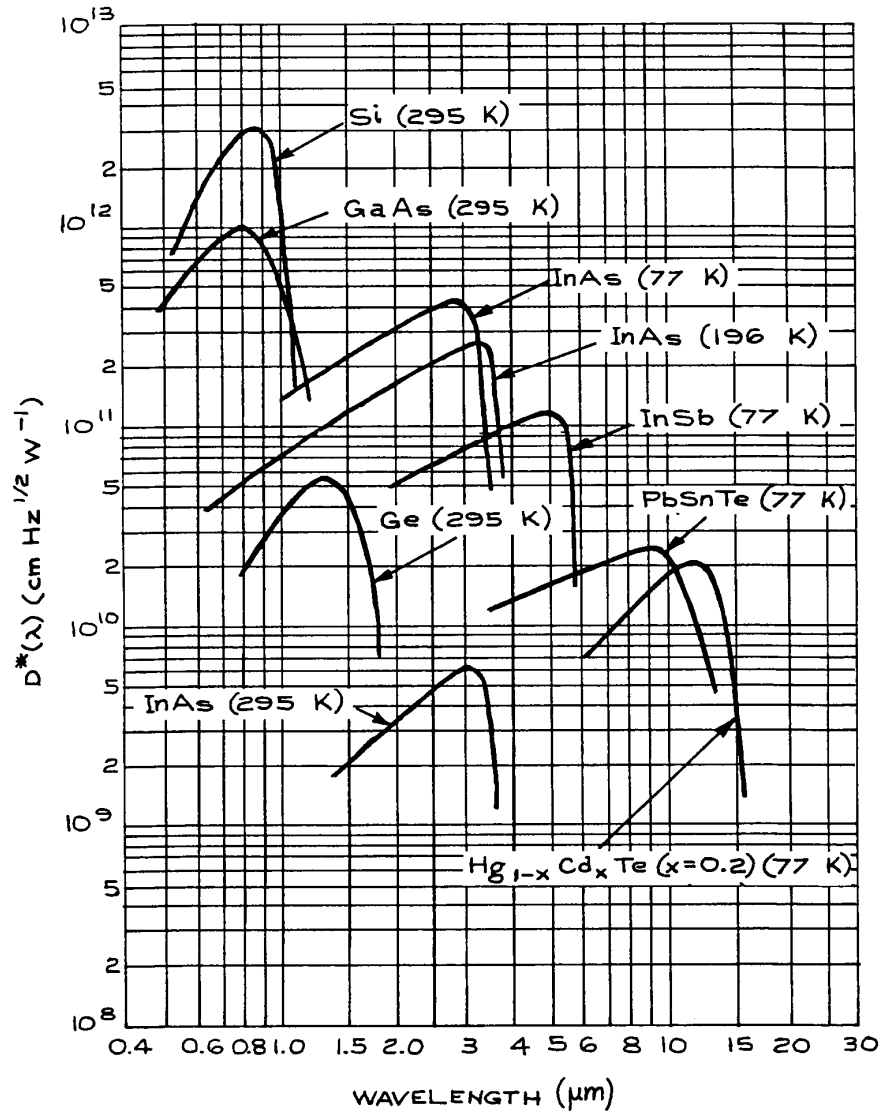
In Figure 7.28, the bias voltage  $V_B$  is not necessary, but for many photodiodes will improve the speed of response, albeit at the expense of an increase in noise. Integrated packages incorporating a photodiode and operational amplifier are available from Centronic (especially UV detectors), Hamamatsu, New Focus, Newport Corporation,

Opto Diode, OSI Optoelectronics, Perkin-Elmer Optoelectronics, and Thorlabs. The  $p$ - $i$ - $n$  structure is most commonly used in these devices because its performance, in terms of quantum efficiency (number of useful carriers generated per photon absorbed) and frequency response, can be readily optimized. These devices have very low noise and fast response. In practice, the limiting sensitivity that can be obtained with them will be determined by the noise of the associated amplifier circuitry.

### 7.6.1 Avalanche Photodiodes

*Avalanche photodiodes* (APD's) are available commercially for operation from about 250 nm to 1800 nm. Silicon APDs operate between 300 to 1100 nm, germanium APDs between 800 and 1600 nm, and InGaAs APD's from 900 to 1700 nm. Although they are more costly than Si or Ge APDs, InGaAs APDs generally have lower noise and have a faster response time for a given active area. If the reverse-bias voltage on a photodiode is increased, photoinduced charge carriers can acquire sufficient energy transverseing

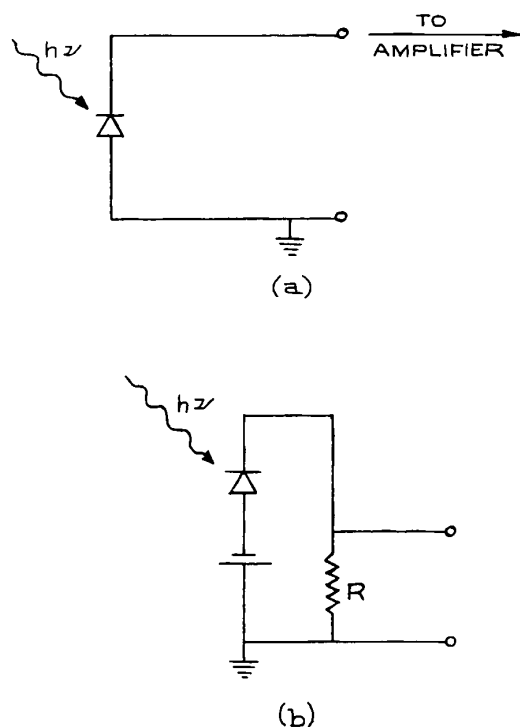




**Figure 7.24**  $D^*$  as a function of wavelength for various photovoltaic detectors. (Courtesy of Hughes Aircraft Company.)

the junction region to produce additional electron-hole pairs. Such a photodiode exhibits current gain and is called an avalanche photodiode (APD). It is, in some respects, the solid-state analog of the photomultiplier. Avalanche photodiodes are noisier than  $p-i-n$  photodiodes, but because they have internal gain, the practical sensitivity that can

be achieved with them is greater. For detection of weak light signals, particularly short pulses, the signal-to-noise ratio will not be dominated by the associated electronics. A good discussion of the design, fabrication, and operating characteristics of various photodiode detectors has been given by Gowar.<sup>34</sup> Avalanche photodiodes are available



**Figure 7.25** Photovoltaic detector operated in: (a) open-circuit mode; (b) reverse-biased mode.

from Advanced Photonix, Laser Components DG, OSI Optoelectronics, Pacific Silicon Sensor, Perkin-Elmer Optoelectronics, and Sensors Unlimited.

### 7.6.2 Geiger Mode Avalanche Photodiodes

If an APD with a very low dark current is operated in the *Geiger mode*,<sup>34–37</sup> where the reverse bias voltage  $V_R$  is set greater than the diode breakdown voltage  $V_{BR}$ , then a single photon can trigger a substantial circuit pulse. In other words, the current or voltage pulse they produce from the detection of a single photon is above their noise levels. The internal gain of the APD can reach  $10^5$ – $10^6$ . To prevent the diode from permanent damage it must be operated with a “quenching” current, which reduces the bias voltage to

stop the avalanche process. Otherwise the diode will conduct a large current when it is biased above the breakdown voltage. It is desirable to reduce the dark current by cooling the APD. APDs operated in this way are true “single photon” detectors, and can replace PMTs in many photon-counting applications.

Photodiodes using silicon, InGaAs/InP, InGaAs/Si and HgCdTe have all been successfully used in Geiger mode. Complete Geiger-mode APD modules are available from Perkin-Elmer Optoelectronics. Dark counts below 1 count per second (CPS) have been reported for Perkin-Elmer SLiK silicon APDs cooled to  $-20^\circ\text{C}$ , and dark counts below 250 cps are typically achievable with a cooled PerkinElmer 0.5 mm C30902S APD.

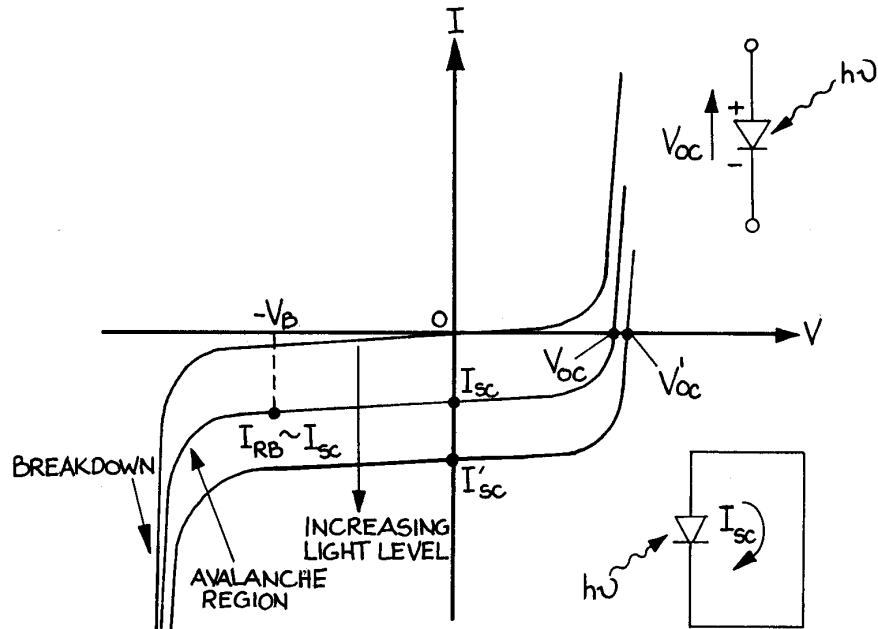
An example of an operating arrangement for an InGaAs cooled Geiger mode APS is shown in Figure 7.29. For operation in Geiger mode, the reverse bias is held just below the breakdown voltage (20–40 V). The bias voltage is briefly gated above the breakdown value synchronously with the excitation process that is expected to generate single photons. Avalanche does not generally occur unless an actual photon is detected. The avalanche process generates a significant current pulse with peak current 1 mA. When the gating voltage applied to the APD is removed, the avalanche is quenched because the APD is now below breakdown, and the current pulse terminates.

## 7.7 DETECTOR ARRAYS

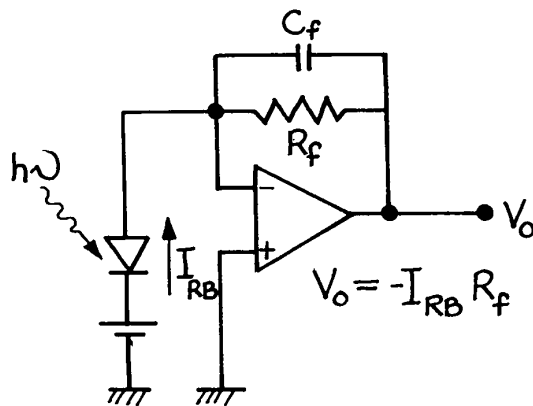
Arrays of individual photodetector elements, based on photodiode, photoconductive, pyroelectric, microbolometers, CMOS, and CCD elements, are widely available for use in both the visible and the infrared. Such arrays are useful in a variety of applications, such as spectral analysis, image analysis, and position sensing.

### 7.7.1 Reticons

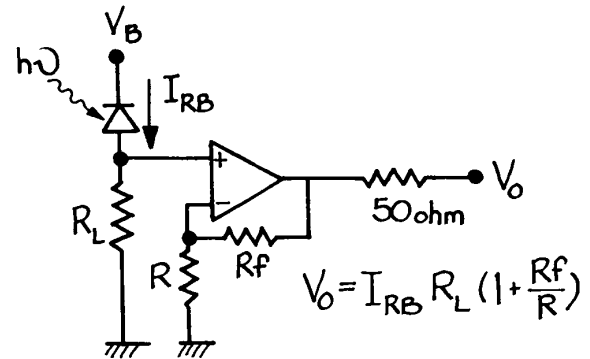
Multiple element linear arrays, often called “reticons,” are widely used in spectrographs. Light from the diffraction grating in such an instrument falls on a such an array instead of on an exit slit. Arrays based on silicon, InGaAs, HgCdTe, PbS, PbSe are available from various



**Figure 7.26** I-V characteristics of a photodiode in the dark and with increasing levels of illumination.



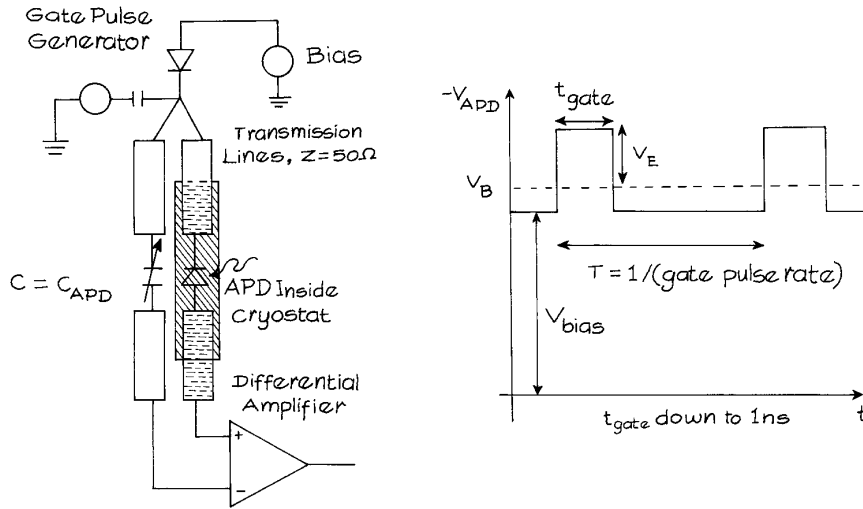
**Figure 7.27** Simple op-amp circuit for reverse-bias operation of a photodiode. If the d.c. bias is not included, then the detector is held at zero voltage and is operated in a "current-mode."



**Figure 7.28** Op-amp circuit for reverse-bias operation of a photodiode.

manufacturers, including Advanced Photonix (Si), Cal-Sensors (PbS, PbSe), Fermionics (HgCdTe, InGaAs), Hamamatsu (Si, InGaAs), Kodak (Si), Perkin-Elmer Optoelectronics (Si), SensArray Infrared (PbS and PbSe), and

Sensors Unlimited (Goodrich)(InGaAs). Silicon-based reticons are linear CCD (charge coupled device) arrays. The number of pixels in a linear detector can be as large as 10 2000, although 1024 or 2048 element arrays are most common. Linear arrays can also be used for position sensing.



**Figure 7.29** Arrangement for operating a Geiger-mode APD.<sup>39</sup>

## 7.7.2 Quadrant Detectors

Quadrant detectors (quads) are particularly useful in laser-beam tracking and similar operations. These detectors have four contiguous active elements with very little inactive area between them, as shown in Figure 7.30(a). If a light beam is centered on the detector, all four elements provide the same signal. By the use of a circuit, such as the one shown in Figure 7.30(b), simultaneous display of the  $X$  and  $Y$  displacements of the beam from the center of the array can be obtained. If the signals from the four quadrants are  $A$ ,  $B$ ,  $C$ , and  $D$  in the circuit shown in Figure 7.30(b), then the relative position of the center of a circular illumination spot can be calculated as:

$$x = \frac{(B + D) - (A + C)}{A + B + C + D} \quad (7.18)$$

$$y = \frac{(A + B) - (C + D)}{A + B + C + D} \quad (7.19)$$

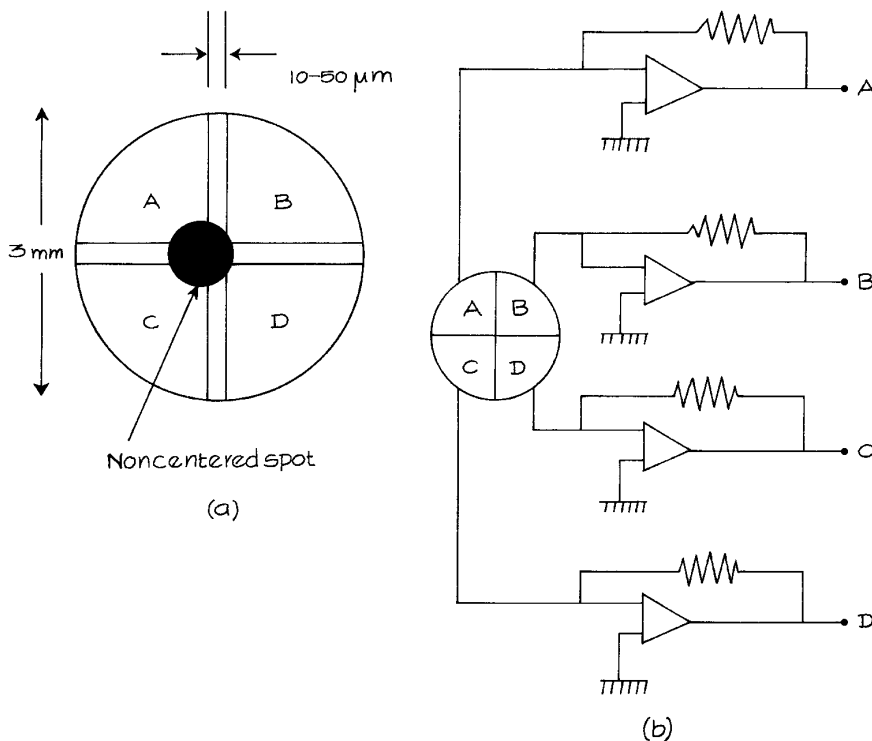
The displacements  $x$  and  $y$  are fractions, which would have to be calibrated by moving the beam centroid by a known distance in a particular application.

Silicon quadrant detectors are available from Advanced Photonix, Electro Optical Components, OSI Optoelec-

tronics, and Silicon Sensors. Ge and InAs quad detectors are available from Judson. Infrared HgCdTe quad detectors are available from Judson and Raytheon Vision Systems (formerly Santa Barbara Research Center). In some simple situations, small one-dimensional deflections of a light spot can be measured by focusing the light spot on the edge of a razor blade, and measuring the amount of light reaching a photodetector placed beyond the sharp edge, as shown in Figure 7.31.

## 7.7.3 Lateral Effect Photodetectors

Lateral effect photodetectors (LEP) also provide output signals that allow determination of the position of the centroid of an incident light spot on the surface of the LEP. They are therefore useful in experiments in which a laser beam is deflected, and the deflection needs to be detected, or they can be used in general laser-beam tracking applications. The LEP, in its simplest form, consists of a surface resistive layer on top of a  $p$ - $n$  junction substrate, and has two lateral electrodes. When a light ray is received, the resulting photo-induced currents diffuse separately along the junction faces and recombine at a distance from their origin of separation. The analytical solution of the diffusion equation, representing the potential on the surface of the LEP on the two sides varies approximately



**Figure 7.30** (a) Schematic diagram showing the construction of a quadrant detector; (b) a circuit suitable for providing  $X$  and  $Y$  beam-deflection readouts in a beam-centering or beam-tracking application.

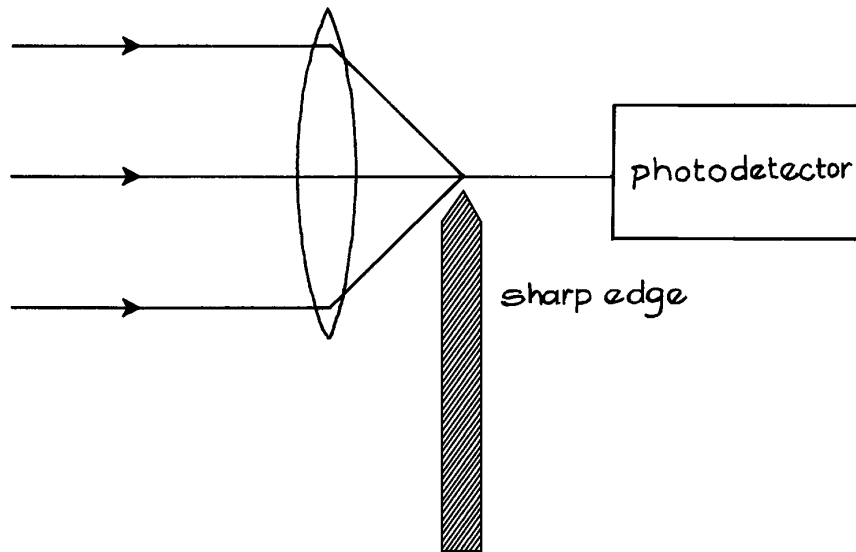
linearly with respect to the incident position of the light and the received power. The dependence on the received power can be eliminated by normalizing with respect to the total received power.

Normally, the fabrication of a two-dimensional LEP can be divided into: (1) Wallmark dual-axis, (2) duo-lateral, and (3) tetra-lateral structures.<sup>40</sup> Most commercial LEPs use the duo-lateral type, shown schematically in Figure 7.32, since it provides the best linearity. Most off-the-shelf LEPs have a 0.1% linearity error near the center region and the error increases to 1% when approaching the edges. These devices can detect the movement of a small light spot with micrometer or smaller precision. Lateral effect photodetectors with preamplifiers are available from Electro-Optical Systems (EOS), On-Trek, and Pacific Sensor. LEP bundles, including post-amplification circuits, are available from Hamamatsu, Newfocus, and Thorlabs. In

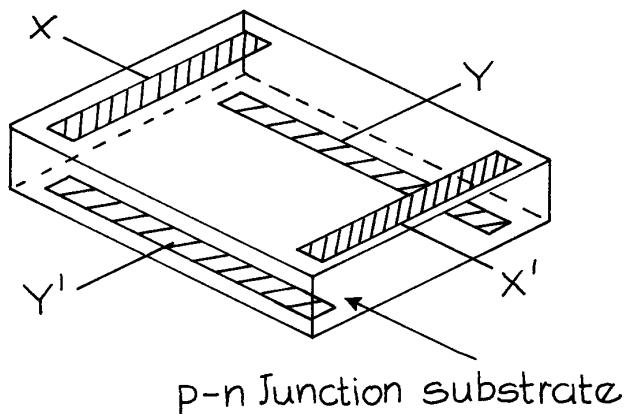
contrast to quadrant detectors, LEPs are superior for their large linear region and small internal resistance. The latter results in a short response time, but a larger noise equivalent power. This implies that quadrant detectors are better in applications that involve weak received power, such as feedback sensors in remote sensing applications. LEPs, on the other hand, are advantageous as local kinematic sensors.

### 7.7.4 Imaging Arrays

Two-dimensional arrays of photodetector generally show up in imaging systems. Complementary Metal-Oxide-Semiconductor (CMOS) or CCD arrays based on silicon are common in visible and near infrared digital camera systems. These array detectors can be thought of as equivalent to an array of “photon-buckets.” The number of



**Figure 7.31** Schematic arrangement for measuring the deflection of a focused light beam by using a knife edge.



**Figure 7.32** Schematic construction of duo-lateral lateral effect photodetector (LEP).

photons detected at each array element produces a corresponding number of electron-hole pairs. The electrons at each pixel are trapped in a potential well, equivalent to a small capacitor, and then read out sequentially by moving these electrons from well to well.<sup>41</sup> InGaAs arrays provide performance between  $0.9\ \mu\text{m}$  and  $2.2\ \mu\text{m}$ . Infra-

red cameras for operation beyond  $10\ \mu\text{m}$  use arrays of microbolometers (see later). A comprehensive list of suppliers of array detectors in different spectral regions is available.<sup>25</sup>

In consumer digital cameras, CCD and CMOS arrays are the standard image detectors. Such cameras generally contain all the electronics to read out the signals from the array elements and store them in digital format, including color and grayscale. The most common image format from such cameras is .jpeg. The number of available pixels in such cameras continues to increase, and formats with more than 16 million pixels are already available. Typical formats are:  $1216 \times 912$  (1 megapixel),  $1600 \times 1200$  (2 megapixel),  $2240 \times 1680$  (4 megapixel), and  $4064 \times 2704$  (11.1 megapixel). The actual number of pixels in these arrays is larger than the number specified by the resolution, because some dark array elements are used for reading out the image. It is worth noting that the effective resolution of a good-quality film camera is at least 20 million pixels, determined by the “grain” size in the film. The image stored by such cameras is a single “frame,” captured as the average illumination of each pixel over the exposure time chosen. Color is generally captured by having a filter

array placed in front of the pixels. The filter array usually has alternating rows of red/green and blue/green filters, a so-called “Bayer pattern.” The image is processed using the signals from adjacent pixels to provide a best estimate of the color at each pixel, a process called *demosaicing*.<sup>42</sup> Advanced digital cameras use Foveon X3 technology (Foveon, Inc., 2820 San Tomas Expressway, Santa Clara, CA 95051), in which the active pixel elements are embedded in a silicon matrix at different depths. Because blue, green, and red light penetrates different distances into silicon, the camera acts as though it has three sequential blue, green, and red imaging arrays. The quantum efficiency, sensitivity, maximum frame rate, saturation light level, and dynamic range of CCD and CMOS imagers vary and manufacturers’ data sheets should be consulted. Of the two, CCD imagers are generally more sensitive, and have lower noise than CMOS imagers; CMOS imagers on the other hand use less power and are less expensive. As CCD imagers cannot be gated electronically, if short exposures are required, then a mechanical shutter is needed. Infrared imaging arrays with response from 0.5–14  $\mu\text{m}$ , based on InSb and HgCdTe, are available from Raytheon Vision Systems; HgCdTe imaging arrays are available from Sofadit.

The experimentalist who wishes to work with CCD or CMOS imaging arrays that are not already integrated into a complete camera system should consult the suppliers of such devices, such as Kodak or Sony. This is a specialized endeavor. If it is desired to have complete control of image capture as a digital file directly into a computer, then a digital camera with compatibility with a framegrabber card is needed. Framegrabber cards usually utilize a special interface standard called Cameralink, which is used by all major scientific camera manufacturers. It allows for data rates in excess of 2 Gb/s. Scientific cameras differ from consumer ones in that they provide uncompressed frames at high framerates, such as 210 fps in the case of the Imperx IPX-210. These uncompressed frames allow for advanced image analysis in Labview, Matlab, and other image-processing applications.<sup>43</sup> The primary manufacturers of framegrabbers are Imperx and National Instruments, while cameras can be acquired from Imperx, Dalsa, Pulnix, and Flir. It is worth noting that some cameras have now come on the market that use a gigabit Ethernet standard to transmit frames, the Imperx IPX-2M30H-GC (1920  $\times$  1080 30 fps color) being one example.

### 7.7.5 Image Intensifiers

In some applications, for example in astronomy or night vision, it is desirable to obtain an image that can be viewed by the human eye, or with a digital camera, when the light levels available are extremely low. Low-light-level imaging can be carried out with an *image intensifier*. Image intensifiers work by using weak visible or near infrared light to liberate photoelectrons from a GaAs photocathode, which has a high quantum efficiency to about 920 nm. The photoelectrons are accelerated onto a microchannel plate, where secondary electron amplification increases their number. The amplified electron stream then falls onto a phosphor screen where the image is rendered visible. Figure 7.33 shows the principle of operation. As image intensifiers have evolved, photocathode materials and the electron multiplication stages have improved. The earliest image intensifiers were called *generation 0*, G0, which used S1 photocathodes, G1 tubes used an improved S-20 photocathode. G2 and G3 tubes used a MCP for electron multiplication, typical amplification gains can be as high as  $10^7$ . G2 tubes use an S-25 photocathode with typical peak sensitivity of 50  $\mu\text{A/W}$ . G3 tubes use GaAs or GaAsP photocathodes with typical sensitivity of 200  $\mu\text{A/W}$ . Nowadays only G2 and G3 tubes should be considered for use. Manufacturers of image-intensifier tubes include Electrophysics Corporation, Hamamatsu, ITT Night Vision, Photek, Photonis, and Thales.

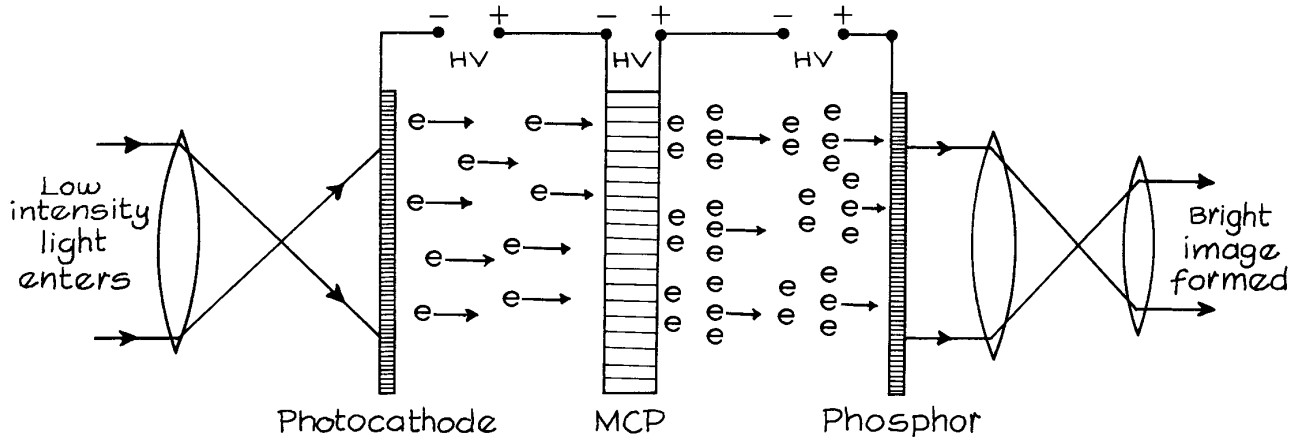
## 7.8 SIGNAL-TO-NOISE RATIO CALCULATIONS

### 7.8.1 Photomultipliers

In many applications using photomultiplier tubes, the anode pulses are integrated to give a fluctuating analog current. The shot noise originating at the photocathode is, from Equation (7.1):

$$\langle i_N^2 \rangle_c = 2e(\langle i_c \rangle + \langle i_d \rangle)\Delta f \quad (7.20)$$

where  $\langle i_c \rangle$  is the average photocathode current produced by a light source and  $\langle i_d \rangle$  is the average photocathode dark current. This noise is multiplied by the amplification



**Figure 7.33** Operating arrangement of a microchannel plate image intensifier.

of the electron number by interaction with the  $N$  dynodes of the tube. If each dynode has a secondary emission multiplication efficiency  $\delta$  the overall gain of the tube is  $G$ :

$$G = \delta^N \quad (7.21)$$

Because of statistical fluctuations in the secondary emission process and, in addition, because electrons can originate thermionically from the dynodes, the noise at the anode is further increased by a noise factor  $F$ . The overall noise appearing at the anode is:

$$\langle i_N^2 \rangle_A = 2eG^2F(\langle i_c \rangle + \langle i_d \rangle)\Delta f \quad (7.22)$$

For  $\delta = 4$ , and a 14 stage tube,  $G = 2.6 \times 10^8$ , and typically  $F \simeq \delta(\delta - 1) = 1.08$ .

A photomultiplier is a current source: to convert this current to a detected voltage the amplified photoelectron current passes through an anode resistor of value  $R$ . This anode resistor may be part of the photomultiplier circuit, or may be provided in whole or in part by the input impedance of a following amplifier stage. The Johnson noise from the resistor is:

$$\langle i_N^2 \rangle_R = \frac{4kT\Delta f}{R} \quad (7.23)$$

To optimize detection of a signal, it is common practice to amplitude modulate the signal at some angular frequency  $\omega_m$  so as to permit synchronous detection at this

frequency (see Chapter 6). The optical power reaching the photocathode can be represented in this case as:

$$P = P_0(1 + m \sin \omega_m t) \quad (7.24)$$

where  $m$  is a modulation parameter. The average photocathode current is:

$$\langle i_c \rangle = \frac{e\eta P_0}{h\nu} \quad (7.25)$$

and:

$$i_c(t) = \langle i_c \rangle(1 + m \sin \omega_m t) \quad (7.26)$$

At the anode, the time-varying part of this current is:

$$i_s(t) = \langle i_c \rangle Gm \sin \omega_m t. \quad (7.27)$$

The signal-to-noise ratio (S/N) at the input to the following electronics is therefore:

$$\frac{\langle i_s^2 \rangle}{\langle i_N^2 \rangle_A + \langle i_N^2 \rangle_R} = \frac{\langle i_c \rangle^2 G^2 m^2 / 2}{2eG^2F(\langle i_c \rangle + \langle i_d \rangle)\Delta f + 4kT\Delta f/R} \quad (7.28)$$

The noise from the photomultiplier tube is usually sufficiently large that the Johnson noise can be neglected. If it is assumed that  $\langle i_d \rangle \gg \langle i_c \rangle$ , then for  $m = 1$  and  $S/N = 1$  we have, from Equations (7.25) and (7.28):

$$(P_0)_{\min} = \frac{2h\nu}{\eta} \sqrt{\frac{F\langle i_d \rangle \Delta f}{e}} \quad (7.29)$$



*Example:* Typical values for a good photomultiplier will be  $\eta = 0.2$ ,  $F \simeq 1$ ,  $\langle i_d \rangle \simeq 10^{-15}$  A. For a 530 nm source, Equation (7.29) gives  $(P_0)_{\min} = 3 \times 10^{-16}$  W.

## 7.8.2 Direct Detection with $p-i-n$ Photodiodes

The shot noise from a  $p-i-n$  photodiode is:

$$\langle i_N^2 \rangle_{\text{SN}} = 2e(\langle i_s \rangle + i_d)\Delta f \quad (7.30)$$

where  $\langle i_s \rangle$  is the average signal current, and  $i_d$  is the dark current (usually specified, for a low-noise photodiode, in units of nA or pA).

The average signal current is, in a similar way to Equation (7.25),

$$\langle i_s \rangle = \frac{e\eta P_0}{h\nu} \quad (7.31)$$

where  $\eta$  is the quantum efficiency, which is much larger than for a photomultiplier: values of 0.7–0.8 are common. For the simple equivalent current of the detector shown in Figure 7.34, there is an additional Johnson noise contribution of magnitude:

$$\langle i_N^2 \rangle_{\text{JN}} = \frac{4kT\Delta f}{R} \quad (7.32)$$

The overall S/N ratio for direct detection of an unmodulated signal is:

$$\frac{\langle i_s^2 \rangle}{\langle i_N^2 \rangle_{\text{SN}} + \langle i_N^2 \rangle_{\text{JN}}} = \frac{(e\eta P_0/h\nu)^2 R}{2eR(e\eta P_0/h\nu + i_d)\Delta f + 4kT\Delta f} \quad (7.33)$$

This S/N ratio would be reduced by a factor  $m^2/2$  for a modulated signal (cf. Equation 7.28). If shot noise dominates over dark current and Johnson noise, then the shot-noise-limited S/N ratio is:

$$\frac{S}{N} = \frac{\eta P_0}{2h\nu\Delta f} \quad (7.34)$$

In a practical application using a photodiode there will be additional stages of electronic amplification that add noise. It is common to characterize the effect of an electronic circuit on the noise by its *noise figure*,  $F_N$ .

The noise figure can be defined conveniently as:

$$F_N = \frac{\text{noise power at output of circuit}}{\text{noise power at the input} \times \text{amplifier power gain}} \quad (7.35)$$

For input Johnson noise the mean-square, amplified, output noise current is:

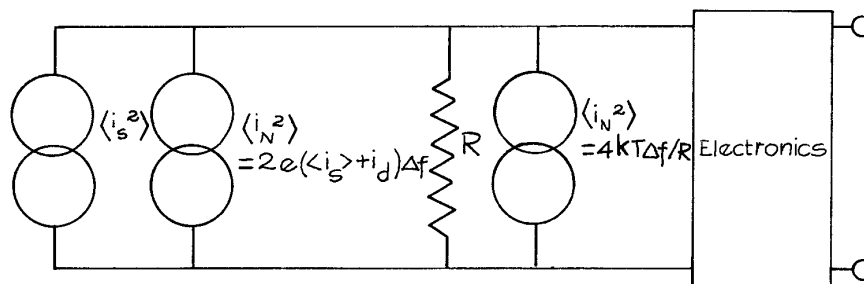
$$\langle i_N^2 \rangle_{\text{AJN}} = \frac{4kTG\Delta f}{R} \quad (7.36)$$

where  $G$  is the power gain of the amplifier within the frequency band being considered. The actual output mean-square noise current is:

$$\langle i_N^2 \rangle_{\text{actual}} = \frac{4kTGF_N\Delta f}{R} \quad (7.37)$$

It is as if the amplifier were noiseless with the input Johnson noise increased by the noise figure.

The noise figure is frequently quoted in dB, a noise figure of 3 dB would represent a doubling of the



**Figure 7.34** Equivalent circuit of a  $p-i-n$  photodiode.

output noise over the value expected from a noiseless amplifier. The term *noise temperature*  $T_i$  is also used, defined by:

$$F_N = 1 + \frac{T_i}{T_{\text{ambient}}} \quad (7.38)$$

*Example:* We illuminate an InGaAs photodiode with a responsivity of 0.8 A/W at 1.3  $\mu\text{m}$  with a power of 10  $\mu\text{W}$ . The system band width is 100 MHz, the dark current is 5 nA (equivalent to 0.5 pA  $\text{Hz}^{-1/2}$ ). We assume that the amplification electronics has a noise figure of 6 dB (relative to a 50 ohm input). We note the following:

Received power  $P_0 = 10^{-3} \times 10 \text{ mW} = 10 \mu\text{W}$

Signal current  $\langle i_s \rangle = 8 \mu\text{A}$

Dark current  $\langle i_d \rangle = 5 \text{ nA}$  (in this case dominated by  $\langle i_s \rangle$ )

Signal Power =  $\langle i_s \rangle^2 R = 3.2 \text{ nW}$

Shot noise power =  $2e(\langle i_s \rangle + \langle i_d \rangle) \delta f R = 1.28 \times 10^{-14} \text{ W}$

Johnson noise power =  $4kT\delta f = 4(1.38 \times 10^{-22})(300)(10^8) = 1.66 \text{ pW}$

Effective Johnson noise power including noise figure =  $10^{0.6} \times 1.66 \text{ pW} = 6.6 \text{ pW}$ .

In this case the Johnson noise is dominant, the effective S/N ratio is:

$$\frac{S}{N} = \frac{3.2 \times 10^{-9}}{6.6 \times 10^{-12}} = 485$$

### 7.8.3 Direct Detection with APDs

In an APD there is a multiplication of the number of charge carriers by a factor  $M$ . This multiplication can result from secondary ionizations produced by both electrons and holes. It is desirable that one or other of these charge carriers should have a significantly greater secondary ionization coefficient<sup>b</sup> than the other.<sup>44</sup> The current in an APD increases by the factor  $M$ , but the associated shot noise increases further, because in a given avalanche process  $M$  will fluctuate, taking values  $M$ ,  $M \pm 1$ ,  $M \pm 2$ , etc. The mean-square noise current thereby increases, not by a factor  $M^2$ , but by a factor  $FM^2$ , where  $F$  is called the *noise factor*. In silicon APDs,  $F$  typically lies in the range 2–20. Therefore, the shot noise becomes:

$$\langle i_N^2 \rangle_1 = 2eFM^2(\langle i_s \rangle + i_d)\Delta f \quad (7.39)$$

Both photo- and dark-generated carriers contribute to the shot noise. The overall S/N ratio is modified from Equation (7.33) and becomes:

$$\frac{\langle i_s^2 \rangle}{\langle i_N^2 \rangle_{\text{SN}} + \langle i_N^2 \rangle_{\text{JN}}} = \frac{M^2(e\eta P_0/h\nu)^2 R}{2eRFM^2(e\eta P_0/h\nu + i_d)\Delta f + 4kT\Delta f} \quad (7.40)$$

Equation 7.40 shows that the S/N ratio improves with increasing multiplication as the Johnson noise contribution becomes less important until the avalanche shot noise becomes dominant. The avalanche-limited S/N ratio is:

$$\frac{S}{N} = \frac{\eta P_0}{2Fh\nu\Delta f} \quad (7.41)$$

a reduction of the signal-noise-ratio from the quantum-limited value by the noise factor.

The NEP can be computed for both  $p-i-n$  photodiodes and APDs by setting  $S/N = 1$  in equations like (7.33) and (7.40) and thereby the minimum input power for  $S/N = 1$  is determined. For  $P_0 = P_{\text{min}}$ , we expect the dark current to be larger than the signal current, so Equation 7.40 becomes:

$$\frac{S}{N} = \frac{(M\eta P_0/h\nu)^2}{(2eFM^2 i_d + 4kT/R)\Delta f} \quad (7.42)$$

For  $S/N = 1$  and a 1 Hz bandwidth,  $P_0 = P_{\text{min}} = \text{NEP}$ , which gives, for an APD:

$$\text{NEP}(\text{W Hz}^{-1/2}) = \frac{h\nu}{M\eta} \sqrt{2eFM^2 i_d + 4kT/R} \quad (7.43)$$

The equivalent result for a  $p-i-n$  diode can be obtained by setting  $M = F = 1$ .

*Example.* For an InGaAs APD with  $F = 10$ ,  $M = 100$ ,  $R = 50 \Omega$ ,  $i_d = 2 \text{ nA}$ , and responsivity 0.8 A/W, the quantum efficiency is:

$$\eta = \frac{0.8(6.626 \times 10^{-34})(3 \times 10^8)}{(1.6 \times 10^{-19})(1.55 \times 10^{-6})} = 0.64$$

From Equation 7.43 the NEP is:

$$\begin{aligned} \text{NEP}(\text{W} = \text{Hz}^{-1/2}) \\ &= \frac{(2 \times 1.6 \times 10^{-19} \times 10^5 \times 2 \times 10^{-9} + 4 \times 1.38 \times 10^{-22} \times 300/50)^{1/2}}{0.8 \times 100} \\ &= 4.7 \times 10^{-13} \text{ W}/\sqrt{\text{Hz}} \end{aligned}$$

Note that in this case the thermal noise is dominant.

### 7.8.4 Photon Counting

For the detection of very low light levels where photomultipliers or Geiger-mode APDs can be used, the technique of *photon counting* is the most sensitive technique. Suppose that the signal to be detected corresponds to a photon flux at the detector of  $N_p$  per second. This corresponds to a detected count rate of  $N_s = \eta N_p$  per second. Along with this goes a dark count rate of  $N_d$  per second. If the detected counts are integrated for  $T$  seconds the total number of counts is:

$$N_T = (N_s + N_d)T \quad (7.44)$$

Since the number of counts observed in a given integration interval fluctuates statistically, then, for a Poissonian distribution, the standard deviation of the number of counts observed is  $N_{SD} = \sqrt{N_T}$ . Thus the counts recorded during “signal” intervals can be written as:

$$N_T = (N_s + N_d)T \pm \sqrt{(N_s + N_d)T} \quad (7.45)$$

If the dark counts alone are integrated over the same time interval  $T$  the total number of counts can be written as:

$$N_D = N_d T \pm \sqrt{N_d T} \quad (7.46)$$

The signal-to-noise ratio of the photon counting process is determined by the difference between  $N_T$  and  $N_D$ , taking into account their respective standard deviations. It can be characterized as:

$$\frac{S}{N} = \frac{N_s T}{\sqrt{(N_s + 2N_d)T}} \quad (7.47)$$

This signal-to-noise ratio can be continuously improved by extending the integration time. The “signal” and “dark” counting intervals can be alternated by chopping or “gating.”<sup>45</sup>

*Example.* Consider a dark count rate of 100/s and a signal count rate of 1/s. A signal-to-noise ratio of 10 will be achieved when  $T/\sqrt{201T} = 10$ , i.e. after an integration period of about 6 hours. It should be noted that if the count rates are too high, so that some counts are missed – so called “pile-up” effects,<sup>46</sup> then these calculations are slightly modified.

## 7.9 PARTICLE AND IONIZING RADIATION DETECTORS

For charged-particle energies up to a few tens of keV there are two detection schemes in wide use. The particles can be collected at a metal surface and the resultant electrical current measured directly, or they can be detected by collecting the slower secondary electrons that are ejected from a metal surface by impact of the primary particle. Charged particles can also be detected with photographic emulsions, scintillators, and various solid-state devices. Emulsions have largely been supplanted by the methods mentioned above, and solid-state devices are only suitable for the detection of high-energy particles (above 30 keV). Table 7.5 lists the properties of common particle detectors.

The collector for the direct detection of a current of charged particles is called a *Faraday cup*. Typical designs are illustrated in Figure 7.35. These simple collectors are connected directly to a current-measuring device and are useful at currents down to the detection limits of modern electrometers – about  $10^{-14}$  A. A properly-designed Faraday cup does not permit secondary electrons to escape. For a positive-ion collector, the loss of a secondary electron appears to the current-measuring instrument as an additional ion, while for an electron collector the loss of each secondary cancels the effect of an incident primary electron. To prevent the escape of secondary electrons, the

**Table 7.5 PARTICLE DETECTORS**

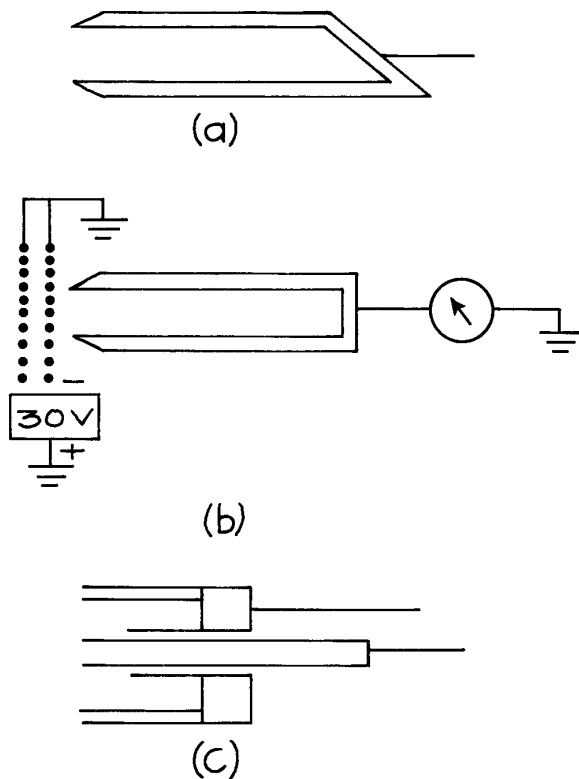
Particle Detector	Output Signal Level	Charge-Collection Time
Ge	2.8 eV/electron hole pair	0.1–10 ns <sup>a</sup>
Si	3.5 eV/electron hole pair	0.1–10 ns <sup>a</sup>
Electron Multiplier	$10^6$ – $10^8$ electrons/ incident electron	0.5–1 ns
Microchannel plate	$10^3$ – $10^4$ electrons/ incident photon	0.1 ns
Scintillator:		
Plastic	~1 photon/3keV	0.1–10 ns
Alkali halide	~1 photon/3keV	1 $\mu$ s
Gas-filled tube	25–35 eV/ion pair	0.01–5 $\mu$ s

<sup>a</sup> For narrow depletion depth.

depth of a Faraday cup should be at least five times its diameter. A suppressor aperture or grid biased to about  $-30$  V in front of a Faraday cup effectively prevents the escape of most secondaries.

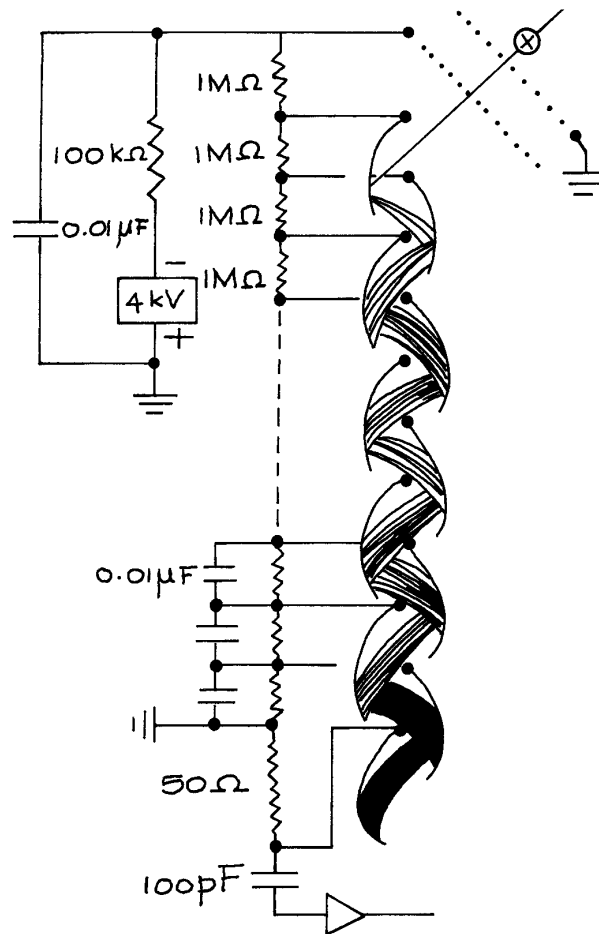
A grounded grid in front of the suppressor as illustrated in Figure 7.35(b) prevents field penetration from the suppressor in the direction of the incident current source. When a particle beam must be aligned and sharply focused, a concentric pair of collectors [Figure 7.35(c)] is useful. With this arrangement the beam-focusing elements are adjusted to maximize the ratio of current to the inner cup in relation to the current to the outer cup.

Charged-particle currents of less than  $10^{-14}$  A can be detected with an electron multiplier, which is like a photomultiplier without the photocathode. As illustrated in



**Figure 7.35** Faraday cup designs. Design (b) has a grid biased to suppress secondary electron emission. Design (c) has a double cup for aligning and focusing a beam.

Figure 7.36, these devices have an internal dynode structure. Secondary electrons produced by impact of a particle on the first dynode are accelerated towards the second dynode through a potential drop of 100–300 V. The impact of these electrons results in a number of secondaries that are in turn accelerated into the third dynode, and so on. The secondary emission coefficient of the dynodes is typically about three electrons per incident particle; the impact of a



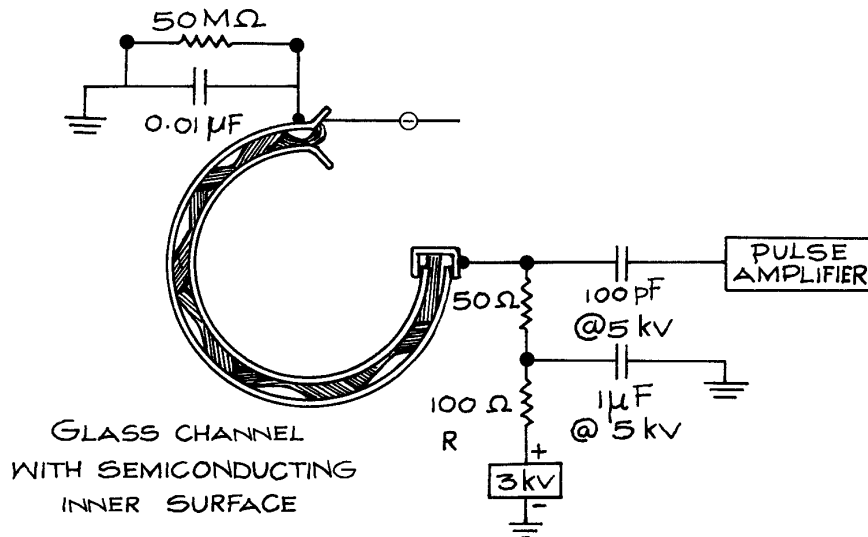
**Figure 7.36** An electron multiplier. Electrical connections shown are appropriate for detecting positive ions. Grids at the entrance of the multiplier prevent the escape of secondary electrons.

single particle on a multiplier with  $n$  electrodes results in a pulse of about  $3^n$  electrons.

Electron multipliers are available with 10 to 20 dynodes to provide gains of  $10^6$  to  $10^9$ . The dispersion in time of the shower of electrons from the last dynode is on the order of 10 ns. The corresponding current through a 50 ohm measuring resistor ranges from 800  $\mu$ V to 800 mV, which is easily detected with modern pulse-counting electronics. Furthermore, if time resolution is not required, the output current from an electron multiplier can be measured directly with an electrometer.

The electrodes of an electron multiplier are fabricated of a material that has a high work function – typically a Be–Cu alloy that has been “activated” by some proprietary process. The high work function is desirable because it inhibits thermionic emission that would result in noise pulses at the multiplier output. On the other hand, incident particles must have energies in excess of a few hundred eV to assure efficient secondary emission. When working with lower-energy particles, it is necessary to provide an acceleration stage at the multiplier input. For incident particles with energies greater than 300 eV, the detection efficiency of an electron multiplier can exceed 90%.

For the detection of positive ions, an electron multiplier is usually operated with the first dynode at a high negative potential and the last dynode near ground potential, as in Figure 7.36. This arrangement provides acceleration of incident ions. For electrons or negative ions, the cathode is usually biased positive by a few hundred volts and the last dynode is at a high positive potential relative to ground, as in Figure 7.37. In this latter configuration, it is only convenient to operate in a pulse-counting mode, since the detection electronics must be electrically insulated from the measuring resistor. Ordinarily, the multiplier output is coupled to the pulse-counting electronics via a high-voltage disc ceramic capacitor of about 100 pF. A problem sometimes encountered with this arrangement is that sparking between dynodes or across dynode resistors gives rise to large high-frequency transients that are transmitted through the coupling capacitor and damage the electronics. This problem can be cured by coupling the output of the multiplier to the electronics through a transformer that will attenuate large pulses because of saturation of the ferrite core material. A suitable transformer can be produced by making a 10-turn bifilar winding of well-insulated wire on a ferrite core (e.g. Ferroxcube),



**Figure 7.37** A channel electron multiplier. Electrical connections shown are for the detection of electrons.

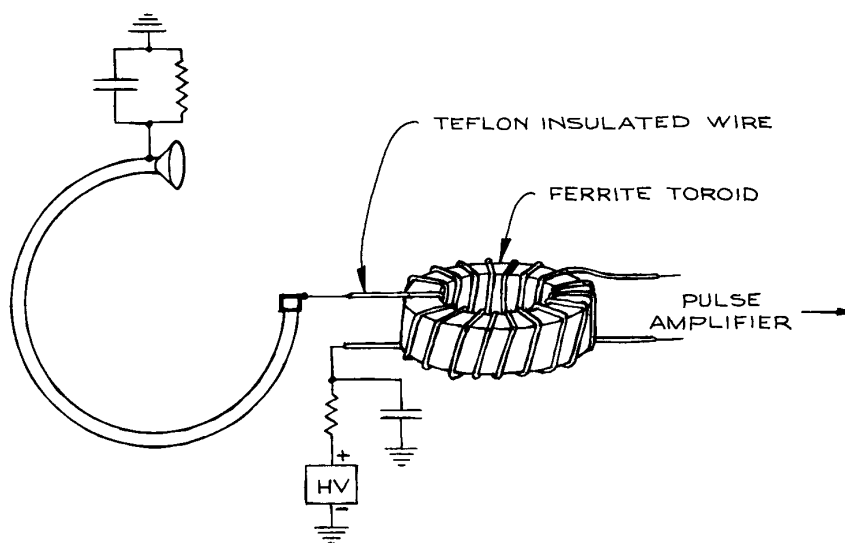
as illustrated in Figure 7.38.<sup>47</sup> The number of turns and their position must be carefully adjusted to assure a proper impedance match between the multiplier and the electronics.

The output current from an electron multiplier is limited by the current available from the resistive voltage divider that establishes the potential of the dynodes. To ensure that the gain of a multiplier is not reduced because of depletion of the dynode charge, the maximum output current should be at least an order of magnitude less than the current drawn by the dynode resistor chain. For CEMs, the output current should be an order of magnitude less than the current from the high-voltage supply through the (typically  $10^9$  ohm) resistance of the channel.

An important variant of the electron multiplier uses a continuous dynode as illustrated in Figure 7.37. These are channel electron multipliers (CEM) or channeltrons. They are glass tubes with semiconducting inner surfaces. The end-to-end resistance is about  $10^9$  ohm. In operation, a potential of about 3 kV across the CEM yields a gain of about  $10^8$ . Channel electron multipliers have the advantage of small size (about 50 mm long), low cost, and rugged-

ness. Recently, monolithic ceramic CEMs have become available from OptoTechnik. These consist of a ceramic block with an interior serpentine channel. Many configurations are available, with entrance apertures as small as a few  $\text{mm}^2$  to more than  $1 \text{ cm}^2$ . The overall volume ranges from 1 to  $5 \text{ cm}^3$ . The monolithic ceramic CEMs are very robust. They have found wide use in rocket-borne instruments, where forces at launch can exceed 30 G.

The gain of a CEM is determined, not by the absolute size, but rather by the ratio of the length of the channel to the diameter. This ratio is about 50 for a conventional CEM. Very small CEMs are available in close-packed arrays. Each channel is a few micrometers in diameter and a few mm long. The arrays, known as microchannel plates or MCPs, are available in sizes up to 10 cm on a side. Because the length-to-diameter ratio is only about 20, microchannel plates are usually stacked in pairs to provide easily detected charge pulses. Owing to the small size of the channel, the charge cloud through a channel is temporally much more compact than in a full-size CEM. The charge cloud is delivered in about 1 ns. With leading-edge discrimination it has been possible to time the arrival of a



**Figure 7.38** A coupling transformer between an electron multiplier and a sensitive pulse amplifier. Properly constructed, a transformer will faithfully transmit signal pulses while attenuating large noise pulses.

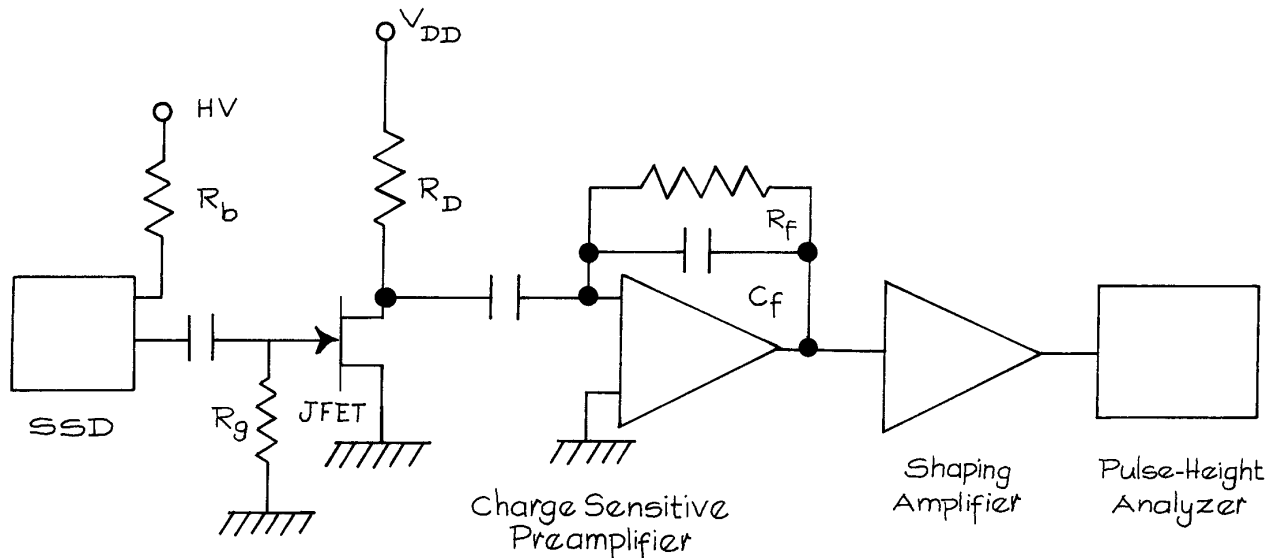
charged particle at the detector with a resolution of a few hundred picoseconds.

Microchannel plates are used as position-sensitive detectors of electrons and ions at energies up to at least 10 keV. A particle striking the front of the plate triggers an electron avalanche in only one channel. The position of the charge cloud emitted at the output of a channel can be located with an accuracy approaching the diameter of the channel. Both one-dimensional and two-dimensional position information can be extracted. A number of methods have been developed to obtain position information.<sup>48</sup> The most straight-forward scheme is to provide an array of collectors each with its own pulse amplifier and counter. This is the most robust method, but expensive and impractical if the full resolution of the microchannel plate is to be realized – a standard 2.5 cm round MCP can potentially provide more than  $10^4$  pixels. Many position-sensitive detectors employ a resistive strip anode across the back of a stacked pair of MCPs. The arrival of a charge pulse at a point on the anode is detected simultaneously as a pulse at each end or each corner of the anode, and the position is determined from the ratio of the pulse heights.

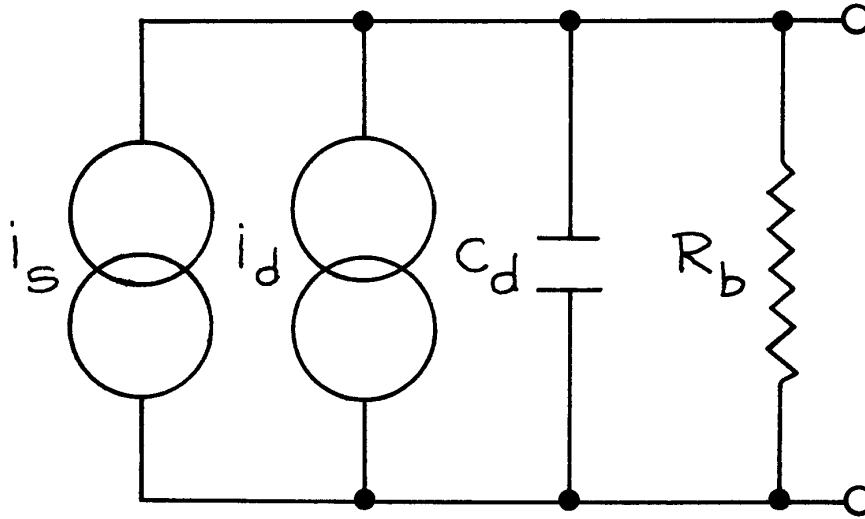
### 7.9.1 Solid-State Detectors

A solid-state detector (SSD) converts particle energy to charge that is then collected and amplified. Silicon detectors convert every 3.6 eV of particle kinetic energy to a single electron–hole pair. The energy resolution of a SSD depends on the charge generated by the incident particle, the noise generated by the detector, and the electronic circuits to which the output of the detector is connected. Resolution is also affected by the charge generation mechanism within the detector. Charge generation falls within a purely statistical process and one determined by a constant energy loss per electron–hole pair creation. There is also the variable energy loss in the thin window of the detector and the efficiency of charge collection as a function of the geometry of the electron–hole pair production. The arrangement of an SSD with FET amplifier followed by a shaping amplifier and pulse-height analyzer is shown in Figure 7.39.

The SSD is essentially a reverse-biased diode with the bias voltage applied through a high resistance  $R_b$  in series with the voltage supply. The FET is used in the common



**Figure 7.39** A solid-state detector with a n FET amplifier followed by a pulse-shaping amplifier and a pulse-height analyzer.



**Figure 7.40** Equivalent circuit of a solid-state detector.

source (CS) configuration as a transconductance amplifier (see also Chapter 6). There is a large gate bias resistor  $R_g$  from the gate to ground. The shaping amplifier, increases the amplitude of the signal from the FET output and also transforms the output pulse into one with an approximate Gaussian shape with rise-time  $\tau_r$  and fall-time  $\tau_f$ . Normally  $\tau_r$  and  $\tau_f$  are set equal to a single time  $\tau$ . The response of the shaping amplifier as a function of frequency  $\omega$  is given by:

$$A(\omega) = A_0 \frac{(\omega/\tau)}{1 + (\omega/\tau)^2} \quad (7.48)$$

where  $A_0$  is the mid-band gain of the amplifier.

The SSD can be modeled as a current source  $i_s$  in parallel with a capacitance  $C_d$  that represents the effective capacitance of the SSD and the high voltage bias resistor  $R_b$ , as shown in Figure 7.40. The current source  $i_d$  represents the leakage current of the detector. The small-signal model for the FET, shown in Figure 7.41, includes the capacitance of the gate-channel junction  $C_g$  and the voltage-controlled current source,  $g_m v_{gs}$ , where  $g_m$  is the transconductance of the JFET and  $v_{gs}$  is the gate-source signal voltage. Current flowing in the reverse-biased gate-channel junction is represented by the current source  $i_g$ . There are various contributions to both shot and Johnson noise in the arrangement shown in Figure 7.41. These are:

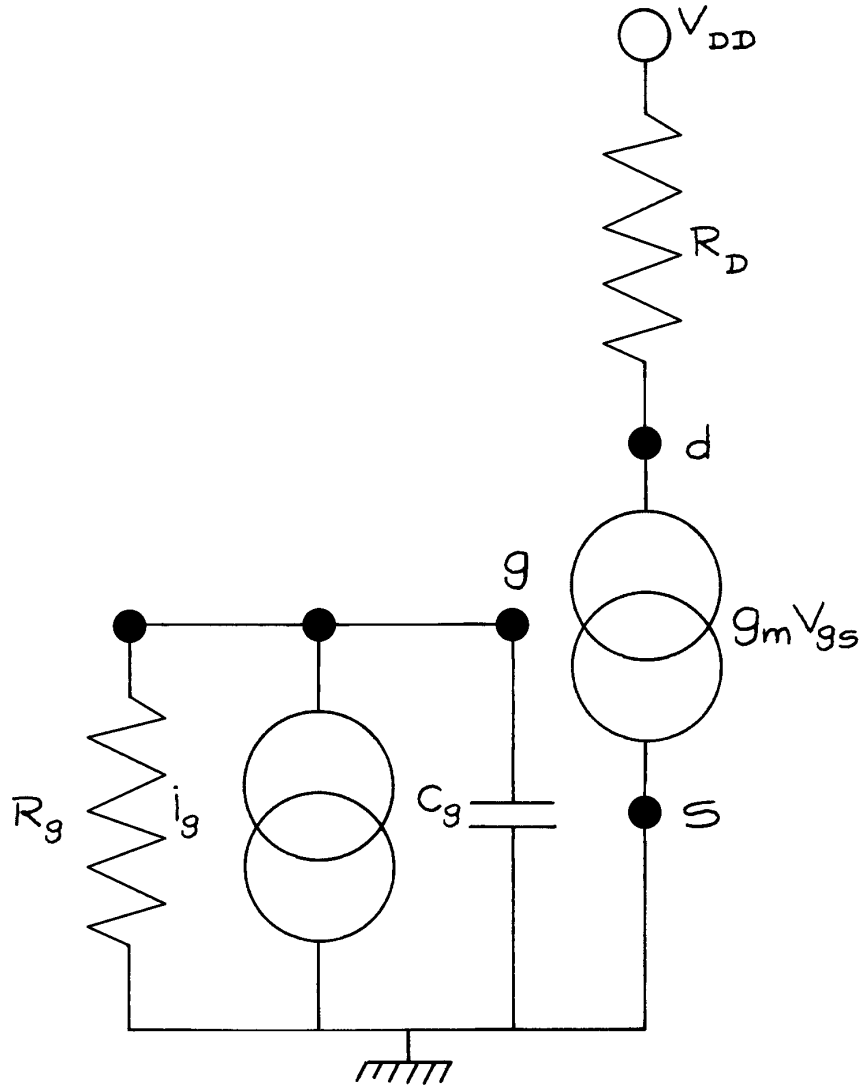
- (1) Shot noise from the reverse-bias leakage current of the SSD  $\langle i_N^2 \rangle_{SN1} = 2ei_d \Delta f$
- (2) Johnson noise from the series bias resistor  $\langle i_N^2 \rangle_{JN1} = 4kT \Delta f / R_b$
- (3) Shot noise from the JFET gate current  $\langle i_N^2 \rangle_{SN2} = 2ei_s \Delta f$
- (4) Johnson noise from the gate bias resistor  $\langle i_N^2 \rangle_{JN2} = 4kT \Delta f / R_g$
- (5) Johnson noise from the channel resistance  $\langle i_N^2 \rangle_{JN3} = (4kT \delta f \gamma g_m)$ , where for a long channel device  $\gamma = 2/3$ .<sup>49</sup>

The current noise can be converted to voltage noise at the gate of the FET by multiplying by the input impedance of the FET. This impedance is dominated by  $C_d$  and  $C_g$ . Because the noise sources are incoherent they are added in quadrature. The total noise voltage is then:

$$\langle i_N^2 \rangle = \left\{ \left( \frac{1}{4\pi^2 f^2 C_T^2} \right) \left[ 2ei_d + 2ei_s + \frac{4kT}{R_b} + \frac{4kT}{R_g} \right] + 4kT \gamma g_m \right\} \Delta f \quad (7.49)$$

where  $C_T = C_d + C_g$ . When this expression is integrated over the response function of the shaping amplifier, the noise voltage referred to the input of the FET is:





**Figure 7.41** Small-signal equivalent circuit of an FET.

$$\langle V_N^2 \rangle = \frac{e\tau}{4C_T^2} (i_d + i_s) + \frac{kT\tau}{2C_T^2} \left( \frac{1}{R_b} + \frac{1}{R_g} \right) + \frac{kT}{2\gamma g_m \tau} \quad (7.50)$$

The equivalent noise charge (ENC)<sup>50</sup> is obtained by multiplying this expression by  $C_T$ , with the result:

$$\text{ENC} = \frac{e\tau}{4} (i_d + i_s) + \frac{kT\tau}{2} \left( \frac{1}{R_b} + \frac{1}{R_g} \right) + \frac{kTC_T^2}{2\gamma g_m \tau} \quad (7.51)$$

As an example, let us evaluate this expression using the experimental parameters of the Canberra fully depleted passivated implanted planar silicon (PIPS) solid-state

detector, model FD 150-16-500 RM. The parameters of the 2SK152 low-noise  $n$ -channel JFET are used to evaluate the noise contribution from the JFET. A time constant of  $10^{-6}$  s was used for the shaping parameter. Below is a list of the magnitudes of the parameters:

$$\begin{aligned}\tau &= 10^{-6} \text{ s} \\ i_d &= 166 \text{ nA} \\ i_s &= 5 \text{ nA} \\ T &= 300 \text{ K} \\ C_d &= 50 \text{ pF}, C_g = 15 \text{ pF}, C_T = 65 \text{ pF} \\ R_b &= 108 \text{ ohm} \\ R_g &= 108 \text{ ohm} \\ g_m &= 30 \text{ mA/V}\end{aligned}$$

**Additional Considerations.** In order to transfer the maximum energy from the detector to the JFET, the input impedance of the JFET, should be matched to the output impedance of the SSD. This means that a transistor with  $C_g = C_d$  should be chosen and stray capacitance eliminated. The ratio of energy transferred to theoretical maximum energy transfer is given by  $4x/(1+x)^2$ , where  $x$  is the ratio of detector to gate capacitance,  $C_d/C_g$ . The choice of the shaping constant  $\tau$  is an important factor in the ENC, but it is also important to recognize that  $\tau$  fixes the maximum count rate. For  $\tau = 10^{-6}$  s, the maximum asynchronous count rate that can be accommodated with 90% efficiency is 100 kHz. The temperatures of the SSD and JFET can be controlled. In this analysis the temperature of the SSD and JFET are assumed equal, however it is possible to independently cool one or the other. The leakage current of the Canberra PIPS detector is a strong function of temperature and is given by

$$i_d = 63 \cdot 10^{-9} \cdot 2^{(T-92)/5} \quad (7.52)$$

## 7.9.2 Scintillation Counters

Scintillation counters allow the detection of ionizing radiation (X-rays,  $\gamma$ -rays, and energetic particles (protons, neutrons, muons, pions, etc.) by conversion of their energy into visible light. A typical scintillation counter uses a fluorescent solid or liquid that converts a nonionizing photon or energetic particle into a flash of visible light. The

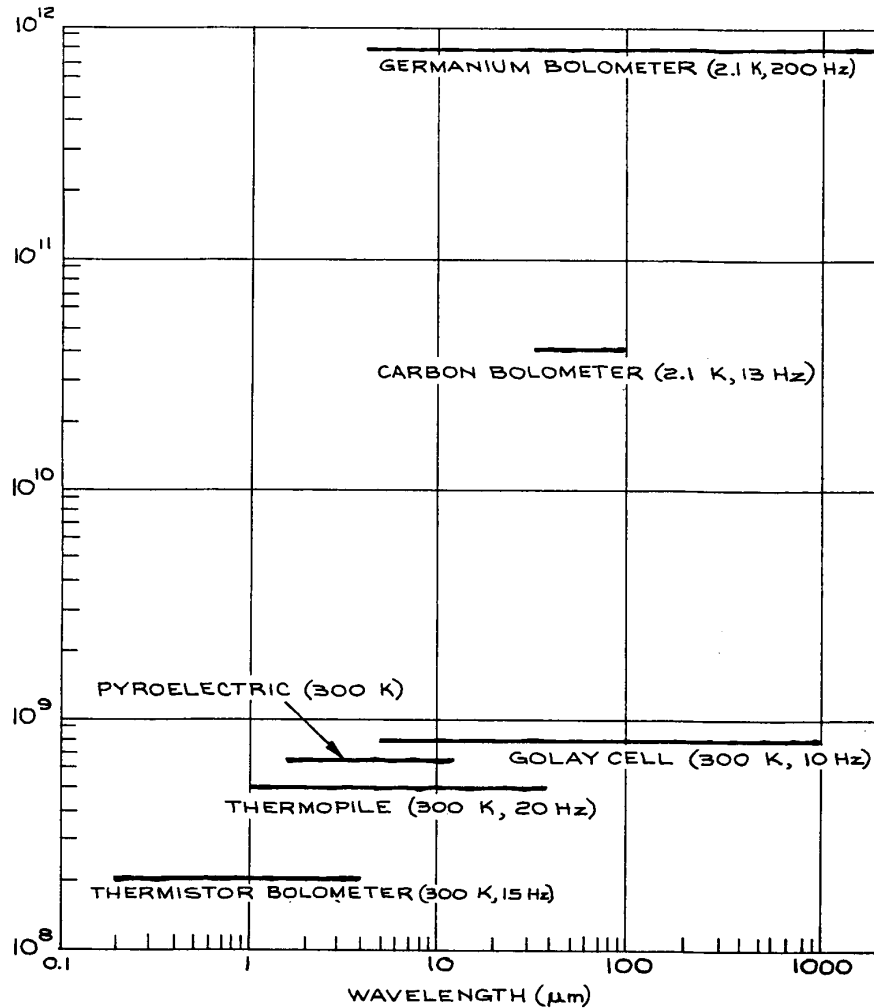
visible light is generally detected with a photomultiplier with a large area photocathode placed in close proximity to the scintillation material. Scintillators are often made of plastics impregnated with a fluorescent dye. The intensity of the light flash produced by detection of an energetic photon or particle is correlated with the energy of the detected photon or particle, so the energy spectrum of the detected radiation or particles can be determined from the flash intensity. This is often carried out using a *multi-channel analyzer*, an instrument that sorts detected electric pulses into a histogram that shows their height distribution. Scintillators are characterized by their efficiency; an efficiency of 0.1 corresponds to the production of one visible or UV photon from the detection of a particle or photon of energy 25 eV.<sup>46</sup> Also of importance are the absorption and emission spectra, signal linearity, and the rise and fall times of the scintillation pulse. These pulses can be as short as 1 ns for a modern plastic scintillator. Further information about scintillation counters can be obtained from the specialist literature.<sup>51-54</sup>

## 7.9.3 X-Ray Detectors

X-rays can be detected with scintillators, for example organic scintillators, single crystals of bismuth germanate (BGO), or single crystals of sodium-activated cesium iodide [CsI(Na)], but are generally detected with silicon or germanium photodiodes. Such detectors are available from Amptek, Canberra, and Hamamatsu.

## 7.10 THERMAL DETECTORS

Thermal detectors, in principle, have a detectivity that is independent of wavelength from the vacuum ultraviolet upward, as shown in Figure 7.42. The absorbing properties of the “black” surface of the detector will, however, show some wavelength dependence, and the necessity for a protective window on some detector elements may limit the useful spectral bandwidth of the device. Most commonly available thermal detectors, although by no means as sensitive as various types of photon detectors, achieve spectral response very far into the infrared – to the microwave region, in fact – while conveniently operating at room temperature. The operating characteristics of some



**Figure 7.42** Typical  $D^*(\lambda)$  curves for various thermal detectors assuming total absorption of incident radiation. The operating temperature, modulation frequency, and typical useful wavelength range are shown for each detector. In each case, the detector is assumed to view a hemispherical surround at a temperature of 300 K.

commercially available thermal detectors are given in Table 7.6. Each of these detectors is discussed briefly below. Putley<sup>55</sup> gives a more detailed discussion.

### 7.10.1 Thermopiles

*Thermopiles*, although they are one of the earliest forms of infrared detector, are still widely used. Their operation is

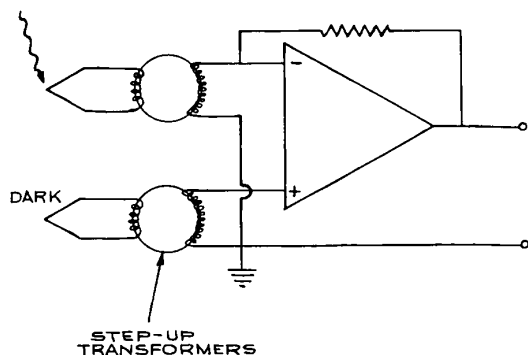
based on the Seebeck effect, where heating the junction between two dissimilar conductors generates a potential difference across the junction. An ideal device should have a large Seebeck coefficient, low resistance (to minimize ohmic heating), and a low thermal conductivity (to minimize heat loss between the hot and cold junctions of the thermopile). These devices are usually operated with an equal number of hot (irradiated) and cold (dark) junctions,

**Table 7.6 Characteristics of commercially available thermal detectors**

Device	$D^*$ ( $\text{cm Hz}^{1/2} \text{ W}^{-1}$ )	Time Constant	Suppliers <sup>a</sup>
Thermopile	$(1-4) \times 10^8$	20 $\mu\text{s}$ –60 ms	Boston Electronics, Eppley Hamamatsu, Laser Probe, Newport, Perkin-Elmer Optoelectronics Scientech, Thorlabs
Pyroelectric	$10^6$ – $10^9$	>100 ps	Perkin-Elmer Optoelectronics, Hamamatsu, Scientech Laser Probe, SensArray, Spiricon, Scitec Instruments
Bolometer	$2.5$ – $10^8$	1 ms	Eltec Eltec, Electro Optical Components
Golay Cell	$<10^{-10}$ NEP ( $\text{W Hz}^{1/2}$ )	$\approx 10$ ms	QMC Instruments SensArray Sofrdir QMC Instruments, Microtech Instruments

the latter serving as a reference to compensate for drifts in ambient temperature. Both metal (copper–constantan, bismuth–silver, antimony–bismuth) and semiconductor junctions are used as the active elements. The junctions can take the form of evaporated films, which improves the robustness of the devices and reduces their time constant, although this is still slow (0.1 ms at best). Because a thermopile has very low output impedance, and generates low voltages, it must be used with a specially designed low-noise amplifier, or with a step-up transformer, as shown in Figure 7.43. Such transformers can be conveniently built in the laboratory by winding the primary and secondary coils on a small ferrite torus.

Although thermopiles, along with other thermal detectors, do not have absolutely flat spectral response over unlimited wavelength regions, they can be remarkably flat



**Figure 7.43** Operating circuit for a thermopile using a dark junction (or junctions) and a differential amplifier for compensation.

within restricted regions – in the visible or from 1 to 10  $\mu\text{m}$ , for example. In addition, such detectors with calibrations traceable to NIST are available. Thus they are invaluable in the absolute calibration (both for radiant sensitivity and for spectral response) of light sources, detectors, and spectrometers. Thermopiles are available from Boston Electronics, The Eppley Laboratory, Hamamatsu, Laser Probe, Perkin-Elmer Optoelectronics, Scientech, and Thorlabs.

### 7.10.2 Pyroelectric Detectors

*Pyroelectric detectors* utilize the change in surface charge that results when certain asymmetric crystals (ones that can possess an internal electric dipole moment) are heated. The crystalline material is fabricated as the dielectric in a small capacitor, and the change in charge is measured when the element is irradiated. Thus, these devices are inherently a.c. detectors. If the chopping frequency of the input radiation is slow compared to the thermal relaxation time of the crystal, the crystal remains close to thermal equilibrium and the current response is small. When the chopping period becomes shorter than the thermal relaxation time, much greater heating and current response results. The responsivity of the detector in this case can be written as:

$$R = \frac{p(T)}{\rho C_p d} \text{ (A/W)} \quad (7.53)$$

where  $p(T)$  is the pyroelectric coefficient at temperature  $T$ ,  $d$  is the spacing of the capacitor electrodes, and  $\rho$  and  $C_p$  are the density and specific heat of the crystal, respectively.

The equivalent circuit of a pyroelectric detector is a current source in parallel with a capacitance, which can range from a few to several hundred picofarads. For optimum performance, the resultant high impedance must be matched to a low-input-impedance, low-output-impedance amplifier. Two examples of such circuits are given in Figure 7.44. Pyroelectric detectors are available from DIAS Infrared GmbH, Hamamatsu, Laser Probe, Perkin-Elmer Optoelectronics, Scientech, Scitec Instruments and SensArray.

Frequently pyroelectric detector packages contain two pyroelectric elements connected with reverse polarity so that the detector only responds to the difference in illumination between the two elements. This is a useful way to reduce the effect of background radiation. For faster response the capacitance of the detector must be shunted with a small resistor (though this reduces the detectivity).

Pyroelectric detectors are robust and have frequency responses extending from a few hertz to 100 GHz or so. Their detectivities are comparable with those of thermopiles, and they also have a flat spectral response. They can replace the thermopile in many applications as a convenient, room temperature, wide-spectral-sensitivity detector of infrared and visible light. Pyroelectric detector arrays are available from Electro-Optical Components, Scitec, SensArray, Spectrum Detector, and Spiricon.

Pyroelectric detectors are widely used for the detection of very short-duration infrared laser pulses. High-energy pulses, however, generate acoustic waves in the detector crystal giving rise to spurious signals. These acoustic signals are more of a problem in the observation of long laser

pulses ( $> 100$  ns) than they are with short pulses. Pyroelectric detectors are much more sensitive than *photon-drag detectors*<sup>56</sup> for this purpose.

Photon-drag detectors operate by utilizing the momentum transfer between photons and photo-produced carriers in a material, creating an electromotive force. They are fast, but relatively insensitive with typical NEP of  $4 \times 10^{-3}$  W Hz<sup>1/2</sup>. They are mainly used for detecting intense pulses from 10  $\mu$ m carbon dioxide lasers.

### 7.10.3 Bolometers

The resistance of a solid changes with temperature according to a relation of the form:

$$R(T) = R_0[1 + \gamma(T - T_0)] \quad (7.54)$$

where  $\gamma$  is the temperature coefficient of resistance, typically about 0.05/K for a metal, and  $R_0$  is the resistance at temperature  $T_0$ .

A *bolometer* is constructed from a material with a large temperature coefficient of resistance. Absorbed radiation heats the bolometer element and changes its resistance. Bolometers utilize metal, semiconductor, or almost superconducting elements. Metal bolometers utilize fine wires (platinum or nickel) or metal films. The mass of the element must be kept small in order to maximize its temperature rise. Even so, the response time is fairly long ( $\geq 1$  ms). Semiconductor bolometer elements (thermistors) have larger absolute values of  $\gamma$  and have largely replaced metals except where very long-term stability is required.

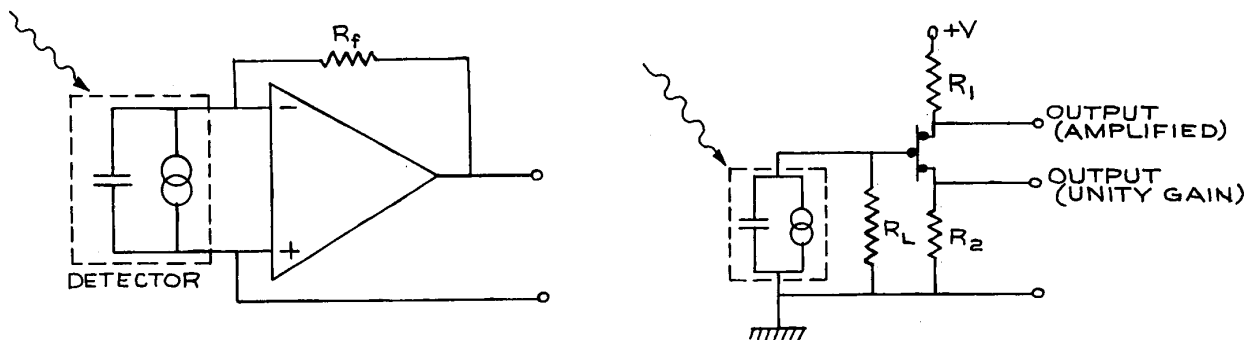
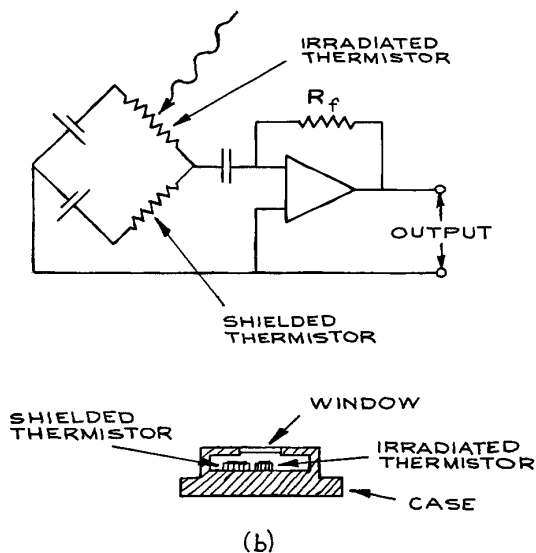
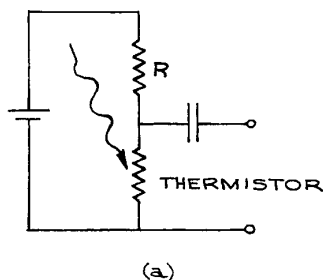


Figure 7.44 Two examples of operating circuits for pyroelectric detectors.

Bolometer elements can be operated in several ways; a simple bias circuit for a single element is shown in Figure 7.45(a). However, it is usual to operate the elements in pairs in a bridge circuit, as shown in Figure 7.45(b) (see also Chapter 6). One element is irradiated, while the second serves as a reference and compensates for changes in ambient temperature.

Thermistors have a negative  $I$ - $V$  characteristic above a certain current and will exhibit destructive thermal runaway unless operated with a bias resistor. It is usually best to operate the thermistor at currents below the negative-



**Figure 7.45** Operating circuits for thermistor bolometers: (a) simple bias circuit; (b) bridge circuit using compensating shielded thermistor, with device construction shown.

resistance part of its  $I$ - $V$  characteristic. Bolometers can operate to wavelengths up to 1000  $\mu\text{m}$ , usually limited by the transmission of their entrance window. These devices are available from B.F. Goodrich (acquired Barnes Engineering from the EDO Corporation), Infrared Laboratories, and Sofradir.

*Microbolometers*, generally used in thermal imaging applications in the 8–16  $\mu\text{m}$  spectral region, use a grid of vanadium oxide or amorphous silicon on top of a silicon grid. Infrared radiation heats the vanadium oxide, or amorphous silicon, and changes its electrical resistance. The array of resistance changes is read out and provides a thermal image of an object. They operate at room temperature and do not require cooling.

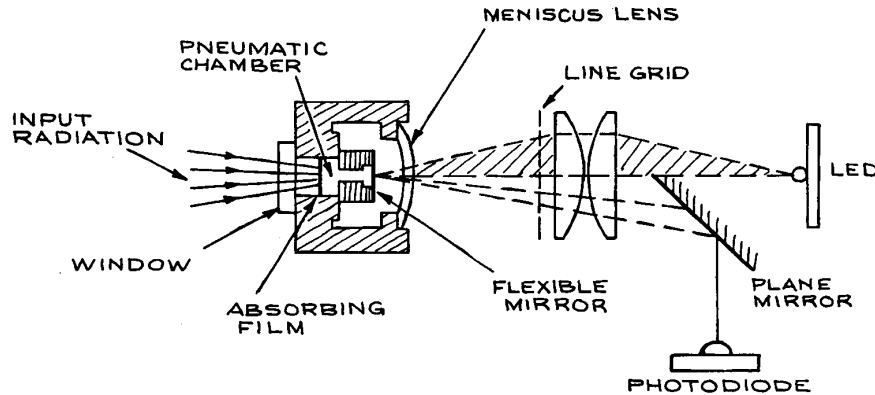
Microbolometer arrays are typically  $640 \times 480$ ,  $320 \times 240$ , or  $160 \times 120$  pixels. Arrays are available from NEC, Honeywell, and ULIS-IR. Complete camera systems are available from BAE, CDIP, FLIR, Fluke Test Instruments, Infrared Solutions, InfraredVision Technologies Corporation (L-3 Communications), ISG Thermal Systems, Jenoptik, Land Infrared, Raytek, and Thermoteknix.

#### 7.10.4 The Golay Cell

In a Golay cell (named for its inventor, M.J.E. Golay),<sup>57</sup> radiation is absorbed by a metal film that forms one side of a small sealed chamber containing xenon (used because of its low thermal conductivity). Another wall of the chamber is a flexible membrane, which moves as the xenon is heated. The motion of the membrane is used to change the amount of light reflected to a photodetector. The operating principle and essential design features of a modern Golay cell are shown in Figure 7.46. Although these detectors are fragile, they are quite sensitive and are still used for far infrared spectroscopy. Golay cells are available from QMC Instruments and Microtech Instruments.

### 7.11 ELECTRONICS TO BE USED WITH DETECTORS

The primary signal from a photodetector may be a time-varying current, as, for example, in the case of a photomultiplier, or a time-varying voltage, as, for example, in the case of a photodiode used in photovoltaic mode. In general, the primary signal from the photodetector will



**Figure 7.46** Schematic design of the Golay detector. The top half of the line grid is illuminated by the LED and imaged back on the lower half of the grid by the flexible mirror and meniscus lens. Any radiation-induced deformation of the flexible mirror moves the image of the line grid and changes the illumination reaching the photodiode.

be processed using electronics that follows the detector. This is most often a current or voltage amplifier, or a current to voltage converter (transimpedance) amplifier. When a detector is being selected, it is often time-saving to buy a detector–amplifier integrated combination that will provide an electrical signal in a useful format for analysis. Manufacturer’s data sheets should be consulted for recommended operational circuits. A few issues are important in designing or specifying circuits to be used in conjunction with the detector. For optimum power transfer, the output impedance of the detector should be matched to the input impedance of the following circuit. For example, if a photomultiplier is used with a  $50\ \Omega$  anode resistor, then the input impedance of the following circuitry should also be  $50\ \Omega$  for optimum high-speed performance. Many voltage amplifiers have a very large input impedance, so if the input to the amplifier is connected to a detector with output impedance  $R$ , the input impedance of the circuit will become  $R$ . For high-speed detector operation the signal from the detector may be connected by  $50\ \Omega$  coaxial cable to the input of the following electronics. The input impedance of this electronics should also be  $50\ \Omega$ . The choice of input and output impedance will affect the noise of the entire detector/amplifier combination. If the noise figure of the electronics is known, then the overall noise performance can be deter-

mined, using the methods discussed in Section 7.2. If the detector/electronics combination has an unknown noise figure then this will have to be determined experimentally. This is a specialized endeavor, but many manufacturers, such as Agilent, Maxim, National Instruments, and Rohde and Schwarz have application notes available to explain how this is done.

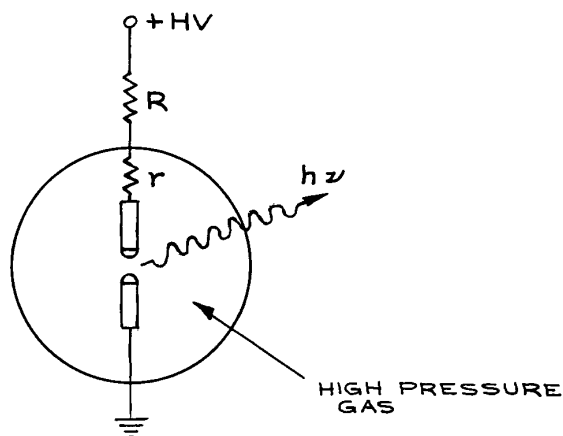
## 7.12 DETECTOR CALIBRATION

In certain cases, the data supplied by a detector manufacturer on parameters such as  $D^*(\lambda)$  or  $\tau$  will only indicate a range of values within which the detector’s characteristics will fall, or the data may not be sufficiently reliable. If necessary, more accurate detector calibration can be carried out.

To determine  $D^*(\lambda)$  for a detector, its response must be measured with a source giving a known irradiance  $E_c(\lambda)$  at the detector. This can be done either with a calibrated source such as a blackbody, or by comparison with another detector, such as a thermopile, which has a calibrated response. To determine the temporal response function of a detector operated in a given configuration, all that is necessary is to irradiate the detector with a very short light pulse and record the output pulse shape on a fast

oscilloscope that has its input appropriately terminated. For response functions below about 1 ns, a sampling oscilloscope should be used in conjunction with repetitive short-light-pulse irradiation of the detector. For detectors with response faster than about 1 ns, sufficiently short light pulses can be obtained from a mode-locked Nd:YAG, Nd:glass, ruby, or dye laser. Pulses from mode-locked argon or CO<sub>2</sub> lasers may not be short enough to calibrate a very fast detector. To obtain short light pulses outside conveniently available wavelength regions, nonlinear harmonic generation or mixing schemes can be used, but the difficulties of such techniques can be severe. To determine the time response of a detector that responds slower than 1–2 ns, any pulsed light source of much shorter duration than the detector response can be used. Short-duration ( $\approx 10$  ns) flashlamps are available from Perkin-Elmer Optoelectronics and Xenon Corporation.

Very-short-duration, low-energy, pulsed light sources are easily made in the laboratory. A very simple design is shown in Figure 7.47. A high-voltage discharge between two electrodes spaced by a few millimeters or less in a high-pressure gas is used. Hydrogen works very well, and even air at atmospheric pressure is satisfactory. To obtain a short-duration flash, only the self-capacitance  $C_s$  of the electrodes must be allowed to discharge. To accomplish this, a small charging resistor must be placed very



**Figure 7.47** Simple design for high-pressure, nanosecond-duration pulsed light source. Generally  $R \gg r$ , where  $r$  ranges from 1 k ohm to 1 M ohm.

close to the high-voltage electrode; a larger series charging resistor can be placed farther from the electrode. The tips from ballpoint pens make excellent electrodes in this application. The flash duration is proportional to  $L/p$ , where  $L$  is the electrode spacing and  $p$  the gas pressure. The breakdown voltage is proportional to  $L$  and also approximately to  $p$ : the capacitance of the electrode gap is proportional to  $L$ , so the flash energy  $\frac{1}{2}C_s V^2$  is proportional to  $L^3 p^2$ . Thus, for short-duration, high-energy flashes,  $p$  should be high. The repetition frequency of the lamp, which is most easily operated in a free-running mode, will depend on the charging time constant.

## Endnotes

- <sup>a</sup> The first dynode often has a larger  $\delta$  value than the others, but we will assume that  $\delta$  is an average value.  
<sup>b</sup> The secondary ionization coefficient is the number of secondary electron-hole pairs produced per unit length by a specific charge carrier (electron or hole) in travelling through the material.

## Cited References

- 1 R. K. Bock and A. Vasilescu, *The Particle Detector Briefbook*, Springer, New York, 1998.
- 2 C. Enss, *Cryogenic Particle Detection*, Topics in Applied Physics, Vol. **99**, Springer, New York, 2005.
- 3 R. C. Fernow *Introduction To Experimental Particle Physics*, Cambridge University Press, Cambridge, 1989.
- 4 K. Kleinknecht *Detectors For Particle Radiation*, 2nd revised edn., Cambridge University Press, Cambridge, 1998.
- 5 P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, 1962.
- 6 C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
- 7 R. D. Hudson, Jr., and J. W. Hudson (Eds.), *Infrared Detectors*, Benchmark Papers in Optics, Vol. **2**, Dowden, Hutchinson and Ross, Stroudsburg, PA, 1975.
- 8 D. Wood, *Optoelectronic Semiconductor Devices*, Prentice-Hall, New York, 1994.
- 9 R. J. Keyes (ed.), *Optical and Infrared Detectors*, Topics in Applied Physics, Vol. **19** Springer, Berlin, 1977.
- 10 R. H. Kingston, *Detection of Optical and Infrared Radiation*, Springer Series in Optical Sciences, Vol. **10**, Springer, Berlin, 1978.



- 11 D. Attwood, *Soft X-Rays and Extreme Ultraviolet Radiation Principles and Applications*, Cambridge University Press, Cambridge, 2007.
- 12 *Infrared Detectors and Emitters: Materials and Devices*, P. Capper and C. T. Elliott (Eds.), Kluwer, Dordrecht, 2001.
- 13 *Handbook of Infra-red Detection Technologies*, M. Henini and M. Razeghi (Eds.), Elsevier, Amsterdam, 2002.
- 14 G. Knoll, *Radiation Detection and Measurement*, 3rd revised edn., John Wiley & Sons, Ltd, Chichester, 2000.
- 15 G. H. Rieke, *Detection of Light: From the Ultraviolet to the Submillimeter*, 2nd edn., Cambridge University Press, Cambridge, 2003.
- 16 J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing*, Wiley-Interscience, New York, 1989.
- 17 RCA photomultiplier tube types formerly manufactured by RCA are supplied by Burle Industries (now merged with Photonis).
- 18 Phillips photomultiplier tubes were formerly available from Amperex, but many of these tube types are now manufactured by Photonis. These tubes are also available from Richardson Electronics.
- 19 L. Brocco, D. Casadei, G. Castellini, *et al.*, Behavior in strong magnetic field of the photomultipliers for the TOF system of the AMS-02 space experiment, *Proceedings of ICRC 2001*: 2193, Copernicus Gesellschaft, 2001.
- 20 Static crossed-field photomultipliers used to be available from ITT. See also V. J. Koester, Improved timing resolution in time-correlated photon-counting with a static crossed-field photomultiplier, *Anal. Chem.*, **51**, 458–459, 1979.
- 21 *Photon Counting Applications, Quantum Optics, antialid Quantum Cryptography*, I. Prochazka, A. L. Migdal I, A. Pauchard, *et al.* (Eds.), *Proceedings of SPIE* – 6583, May 10, 2007.
- 22 H. R. Zwicker, Photoemissive detectors, in *Optical and Infrared Detectors*, R. S. Keyes (Ed.), Topics in Applied Physics, Vol. **19**, 149–196, Springer, Berlin, 1977.
- 23 *Photomultiplier Tubes: Basics and Applications*, 2nd edn., Hamamatsu Photonics, Hamamatsu City, Japan, 1999.
- 24 “Window” discriminators are available from EG&G and LeCroy.
- 25 *Laser Focus Buyers Guide*, published annually by Pennwell Publishing Co., 1421 South Sheridan, Tulsa, OK 74112 01460, lists a large number of suppliers of a wide range of optical detectors as does the *The Photonics Buyers’ Guide*, published annually by Photonics Spectra, Laurin Publishing Co., Berkshire Common, P.O. Box 4949, Pittsfield, MA 01202–4949.
- 26 D. Long, *Photovoltaic and photoconductive infrared detectors*, in *Optical and Infrared Detectors*, R. J. Keyes (Ed.), Topics in Applied Physics, **19**, 101–147, Springer, Berlin, 1977.
- 27 W. J. Moore and H. Shenker, A high-detectivity gallium-doped germanium detector for the 40–120 region, *Infrared Phys.*, **5**, 99–106, 1965.
- 28 F. J. Low, Low temperature germanium bolometer, *J. Opt. Soc. Am.*, **51**, 1300–1304, 1961. See also B. T. Draine and A. J. Sievers, High responsivity, low-noise germanium bolometer for far infrared, *Opt. Commun.*, **16**, 425–429, 1976.
- 29 H. Shenker, W. J. Moore, and E. M. Swiggard, Infrared photoconductive characteristics of boron-doped germanium, *J. Appl. Phys.*, **35**, 2965–2970, 1964.
- 30 E. H. Putley, Indium antimonide submillimeter photoconductive detectors, *Appl. Opt.*, **4**, 649–656, 1965.
- 31 M. A. C. S. Brown and M. F. Kimmitt, Far-infrared resonant photoconductivity in indium antimonide, *Infrared Phys.*, **5**, 93–97, 1965.
- 32 P. Norton, HgCdTe infrared detectors, *Opto-Elect. Rev.*, **10**(3), 159–174, 2002.
- 33 G. L. Hansen, J. L. Schmit, and T. N. Casselman, Energy gap versus alloy composition and temperature in  $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ , *J. Appl. Phys.*, **53**, 7099–7101, 1982.
- 34 J. Gowar, *Optical Communication Systems*, 2nd edn., Prentice-Hall, Englewood Cliffs, NJ, 1993.
- 35 H. Dautet, P. Deschamps, B. Dion, *et al.*, Photon-counting techniques with silicon avalanche photodiode, *Appl. Opt.*, **32**(21), 3894–3900, 1993.
- 36 S. Cova, A. Lacaita, F. Zappa, and P. Lovati, Avalanche photodiodes for near-infrared photon counting in *Advances in Fluorescence Sensing Technology II*, J. Lakowicz (Ed.), Proc. SPIE 2388, paper 09, SPIE Conf. Photonics West, San Jose, CA, USA, Feb. 6–7, 56–66, 1995.
- 37 A. Lacaita, F. Zappa, S. Cova, and P. Lovati, Single-photon detection beyond 1  $\mu\text{m}$ : performance of commercially available InGaAs/InP detectors, *App. Opt.*, **35**, 2986–2996, 1996.
- 38 S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, Avalanche photodiodes and quenching circuits for single-photon detection, *Appl. Opt.* **35**, 1956–1976, 1996.
- 39 T. Nesheim, *Single Photon Detection Using Avalanche Photodiode*, M.Sc. Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 1999.
- 40 H. J. Woltring, Single- and dual-axis lateral photodetectors of rectangular shape, *IEEE Transactions on Electron Devices*, **22**(8), 581–590, 1975.

- 41 *Solid-State Imaging with Charge-Coupled Devices* (Solid-State Science and Technology Library), A. J. Theuwissen (Ed.), Kluwer, Dordrecht, 2002.
- 42 Tracy V. Wilson, K. Nice, and G. Gurevich, How digital cameras work. November 29, 2006 <http://www.howstuffworks.com/digital-camera.htm> (June 02, 2007).
- 43 T. Klinger, *Image Processing with Lab VIEW and IMAQ Vision*, Prentice-Hall, Englewood Cliffs, NJ, 2003.
- 44 R. J. McIntyre, Multiplication noise in uniform avalanche diodes, *IEEE Trans. Electron. Devices*, **ED-13**, 164–169, 1966.
- 45 Hamamatsu Photon Counting Application Note. [http://sales.hamamatsu.com/assets/applications/ETD/PhotonCounting\\_TPHO9001E04.pdf](http://sales.hamamatsu.com/assets/applications/ETD/PhotonCounting_TPHO9001E04.pdf)
- 46 C. C. Davis and T. A. King, Correction methods for photon pile-up in lifetime determination by single-photon counting, *J. Phys. A*, **3**, 101–109, 1970.
- 47 J. Milman and H. Taub, *Pulse, Digital, and Switching Waveforms*, Chap. 3, McGraw-Hill, New York, 1965; C. N. Winningsstad, *IRE Trans. Nucl. Sci.*, **NS43**, 26, 1959; C. L. Ruthroff, *Proc IRE*, **47**, 1337, 1959.
- 48 L. J. Richter and W. Ho, Design and performance of a double pass high resolution electron energy loss spectrometer, *Rev. Sci. Instrum.*, **57**, 1469, 1986; R. M. Tromp, M. Copel, M. C. Reuter, A new 2D particle detector for a toroidal electrostatic analyzer, *Rev. Sci. Instrum.*, **62**, 2679, 1991.
- 49 P. Horowitz and W. Hill, *The Art of Electronics*, 2nd edn., Cambridge University Press, Cambridge, 1989.
- 50 V. Radeka, Low noise techniques in detectors, *Annual Review of Nuclear and Particle Science*, J. D. Jackson (Ed.), Vol. **38**, 1988, pp. 217–278.
- 51 R. L. Heath, R. Hofstadter and E. B. Hughes, Inorganic scintillators: a review of techniques and applications, *Nucl. Instr. Meth.*, **162**, 431–476, 1979.
- 52 J. B. Birks, *Theory and Practice of Scintillation Counting*, Pergamon Press, New York, 1964.
- 53 R. K. Bock and A. Vasilescu, *The Particle Detector Briefbook*, Springer, Berlin, 1998.
- 54 G. Knoll, *Radiation Detection and Measurement*, 3rd revised edn., John Wiley & Sons, Ltd, Chichester, 2000.
- 55 E. H. Putley, Thermal Detectors, in *Optical and Infrared Detectors*, R. J. Keyes (Ed.), Topics in Applied Physics, Vol. **19**, Springer, Berlin, 1977.
- 56 A. F. Gibson, M. F. Kimmitt, and A. C. Walker, Photon drag in germanium, *Appl. Phys. Lett.*, **17**, 75–77, 1970. See also

- R. Kesselring, A. W. Kalin, and F. K. Kneubuhl, Fast midinfrared detectors, *Infrared Phys.*, **33**, 423–436, 1992.
- 57 M. J. E. Golay, A pneumatic infra-red detector, *Rev. Sci. Instrum.*, **18**, 357–362, 1947.

## General References

- D. Attwood, *Soft X-Rays and Extreme Ultraviolet Radiation Principles and Applications*, Cambridge University Press, Cambridge 2007.
- R. K. Bock and A. Vasilescu, *The Particle Detector Briefbook*, Springer, Berlin, 1998.
- P. Capper and C. T. Elliott (Eds.), *Infrared Detectors and Emitters: Materials and Devices*, Kluwer, Dordrecht, 2001.
- C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
- C. F. G. Delaney and E. C. Finch, *Radiation Detectors*, Oxford Science Publications, Clarendon Press, Oxford, 1992.
- C. Enss, *Cryogenic Particle Detection*, Topics in Applied Physics, Vol. **99**, Springer, Berlin, 2005.
- R. C. Fernow *Introduction To Experimental Particle Physics*, Cambridge University Press, Cambridge, 1989.
- M. Henini and M. Razeghi (Eds.), *Handbook of Infra-Red Detection Technologies*, Elsevier, Amsterdam, 2002.
- R. D. Hudson, Jr., and J. W. Hudson (Eds.), *Infrared Detectors*, Benchmark Papers in Optics, Vol. **2**, Dowden, Hutchinson and Ross, Stroudsburg, PA, 1975.
- K. Kleinknecht *Detectors For Particle Radiation* 2nd revised edn., Cambridge University Press, Cambridge, 1998.
- R. H. Kingston, *Detection of Optical and Infrared Radiation*, Springer Series in Optical Sciences, Vol. **10**, Springer, Berlin, 1978.
- G. Knoll, *Radiation Detection and Measurement*, 3rd revised edn., John Wiley & Sons, Inc., 2000.
- P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, 1962.
- R. J. Keyes (Ed.), Topics in Applied Physics, *Optical and Infrared Detectors*, Springer, Berlin, 1977.
- G. H. Rieke, *Detection of Light: From the Ultraviolet to the Submillimeter*, 2nd edn., Cambridge University Press, Cambridge, 2003.
- J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing*, Wiley-Interscience, New York, 1989; D. Wood, *Optoelectronic Semiconductor Devices*, Prentice-Hall, New York, 1994.

## MEASUREMENT AND CONTROL OF TEMPERATURE

There are two levels of concern with temperature in scientific experiments. One is the control of temperature to achieve some secondary (but essential) aim in the apparatus. Examples are the use of a cold trap in a vacuum line and the use of heaters and coolants in a distillation column. For such needs, the temperature needs only to be known and kept constant to within a few kelvins. In the second case, the measurement of the dependence of physical parameters on temperature is a primary aim of the experiment. A physicist learns about the nature of a material by measuring such properties as density or heat capacity as a function of temperature. An organic chemist studies the kinetics of a chemical reaction by measuring its rate of reaction as a function of temperature. For these experiments, the temperature must be varied over a range and controlled at any point in that range, at resolutions better than 1 K.

Sometimes the temperature must be known *accurately*. That is, the measurement must be closely calibrated to the International Temperature Scale. Accurate measurements are necessary if the new data are to be used with other measurements on the system under study. If measurements of the density of water are to be combined with measurements of its kinematic viscosity to calculate its shear viscosity as a function of temperature, then the temperature must be measured with the same accuracy in both the density measurements and the kinematic viscosity measurements. The accuracy also matters if measurements in one laboratory are to be compared to measurements in another laboratory of that same quantity (e.g., the shear viscosity as a function of temperature).

For some purposes, the *precision* is more important than the accuracy. Such is the case when it is the derivative of a property with respect to temperature that is of interest. Then the changes in temperature need to be measured with very fine resolution, but the accuracy, the exact temperature on the international scale, is not as important. A physicist wishing to test a theoretical prediction for the thermal expansion (the change of volume with temperature) of a superconducting solid will be more concerned with precision than with accuracy, even though he or she will have some requirements on the accuracy.

In this chapter we will describe how to measure and control the temperature for simple needs, usually meaning a resolution of 1 K. We will also describe more precise and accurate methods for temperature measurement and control, methods that achieve a resolution of 1 mK and better. Since we must be able to measure the temperature in order to control it, we will first discuss thermometry. For both measurement and control, we will focus on the temperature range from 10 K to about 1800 K, and on methods that are practical for most laboratories. We will briefly discuss thermometry at very low and very high temperatures.

### 8.1 THE MEASUREMENT OF TEMPERATURE

A thermometer can be constructed from a system with a measurable property that depends strongly on temperature. We will discuss here the “working” thermometers useful

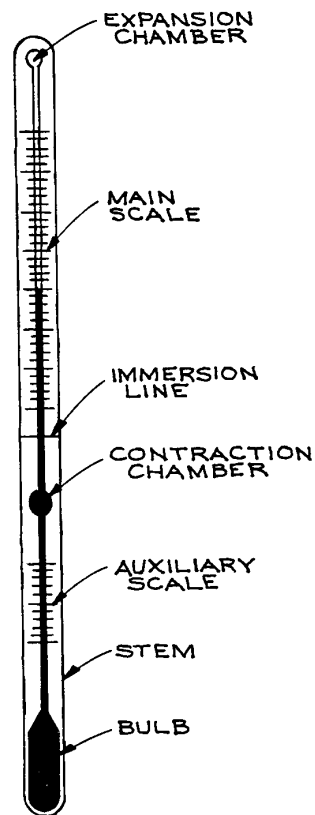
in most laboratories, as opposed to special “standard” thermometers, such as gas thermometers, used in calibration laboratories.<sup>1-4</sup>

### 8.1.1 Expansion Thermometers

*Expansion thermometers* use as the thermometric property the volume of the thermometric system. For example, the common mercury thermometer uses the change in volume (as reflected in the change in length) of a column of mercury confined to a cylindrical capillary tube. There exist expansion thermometers using the thermal expansion of a metal that have industrial applications.<sup>2</sup> We will limit ourselves to the *liquid-in-glass thermometers* used in scientific research.<sup>5,6</sup> Liquid-in-glass thermometers date to the seventeenth century<sup>1</sup> and are still common in research laboratories. They do not lend themselves readily to computer automation of the measurement process, but they are useful for approximate measurements that do not need to be automated, or as laboratory standards for the calibration of other thermometers.

Such thermometers are limited in range by the freezing and boiling points of the liquids used and by the softening point of the glass in which the liquid is contained. Mercury has been the most common thermometer fluid because of its nearly linear coefficient of thermal expansion, but organic liquids are also used and are safer. Design considerations also include the choice of the glass and the relative dimensions of the parts of the thermometer. Thermometers can be designed for high sensitivity in a particular range by including in the liquid column a section of larger diameter, a *contraction chamber*, so that the liquid expands in that volume before it reaches the range of interest (see Figure 8.1). Mercury thermometers can operate in the range 235–923 K. Organic liquids have the disadvantages that they do not have very linear thermal expansions and they must be dyed to be visible, but they are still useful: toluene (183 to 373 K), ethanol (163 to 373 K), pentane (73 to 293 K). Thermometers made with mercury and with organic liquids are available with resolutions as good as 0.02 K (Brooklyn Thermometer).

There are a number of difficulties in obtaining high accuracy and precision with liquid-in-glass thermometers. Good thermal contact between the thermometer and the system of interest is difficult to achieve for systems other



**Figure 8.1** Mercury-in-glass thermometer. The auxiliary scale is designed to allow calibration at the ice point. The contraction chamber permits high sensitivity (a small capillary diameter) without having a very long stem. The immersion line indicates the depth to which the thermometer was immersed in the calibration bath. The expansion chamber prevents the buildup of pressure and avoids breakage on overheating.

than fluids. The thermometer can be sensitive to pressure changes. If the thermometer is heated to its highest temperatures, it will require days to return to its original calibration because of hysteresis in the glass.

Mercury thermometers designed to be accurate and stable to about 0.01 K over temperature ranges of 3 to 5 K are commercially available (Brooklyn Thermometer, Omega Engineering) and make good, cheap working temperature standards. Mount the mercury thermometer in a stirred bath, the temperature of which is under control (Section 8.2), along with any other thermometers that you want to

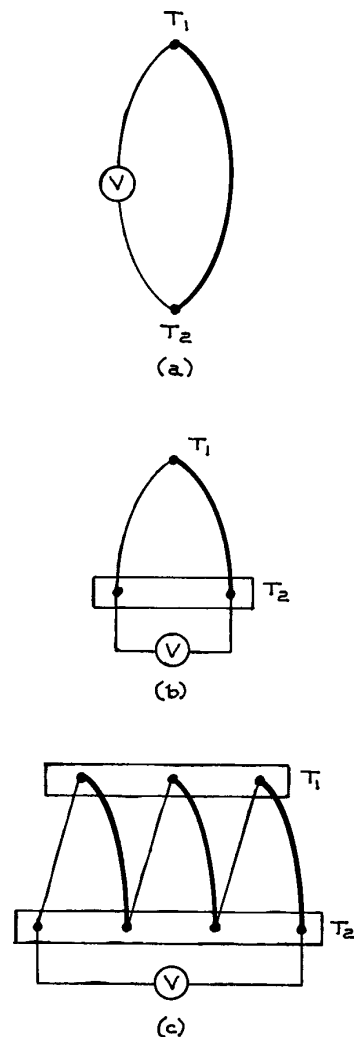
calibrate against the mercury thermometer. Be careful in reading the meniscus position to avoid parallax error; a telescope makes this easier. Use the thermometer at the same immersion depth at which it was originally calibrated; a difference in immersion depth will necessitate stem corrections. Tap the thermometer before each measurement to avoid sticking of the mercury column to the stem wall. Monitor the stability of the thermometer by regularly measuring the ice point (Section 8.1.9).

A final hint: if the liquid in a thermometer becomes separated, it can be rejoined by cooling the thermometer bulb with dry ice shavings until all the liquid is back in the bulb. Never heat a liquid-in-glass thermometer to expand the liquid to the full length of the capillary tube in an effort to rejoin the liquid: the thermometer will break.

### 8.1.2 Thermocouples<sup>7-10</sup>

*Thermocouples* use as the thermometric property the voltage that develops between two junctions of two different metals when the junctions are at different temperatures.<sup>11</sup> Thermocouple thermometers have been used since about 1885,<sup>12</sup> and still have uses ranging from hand-held digital thermometers to sophisticated calorimeters.

When there exists a difference in electron density between two points in a metallic conductor, a voltage develops between those two points. Such a difference in electron density can arise even in a conductor made of a homogeneous metal, because the electron density varies when the temperature varies. Thus, a temperature gradient across any metallic material causes a voltage to develop across that material. If we try to measure this voltage by attaching leads made of the same metal, then the voltage difference will also develop in the leads and the effects will cancel. If the leads are of a different metal, as shown in Figure 8.2(a), then the difference in the thermally induced voltages between the two metals will cause a net voltage between the junctions: the *Seebeck effect*.<sup>13</sup> This thermoelectric voltage, or “*thermal emf*,” depends upon the particular choice of metals and upon the difference in temperature between the junctions, but depends only slightly on the pressure. The Seebeck effect is a bulk property of metals in a temperature gradient. It is not a property of the junctions themselves: it is not a “contact potential,” and treating it as such can lead to mistakes in the use of



**Figure 8.2** Thermocouple configurations. Lines of different thicknesses represent wires of different metals or alloys.  $T_1$  and  $T_2$  are two different temperatures. V is a voltmeter.

thermocouples.<sup>14</sup> A related property is the *Peltier effect*:<sup>15</sup> when a current is passed through a circuit such as Figure 8.2(a), by replacing the voltmeter with a battery, then one junction will be heated and the other junction will be cooled. The Peltier effect can be used to construct electrical heating and cooling devices.

The connection of two wires of *two different metals* at junctions at *two different temperatures*, as in Figure 8.2(a),

is called a *thermocouple*. The voltage developed across the thermocouple is a measure of the difference in temperature,  $(T_1 - T_2)$ . Thus, the measurement of the temperature with a thermocouple requires only a voltmeter of sufficient resolution. The relationship between the thermoelectric voltage and the temperature difference, while nearly a direct proportionality, cannot be predicted, but must be measured for a given pair of metals. The nature of the Seebeck effect leads to several principles useful in the design of such thermocouple thermometers:<sup>2,16</sup>

- (1) *The law of homogeneous metals:* a thermocouple requires two different metals. If the two wires are of the same material, then no net thermal voltage will develop, even in the presence of a temperature gradient. Therefore, voltmeter lead wires, as shown in Figure 8.2(b), will not contribute to the measured voltage if they are made of the same metal, and if their ends (at  $T_2$  and at the voltmeter) are at the same temperatures so that they are under the same temperature gradient. On the other hand, inhomogeneities due to strains or impurities in a metal conductor will be, in effect, regions of a different metal, and will, in the presence of a temperature gradient, act as “mini-thermocouples,” generating spurious thermal voltages and leading to inaccuracies in the thermometry.
- (2) *The law of intermediate metals:* a thermocouple requires two different temperatures. If there is no difference in temperature between the ends of a pair of wires, then no net thermal voltage will develop between them, even if the wires are of different metals. That is, in Figure 8.2(a), if  $T_1 = T_2$ , no voltage develops. A consequence of this law is that another wire may be inserted into the thermocouple circuit, so long as both ends of that wire are at a common temperature, that need not be either of the temperatures involved in the actual thermocouple ( $T_1$  or  $T_2$ ). This means that thermocouple junctions may be made with any kind of solder or any other method of attachment, so long as each junction is itself all at the same temperature.
- (3) *The law of successive metals:* if thermocouples are made of pairs of metal A with metal B, metal B with metal C, and metal A with metal C, then the voltages developed by each of the thermocouples between a given pair of junction temperatures will be related as

$V_{AC} = V_{AB} + V_{BC}$ . This equation allows the calculation of tables of thermocouple voltages for any pair of metals A and C if the tables are already available for A and C with respect to a reference metal B.

- (4) *The law of successive temperatures:* if a given thermocouple develops a voltage  $V_a$  between  $T_1$  and  $T_2$  and a voltage  $V_b$  between  $T_2$  and  $T_3$ , then it will develop a voltage  $V_c = V_a + V_b$  between  $T_1$  and  $T_3$ . It follows that only the junction temperatures matter: intermediate temperatures along homogeneous wires do not matter. It also follows that if thermocouple tables are available in which voltages are given for each temperature  $T_1$  with respect to a reference temperature  $T_2$ , then voltages may be calculated with respect to a different reference temperature  $T_3$  if the voltage  $V_b$  is known.

Thermocouple junctions may be placed in series in order to amplify the voltage signal. Such a *thermopile* is shown in Figure 8.2(c), where there are three identical thermocouples in series, producing a voltage three times that of a single thermocouple (since voltages in series add).

There are a great many possible thermocouple pairs, since any two metals or alloys can form a thermocouple. The choice of a particular pair will depend upon its sensitivity in the temperature range of interest and upon such considerations as corrosion resistance. Several thermocouple pairs have come into common use; their properties are well known, and good-quality wire is readily available (Omega Engineering; Watlow).

Table 8.1 gives the most common thermocouple types and their properties. Types C, D, and G are all various combinations of alloys of tungsten and rhenium, and are useful at high temperatures because these metals do not readily vaporize. Types R and S are of platinum with platinum–rhodium alloy and have low sensitivity, but have high stability and accuracy. Types K, N, and E all use nickel alloys. Type K has a wide range of usefulness, low thermal conductivity, and good resistance to oxidation, but cannot be used in a reducing atmosphere above 1100 K and shows instability in the calibration due to a phase transition in Chromel at 570 K. Type N is a more stable substitute for Type K. Type E is common in low-temperature applications because the thermal conductivities of the materials are low, and thus the wires do not carry heat into the apparatus. Types E, J, and T all use the

**Table 8.1 Properties of common thermocouples**

Type <sup>a</sup>	Materials	Useful Range (K)	Sensitivity ( $\mu\text{V/K}$ )	
			20 K	300 K
C, D, G	Tungsten/Rhenium	32–2600	—	20
R, S	Platinum/PtRh	220–1800	—	6
K	Chromel <sup>b</sup> /Alumel <sup>c</sup>	10–1500	4.1	41
N	Nicrosil/Nisil <sup>e</sup>	10–1500	3	26
E	Chromel <sup>b</sup> /Constantan <sup>d</sup>	10–700	8.5	61
J	Iron/Constantan <sup>d</sup>	60–1000	—	52
T	Copper/Constantan <sup>d</sup>	20–700	4.6	41
	Chromel <sup>b</sup> /AuFe	10–1000	15	20
	Chromel <sup>b</sup> /CuFe	10–300	>11	

<sup>a</sup> The type designation is that of the Instrument Society of America

<sup>b</sup> Chromel is an alloy of nickel with 10% chromium

<sup>c</sup> Alumel is an alloy of nickel with 2% aluminum, 2% manganese, and 1% silicon

<sup>d</sup> Constantan is an alloy of copper with 43% nickel

<sup>e</sup> Nicrosil is a alloy of nickel, chromium, and silicon; Nisil is an alloy of nickel, magnesium, and silicon.

copper–nickel alloy Constantan. Type J has the advantage of high sensitivity and the disadvantage that iron is readily oxidized. Type T, probably used most often, has the advantage that one material is copper, from which the voltmeter leads are also likely to be made, so that the chance of extraneous thermal voltages is reduced. On the other hand, copper oxidizes above 620 K and has a very high thermal conductivity (see Table 8.5 below). The last two Chromel thermocouples in Table 8.1, one with gold–iron alloy and one with copper–iron alloy, are used at cryogenic temperatures, along with types E, K, and T (Section 8.1.5). Thermocouples made of wires from two different elemental metals (gold, silver, platinum, etc.) can be used for cases requiring very high stability and precision.<sup>4,17</sup>

The construction of reliable thermocouple thermometers requires care. Use good-quality, annealed thermocouple wire. Each wire can be individually tested for homogeneity by connecting its ends to a voltmeter and passing the wire slowly through liquid nitrogen; wires that do not show any measurable thermoelectric voltage are suitable for use. Use the largest-diameter wire that you

can tolerate: large wire is more likely to be homogeneous and is less susceptible to strains. Make the junctions mechanically sound and electrically and thermally continuous. In principle, thermocouple wires can be twisted or clamped to make a junction.<sup>9</sup> In practice, better results are obtained if they are soldered or welded (see Section 1.4). For work below 443 K, Type T can be joined with ordinary tin–lead solder, using a rosin flux. All can be soldered with low melting tin–silver solder. For work at higher temperatures and for better mechanical strength, all can be brazed with a high-melting silver alloy (Section 1.4.6). All but Type T can be spot-welded. All can be welded with a torch, using a reducing flame of oxygen with either natural gas or acetylene to melt the two wires together into a firm bead. Ready-made thermocouples, thermopiles, and thermopile arrays can be purchased from Omega Engineering, SensArray, InstruLab, Watlow, *et al.*

Make good thermal contact between the junctions and the points at which the temperatures are to be measured. Junctions can be mechanically attached to solids using a thermally conducting grease, or bonded with a thermally conducting adhesive (Cotronics, Omega Engineering, Wakefield). Thermocouples are often placed in protective metal or ceramic sheaths (*thermowells*),<sup>16</sup> but such arrangements increase the response time. Protect the wires from mechanical stresses. The wires can be electrically insulated with Teflon (to 553 K) or Kapton (to 700 K); fiberglass or ceramic is required for higher temperatures. Avoid air drafts and temperature gradients in the vicinity of a thermocouple circuit.

A thermocouple or thermopile produces a low-level d.c. voltage (10–100  $\mu\text{V}$ ) that requires particular care in measurement (see Keithley handbook<sup>18</sup> and application notes). A digital voltmeter of appropriate sensitivity suffices for the measurement of the thermocouple voltage (Keithley, Agilent). Follow appropriate procedures for shielding and grounding in the measurement circuitry (Section 6.9). Commercial thermocouple amplifiers are available for \$2 to \$20 (Analog Devices, Linear Technology). Sometimes (e.g., for control purposes) it is desirable to linearize the output signal, and such circuits are also available (Analog Devices).

Thermocouples measure the temperature differences between junctions. This property can be used to advantage when a difference is just what is required, as in some

temperature-control circuits. If, instead, an absolute temperature measurement is required, then one junction (the *reference junction*) must be kept at a constant, known temperature. For example, let  $T_2$  in Figure 8.2 be the reference junction of known  $T_2$ . Then the absolute temperature at  $T_1$  is obtained by combining the known  $T_2$  and the difference ( $T_1 - T_2$ ) as obtained from the thermocouple voltage (see below). A simple (but inconvenient) way to do this is to keep the reference junction at 273.15 K by immersing it in a slurry of ice and water (all from distilled water). The reference temperature can also be established with commercially available fixed point cells that are made from Peltier devices (Hart Scientific, Omega Engineering; \$1200).

Two less-expensive methods are either to make an *isothermal block*, or to use an electronic compensation circuit. An isothermal block consists of a material of high thermal conductivity (Table 8.5) onto which the reference junction and an independent thermometer are mounted with good thermal contact, and then surrounded by thermal insulation (Table 8.5). Then the independent thermometer is used to measure the reference temperature. This is useful, even though it requires an independent thermometer, because that thermometer only needs to function near room temperature, but then permits thermocouple measurements of temperatures far from room temperature.

An *electronic compensation circuit* or “electronic ice point” is an electronic circuit that measures the temperature at the reference junction and then adds a voltage to the thermocouple circuit that compensates for the fact that the reference junction is not actually at 273.15 K. These “cold junction compensation” circuits can be made from components, or purchased. Analog Devices and Linear Technology make integrated circuits (AD594, AD595, LT1025) that include both compensator and amplifier for \$15–\$40, and Omega makes complete devices for \$200–\$600. For the best thermocouple measurements of temperature, reference junctions made from ice baths or other fixed-point baths (Sections 8.1.9; 8.2.1) permit more accurate measurements ( $\pm 0.01$  K) than do these isothermal blocks and electronic ice points ( $\pm 0.4$  to  $\pm 1$  K).

In order to convert the thermocouple voltage to a temperature, it is necessary to use conversion tables or to calibrate directly. The National Institute of Standards and Technology has published extensive tables of thermo-

couple voltages. A calibration using the tables is likely to be accurate only to a few degrees, due to variations among wires. For more accurate measurements, the thermocouple must be compared with a standard thermometer (usually a platinum resistance thermometer; see Section 8.1.3 below) over the temperature range of interest. Voltage-temperature points either from the tables or from a direct calibration can be fitted by a polynomial:

$$T = a_0 + a_1V + a_2V^2 + \dots \quad (8.1)$$

where  $T$  is the temperature,  $V$  is the voltage, and enough terms are included to describe the data. The calibration determines the accuracy of the temperature measurement on the International Temperature Scale (Section 8.1.9). The precision of temperature measurements with thermocouples is considerably better than the accuracy: temperature differences can be resolved to 1 mK, using thermopiles and sensitive amplifiers/voltmeters.

The advantages of thermocouples as thermometers are: they are simple, rugged, and inexpensive; thermopiles can have resolutions of 1 mK; they can be used over a very broad range of temperature (see Table 8.1); they are small; they respond quickly. The disadvantages are: the voltage signals are small and hard to measure; accuracy is hard to attain and maintain; a reference junction is necessary for absolute temperature measurements; they are very sensitive to magnetic fields.

### 8.1.3 Resistance Thermometers

*Resistance thermometers* use as the thermometric property the dependence of electrical resistance on temperature. There are two main kinds of *resistance temperature detectors* (RTDs): conducting metal RTDs and semiconductor RTDs. Metal RTDs have been in use since the late nineteenth century,<sup>19</sup> whereas semiconductor RTDs came into use only in the twentieth century. Resistance temperature detectors are invaluable in modern thermometry because they are small and stable, and provide an easily measured electrical quantity.

**Metal Resistance Thermometers.**<sup>6</sup> Metal resistance thermometers consist of wire coils or thin films of electrically conducting metals or alloys, together with the



instrumentation needed to measure the electrical resistance. As the temperature increases, the increase in kinetic energy of the metal atoms leads to an increase in the electrical resistivity. The larger the initial resistance, the larger the change of resistance with temperature, and thus the greater the sensitivity.

While a number of metals can be used for thermometry (Table 8.2), platinum is most commonly used because it is resistant to corrosion and its electrical resistance has a nearly linearly increase with temperature of 0.4%/K.<sup>20</sup> Indeed, the *platinum resistance thermometer (PRT)* is the standard device for the International Temperature Scale of 1990 between 13.8033 K and 961.78 K (see Section 8.1.7).<sup>21</sup> Some typical platinum resistance thermometers are shown in Figure 8.3. Usually, PRTs have resistances of 25 or 100  $\Omega$  at room temperature, but thin-film PRTs can have resistances of 1000 to 2000  $\Omega$ . The resistance as a function of temperature behaves as (see Figure 8.4):

$$R = R_0 + R_1T + R_2T^2 + \dots \quad (8.2)$$

where the  $T^2$  term is important for measurements with a resolution of millikelvins. With care in the resistor construction and in the measurement circuitry, a resolution of 0.1 mK can be achieved over a very broad range of temperatures with platinum resistance thermometers (see Figure 8.4 and Table 8.2), available from suppliers such as Barnant, Hart, Lake Shore Cryotronics, Minco, Omega Engineering, and Rosemount. Thin, surface-mount foil resistance thermometers made from platinum are also available (Minco).

Resistance thermometers are also made of other metals (see Table 8.2). Nickel has a high sensitivity to temperature change, but the dependence of resistance on temperature is nonlinear, making its use over a large temperature range more difficult.<sup>22–24</sup> Copper has a low resistivity, so a longer wire is required to achieve a useful resistance, but its resistance is fairly linear in temperature.<sup>25,26</sup> Copper and nickel thermometers are made by Thermometrics. The other materials in Table 8.2 are used for cryogenic thermometry and will be discussed below in Section 8.1.5. New materials, such as niobium thin films, are under development for resistance thermometry.<sup>27</sup>

### Semiconductor Resistance Thermometers.

*Thermistors*<sup>12,16</sup> are small beads of semiconducting material (metal oxides), the resistance of which depends on temperature (Table 8.3). While electrons flow freely in a metal, in a semiconductor, the electrons must first be excited into a conducting state; the electrons will thus flow more freely and the resistance will decrease as the temperature increases. The term “negative temperature coefficient” or “NTC” is used. The resistance of a thermistor decreases at a rate of about 4%/K; a thermistor is 10 times more sensitive than a PRT. A thermistor differs from a metal resistance thermometer in that its electrical resistance is much more sensitive to temperature change, its resistance decreases as temperature increases, and its resistance is not linear in its temperature dependence (see Figure 8.4). There exist thermistors that have positive temperature coefficients of resistance, for which the

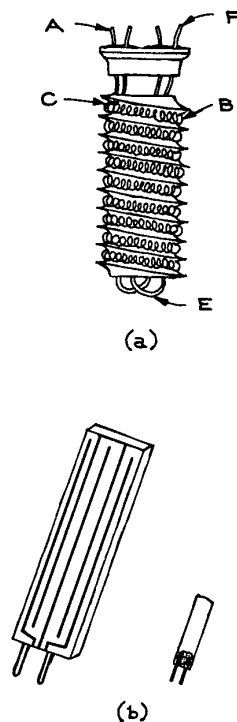
**Table 8.2 Properties of common resistance thermometers<sup>3</sup>**

Materials	Resistivity ( $\Sigma m$ )H $10^8$ at 273 K	Temperature Coefficient of Resistance ( $K^{-1}$ )H $10^3$	Useful Range (K)	Sensitivity <sup>a</sup> (mk)	
				20 K	300 K
Platinum	11	3.9	20–1000	10	0.1
Nickel	59	6.0	213–423		
Copper	1.7	3.9	173–373		
Ruthenium oxide			0.04–40	5 <sup>c</sup>	—
Rhodium-iron <sup>b</sup>	42 <sup>c</sup>		0.5–300	0.05	5

<sup>a</sup> Approximate resolution under typical conditions

<sup>b</sup> Not for use in magnetic field

<sup>c</sup> At 4.2 K



**Figure 8.3** Platinum resistance thermometers: (a) standard, high-resolution model made by Minco Products, Inc., in which the platinum is wound as a wire coil on a ceramic base. Two leads (A) pass through the lid and attach to one end of the coil (B) at C. At the other end, two leads are again attached (E) and pass through the center of the base to the lid at F. The assembly is then enclosed in an inert atmosphere. This PRT is 3/8" long; (b) Film RTD such as those made by Omega Engineering, Inc. These are called Athick@ and Athin@ film RTDs, respectively. They can be very small, 1/8" or less in length.

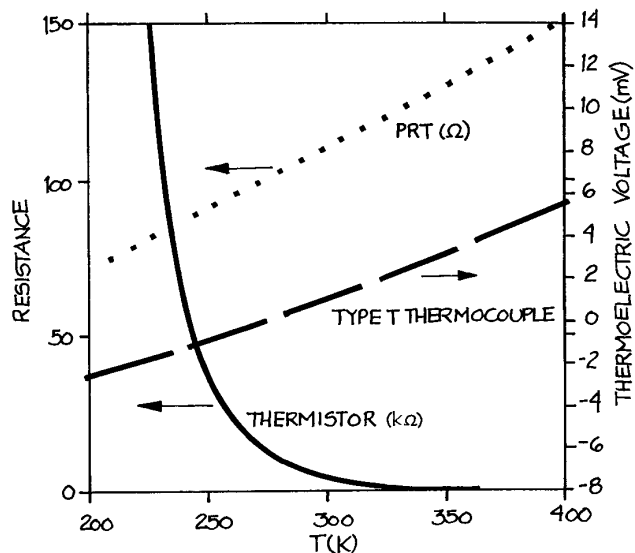
resistance increases with temperature as for metals; they are not useful for thermometry, but are used for circuit protection devices.<sup>16</sup>

The precision/accuracy is that to be expected without extraordinary attention to purity and pressure control.

The Steinhart–Hart equation provides a good empirical description of a thermistor (see Figure 8.4):<sup>28,29</sup>

$$1/T = A + B(\ln R) + C(\ln R)^3 + \dots \quad (8.3)$$

where  $A$ ,  $B$ , and  $C$  are empirical constants and where higher-order terms of odd power may be required for a



**Figure 8.4** The left ordinate gives the resistance as a function of temperature  $T$  for a thermistor (solid line) which has a room temperature resistance of about 4 k $\Omega$ , and for a platinum resistance thermometer, PRT, (dotted line) which has a room temperature resistance of about 100  $\Omega$ . Note that the scales differ by a factor of 1000. The right ordinate gives the thermoelectric voltage referenced to 273 K of a Type T copper/constantan thermocouple (dashed line) as a function of temperature.

broad temperature range or for high resolution. If a linearized resistance is required, then linearizing circuits can be made<sup>12,16</sup> or bought. For example, the YSI model 4800 linearizes thermistor output to  $\pm 0.1$  K from 273 K to 373 K (\$75). (YSI and Betatherm are now subdivisions of Measurement Specialties.) Linearizing circuits for thermistors are also available from Omega Engineering.

Typical thermistor resistances are 1 to 10 k $\Omega$ . Thermistors can readily resolve a millikelvin and can be made to resolve a microkelvin. Thermistors can be very small (1 mm), and very stable with time and temperature cycling.<sup>30</sup> Their disadvantage is that their resistance rapidly decreases with increasing temperature and thus their usefulness decreases above about 373 K, but thermistors designed for use up to 573 K are available (Measurement Specialties, Adsem).

**Table 8.3 Properties of common semiconductor thermometers**

Type	Materials	Useful Range (K)	Sensitivity <sup>a</sup> (mK)	
			20 K	300 K
Semiconductor	Thermistor	77–600	—	0.01
	Germanium <sup>b</sup>	1–30	0.5	—
	Carbon	0.04–300	1	10 <sup>3</sup>
	Carbon-in-glass	1–100	0.75	—
Diode	Silicon <sup>b</sup>	0.01–300	20	15
	Gallium arsenide <sup>b</sup>	0.01–300	15	100

<sup>a</sup> Approximate resolution under typical conditions

<sup>b</sup> Not for use in magnetic field

Interchangeable thermistors (matched as well as  $\pm 0.1$  K over 50 K), ultrastable thermistors (changing  $< 0.01$  K in 100 months at 298 K), and thin surface-mount thermistors are also available (Measurement Specialties 46000 series; Thermometrics). Thermistor-based thermometers have been developed for measurements in radio-frequency-field environments.<sup>31</sup>

**Measurement Systems For Resistance Thermometry.**<sup>32</sup> Complete resistance thermometry systems are commercially available for lower resolution (0.01 to 1 K) measurements (Anton Paar, Barnant, Hart Scientific, Omega Engineering, Techno, Yokogawa, Newport Instruments) and for higher resolution measurements to 1 mK (Anton Paar, Guildline Instruments, Hart Scientific, Instrulab), or to 0.25 mK (Hart Scientific). Nevertheless, it is useful to understand the considerations in constructing the resistance-measuring equipment so that one may take advantage of equipment one already has, or make a thermometer for a special need, such as for very high resolution in a particular temperature range.

Resistance thermometry requires the measurement of electrical resistance, usually by putting a known current through the resistance thermometer and then measuring the voltage developed across it. The first consideration for a resistance thermometer is the stability over time of the resistor itself (RTD or thermistor). The stability requirements depend on the resolution required and deter-

mine the price. As discussed above, thermistors and RTDs have been developed that drift less than a few mK per year. Mechanical stresses caused by mounting (e.g., use of hard adhesives) can, however, cause significant drifts in resistance thermometers.

The second consideration for a resistance thermometer is its *self-heating*.<sup>14</sup> When a current flows through a resistor, there occurs Joule heating of  $I^2R$ , where  $I$  is the current and  $R$  is the resistance. The current in the thermometric resistor must be sufficiently low that the resistor itself is not heated above the temperature of its surroundings. The *dissipation factor* of the resistor specifies the self-heating. For example, a typical dissipation factor for a thermistor is 4  $\mu\text{W}/\text{mK}$ . This means that a current resulting in 4  $\mu\text{W}$  of power will heat the thermistor by 1 mK. To obtain a resolution of 1 mK with a 4 k $\Omega$  thermistor, the current is thus limited to  $3 \times 10^{-5}$  A. For a 25  $\Omega$  PRT, the current limit is about 3 mA for a resolution of 1 mK.

The third consideration is the possibility of thermal emfs in the thermometer circuit. These are the very effects that make thermocouple thermometers possible (Section 8.1.2), but as extraneous voltages they can interfere with resistance thermometry when the resistance is determined by measuring a d.c. voltage. Thermal emfs may be evaluated by reversing the current in the resistance circuit: the effect will add for one current direction and subtract for the other. If such effects are present, they may be corrected by averaging the two readings or by using an a.c. resistance-measuring method.

The fourth consideration is the effect of the resistor leads. If the resistance of the lead wires enters into the resistance measurement, then that resistance and its temperature dependence can cause an error in the thermometry. Lead effects can be minimized by using short wires of large diameter. They can be eliminated by using *four-wire measurement* techniques, in which two wires carry the current through the resistor and two other wires are used to measure the voltage across it.<sup>14</sup> If the voltage is measured with a high-impedance device, then no significant current is carried in the voltage leads, no significant voltage drop occurs across those leads, and there is no lead error.

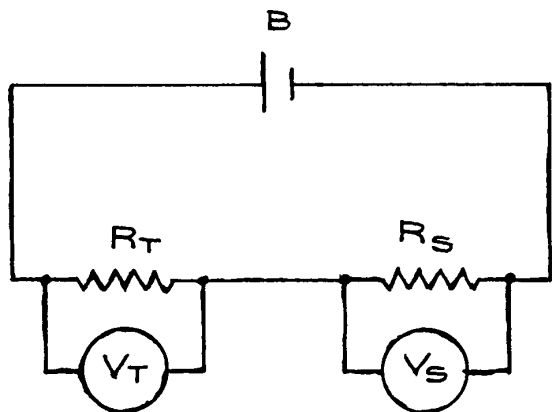
The design of precision resistance-measuring circuits is an art that changes as technology advances. We will not discuss the details of such designs here, but refer the reader to the literature<sup>1,12,14,16</sup> and to the manufacturers of

resistance instruments (Guildline, Lake Shore Cryotronics, Stanford Research; \$2500–\$8000), and of digital multimeters (Agilent, BK Precision, Keithley; \$700–\$4500).

Figure 8.5 shows a simple circuit that requires only a digital voltmeter, a battery, and a standard resistor for the assembly of a resistance thermometer with a resolution of 1 mK. A 1.5 V silver oxide battery is a cheap, readily available, and stable source of a 1.5 V voltage. The thermometer resistance  $R_T$  will indicate the temperature. For a 4 k $\Omega$  thermistor, a standard resistor  $R_S$  of about 50 k $\Omega$  is required to keep the current below the level of self-heating. The standard resistor must be stable with time and temperature (even though it will remain at room temperature); Vishay makes such stable resistors.  $R_T$  equals  $V_T/I$ , where  $V_T$  is the voltage drop across  $R_T$  and  $I$  is the current through  $R_T$ . The current  $I$  is the same in  $R_S$ , thus  $I$  equals  $V_S/R_S$ , where  $V_S$  is the voltage drop across the standard resistor and  $R_S$  is its resistance. Thus:

$$R_T = R_S(V_T/V_S) \quad (8.4)$$

If  $R_S$  is known, then only  $V_T$  and  $V_S$  need to be measured to obtain  $R_T$ . Actually,  $R_S$  need not be known exactly, since it is a constant. The ratio  $V_T/V_S$  can be used as the thermometric variable when the thermometer is calibrated, and Equation (8.3) can be written with this ratio replacing  $R$ . In order to measure temperature to 1 mK, the voltages across



**Figure 8.5** A simple resistance thermometry circuit: a series arrangement of a thermistor or a platinum resistance thermometer ( $R_T$ ), a standard resistor ( $R_S$ ), and a battery (B).

the thermistor and resistor of 0.1 to 1 V need to be measured to six significant figures; some voltmeters can directly measure the ratio  $V_T/V_S$ .

### 8.1.4 Semiconductor Thermometers<sup>12,16</sup>

The thermistors discussed above are made from semiconductors and in that sense, thermistors are semiconductor (resistance) thermometers. Semiconducting elements such as germanium are also used to make resistance thermometers (see Section 8.1.5 below on cryogenic thermometry).

A different category of semiconductor thermometers uses a junction semiconductor (diode or transistor), and takes as the thermometric property the temperature dependence of the current–voltage characteristics of the junction itself. These devices produce currents or voltages that are linear (within  $\pm 1$  K) in temperature. Typical sensitivities are 1  $\mu$ A/K and 10 mV/K, leading to typical temperature resolutions of 0.1 to 1 K over a typical range of 223–423 K.<sup>33</sup> The diodes or transistors can be combined into more complex circuits that amplify and linearize or even digitize the signals. Such integrated circuit (IC) devices can require just a d.c. supply voltage to produce a temperature transducer. These IC thermometers are cheap and simple and easily automated. They can be used as the independent thermometers in the thermocouple reference junction circuits discussed above,<sup>12</sup> or wherever a measurement of  $\pm 1$  K is required. They are not yet capable of refined temperature measurements. For measurements of the output voltages or currents at these resolutions, digital multimeters suffice (Agilent, BK Precision, Keithley).

Integrated-circuit thermometers that use diodes include the LM95233–LM95235 series and the LM35 (National Semiconductor). Other diode thermometers that are widely used in cryogenic thermometry will be discussed below in Section 8.1.5. Integrated-circuit thermometers that use transistors include the AD590<sup>12</sup> and the AD7416–AD7418 series of 10-bit temperature-to-digital converters (Analog Devices). Costs are only \$1 to \$7. Other suppliers are Andigilog and Adsem. Data sheets for IC thermometers are readily found on the internet.

Some IC temperature sensors (e.g., AD22100, Analog Devices) operate on the temperature dependence of the resistance of a thin metal film, but they are packaged in integrated circuits with voltage outputs that are nearly

**Table 8.4 Fixed temperatures from phase transitions<sup>a</sup>**

Temperature		System <sup>b</sup>	Phase Transition	Precision (K)
(K)	(°C)			
356	183	67% tin, 33% lead	Eutectic	2
273	100.0	Water	Vaporization	0.1
302.92	29.77	Gallium	Melting	0.05
286	13	<i>p</i> -Xylene	Melting	1
273	0.0	Water	Melting	0.002
263	-10	Diethylene glycol	Melting	1
252	-21	23% sodium chloride/77% water	Eutectic	2
243	-30	Bromobenzene	Melting	1
232	-41	Acetonitrile	Melting	1
221	-52	Benzyl acetate	Melting	1
210	-63	Chloroform	Melting	1
200	-73	Trichloroethylene	Melting	1
195	-78	Carbon dioxide	Sublimation	1
189	-84	Ethyl acetate	Melting	1
182	-91	Heptane	Melting	1
175	-98	Methanol	Melting	1
166	-107	Isooctane	Melting	1
157	-116	Ethanol	Melting	1
147	-126	Methylcyclohexane	Melting	1
142	-131	Pentane	Melting	1
132	-141	1,5-Hexadiene	Melting	1
113	-160	Isopentane	Melting	1
77	-196	Nitrogen	Vaporization	1

<sup>a</sup> See references 1,21,62,63

<sup>b</sup> Percentages are given as weight per cent

<sup>c</sup> The precision/accuracy is that to be expected without extraordinary attention to purity and pressure control

linear in temperature (+22.5 mV/K) and that allow a resolution of about 0.1 K. Thus they are “IC resistance thermometers.”

### 8.1.5 Temperatures Very Low: Cryogenic Thermometry

*Cryogenic temperatures* can be taken as those below 90 K, the boiling point of oxygen. The lowest temperatures so far measured are in the 100 pK range.<sup>34</sup> Cryogenic experiments require fastidious attention to the control of the flow

of heat into the apparatus (cryostat), and there is a considerable literature on these techniques.<sup>35–41</sup> Some of the thermometers discussed above can be used at low temperatures, and some special methods have been developed for use at low temperatures.

Thermocouples can be used at low temperatures, above about 10 K (for example, copper–constantan, Table 8.1). The thermocouple leads must be thermally anchored at each stage of the cryostat in order to avoid thermal emfs. See Section 8.1.2 for suppliers of thermocouples. Resistance thermometers are also valuable at low temperatures:<sup>42</sup> platinum, nickel,<sup>24</sup> copper, ruthenium oxide,<sup>43</sup> rhodium–iron, constantan, and manganin (Table 8.2). Platinum resistance thermometers lose sensitivity below about 20 K. Rhodium–iron resistors are fairly linear in the dependence of resistance on temperature, but have some dependence of the resistance on magnetic field (*magnetoresistance*). Manganin and constantan have very low thermal expansions and temperature coefficients of resistance at room temperature (Table 8.5) and are often used to make stable resistors, but, at temperatures below about 200 K, their resistances begin to decrease with temperature and then they can be used as thermometer elements. Lake Shore Cryotronics makes the Cernox (zirconium oxynitride) thin-film resistance thermometer, which has low magnetoresistance. Carbon resistors were traditional in cryogenic thermometry, but they are unstable on temperature cycling, exhibit magnetoresistance, and are now hard to obtain. Carbon-in-glass resistors are more stable and more available. There are semiconductor thermometers available for low temperatures;<sup>41,44</sup> Table 8.3 gives typical ranges and sensitivities for silicon diodes and gallium arsenide diodes. Lake Shore Cryotronics, Scientific Instruments, Inc., and Minco are sources of the above thermometers for low temperatures.

In addition to the thermometers that are low-temperature adaptations of thermometric devices used at higher temperatures, there are thermometric devices that are used only at low temperatures. The details of these technologies are outside the scope of this chapter, but an introduction will be provided, and the same will be done for high temperature thermometers in the next section. The thermometer types and their operating principles will be listed, with references for further information.

Table 8.5 Properties of thermostat materials at 298 K

<i>Conductor</i>	<i>Thermal Conductivity (W/cm K)</i>	<i>Specific Heat (J/g K)</i>	<i>Density</i>	<i>Emissivity (polished) (g/cm<sup>3</sup>)</i>
Aluminum	2.4	0.90	2.7	0.05
Brass	1.2	0.38	8.5	0.03
Constantan	0.22	0.4	8	—
Copper	4.0	0.39	8.9	0.02
Manganin	0.21	0.41	8.5	—
Nickel	0.91	0.44	8.9	0.05
Platinum	0.69	0.13	21.5	0.04
Stainless steel	0.14	0.4	7.8	0.1
<i>Insulator</i>	<i>Thermal Conductivity (W/cm K)</i>	<i>Specific Heat (J/g K)</i>	<i>Density (g/cm<sup>3</sup>)</i>	<i>Useful Temperature Range (K)</i>
Air (1 atm)	0.00025	1.0	0.0012	—
Alumina	0.06 <sup>a</sup>	1.0	4.0	3 to 2200
Bakelite	0.0023	1.6	1.3	3 to 390
Fiber@s	0.00002–0.004	0.8	0.06	3 to 1100
Macor	0.0017	0.75	2.5	3 to 1300
Magnesia	0.07	1.2	3.6	3 to 2700
Nylon	0.0024	1.7	1.1	3 to 420
Pyrex	0.001	0.8	2.2	3 to 1100
Styrofoam	0.00003–0.0003	1.3	0.05	3 to 350
Teflon TFE	0.004	1.0	2.1	3 to 350

<sup>a</sup> At about 1200 EC

Thermometers specific to cryogenic temperatures can be divided into primary thermometers, which give directly the thermodynamic temperature, and secondary thermometers, which require calibration against a primary thermometer.<sup>45</sup> Each technique requires its own particular measurement system external to the cryostat. The primary thermometers are based on several different principles:

- (1) *Gas thermometers* use the pressure of a gas (helium) under conditions such that the ideal gas law  $PV = nRT$  is applicable. For known, fixed volume  $V$ , number of moles of gas  $n$ , and gas constant  $R$ , the measurement of the pressure allows the temperature to be computed. Gas thermometry requires careful design, but can be used in the range 4.2 to 100 K.<sup>2,46</sup> Gas thermometry, as well as vapor pressure and melting curve thermometry (below), requires means of measuring pressure.<sup>36,40,41</sup>
- (2) *Vapor-pressure thermometers* use the pressure of a gas under conditions such that liquid and vapor coexist.<sup>40–42,46</sup> For a given pressure, the temperature is thermodynamically determined. Equations for these vapor pressure curves are known and allow the calculation of the temperature. The useful temperature range is determined by the gas chosen: oxygen (54–90 K), nitrogen (63–84 K), neon (24–40 K), hydrogen (14–20 K), <sup>3</sup>He and <sup>4</sup>He (0.3–5 K).
- (3) *Helium melting-curve thermometers* use the pressure of a gas under conditions such that solid and liquid coexist.<sup>36,41,42,46–48</sup> Again, the temperature is determined by a pressure measurement. The useful range is from 1 to 230 mK for <sup>3</sup>He.
- (4) *Noise thermometers* are based on the temperature dependence of the noise (fluctuations) in the Brownian motion of conducting electrons, which results in a

noise in the voltage across a resistor.<sup>1,36,41,49,50</sup> The noise in the voltage is measured to determine the temperature, but these voltages are at the  $\mu\text{V}$  level and are hard to measure. Noise thermometers have useful signals only below 1 K. Noise and nuclear orientation thermometers (below) are not sold commercially, but must be “home-made.”

- (5) *Nuclear orientation thermometers* depend on the occupation of nuclear energy levels as a function of temperature.<sup>2,36,42</sup> The energy levels are split by a magnetic field. At low temperatures, only the lowest state is occupied. At high temperatures, all states are occupied. At intermediate temperatures, the occupied levels emit  $\gamma$ -rays anisotropically, and the  $\gamma$ -ray emission rate parallel to the magnetic field is a measure of temperature. The materials usually used are  $^{60}\text{Co}$  and  $^{54}\text{Mn}$ . The useful range is 0.003 to 0.03 K.<sup>2</sup>
- (6) *Coulomb blockade thermometers* use the temperature dependence of the conductance of a tunnel junction array in the range 20 mK to 30 K.<sup>51</sup>

*Secondary thermometers* specific to low temperatures are:

- (1) *Magnetic thermometers* are based on the temperature dependence of the paramagnetic susceptibility.<sup>1,42</sup> The most widely used material is cerium magnesium nitrate (CMN), used from a 10 mK to 4.2 K.<sup>41</sup> Magnetic and NMR thermometers (below) must be home-made.
- (2) *NMR thermometers* use the temperature dependence of the nuclear magnetic resonance.<sup>36,41</sup> An NMR thermometer made from platinum wire has been used from 0.5 mK to 30 K.<sup>52</sup>
- (3) *Capacitance thermometers* depend on the temperature dependence of capacitors made of layered glass and ceramic, and operate between 1.4 K and 290 K (Lake Shore Cryogenics).<sup>44</sup> They are not reproducible through heating/cooling cycles, but are useful as relative or control thermometers.<sup>39</sup>

### 8.1.6 Temperatures Very High

As for low temperature thermometers, some “room temperature” devices can also operate into the higher range of temperatures. Thermocouples are used up to 2600 K

(Section 8.1.2; Table 8.1). Metal resistance thermometers are also of value: platinum is useful up to about 1300 K (Section 8.1.3; Table 8.2).

Research at high temperatures often requires a method that does not involve contact between the probe and the system of interest. *Pyrometers* rely on thermal radiation (which includes ultraviolet, visible, and infrared wavelengths, 7 to 20  $\mu\text{m}$ ) and operate in the range 373 K to 4000 K.<sup>3,53</sup> For example, astronomers use thermal radiation to measure the temperatures of bodies such as the sun (6000 K). The total radiation intensity per unit surface area, per unit time, of a body is  $M(T) = \epsilon/\Phi T^4$ , where  $\Phi$  is the known Stefan–Boltzmann constant and  $\epsilon$  is the emissivity of the body.<sup>1</sup> If the radiating body is a “blackbody,” then  $\epsilon = 1$  and the pyrometer is a primary thermometer. The emissivity  $\epsilon$  depends on wavelength for nonblackbodies. In practice, the conditions needed for primary, blackbody pyrometry are rarely met (or needed) outside of standards laboratories.

A practical pyrometer requires an optical detector to collect the radiation, together with a detector/transducer system to convert the radiation intensity to an electrical signal and/or read-out.<sup>1,54</sup> Some devices use fiber optics to collect the radiation and delivery it to the detector. Thermal detectors convert the collected radiation into heat that can then be measured by thermocouples, resistance thermometers (*bolometers*), or pyroelectric devices.<sup>3</sup> Photon detectors (photovoltaic, photoemissive, or photoconductive) convert the radiation directly into an electrical signal. Pyrometers with 1% accuracy can be constructed in the laboratory,<sup>55</sup> but there are good commercial products available and making one’s own pyrometer is not likely to be necessary. Commercial pyrometers of the “broadband” type operate over a span of wavelengths (0.3 to 20  $\mu\text{m}$ ), use the total emissivity, and are used up to 1300 K with 1% accuracy (Omega Engineering). Narrow-band pyrometers, such as infrared pyrometers, operate over a wavelength of about 1  $\mu\text{m}$ , and can measure up to 3300 K with 0.2 to 2% accuracy. The Omega Engineering website has a lengthy discussion of the design and calibration of radiation thermometers. The costs, depending on the resolution and on the level of the instrumentation, range from \$100 to \$5000 (Omega Engineering, Land Infrared). Sears sells a Craftsman hand-held infrared thermometer for the range 255 to 800 K at 2% accuracy for \$80.

### 8.1.7 New, Evolving, and Specialized Thermometry

*Fluorescence (or phosphor) thermometry* uses the temperature dependence of fluorescent intensity: a “reporter” molecule that will fluoresce is embedded in the system of interest; it is excited by light (usually a laser), and then the “after-glow” intensity is measured to determine the temperature. This technique is useful from 200 to 3000 K, at 5% accuracy. Such devices are not available commercially, but can be of value in the laboratory.<sup>3,56</sup> For example, fluorescence thermometry has been used to measure temperatures on nanoscale devices.<sup>57</sup>

The temperature dependence of the *infrared absorption spectrum* of a molecule (e.g., H<sub>2</sub>O) can be used to measure temperature, so-called “molecular thermometers.”<sup>58</sup> This is one of the methods of thermometry that can be used at the nanoscale.<sup>59</sup>

### 8.1.8 Comparison of Main Categories of Thermometers

Liquid-in-glass thermometers are cheap and easy to use, but are limited in range, in precision (0.01 K), and in means of automation. Thermocouples are cheap, are useful over a wide range, have good precision (mK), and can be automated; however, spurious voltages can cause problems. Platinum resistance thermometers offer broad range and very good precision (0.1 mK), at higher cost. Thermistors allow extremely high precision (1  $\mu$ K), but for limited range in temperature. Resistance thermometers are readily automated. Integrated circuit thermometers offer simple, easily automated temperature measurement when a low resolution of 0.1 to 1 K is all that is required.

### 8.1.9 Thermometer Calibration

Ideally, the working scientist would like to have temperature measurements that are on the fundamental, thermodynamic temperature scale. Primary measurements on the thermodynamic scale are usually made only in standards laboratories with very expensive and very complex thermometers, such as ideal-gas thermometers. The exceptions are cryogenics laboratories, where primary thermometers

are often integrated into the apparatus. The thermodynamic temperature is generally conveyed to the working laboratory by means of the International Temperature Scale, which was last defined in 1990 (ITS-90). The ITS approximates the thermodynamic temperature scale as closely as is technically and practically possible by defining the temperatures for a set of fixed points and by specifying the means for interpolation between those fixed points. *Fixed points* are temperatures that can always be reproduced, and are usually the temperatures at which phase transitions occur. The ITS is maintained in the United States at the National Institute of Standards and Technology (NIST).<sup>60</sup> Differences between the thermodynamic temperature scale and ITS-90 are 0.020 K at 373 K and 0.150 K at 903 K.

Other laboratories may send a thermometer (usually a platinum resistance thermometer) to NIST for calibration on the ITS-90, and then use that *working standard* to calibrate still other thermometers. Such a calibration can cost \$1000 or more. Most companies that make and sell thermometers maintain such secondary standards, with which they can perform calibrations “traceable to the National Institute of Standards and Technology” (Guildline, Hart Scientific).

For most scientific laboratories, the best procedure is to acquire a thermometer calibrated to the accuracy needed for that laboratory, either by purchasing a calibrated thermometer or by sending a thermometer to a calibration laboratory (NIST or a commercial service). Once calibrated, that “working standard” thermometer should be regularly checked for shifts or drifts in its calibration by measuring a fixed point. The most reliable and available such fixed point is the triple point of water (where ice, liquid water, and water vapor coexist), which can be reproduced to better than 1 mK. Water-triple-point cells are commercially available (Hart), but triple-point measurements are difficult. It is easier to use the ice point, the freezing point of water, as a fixed point. The temperature at which ice and water coexist at one atmosphere of pressure is 273.15 K and can be reproduced in most laboratories to 0.002 K. First, distilled water is degassed by several cycles of freezing it, evacuating the gas above it, and then remelting it. The ice point can then be achieved by mixing distilled, degassed water with small pieces of ice made from distilled, degassed water, and then continuously



stirring that slurry while the temperature measurement is made on the thermometer immersed in the slurry.

A variety of thermometric standard reference materials, are sold by NIST including calibrated thermometers and fixed-point cells, covering the range from 14 to 2040 K. Many fixed-point cells are available commercially (Hart Scientific).

Other thermometers can be compared to the working standard by placing both in a temperature-controlled environment (see Section 8.2), close together and under the same conditions at which the working standard was calibrated. The temperature is then varied over the range of interest, and measurements are made on the working standard and on the thermometer under calibration. The temperature is taken from the working standard, and an appropriate equation (see above) is fitted to the thermometric property of the second thermometer as a function of temperature. Keep in mind E. B. Wilson's admonition that,<sup>61</sup> "An experimenter of experience would as soon use calibrations carried out by others as he [or she] would use a stranger's toothbrush."

## 8.2 THE CONTROL OF TEMPERATURE

We will first consider cases for which the temperature is to be controlled at a single fixed value. Usually such cases require only rough control, to a degree or two. For example, we may want a trap in a vacuum system that will condense one vapor, but not others. Precise control at fixed temperatures would be required for thermometer calibrations using fixed points (see Section 8.1.9). We will also consider the control of temperature over a range within which any value may be selected. We will discuss how to achieve such control at various levels of precision.

### 8.2.1 Temperature Control at Fixed Temperatures

Control at fixed temperatures is most easily achieved at points of phase equilibrium.<sup>1,2</sup> For example, the point at which the solid, liquid, and gas phases of a pure substance coexist (the triple point) has a completely determined pressure and a completely determined temperature. Triple

points thus make excellent temperature reference points, with accuracies of about 1 mK. As discussed in Section 8.1.9, several triple points are used to define the International Temperature Scale. They are not appropriate for temperature control in most laboratories, however, since triple points can be difficult to achieve and to maintain.

The temperatures of vaporization, sublimation, and melting of pure substances are fully determined when the pressure is fixed. These phase transitions are readily achieved and maintained, and are therefore very useful for temperature control. The *eutectic temperature*, the temperature of a two-component mixture at which a liquid solution and both pure solids coexist at a fixed pressure, can also serve as a temperature control point. For all these phase equilibria, the accuracy and stability of the temperature depend on the accuracy and stability of the pressure, and on the purity of the materials used.

Melting-point baths are easily made and are therefore in common use. The making of an ice-point bath is discussed in Section 8.1.2, where the bath is used for a thermocouple reference junction. Melting-point baths using organic liquids or aqueous solutions, with temperatures ranging from 113 to 286 K, can be prepared.<sup>62-64</sup> These *slush baths* are made by cooling the organic liquid with either dry ice or liquid nitrogen until the bath liquid begins to freeze and a slush of solid and liquid is formed. The slush should be contained in a Dewar flask and frequently stirred. Care must be taken in using liquid nitrogen, which can condense air and possibly create explosive mixtures.

Some useful phase-transition baths are given in Table 8.6. Coyne gives a table of melting-points for aqueous solutions.<sup>64</sup>

### 8.2.2 Temperature Control at Variable Temperatures

Control at variable temperatures requires that the temperature be measured by a control thermometer and compared with the desired temperature (the *set point*  $T_S$ ). The difference in temperature from the set point indicates whether the system needs to be heated or cooled. If the control thermometer provides an electrical signal, then that signal can be used to drive a heater or a cooler, and thus to control the temperature.<sup>65</sup>

Table 8.6 Properties of bath liquids at 298 K

Liquid	Useful Temperature Range		Viscosity (cP)	Thermal Conductivity (W/cm K)	Density (g/cm <sup>3</sup> )	Specific Heat (J/g K)	Price (\$/L)
	(°C)	(K)					
Silicone oils: <sup>a</sup>							
200	-40 to 180	233 to 450	1	0.001	0.96	1.7	20
510	-50 to 180	223 to 450	50	0.001	0.62	1.7	20
550	-40 to 200	233 to 470	100	0.001	1.1	1.7	20
710	-10 to 230	263 to 500	500	0.001	1.1	1.7	50
Hydrocarbon oils: <sup>b</sup>							
74971	25 to 160	298 to 430	~10	~0.001	0.8	2	20
74972	25 to 150	298 to 420	~10	~0.001	0.8	2	20
74974	25 to 110	298 to 380	~10	~0.001	0.8	2	20
Water	5 to 90	278 to 360	1.0	0.0060	1.0	4.2	0
Ethylene glycol:							
100%	-11 to 180	260 to 450	32	0.0026	1.1	2.4	10
50 wt% water	-33 to 105	240 to 380	3.7	—	1.1	3.4	5
Ethanol:							
100%	-117 to 60	160 to 330	1.0	0.0017	0.79	2.5	4
59 wt% water	-30 to 75	240 to 350	2.4	—	0.92	4.0	2
Methanol:							
100%	-90 to 60	180 to 330	0.55	0.0014	0.80	2.6	5
50 wt% water	-50 to 65	220 to 340	1.8	—	0.92	3.6	2

<sup>a</sup> Dow Corning Corp

<sup>b</sup> Precision Scientific, Inc.

Figure 8.6 shows the main elements of such a *thermostat*. The thermostat will include the control thermometer, from which an electrical signal goes to the signal conditioner, where the signal is compared with the signal expected at the set point, conditioned further in ways to be described below, and then amplified to drive the thermostat heater or cooler. Although it is possible to control a cooling device such as Peltier junction,<sup>66</sup> it is usually easier to control a heating element. Thus, the thermostat may include a cooler operating at a constant power to cool the system to a temperature below the set point, from which the set point is reached by means of the control heater. At higher temperatures, the thermostat may require an auxiliary heater, operating at constant power, to raise the temperature to near the set point, from which temperature the control heater achieves the set point. The thermostat will

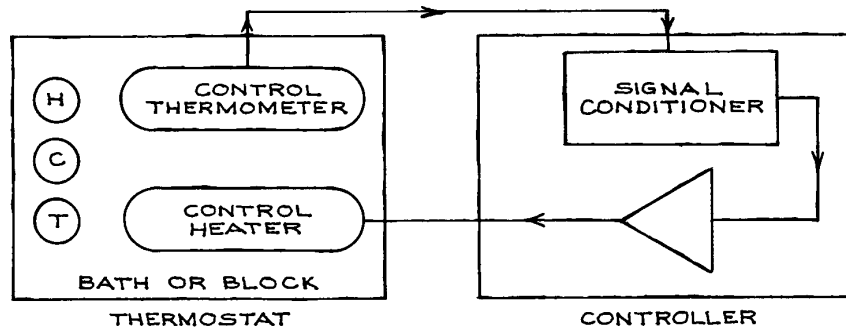
also include a measurement thermometer, independent of the control thermometer.

The *thermal circuit* may be modeled in analogy to an electrical circuit.<sup>67-71</sup> The flow of heat is analogous to a flow of electrical current. There is associated with each element a *thermal resistance*,  $R$ . For a wall of thickness  $d$ , area  $A$ , and thermal conductivity  $k$ ,  $R$  is given by:

$$R = d/Ak \quad (8.5)$$

Table 8.5 gives values of the thermal conductivity for various materials used in thermostat construction. In analogy with Ohm's law for current flow, the heat flow  $J$  induced by a temperature difference  $\Delta T$  across the wall is:

$$J = \Delta T/R \quad (8.6)$$



**Figure 8.6** The elements of a thermostat. The controller uses the signal from the control thermometer to drive the control heater. The thermostat can also include an auxiliary heater (H), a cooling element (C), and a measurement thermometer (T).

The *thermal capacitance*  $C$  of a thermostat element is its specific heat, values of which are also given in Table 8.5. The *thermal time constant* for an element is  $RC$ . The analysis of such thermal circuits can, in principle, be made in the same way as that for electrical circuits.<sup>68,69,71</sup>

If the required level of temperature control is not high ( $\pm 0.1$  K or more), then the controller can simply turn the control heater on or off, depending on whether the thermostat is too hot or too cold. This *on-off control* will, of course, produce large oscillations around the set point. For better temperature control, the controller must have a more sophisticated feedback loop (see Section 6.7.9).<sup>12,69,71,72</sup>

First, the thermostat will require a constant “background” heating. Most of the background heating should be done by the auxiliary heater rather than by the control heater, but the control heater could have a level  $P_B$  of background power. Second, the controller will provide to the heater a power  $P_p$  proportional to the deviation from the desired set point:

$$P_p = G(T_S - T) \quad (8.7)$$

where  $G$  is the *proportional gain* and  $T$  is the current temperature. As  $G$  is increased, a critical value will be reached at which temperature oscillations will occur; i. e., the control will revert to the on-off mode. The total heater power  $P$  is then:

$$P = P_B + G(T_S - T) \quad (8.8)$$

Third, the need for background power is likely to change, since the heat losses to the environment will change as the temperature of the thermostat is varied and as the temperature of the environment changes. The controller can adjust for this changing requirement by providing to the heater a power related to the integral over time of the deviation from the set point. The total heater power then becomes:

$$P = P_B + G(T_S - T) + RI(T_S - T)dt \quad (8.9)$$

This *integral control* is also called *reset control*, and  $R$  is the reset constant. The combination of proportional control and integral control is called *two-term control* or *PI control*.

Sometimes another control term, proportional to the derivative of the temperature with time, is included:

$$P = P_B + G(T_S - T) + RI(T_S - T)dt + D(dT/dt) \quad (8.10)$$

The derivative term alters the response when there are fast changes in temperature. Control using proportional, integral, and derivative terms is called *three-term control* or *PID control*. Laboratory thermostats rarely change temperature rapidly, and derivative control is not usually needed.

Detailed analysis of a thermostat and its controller is rarely possible, because the parameters  $R$  and  $C$  are not well known, because the thermal resistances at interfaces (e.g., between the heater and the thermostat) are very hard to assess, and because effects due to radiation and

convection enter. Nonetheless, the study of model circuits has led to some generalizations about thermostat design that are of practical use.<sup>68,69,72</sup>

- (1) Heat distribution should be maximized. This is done by using materials of high conductivity and by distributing the heaters. For fluid baths (see below), the heat is distributed by stirring.
- (2) The bath or block should have a long time constant with respect to the exchange of heat with the environment. Conductive and convective losses are reduced by using insulating materials or by enclosing the thermostat in a vacuum. Radiation can be reduced by using reflective materials.
- (3) The heater should have a short time constant for transmitting heat to the bath. Therefore it should be of high-conductivity material and should be in good contact with the block or bath fluid.
- (4) The control thermometer should have a very short time constant for response to the bath or block temperature. Therefore it should be small and should be in good contact with the block or fluid.
- (5) The proportional gain should be set at about half the value at which oscillations begin.
- (6) The reset constant should equal the thermal resistance between the thermostat bath or block and the environment, multiplied by the thermal capacitance of the bath or block.

The control loops implementing Equations (8.9) or (8.10) can be done by either analog<sup>73–76</sup> or digital circuits.<sup>66,77–80</sup> The thermometers discussed in Section 8.1 are inherently analog, but most are easily digitized (i.e., a digital voltmeter that reads a resistance or a corresponding voltage). The discussion in Section 8.1.2 noted that it is advantageous to have thermometers that produce outputs linear in temperature for control circuits. Analog Devices has a number of linearized IC thermometers and IC controllers that operate at levels of  $\pm 0.1$  to  $\pm 1$  K. For example, the Analog Devices TMP01 is a sensor and comparator with “a control signal from one of two outputs when the device is either above or below a specific temperature range” (Analog Devices webpage).

Other complete temperature-control instruments that operate on signals from analog thermometers are commercially available. Simple on-off or proportional controllers

designed to control to 0.1 to 2.0 K cost \$40–\$1000 (Minco, Omega Engineering, Cole-Parmer). Instruments are available with proportional control (Omega Engineering, Cole-Parmer) or with full PID control: Lake Shore Cryotronics,  $\pm 1$  mK; Wavelength Electronics,  $\pm 3$  mK; both around \$2000.

For digital control circuits, the analog thermometer signal is converted to a digital signal, operated on by a computer program, then converted back to an analog heater power.<sup>66,78–80</sup> A digital control system can treat thermometer output in the software, by using fits to Equations (8.1)–(8.3), rather than by using hardware linearization. Such a digital controller is especially convenient as a part of a computerized data-collection system, to change the temperature at programmed time or temperature intervals. Commercial digital controllers include those from Wavelength Electronics (see above), and some that are intended for cryogenic temperatures, but that are useful to about 450 K, with resolutions as good as  $\pm 1$  mK (Scientific Instruments, Inc.; Lake Shore Cryotronics).

**Stirred Liquid Baths.** Stirred liquid baths, in which the temperature of a working liquid is controlled, are the most common type of thermostat because they are simple, versatile, and inexpensive. Three kinds of liquid bath devices are available commercially. First, there are controller-stirrer devices that can be inserted into tanks of water to control within  $\pm 10$  to 100 mK, from room temperature to 363 K, for a few hundred dollars (Omega Engineering, VWR, Cole-Parmer). Second, there are complete bath systems that are designed to circulate a fluid to an external experimental assembly; the control is  $\pm 0.01$  to 0.1 K and they can have cooling capacity, so that a typical range is 253 to 473 K, and some are programmable through an RS-232 interface (Neslab, Brookfield, Thermo Fisher Scientific, Lauda, Hart, J-Kem Scientific). Third, there are complete bath systems that are designed for an experimental assembly to be inserted within their enclosed tanks. The suppliers of the circulating baths also make the enclosed baths. Hart Scientific and Guildline make precision enclosed fluid baths stable at levels of  $\pm 0.7$  mK to  $\pm 1$  mK.

Often the researcher decides to construct a fluid bath: to save money, to achieve better control, or to provide a particular configuration. In fact, there is no commercial bath with control to  $\pm 1$  mK or better and with a port through which the experimenter can view the experiment. If you

need such a bath (for example, for phase-separation studies), you must make it.<sup>81</sup> Figure 8.7 shows a typical fluid bath, designed to control to  $\pm 1$  mK or better. The bath vessel can be any large container; 36 L cylindrical Pyrex jars are available that can be used over a large temperature range and that are transparent. Radiative heat losses can be reduced with a layer of aluminum foil, or reflective plastic (such as is used for camping blankets) placed next to the vessel. Fiberglass batting, as is used to insulate buildings, is easily wrapped around the vessel to insulate it. Styrofoam sheets about 3 in. thick can be cut to fit around the vessel. Styrofoam can also be formed *in situ* if a mold is placed around the vessel (Read Plastics, 800-638-6651). The bath is supported on three sections of glass or plastic tubing that are filled with insulation. Much of the heat loss or gain of the bath is through its top, so it is important that the top be well insulated. An insulated lid, cut to allow the various components to be inserted, is one solution. Figure 8.7 shows instead an insulating layer of hollow 3/4 in. O.D. polypropylene balls; the balls insulate while allowing easy access to the bath, but can be used only below about 400 K (Cole-Parmer, VWR).

The choice of a liquid for the bath depends on temperature range, cost, and convenience. Table 8.6 gives the properties of some common bath liquids. Sometimes an oil will be preferred because the electrical conductivity of aqueous solutions makes electrical connections within the bath awkward and unsafe. Silicone oils can be ordered with a rust inhibitor added, but those inhibitors turn brown and make the bath opaque.

The control heater, as discussed above, should be made from material of high thermal conductivity. A commercial heater with a copper sheath and with a power of about 150 to 250 W is satisfactory (Omega Engineering). An ordinary 60 W light bulb is an easy and cheap control heater: a power cord is soldered to the bulb and the connections electrically insulated with a silicone adhesive sealant. The background heater need not respond quickly, and can therefore be a glass or stainless-steel immersion heater (Omega Engineering); its power will depend on the temperature range, the heat capacity of the bath, and the heat losses of the bath. The background heater can be driven by a manually variable transformer (Variac).

There are several ways of providing a cooling element to the bath. Figure 8.7 shows a coil of copper tubing, through

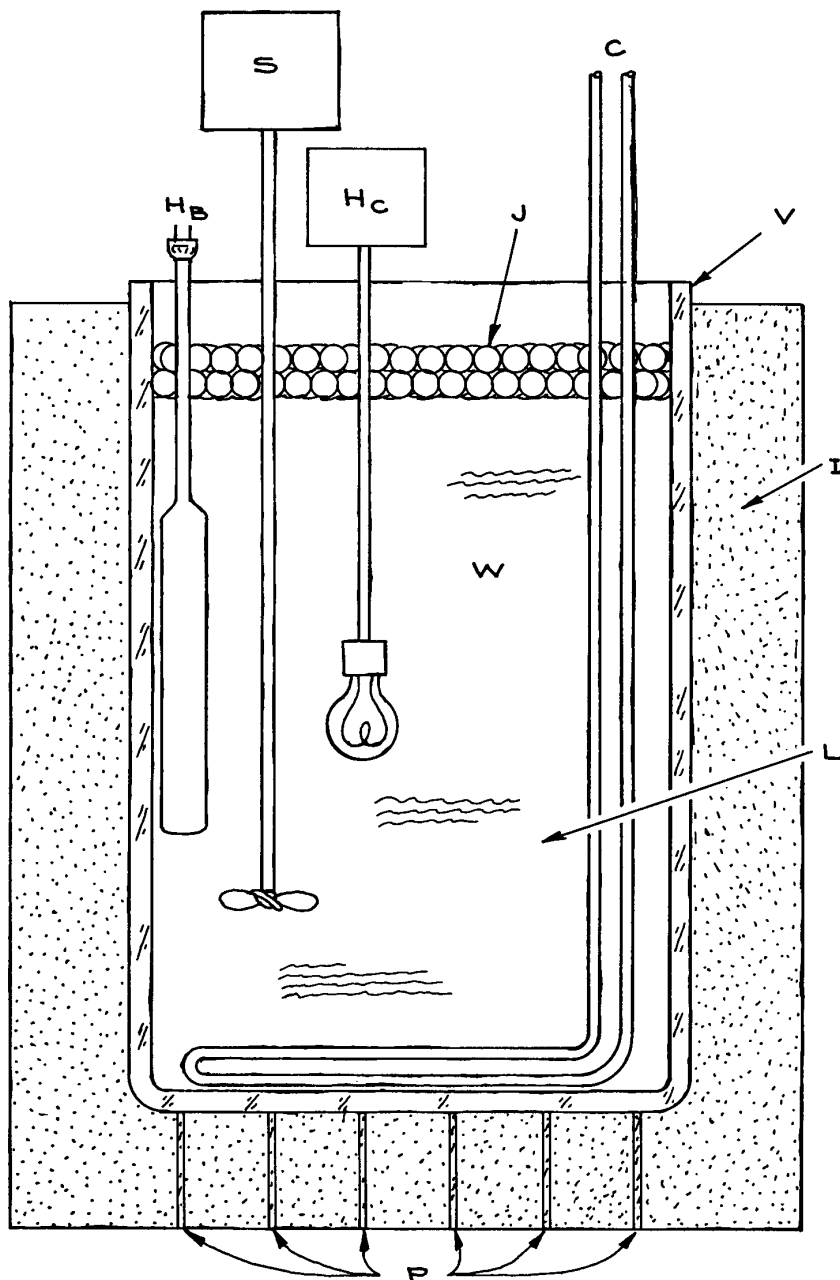
which cold liquid is circulated from an external source. That source can be a commercial refrigerated bath or circulating chiller (Neslab, Cole Parmer). Cold liquid can also be obtained by inserting a commercial immersion cooler into a container of fluid (e.g., diethylene glycol) and pumping that liquid through the bath cooling coil (Neslab, Thermo Fisher, Cole Parmer). An immersion cooler could be inserted directly into the stirred liquid bath, but these coolers are not designed to work at elevated temperatures and such a practice will eventually damage them.

The bath stirrer is an ordinary laboratory stirrer of 1/20 or 1/10 hp. The shaft is of stainless steel to reduce heat transfer from the motor. Stirrers can transmit electrical noise, which can interfere with the controller and thermometers, so it may be necessary to provide the stirrer with an electrical shield.

The configuration of the bath elements is crucial. Figure 8.7 shows the heaters placed near the stirrers, so that the heat is readily distributed. The cooling coil is itself somewhat distributed. The point is to arrange the elements so that the stirrer will prevent temperature gradients.

Safety precautions are in order. The possibility that the bath will overheat can be precluded by installing a device that will turn off all power if a set temperature is exceeded. Omega Engineering, Cole-Parmer, Guildline, and Instruments for Research and Industry sell such devices for \$85–\$300, but cheaper (<\$10) thermal switches are available from Newark Electronics. A pan made of sheet metal or plywood, sealed at the joints and placed under the bath, will contain the bath liquid should the vessel break. Above 200 K, the bath liquids may fume and require the installation of a fume hood. Of course, one should never put hands or tools into a fluid bath when any of its components are powered.

**Enclosure Thermostats.**<sup>82</sup> Enclosure thermostats control the temperature of an experimental block or stage held in an environment of vacuum or of stagnant gas. Such thermostats are used rather than stirred liquid thermostats if state-of-the-art temperature control ( $\pm 0.01$  mK) is required, if the temperature range needed is above or below that of available bath liquids, or if the experimental apparatus cannot be immersed in a stirred liquid. For example, optical experiments are difficult in stirred baths because bubbles and debris get into the optical path. For many experiments,

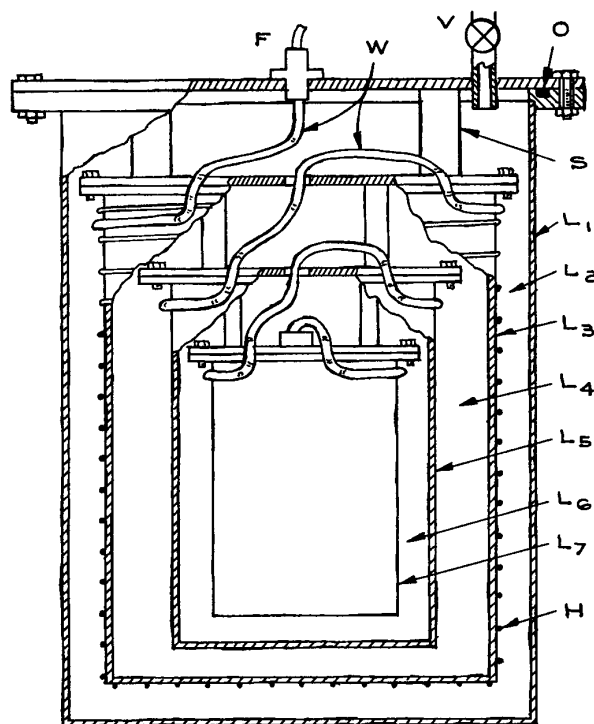


**Figure 8.7** A stirred liquid thermostat. The bath elements are: C, a coil through which cold liquid flows; H<sub>B</sub>, a commercial blade heater serving as a background heater; H<sub>C</sub>, a 60 W light bulb serving as the control heater; I, styrofoam insulation; J, polypropylene balls serving as the top insulation; L, the bath liquid; P, plastic pipes supporting the vessel; S, the stirrer; V, a Pyrex jar serving as the bath vessel; and W, the workspace. The control thermometer and the measurement thermometer are not shown.

the vibrations that are inevitable in a stirred bath cannot be tolerated. Much of the work on enclosure thermostats is in the literature on cryogenics (“cryostats”)<sup>35,36,39,40</sup> and on calorimetry.<sup>83–85</sup>

In an enclosure thermostat, sometimes called an “onion” thermostat, the temperature is controlled by means of concentric layers that are in sequence thermally insulating, thermally conducting, and thermally controlled.<sup>75,86</sup> The insulating layers are usually vacuum regions. The conducting layers are made of metal. A controlled (conducting) layer has a heater and a thermometer, connected to a controller (see above). The innermost conducting layer is the experimental stage or block. The result is that the alternating layers of conductors and insulators attenuate thermal gradients and fluctuations, making them smaller and smaller as the central experimental stage is neared. There are two design concepts for such a thermostat. The first maximizes the heat capacities of the layers in order to reduce the thermal gradients.<sup>76</sup> The disadvantage of this approach is that the time constants are very long, and the changes in temperature are very slow. The second design concept maximizes the number of stages in order to attenuate gradients; this approach has the advantage of short time constants, but is limited by practical size considerations.<sup>73</sup>

The only commercially available enclosure thermostats are cryostats for low-temperature regimes,<sup>39</sup> so it is useful to discuss designs for higher temperatures and for more general purposes. Figure 8.8 shows a configuration for a simple enclosure that will control temperature to 1 mK or better. There are four conducting layers and three insulating layers (all vacuum) in this thermostat. This thermostat can operate only at temperatures above that of its outermost layer. Thus that outer layer must be provided with a means of achieving the required temperature, usually about 0.5 K below the temperature at the experimental stage. For temperatures between room temperature and about 20 K above room temperature, the outer layer can be left at room temperature. For temperatures below room temperature, the outer layer can be immersed in a cold bath or can have copper tubing soldered around it, through which cold liquid can be circulated. For example, to operate between 77 and 127 K, the vacuum container could be immersed in liquid nitrogen. Likewise, for temperatures more than 20 K above room temperature, a hot bath or hot circulating liquid can be used.



**Figure 8.8** A simple enclosure thermostat. The thermostat is made of seven layers. The outer layer ( $L_1$ ) is a vacuum container made of thermally conducting metal and held at a temperature slightly below the desired set point. The lid is sealed with the O-ring seal at O, and the can is evacuated by means of valve V. Layers  $L_2$ ,  $L_4$ , and  $L_6$  are insulating vacuum layers. Layer  $L_3$  is a thermally conducting controlled layer, wound with heater H and fitted with a control thermometer (not shown). Layer  $L_5$  is a conducting layer.  $L_7$  is the experimental stage or block. Supports (S) are made of an insulator such as nylon. Wires (W) to the thermometers, heaters, and experimental transducers enter through the feedthrough F and are thermally anchored on each successive stage.

For some designs, it may be possible to use thermoelectric coolers (TEC or Peltier devices)<sup>66,77,79</sup> or miniature refrigerators to provide cooling below ambient. Recall that for a Peltier device, a current through metal junctions induces a temperature difference. A TEC device has the advantage that it can be either a heater or a cooler, depending on the direction of the current through it; it has the disadvantage that in the cooling mode, the heat removed from the system must be disposed of by some kind of heat

exchanger (flowing air or water). One TEC can pump up to 125 W, and they can be cascaded for increased cooling. Devices cost \$10 to \$100 (Marlow, Melcor, Ferro Tec). The Ferro Tec webpage has an introduction and design guide. Miniature refrigerators operate on the principle of Joule–Thomson cooling. They require flowing gas as the heat exchanger and, depending on the choice of gas, can operate at temperatures between 83 and 303 K, with a cooling capacity of 0.5 to 10 W/cm<sup>2</sup> per stage and control of  $\pm 0.1$  K (MMR Technologies, starting at \$1690).

Of course, the choice of solders, insulating materials, etc., will depend on the temperature range of operation. Copper is the best choice for conducting layers (see Table 8.5), but aluminum is not bad and is much lighter. Isolation between stages can be further increased if radiative losses are reduced by having conducting stages of low emissivity (Table 8.5).

While Figure 8.8 shows an evacuated thermostat, it is often convenient for the insulating layer to be air. In fact, the thermal conductivity of air is not significantly decreased as the pressure decreases until the pressure reaches about  $10^{-4}$  torr. Thus air at 1 atm is just as good an insulator as an evacuated space, unless low pressures can be achieved. At low temperatures, a disadvantage of air is that water can condense from it, but the air can be dried or dry nitrogen can be used. A disadvantage of a vacuum is that vibrations from the vacuum pump are transmitted to the thermostat.

The heater in Figure 8.8 is a wire, wound symmetrically and bifilarly (to reduce inductive effects) on the outside of the controlled layer. The heater is shown as simply glued to the outside of the layer, but better thermal contact is obtained if the wire is glued into grooves machined into the metal. Manganin wire is good for heaters because it has a high electrical resistivity that does not change significantly with temperature (MWS Wire Industries). Foil heaters that can be glued on where needed are often useful for the end plates (Omega Engineering, Minco). It is important that the heater power be distributed as evenly as possible over the controlled layer. The total heater power must be calculated from the estimated heat losses for the thermostat. A thermostat like that in Figure 8.8, 18 cm high and 18 cm in diameter and operating between 273 and 313 K, requires a heater of 315  $\Omega$ , operated at 0–15 V.

Since the heat distribution is entirely passive, attention must be given to all unsymmetrical heat losses that could lead to temperature gradients. All support pieces connect-

ing one layer to another must be of an insulating material (such as nylon or glass) and should present a minimum cross-section for heat transfer. All wires entering the thermostat should be of the smallest possible diameter. All wires must be thermally anchored at each stage: they must be wrapped around the stage in good thermal contact with it, so that they will come to the temperature of that stage. Since windows in a thermostat easily cause gradients, they must be kept small and measures taken to compensate for losses through them.

Cryostats or dewars for low temperatures are often designed and constructed by scientists,<sup>36,39</sup> but are also commercially available (Cryomech, Advanced Research Systems, Janis).

## Cited References

1. T. J. Quinn, *Temperature*, 2nd edn., Academic Press, New York, 1990.
2. J. F. Schooley, *Thermometry*, CRC Press, Inc., Boca Raton, FL, 1986.
3. P. R. N. Childs, J. R. Greenwood, and C. A. Long, Review of temperature measurement, *Rev. Sci. Instrum.*, **71**, 2959–2978, 2000.
4. J. Fischer and B. Fellmuth, Temperature metrology, *Rep. Prog. Phys.*, **68**, 1043–1094, 2005.
5. J. A. Wise, *Liquid-in-Glass Thermometry*, US Government Printing Office, Washington, DC, 1976.
6. *Resistance and Liquid-in-Glass Thermometry*, Vol. 2, R. E. Bentley (Ed.), Springer, Singapore, 1998.
7. D. D. Pollack, *Thermocouples: Theory and Properties*, Chemical Rubber Co., Boca Raton, FL, 1991.
8. T. W. Kerlin, *Practical Thermocouple Thermometry*, Instrument Society of America, Research Triangle Park, NC, 1999.
9. *Manual on the Use of Thermocouples in Temperature Measurement*, 4th edn., American Society for Testing Materials International, West Conshohocken, PA, 1993.
10. *Theory and Practice of Thermometric Thermometry*, Vol. 3 R. E. Bentley (Ed.), Springer, Singapore, 1998.
11. D. D. Pollack, *Thermoelectricity: Theory, Thermometry, Tool*; *ASTM Special Publication 852*, American Society for Testing and Materials, Philadelphia, 1985.
12. T. D. McGee, *Principles and Methods of Temperature Measurement*, John Wiley & Sons, Inc., New York, 1988.



13. T. J. Seebeck, Evidence for the thermal current from the combination Bi-Cu by its action on the magnetic needle, *Ab. K. Akad. Wiss. Berlin* **265**, 1822–1823, 1821.
14. J. V. Nicholas and D. R. White, *Traceable Temperatures: An Introduction to Temperature Measurement and Calibration*, 2nd edn., John Wiley & Sons, Inc., New York, 2001.
15. J. C. A. Peltier, Nouvelle experiences sur la caloricité des courans electriques, *Ann. Chim. Phys.*, **56**, 371, 1834.
16. L. Michalski, K. Eckersdorf, J. Kucharski, and J. McGhee, *Temperature Measurement*, 2nd edn., John Wiley & Sons, Inc., New York, 2001.
17. N. P. Moiseeva, The prospects for developing standard thermocouples of pure metals, *Meas. Tech.*, **47**, 915–919, 2004.
18. K. Instruments, *Low Level Measurements Handbook: Precision DC Current, Voltage, and Resistance Measurements*, 6th edn., Keithley Instruments, Cleveland, OH, 2007.
19. H. L. Callendar, On the practical measurement of temperature: experiments made at the Cavendish laboratory, *Phil. Trans. Royal Soc. London*, **178**, 161–230, 1887.
20. J. L. Riddle, G. T. Furukawa, and H. H. Plumb, *Platinum Resistance Thermometry, NBS Monograph 126*, US Government Printing Office, Washington, DC, 1973.
21. H. Preston-Thomas, The International Temperature Scale of 1990, *Metrologia*, **2–10 + 107**, 1990.
22. G. A. Lushchayev, V. A. Karchkov, E. I. Fandeev, and I. K. Sologyan, Linearization of the characteristic of a nickel resistance thermometer, *Measurement Techn.*, **24**, 475–479, 1981.
23. A. A. Khan, A. R. M. Alamoud, and M. A. Al-Turaigi, Linearization of a nickel resistance detector and its application in temperature measurement, *Intl. J. Elec.*, **67**, 931–936, 1989.
24. O. B. Verbeke, J. Spinnewijn, and H. Strauven, Electroformed nickel for thermometry and heating, *Rev. Sci. Instrum.*, **58**, 654–656, 1987.
25. E. A. Tombasov, Z. P. Chepurayaya, and V. V. Yakunin, Calibration characteristics of copper resistance thermometers in the range from 77 to 273 K, *Measurement Techn.*, **34**, 935–938, 1991.
26. T. M. Dauphinee and H. Preston-Thomas, Copper-resistance temperature scale, *Rev. Sci. Instrum.*, **25**, 884–886, 1954.
27. O. Bourgeois, E. André, C. Macovei, and J. Chaussy, Liquid nitrogen to room-temperature thermometry using niobium nitride thin films, *Rev. Sci. Instrum.*, **77**, 126108/126101–126108/126103, 2006.
28. J. S. Steinhart and S. R. Hart, Calibration curves for thermistors, *Deep-Sea Res.*, **15**, 497–503, 1968.
29. H. J. Hoge, Useful procedure in least squares, and tests of some equations for thermistors, *Rev. Sci. Instrum.*, **59**, 975–979, 1988.
30. T. J. Edwards, Observations on the stability of thermistors, *Rev. Sci. Instrum.*, **54**, 613–617, 1983.
31. J. Schuderer, T. Schmid, G. Urban, T. Samaras, and N. Kuster, Novel high-resolution temperature probe for radiofrequency dosimetry, *Phys. Med. Biol.*, **49**, N83–N92, 2004.
32. J. J. Connelly, Resistance thermometer measurement, in *Handbook of Temperature Measurement: Resistance and Liquid-in-Glass Thermometry*, Vol. 2, R. E. Bentley (Ed.), Springer, Singapore, 1998, pp. 55–82.
33. M. J. Pertijs and J. H. Huijsing, *Precision Temperature Sensors in CMOS Technology*, Springer, Dordrecht, 2006.
34. J. T. Tuoriniemi and T. A. Knuutila, Nuclear cooling and spin properties of rhodium down to picokelvin temperatures, *Physica B*, **280**, 474–478, 2000.
35. A. C. Rose-Innes, *Low Temperature Techniques*, The English Universities Press, Inc., London, 1964.
36. *Experimental Techniques in Condensed Matter Physics at Low Temperatures*, R. C. Richardson and E. N. Smith (Eds.), Addison Wesley, Reading, MA, 1998.
37. F. Pobell, *Matter and Methods at Low Temperatures*, 3rd edn., Springer, New York, 2007.
38. A. Kent, *Experimental Low-Temperature Physics*, AIP Press, New York, 1993.
39. J. Ekin, *Experimental Techniques: Cryostat Design, Material Properties and Superconductor Critical-Current Testing*, Oxford University Press, Oxford, 2006.
40. G. K. White and P. J. Meeson, *Experimental Techniques in Low-Temperature Physics*, 4th edn., Clarendon Press, Oxford, 2002.
41. D. S. Betts, *An Introduction to Millikelvin Technology*, Cambridge University Press, Cambridge, 1989.
42. G. Schuster, D. Hechtischer, and B. Fellmuth, Thermometry below 1 K, *Rept Prog. Phys.*, **57**, 187–230, 1994.
43. R. Sahul, V. Tasovski, and T. S. Sudarshan, Ruthenium oxide cryogenic temperature sensors, *Sens. Actuat. A: Physical*, **125**, 358–362, 2006.
44. Y. Hu, Other secondary thermometers, in *Experimental Techniques in Condensed Matter Physics at Low Temperatures*, R. C. Richardson, R. P. Feynman, and E. N. Smith (Eds.), Westview Press, New York, 1998, pp. 308–320.
45. E. L. Ziercher, K. I. Blum, and Y. Hu, Thermometry, in *Experimental Techniques in Condensed Matter Physics at*

- Low Temperatures*, R. C. Richardson, R. P. Feynman, and E. N. Smith (Eds.), Westview Publisher, New York, 1998.
46. F. Pavese and G. Molinar, *Modern Gas-Based Temperature and Pressure Measurements*, Plenum Press, New York, 1982.
47. W. Ni, J. S. Xia, E. D. Adams, P. S. Haskins, and J. E. McKisson, <sup>3</sup>He melting pressure thermometry, *J. Low Temp. Phys.*, **101**, 305–310, 1995.
48. E. Pentii, J. Tuoriniemi, A. Salmela, and A. Sebedash, Melting pressure thermometry of the saturated helium mixture at millikelvin temperatures, *Rev. Sci. Instrum.*, **146**, 71–83, 2007.
49. R. J. Soulen, W. E. Fogle, and J. H. Colwell, Measurements of absolute temperature below 0.75 K using a Josephson-junction noise thermometer, *J. Low Temp. Phys.*, **94**, 385–487, 1994.
50. L. Spietz, R. J. Schoelkopf, and P. Pari, Shot noise thermometry down to 10 mK, *Appl. Phys. Lett.*, **89**, 183123/10183123/10183123 (2006).
51. J. P. Kauppinen, K. T. Loberg, A. J. Manninen, and J. P. Pekola, Coulomb blockade thermometer: tests and instrumentation, *Rev. Sci. Instrum.*, **69**, 4166–4175, 1998.
52. C. Buchal, J. Hanssen, R. M. Mueller, and F. Pobell, Platinum wire NMR thermometer for ultralow temperatures, *Rev. Sci. Instrum.*, **49**, 1360–1361, 1978.
53. D. P. DeWitt and G. D. Nutter, *Theory and Practice of Radiation Thermometry*, Wiley Interscience, New York, 1988.
54. M. J. Ballico, Radiation thermometry, in *Handbook of Temperature Measurement: Temperature and Humidity Measurement*, Vol. 2, R. E. Bentley (Eds.), Springer, Singapore, 1998, pp. 67–98.
55. P. Poulsen and S. K. Ault, New method of high-precision thermometry, *Rev. Sci. Instrum.*, **77**, 094901094906, 2006.
56. K. T. V. Grattan and Z. Y. Zhang, *Fiber Optic Fluorescence Thermometry*, Chapman and Hall, London, 1995.
57. T. Jigami, M. Kobayashi, Y. Taguchi, and Y. Nagasaka, Development of nanoscale temperature measurement technique using near-field fluorescence, *Intl. J. Thermophys.*, **28**, 968–979, 2007.
58. G. A. Marcus and H. A. Schwettman, Picosecond optical thermometry of protein in H<sub>2</sub>O, *J. Phys. Chem. B*, **111**, 3048–3054, 2007.
59. J. Lee, N. A. Kotov, Thermometer design at the nanoscale, *Nano Today*, **2**, 48–51, 2007.
60. B. W. Mangum, G. T. Furukawa, K. G. Kreider, *et al.* The kelvin and temperature measurements, *J. Res. Natl. Inst. Stand. Tech.*, **106**, 105–149, 2001.
61. E. B. Wilson, *An Introduction to Scientific Research*, McGraw-Hill, New York, 1952.
62. R. E. Rondeau, Slush baths, *J. Chem. Eng. Data*, **11**, 124, 1966.
63. A. M. Phipps and D. N. Hume, General purpose low temperature dry-ice baths, *J. Chem. Ed.*, **45**, 664, 1968.
64. G. S. Coyne, *The Laboratory Companion: A Practical Guide to Materials, Equipment, and Technique*, 2nd edn., Wiley-Interscience, New York, 2005.
65. J. Tapping, Temperature control, in *Handbook of Temperature Measurement: Temperature and Humidity Measurement*, Vol. 1, R. E. Bentley (Ed.), Springer, Singapore, 1998, pp. 203–214.
66. A.W. Sloman, P. Buggs, J. Malloy, and D. Stewart, A microcontroller-based driver to stabilize the temperature of an optical stage to within 1 mK in the range 4–38 °C, using a Peltier heat pump and a thermistor sensor, *Meas. Sci. Tech.*, **7**, 1653–1664, 1996.
67. E. D. West, Constant temperature baths, in *Treatise in Analytical Chemistry*, I. M. Kolthoff and P. J. Elving (Eds.), John Wiley & Sons, Inc., New York, 1967.
68. M. Kutz, *Temperature Control*, John Wiley & Sons, Inc., New York, 1968.
69. M. V. Swaay, Control of temperature – Part one, *J. Chem. Ed.* **46**, A515–A518, 1969.
70. W. K. Roots, *Fundamentals of Temperature Control*, Academic Press, New York, 1969.
71. E. M. Forgan, On the use of temperature controllers in cryogenics, *Cryogenics*, **14**, 207–214, 1974.
72. J. R. Leigh, *Temperature Measurement and Control*, Peter Peregrinus Ltd., London, 1988.
73. D. Sarid and D. S. Cannell, A ±15 microdegree temperature controller, *Rev. Sci. Instrum.*, **45**, 1082–1088, 1974.
74. L. Bruschi, R. Storti, and G. Torzo, Precise temperature controller for resistance thermometers, *Rev. Sci. Instrum.*, **56**, 427–429, 1985.
75. P. K. M. Unni, M. K. Gunasekaran, and A. Kumar, +/- 30 Microkelvin temperature controller from 25 to 103 °C: Study and analysis, *Rev. Sci. Instrum.*, **74**, 231–242, 2003.
76. B. J. Thyssse, The dielectric constant of SF<sub>6</sub> near the critical point, *J. Chem. Phys.*, **74**, 4678–4692, 1981.
77. X. Zhu, E. Krochmann, and J. Chen, Microcomputer-based Peltier thermostat for precision optical radiation measurements, *Rev. Sci. Instrum.*, **63**, 1999–2003, 1992.
78. R. B. Strem, B. K. Das, and S. C. Greer, A digital temperature control and measurement system, *Rev. Sci. Instrum.*, **52**, 1705–1708, 1981.
79. A. Kojima, C. Ishii, K. Tozaki, *et al.* Fine temperature stabilizer for X-ray diffraction measurements, *Rev. Sci. Instrum.*, **68**, 2301–2304, 1997.

80. P. Cofrancesco, U. Ruffina, M. Villa, P. Grossi, and R. Scattolini, A digital temperature control system, *Rev. Sci. Instrum.*, **62**, 1311–1316, 1991.
81. H. Ogasawara, Method of precision temperature control using flowing water, *Rev. Sci. Instrum.*, **57**, 3048–3052, 1986.
82. E. C. Harrigan, Calibration enclosures, in *Handbook of Temperature Measurement: Resistance and Liquid-in-Glass Thermometry*, Vol. 2, R. E. Bentley (Ed.), Springer, Singapore, 1998, pp. 145–160.
83. *Experimental Thermodynamics*, Vol. 1, J. P. McCullough and D. W. Scott (eds.), Butterworth's, London, 1969.
84. D. C. Ginnings, *Precision Measurement and Calibration: Selected NBS Papers on Heat*, US Government Printing Office, Washington, DC, 1970.
85. J. L. Hemmerich, J.-C. Loos, A. Miller, and P. Milverton, Advances in temperature derivative control and calorimetry, *Rev. Sci. Instrum.*, **67**, 3877–3884, 1996.
86. T. M. Sporton, The design of a general purpose air thermostat, *J. Phys. E. Sci. Instrum.*, **5**, 317–321, 1972.
- Superconductor Critical Current Testing*, Oxford University Press, Oxford, 2006.
- Thomas D. McGee, *Principles and Methods of Temperature Measurement*, John Wiley & Sons, Inc., New York, 1988.
- L. Michalski, K. Eckersdorf, J. Kucharski, and J. McGhee, *Temperature Measurement*, 2nd edn., John Wiley & Sons, Inc., New York, 2001.
- National Institute of Standards and Technology, Gaithersburg, MD, Thermometry Group Website.
- J. V. Nicholas and D. R. White, *Traceable Temperatures: An Introduction to Temperature Measurement and Calibration*, 2nd edn., John Wiley & Sons, Inc., New York, 2001.
- Frank Pobell, *Matter and Methods at Low Temperatures*, Springer, Berlin, 1996.
- Daniel D. Pollack, *Thermocouples: Theory and Properties*, CRC Press, Boca Raton, FL 1991.
- Robert C. Richardson and Eric N. Smith (Eds.), *Experimental Techniques in Condensed Matter Physics at Low Temperatures*, Addison-Wesley, Reading, MA, 1998.
- T. J. Quinn, *Temperature*, 2nd edn., Academic, New York, 1990.
- J. F. Schooley (Ed.), *Temperature: Its Measurement and Control in Science and Industry*, Vol. 5, American Institute of Physics, New York, 1982. See also other volumes in this series.
- J. F. Schooley, *Thermometry*, CRC Press, Boca Raton, FL, 1986.
- J. F. Swindells (Ed.), *Precision Measurement and Calibration, Selected NBS Papers on Temperature*, NBS Special Publication 300, Vol. 2, US Government Printing Office, Washington, DC, 1968.
- Temperatures.com, a website with information and resources on thermometry.
- J. A. Wise, *Liquid-in-Glass Thermometry, NBS Monograph 150*, US, Government Printing Office, Washington, DC, 1976.

## General References

- American Society for Testing Materials International, *Manual on the Use of Thermocouples in Temperature Measurement*, West Conshohocken, PA, 1993.
- R. E. Bentley, *Handbook of Temperature Measurement*, Springer-Verlag, Singapore, 1998.
- Peter R. N. Childs, *Practical Temperature Measurements*, Butterworth-Heinemann, Oxford, 2001.
- Jack W. Ekin, *Experimental Techniques for Low-Temperature Measurements: Cryostat Design, Material Properties, and*

