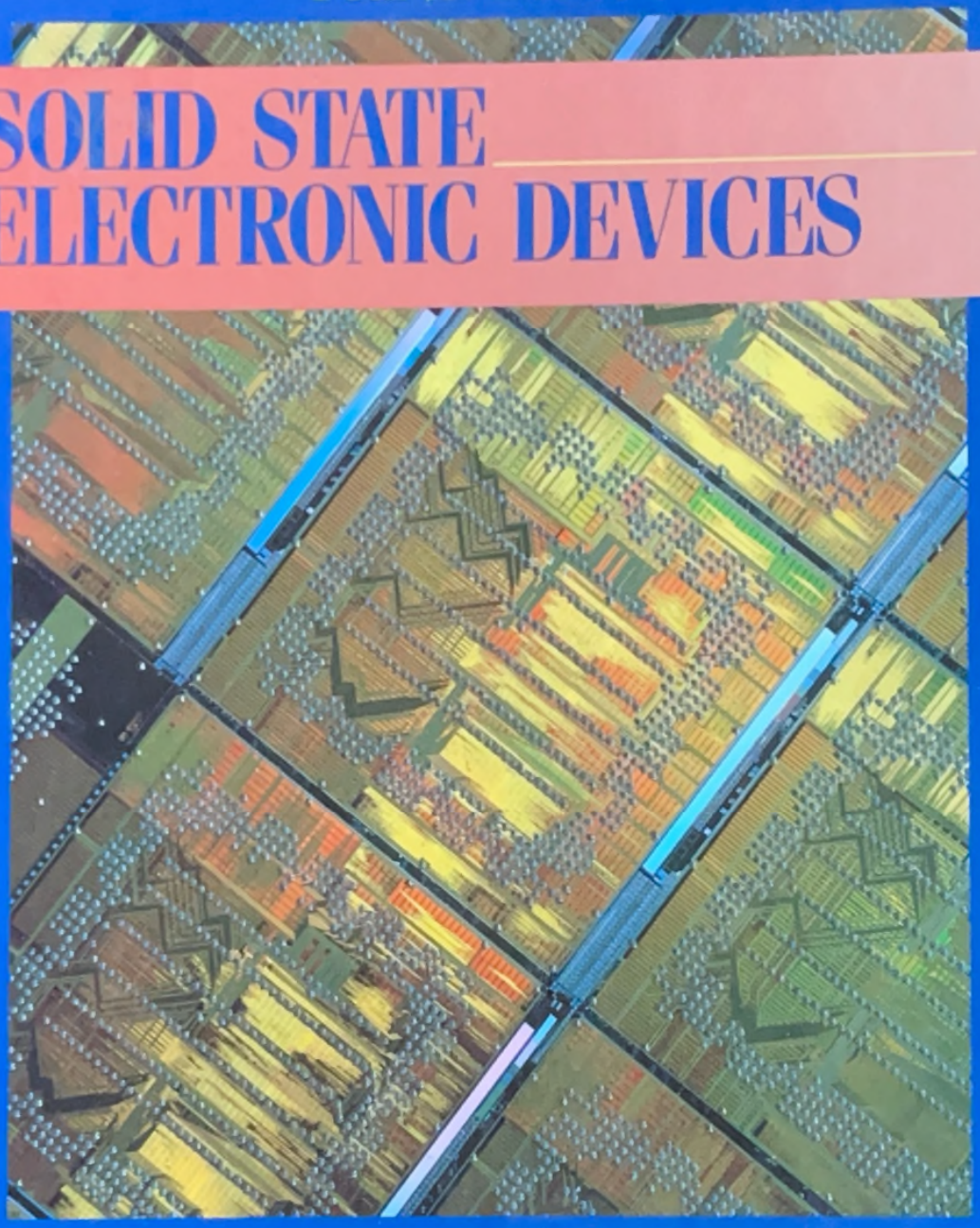
Fourth Edition

SOLID STATE ELECTRONIC DEVICES



Ben G. Streetman

Prentice Hall Series in Solid State Physical Electronics, Nick Holonyak, Jr., Series Editor

Library of Congress Cataloging-in-Publication Data
Streetman, Ben G.
Solid state electronic devices / Ben G. Streetman. – 4th ed.
p. cm. – (Prentice Hall series in solid state physical electronics)
Includes bibliographical references and index.
ISBN 0–13–158767–6 (hard cover)
1. Semiconductors. I. Title. II. Series.
TK7871.85.S77 1995
621.3815'2–dc20

94-37720
CIP

Acquisitions Editor: **Alan Apt**
Production Editor: **Joe Scordato**
Interior Design: **Lee Cohen**
Cover design: **Amy Rosen**
Cover: **Power PC RISC microprocessor chip courtesy of IBM Corp. Photograph by Tom Way.**
Manufacturing buyer: **Lori Bulwin**

© 1995, 1990, 1980, 1972 by Prentice-Hall, Inc.
A Simon & Schuster Company
Englewood Cliffs, New Jersey 07632

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-158767-6

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

CONTENTS

PREFACE xiii

ABOUT THE AUTHOR xvii

1 CRYSTAL PROPERTIES AND GROWTH OF SEMICONDUCTORS 1

1.1 Semiconductor Materials 1
1.2 Crystal Lattices 3
1.2.1 Periodic Structures 3
1.2.2 Cubic Lattices 5
1.2.3 Planes and Directions 7
1.2.4 The Diamond Lattice 9
1.3 Bulk Crystal Growth 12
1.3.1 Growth from the Melt 12
1.3.2 Zone Refining and Floating-Zone Growth 15
1.4 Epitaxial Growth 17
1.4.1 Lattice Matching in Epitaxial Growth 17
1.4.2 Liquid-Phase Epitaxy 19
1.4.3 Vapor-Phase Epitaxy 21
1.4.4 Molecular Beam Epitaxy 24

2 ATOMS AND ELECTRONS 28

2.1 Introduction to Physical Models 29
2.2 Experimental Observations 30
2.2.1 The Photoelectric Effect 30
2.2.2 Atomic Spectra 31
2.3 The Bohr Model 33
2.4 Quantum Mechanics 36
2.4.1 Probability and the Uncertainty Principle 36
2.4.2 The Schrödinger Wave Equation 37
2.4.3 Potential Well Problem 39
2.4.4 Tunneling 41
2.5 Atomic Structure and the Periodic Table 42
2.5.1 The Hydrogen Atom 42
2.5.2 The Periodic Table 45

IPR2025-00238 - BOE
Exhibit 1007 - Page 3

3 ENERGY BANDS AND CHARGE CARRIERS IN SEMICONDUCTORS 51

3.1 Bonding Forces and Energy Bands in Solids 51
3.1.1 Bonding Forces in Solids 52
3.1.2 Energy Bands 54
3.1.3 Metals, Semiconductors, and Insulators 55
3.1.4 Direct and Indirect Semiconductors 56
3.1.5 Variation of Energy Bands with Alloy Composition 58

3.2 Charge Carriers in Semiconductors 61
3.2.1 Electrons and Holes 61
3.2.2 Effective Mass 63
3.2.3 Intrinsic Material 65
3.2.4 Extrinsic Material 66
3.2.5 Electrons and Holes in Quantum Wells 69

3.3 Carrier Concentrations 71
3.3.1 The Fermi Level 71
3.3.2 Electron and Hole Concentrations at Equilibrium 74
3.3.3 Temperature Dependence of Carrier Concentrations 78
3.3.4 Compensation and Space Charge Neutrality 80

3.4 Drift of Carriers in Electric and Magnetic Fields 81
3.4.1 Conductivity and Mobility 81
3.4.2 Drift and Resistance 85
3.4.3 Effects of Temperature and Doping on Mobility 86
3.4.4 High-Field Effects 88
3.4.5 The Hall Effect 89

3.5 Invariance of the Fermi Level at Equilibrium 91

4 EXCESS CARRIERS IN SEMICONDUCTORS 96

4.1 Optical Absorption 96
4.2 Luminescence 99
4.2.1 Photoluminescence 100
4.2.2 Cathodoluminescence 102
4.2.3 Electroluminescence 103

4.3 Carrier Lifetime and Photoconductivity 103
4.3.1 Direct Recombination of Electrons and Holes 103
4.3.2 Indirect Recombination; Trapping 106
4.3.3 Steady State Carrier Generation; Quasi-Fermi Levels 108
4.3.4 Photoconductive Devices 111
4.4 Diffusion of Carriers 112
4.4.1 Diffusion Processes 112
4.4.2 Diffusion and Drift of Carriers; Built-in Fields 115
4.4.3 Diffusion and Recombination; The Continuity Equation 118
4.4.4 Steady State Carrier Injection; Diffusion Length 120
4.4.5 The Haynes–Shockley Experiment 122
4.4.6 Gradients in the Quasi-Fermi Levels 124

5 JUNCTIONS 130
5.1 Fabrication of p-n Junctions 130
5.1.1 Grown Junctions 131
5.1.2 Alloyed Junctions 131
5.1.3 Diffused Junctions 132
5.1.4 Ion Implantation 136
5.2 Equilibrium Conditions 138
5.2.1 The Contact Potential 139
5.2.2 Equilibrium Fermi Levels 144
5.2.3 Space Charge at a Junction 144
5.3 Forward- and Reverse-Biased Junctions; Steady State Conditions 149
5.3.1 Qualitative Description of Current Flow at a Junction 149
5.3.2 Carrier Injection 154
5.3.3 Minority and Majority Carrier Currents 160
5.4 Reverse-Bias Breakdown 162
5.4.1 Zener Breakdown 163
5.4.2 Avalanche Breakdown 164
5.5 Transient and a-c Conditions 166
5.5.1 Time Variation of Stored Charge 167
5.5.2 Reverse Recovery Transient 170
5.5.3 Capacitance of p-n Junctions 173

5.6 Deviations from the Simple Theory 176
5.6.1 Effects of Contact Potential on Carrier Injection 176
5.6.2 Recombination and Generation in the Transition Region 178
5.6.3 Ohmic Losses 180
5.6.4 Graded Junctions 182
5.7 Metal-Semiconductor Junctions 183
5.7.1 Schottky Barriers 184
5.7.2 Rectifying Contacts 186
5.7.3 Ohmic Contacts 187
5.7.4 Typical Schottky Barriers 189
5.8 Heterojunctions 190

6 p-n JUNCTION DIODES 201
6.1 The Junction Diode 201
6.1.1 Rectifiers 201
6.1.2 Switching Diodes 205
6.1.3 The Breakdown Diode 206
6.1.4 The Varactor Diode 207
6.2 Tunnel Diodes 208
6.2.1 Degenerate Semiconductors 208
6.2.2 Tunnel Diode Operation 209
6.2.3 Circuit Applications 212
6.3 Photodiodes 212
6.3.1 Current and Voltage in an Illuminated Junction 212
6.3.2 Solar Cells 215
6.3.3 Photodetectors 217
6.3.4 Noise and Bandwidth of Photodetectors 220
6.4 Light-Emitting Diodes and Lasers 222
6.4.1 Light-Emitting Materials 222
6.4.2 Fiber Optic Communications 224
6.4.3 Multilayer Heterojunctions for LEDs 226

7 BIPOLAR JUNCTION TRANSISTORS 232
7.1 Amplification and Switching 233
7.1.1 The Load Line 233
7.1.2 Amplification 234
7.1.3 Switching 235

7.2 Fundamentals of BJT Operation 235
7.2.1 Charge Transport in a BJT 235
7.2.2 Amplification with BJTs 238
7.3 BJT Fabrication 242
7.4 Minority Carrier Distributions and Terminal Currents 244
7.4.1 Solution of the Diffusion Equation in the Base Region 244
7.4.2 Evaluation of the Terminal Currents 246
7.4.3 Approximations of the Terminal Currents 248
7.4.4 Current Transfer Ratio 250
7.5 Generalized Biasing 251
7.5.1 The Coupled-Diode Model 252
7.5.2 Charge Control Analysis 255
7.6 Switching 257
7.6.1 Cutoff 258
7.6.2 Saturation 259
7.6.3 The Switching Cycle 260
7.6.4 Turn-On Transient 261
7.6.5 Turn-Off Transient 262
7.6.6 Schottky Diode Clamp 264
7.6.7 Specifications for Switching Transistors 264
7.7 Other Important Effects 265
7.7.1 Drift in the Base Region 266
7.7.2 Base Narrowing 268
7.7.3 Avalanche Breakdown 269
7.7.4 Injection Level; Thermal Effects 271
7.7.5 Base Resistance and Emitter Crowding 272
7.8 Frequency Limitations of Transistors 276
7.8.1 Capacitance and Charging Times 276
7.8.2 Transit Time Effects 279
7.8.3 High-Frequency Transistors 279
7.9 Heterojunction Bipolar Transistors 281

8 FIELD-EFFECT TRANSISTORS 288
8.1 The Junction FET 288
8.1.1 Pinch-off and Saturation 290
8.1.2 Gate Control 291
8.1.3 Current–Voltage Characteristics 293

8.2 The Metal–Semiconductor FET 295
8.2.1 The GaAs MESFET 295
8.2.2 The High Electron Mobility Transistor (HEMT) 296
8.2.3 Short Channel Effects 298
8.3 The Metal–Insulator–Semiconductor FET 299
8.3.1 Basic Operation 300
8.3.2 The Ideal MOS Capacitor 301
8.3.3 Effects of Real Surfaces 310
8.3.4 Threshold Voltage 312
8.3.5 The MOS Field-Effect Transistor 314
8.3.6 Control of Threshold Voltage 317
8.3.7 Substrate Bias Effects 320
8.3.8 Capacitance Effects and Self-Aligned Transistors 322
8.3.9 Short-Channel Effects 325

9 INTEGRATED CIRCUITS 331
9.1 Background 332
9.1.1 Advantages of Integration 332
9.1.2 Types of Integrated Circuits 334
9.1.3 Monolithic and Hybrid Circuits 334
9.2 Fabrication of Monolithic Circuits 336
9.2.1 Masking and Selective Doping 337
9.2.2 Fine-Line Lithography 339
9.2.3 Isolation 339
9.3 Monolithic Device Elements 341
9.3.1 Bipolar Transistors 342
9.3.2 Merged Transistors 344
9.3.3 MOS Transistors 346
9.3.4 Integration of Other Circuit Elements 353
9.4 Charge Transfer Devices 357
9.4.1 Dynamic Effects in MOS Capacitors 357
9.4.2 The Basic CCD 358
9.4.3 Improvements on the Basic Structure 359
9.4.4 Applications of CCDs 362
9.5 Very Large-Scale Integration (VLSI) 363
9.6 Testing, Bonding, and Packaging 367
9.6.1 Testing 367
9.6.2 Wire Bonding 368
9.6.3 Flip-chip and Beam-lead Techniques 371
9.6.4 Packaging 372

10 LASERS 377

10.1 Stimulated Emission 377
10.2 The Ruby Laser 380
10.2.1 The Resonant Cavity 380
10.2.2 Population Inversion in Ruby 382
10.2.3 Giant Pulse Lasers 383
10.3 Other Laser Systems 384
10.3.1 Rare Earth Systems 385
10.3.2 Gas Lasers 385
10.4 Semiconductor Lasers 387
10.4.1 Population Inversion at a Junction 387
10.4.2 Emission Spectra for p-n Junction Lasers 390
10.4.3 The Basic Semiconductor Laser 391
10.4.4 Heterojunction Lasers 392
10.4.5 Materials for Semiconductor Lasers 396

11 p-n-p-n SWITCHING DEVICES 399

11.1 Switching Mechanisms 400
11.1.1 The p-n-p-n Diode 400
11.1.2 The Two-Transistor Analogy 401
11.1.3 Variation of α with Injection 402
11.1.4 Forward-Blocking State 402
11.1.5 Conducting State 404
11.1.6 Triggering Mechanisms 404
11.2 The Semiconductor Controlled Rectifier 406
11.2.1 Gate Control 406
11.2.2 Turning off the SCR 407
11.2.3 Bilateral Devices 408
11.2.4 Fabrication and Applications 409

12 NEGATIVE CONDUCTANCE MICROWAVE DEVICES 412

12.1 Transit Time Devices 413
12.1.1 The IMPATT Diode 413
12.1.2 The QWITT Diode 416
12.1.3 The TRAPATT Diode 418
12.2 The Gunn Effect and Related Devices 419
12.2.1 The Transferred Electron Mechanism 419
12.2.2 Formation and Drift of Space Charge Domains 422
12.2.3 Modes of Operation in Resonant Circuits 425
12.2.4 Fabrication 429

APPENDICES 433

I. Definitions of Commonly Used Symbols 433
II. Physical Constants and Conversion Factors 438
III. Properties of Semiconductor Materials 439
IV. Derivation of the Density of States in the Conduction Band 440
V. Solid Solubilities of Impurities in Si and Ge 443
VI. Common Circuit Symbols for Solid State Devices 445

INDEX 449

chapter 4

EXCESS CARRIERS IN SEMICONDUCTORS

Most semiconductor devices operate by the creation of charge carriers in excess of the thermal equilibrium values. These excess carriers can be created by optical excitation or electron bombardment, or as we shall see in Chapter 5, they can be injected across a forward-biased p-n junction. However the excess carriers arise, they can dominate the conduction processes in the semiconductor material. In this chapter we shall investigate the creation of excess carriers by optical absorption and the resulting properties of photoluminescence and photoconductivity. We shall study more closely the mechanism of electron–hole pair recombination and the effects of carrier trapping. Finally, we shall discuss the diffusion of excess carriers due to a carrier gradient, which serves as a basic mechanism of current conduction along with the mechanism of drift in an electric field.

4.1 OPTICAL ABSORPTION†

An important technique for measuring the band gap energy of a semiconductor is the absorption of incident photons by the material. In this experiment photons of selected wavelength are directed at the sample, and the relative transmission of the various photons is observed. Since photons with energies greater than the band gap energy are absorbed while photons with energies less than

†In this context the word "optical" does not necessarily imply that the photons absorbed are in the visible part of the spectrum. Many semiconductors absorb photons in the infrared region, but this is included in the term "optical absorption."

IPR2025-00238 - BOE
Exhibit 1007 - Page 11

the band gap are transmitted, this experiment gives an accurate measure of the band gap energy.

It is apparent that a photon with energy $h\nu \geq E_g$ can be absorbed in a semiconductor (Fig. 4-1). Since the valence band contains many electrons and the conduction band has many empty states into which the electrons may be excited, the probability of photon absorption is high. As Fig. 4-1 indicates, an electron excited to the conduction band by optical absorption may initially have more energy than is common for conduction band electrons (almost all electrons are near E_c unless the sample is very heavily doped). Thus the excited electron loses energy to the lattice in scattering events until its velocity reaches the thermal equilibrium velocity of other conduction band electrons. The electron and hole created by this absorption process are *excess carriers;* since they are out of balance with their environment, they must eventually recombine. While the excess carriers exist in their respective bands, however, they are free to contribute to the conductivity of the material.

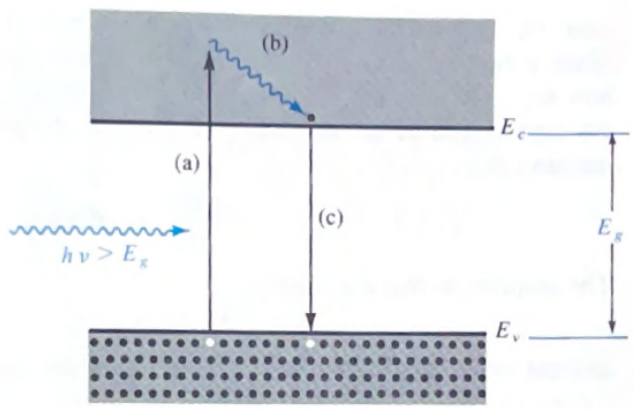


Figure 4-1
Optical absorption of a photon with $h\nu > E_g$: (a) an EHP is created during photon absorption; (b) the excited electron gives up energy to the lattice by scattering events; (c) the electron recombines with a hole in the valence band.

A photon with energy less than E_g is unable to excite an electron from the valence band to the conduction band. Thus in a pure semiconductor, there is negligible absorption of photons with $h\nu < E_g$. This explains why some materials are transparent in certain wavelength ranges. We are able to "see through" certain insulators, such as a good NaCl crystal, because a large energy gap containing no electron states exists in the material. If the band gap is about 2 eV wide, only long wavelengths (infrared) and the red part of the visible spectrum are transmitted; on the other hand, a band gap of about 3 eV allows infrared and the entire visible spectrum to be transmitted.

If a beam of photons with $h\nu > E_g$ falls on a semiconductor, there will be some predictable amount of absorption, determined by the properties of the material. We would expect the ratio of transmitted to incident light intensity to depend on the photon wavelength and the thickness of the sample. To calculate this dependence, let us assume that a photon beam of intensity $\mathbf{I}_0$ (photons/cm^2-s) is directed at a sample of thickness l (Fig. 4-2). The beam contains only photons of wavelength λ, selected by a monochromator. As the beam passes through the sample, its intensity at a distance x from the surface can be calcu-

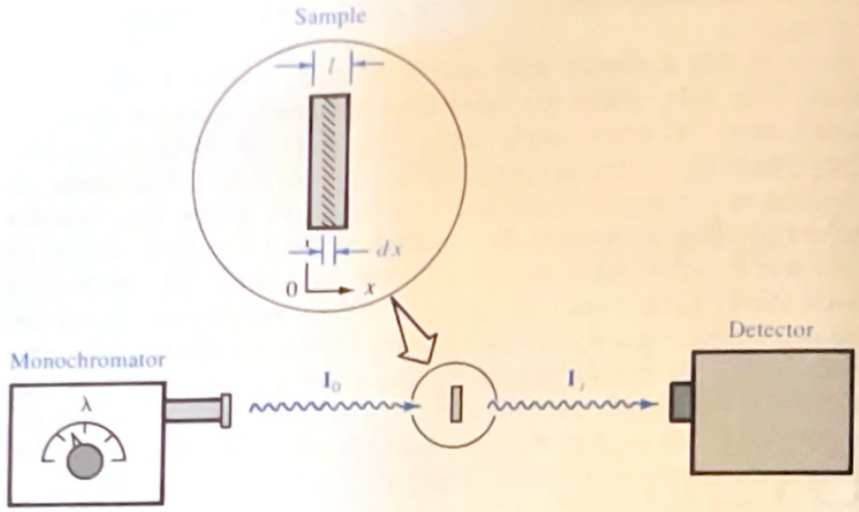


Figure 4-2 Optical absorption experiment.

lated by considering the probability of absorption within any increment dx. Since a photon which has survived to x without absorption has no memory of how far it has traveled, its probability of absorption in any dx is constant. Thus the degradation of the intensity $-d\mathbf{I}(x)/dx$ is proportional to the intensity remaining at x:

$$-\frac{d\mathbf{I}(x)}{dx} = \boldsymbol{\alpha}\mathbf{I}(x) \tag{4-1}$$

The solution to this equation is

$$\mathbf{I}(x) = \mathbf{I}_0 e^{-\boldsymbol{\alpha}x} \tag{4-2}$$

and the intensity of light transmitted through the sample thickness l is

$$\mathbf{I}_t = \mathbf{I}_0 e^{-\boldsymbol{\alpha}l} \tag{4-3}$$

The coefficient $\boldsymbol{\alpha}$ is called the *absorption coefficient* and has units of cm^{-1}. This coefficient will of course vary with the photon wavelength and with the material. In a typical plot of $\boldsymbol{\alpha}$ vs. wavelength (Fig. 4-3), there is negli-

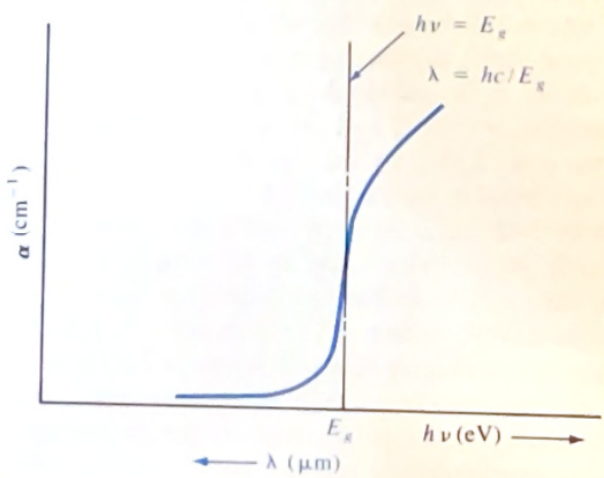


Figure 4-3 Dependence of optical absorption coefficient $\boldsymbol{\alpha}$ for a semiconductor on the wavelength of incident light.

gible absorption at long wavelengths ($h\nu$ small) and considerable absorption of photons with energies larger than E_g. According to Eq. (2-2), the relation between photon energy and wavelength is $E = hc/\lambda$. If E is given in electron volts and λ in micrometers, this becomes $E = 1.24/\lambda$.

Figure 4-4 indicates the band gap energies of some of the common semiconductors, relative to the visible, infrared, and ultraviolet portions of the spectrum. We observe that GaAs, Si, Ge, and InSb lie outside the visible region, in the infrared. Other semiconductors, such as GaP and CdS, have band gaps wide enough to pass photons in the visible range. It is important to note here that a semiconductor absorbs photons with energies equal to the band gap, *or larger*. Thus Si absorbs not only band gap light (~1 μm) but also shorter wavelengths, including those in the visible part of the spectrum.

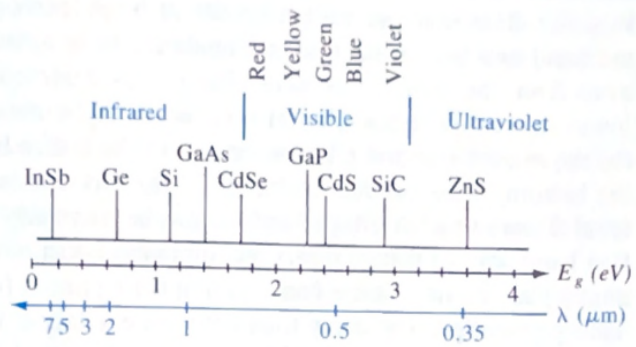


Figure 4-4 Band gaps of some common semiconductors relative to the optical spectrum.

4.2 LUMINESCENCE

When electron-hole pairs are generated in a semiconductor, or when carriers are excited into higher impurity levels from which they fall to their equilibrium states, light can be given off by the material. Many of the semiconductors are well suited for light emission, particularly the compound semiconductors with direct band gaps. The general property of light emission is called *luminescence*.† This overall category can be subdivided according to the excitation mechanism: If carriers are excited by photon absorption, the radiation resulting from the recombination of the excited carriers is called *photoluminescence;* if the excited carriers are created by high-energy electron bombardment of the material, the mechanism is called *cathodoluminescence;* if the excitation occurs by the introduction of current into the sample, the resulting luminescence is called *electroluminescence*. Other types of excitation are possible, but these three are the most important for device applications.

†The emission processes considered here should not be confused with radiation due to incandescence which occurs in heated materials. The various luminescent mechanisms can be considered "cold" processes as compared to the "hot" process of incandescence, which increases with temperature. In fact, most luminescent processes become more efficient as the temperature is lowered.

4.2.1 Photoluminescence

The simplest example of light emission from a semiconductor occurs for direct excitation and recombination of an EHP, as depicted in Fig. 3-5a. If the recombination occurs directly rather than via a defect level, band gap light is given off in the process. For steady state excitation, the recombination of EHPs occurs at the same rate as the generation, and one photon is emitted for each photon absorbed. Direct recombination is a fast process; the mean lifetime of the EHP is usually on the order of 10^{-8} s or less. Thus the emission of photons stops within approximately 10^{-8} s after the excitation is turned off. Such fast luminescent processes are often referred to as *fluorescence*. In some materials, however, emission continues for periods up to seconds or minutes after the excitation is removed. These slow processes are called *phosphorescence*, and the materials are called *phosphors*. An example of a slow process is shown in Fig. 4-5. This material contains a defect level (perhaps due to an impurity) in the band gap which has a strong tendency to temporarily capture (*trap*) electrons from the conduction band. The events depicted in the figure are as follows: (a) An incoming photon with $h\nu_1 > E_g$ is absorbed, creating an EHP; (b) the excited electron gives up energy to the lattice by scattering until it nears the bottom of the conduction band; (c) the electron is *trapped* by the impurity level E_t and remains trapped until it can be thermally reexcited to the conduction band (d); (e) finally direct recombination occurs as the electron falls to an empty state in the valence band, giving off a photon ($h\nu_2$) of approximately the band gap energy. The delay time between excitation and recombination can be relatively long if the probability of thermal reexcitation from the trap (d) is small. Even longer delay times result if the electron is retrapped several times before recombination. If the trapping probability is greater than the probability of recombination, an electron may make several trips between the trap and the conduction band before recombination finally occurs. In such a material the emission of phosphorescent light persists for a relatively long time after the excitation is removed.

The color of light emitted by a phosphor such as ZnS depends primarily on the impurities present, since many radiative transitions involve impurity levels

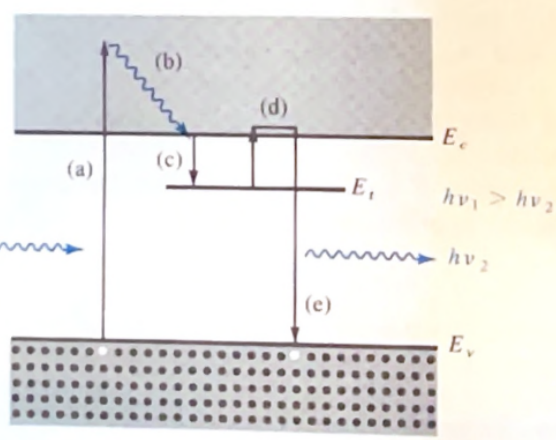


Figure 4-5 Excitation and recombination mechanisms in photoluminescence with a trapping level for electrons.

within the band gap. This selection of colors is particularly useful in the fabrication of a color television screen, as discussed in the next section.

One of the most common examples of photoluminescence is the fluorescent lamp. Typically such a lamp is composed of a glass tube filled with gas (e.g., a mixture of argon and mercury), with a fluorescent coating on the inside of the tube. When an electric discharge is induced between electrodes in the tube, the excited atoms of the gas emit photons, largely in the visible and ultraviolet regions of the spectrum. This light is absorbed by the luminescent coating, and visible photons are emitted. The efficiency of such a lamp is considerably better than that of an incandescent bulb, and the wavelength mixture in light given off can be adjusted by proper selection of the fluorescent material.

EXAMPLE 4-1

A 0.46-μm-thick sample of GaAs is illuminated with monochromatic light of $h\nu = 2$ eV. The absorption coefficient $\boldsymbol{\alpha}$ is 5×10^4 cm^{-1}. The power incident on the sample is 10 mW.

(a) Find the total energy absorbed by the sample per second (J/s).

(b) Find the rate of excess thermal energy given up by the electrons to the lattice before recombination (J/s).

(c) Find the number of photons per second given off from recombination events, assuming perfect quantum efficiency.

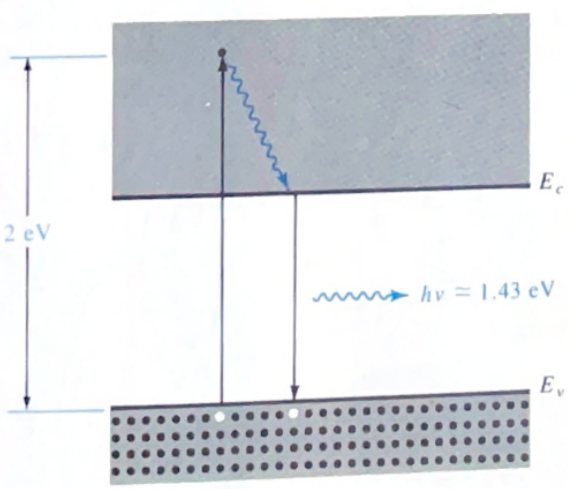
SOLUTION

(a) From Eq. (4-3),

$$\mathbf{I}_t = \mathbf{I}_0 e^{-\alpha l} = 10^{-2} \exp(-5 \times 10^4 \times 0.46 \times 10^{-4})$$
$$= 10^{-2} e^{-2.3} = 10^{-3} \text{ W}$$

Thus the absorbed power is

$$10 - 1 = 9 \text{ mW} = 9 \times 10^{-3} \text{ J/s}$$



(b) The fraction of each photon energy unit which is converted to heat is

$$\frac{2 - 1.43}{2} = 0.285$$

Thus the amount of energy converted to heat per second is

$$0.285 \times 9 \times 10^{-3} = 2.57 \times 10^{-3}\ \text{J/s}$$

(c) Assuming one emitted photon for each photon absorbed (perfect quantum efficiency), we have

$$\frac{9 \times 10^{-3}\ \text{J/s}}{1.6 \times 10^{-19}\ \text{J/eV} \times 2\ \text{eV/photon}} = 2.81 \times 10^{16}\ \text{photons/s}$$

Alternative solution: Recombination radiation accounts for $9 - 2.57 = 6.43$ mW at 1.43 eV/photon.

$$\frac{6.43 \times 10^{-3}}{1.6 \times 10^{-19} \times 1.43} = 2.81 \times 10^{16}\ \text{photons/s}$$

4.2.2 Cathodoluminescence

The most common example of the excitation of luminescent materials by energetic electrons is the cathode-ray tube (CRT). This light-emitting tube is the basis of the oscilloscope, television set, and other visual-display systems. The basic principle of the CRT is the selective excitation of a phosphorescent screen by a beam of energetic electrons within a vacuum tube (Fig. 4-6). Electrons are emitted from a heated cathode and are accelerated by an electric field to a positively charged anode. A beam of electrons passes through a hole in the anode and proceeds to the deflection system where the path of the electrons is bent by an electric or a magnetic field. By varying the deflection, the electron beam can be directed to any spot on the screen. As the energetic electrons collide with the phosphor coating, energy is given up to the phosphor as electrons are excited to higher states within the material. When these electrons recombine, the screen gives off light at the desired spot. In a television picture tube, the intensity and position of the beam is varied to give a visible replica of the image being received.

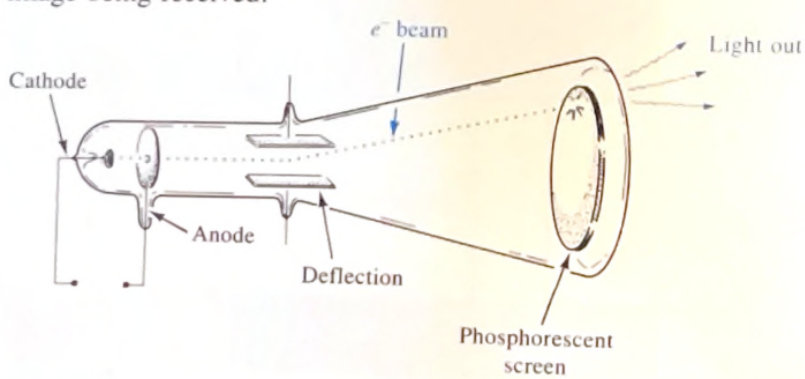


Figure 4-6 Simplified view of a cathode-ray tube.

In a color television tube the screen is coated with a pattern of phosphor dots, which are doped to emit the primary additive colors—red, green, and blue. To excite the proper dots, three electron beams are swept together through the pattern. Each beam controls the excitation of one type of phosphor dot. By varying the relative intensity of light emitted by each of the dots, a broad range of colors can be synthesized.

4.2.3 Electroluminescence

There are many ways by which electrical energy can be used to generate photon emission in a solid. In LEDs an electric current causes the injection of minority carriers into regions of the crystal where they can recombine with majority carriers, resulting in the emission of recombination radiation. This important effect (*injection electroluminescence*) will be discussed in Chapters 6 and 10 in terms of p-n junction theory.

The first electroluminescent effect to be observed was the emission of photons by certain phosphors in an alternating electric field (the Destriau effect). In this device, a phosphor powder such as ZnS is held in a binder material (often a plastic) of a high dielectric constant. When an a-c electric field is applied, light is given off by the phosphor. Such cells can be useful as lighting panels, although their efficiency has thus far been too low for most applications and their reliability is poor.

4.3 CARRIER LIFETIME AND PHOTOCONDUCTIVITY

When excess electrons and holes are created in a semiconductor, there is a corresponding increase in the conductivity of the sample as indicated by Eq. (3-43). If the excess carriers arise from optical excitation, the resulting increase in conductivity is called *photoconductivity*. This is an important effect, with useful applications in the analysis of semiconductor materials and in the operation of several types of devices. In this section we shall examine the mechanisms by which excess electrons and holes recombine and apply the recombination kinetics to the analysis of photoconductive devices. However, the importance of recombination is not limited to cases in which the excess carriers are created optically. In fact, virtually every semiconductor device depends in some way on the recombination of excess electrons and holes. Therefore, the concepts developed in this section will be used extensively in the analyses of diodes, transistors, lasers, and other devices in later chapters.

4.3.1 Direct Recombination of Electrons and Holes

It was pointed out in Section 3.1.4 that electrons in the conduction band of a semiconductor may make transitions to the valence band (i.e, recombine with holes in the valence band) either directly or indirectly. In direct recombination, an excess population of electrons and holes decays by electrons falling from the conduction band to empty states (holes) in the valence band. Energy lost by an electron in making the transition is given up as a photon. Direct recombina-

tion occurs *spontaneously;* that is, the probability that an electron and a hole will recombine is constant in time. As in the case of carrier scattering, this constant probability leads us to expect an exponential solution for the decay of the excess carriers. In this case the rate of decay of electrons at any time t is proportional to the number of electrons remaining at t and the number of holes, with some constant of proportionality for recombination, α_r. The *net* rate of change in the conduction band electron concentration is the thermal generation rate $\alpha_r n_i^2$ from Eq. (3-7) minus the recombination rate

$$\frac{dn(t)}{dt} = \alpha_r n_i^2 - \alpha_r n(t)p(t) \tag{4-4}$$

Let us assume the excess electron-hole population is created at $t = 0$, for example by a short flash of light, and the initial excess electron and hole concentrations Δn and Δp are equal.† Then as the electrons and holes recombine in pairs, the instantaneous concentrations of excess carriers $\delta n(t)$ and $\delta p(t)$ are also equal. Thus we can write the total concentrations of Eq. (4-4) in terms of the equilibrium values n_0 and p_0 and the excess carrier concentrations $\delta n(t) = \delta p(t)$. Using Eq. (3-24) we have

$$\begin{aligned}\frac{d\delta n(t)}{dt} &= \alpha_r n_i^2 - \alpha_r[n_0 + \delta n(t)][p_0 + \delta p(t)] \\ &= -\alpha_r[(n_0 + p_0)\delta n(t) + \delta n^2(t)]\end{aligned} \tag{4-5}$$

This nonlinear equation would be difficult to solve in its present form. Fortunately, it can be simplified for the case of low-level injection. If the excess carrier concentrations are small, we can neglect the δn^2 term. Furthermore, if the material is extrinsic, we can usually neglect the term representing the equilibrium minority carriers. For example, if the material is p-type ($p_0 \gg n_0$), Eq. (4-5) becomes

$$\frac{d\delta n(t)}{dt} = -\alpha_r p_0 \delta n(t) \tag{4-6}$$

The solution to this equation is an exponential decay from the original excess carrier concentration Δn:

$$\delta n(t) = \Delta n e^{-\alpha_r p_0 t} = \Delta n e^{-t/\tau_n} \tag{4-7}$$

Excess electrons in a p-type semiconductor recombine with a decay constant $\tau_n = (\alpha_r p_0)^{-1}$, called the *recombination lifetime*. Since the calculation is made in terms of the minority carriers, τ_n is often called the *minority carrier lifetime*. The decay of excess holes in n-type material occurs with $\tau_p = (\alpha_r n_0)^{-1}$. In the case of direct recombination, the excess majority carriers decay at exactly the same rate as the minority carriers.

---

†We will use $\delta n(t)$ and $\delta p(t)$ to mean instantaneous excess carrier concentrations, and Δn, Δp for their values at $t = 0$. Later we will use similar symbolism for spatial distributions, such as $\delta n(x)$ and $\Delta n(x = 0)$.

EXAMPLE 4-2

A numerical example may be helpful in visualizing the approximations made in the analysis of direct recombination. Let us assume a sample of GaAs is doped with 10^{15} acceptors/cm^3. The intrinsic carrier concentration of GaAs is approximately 10^6 cm^{-3}; thus the minority electron concentration is $n_0 = n_i^2/p_0 = 10^{-3}$ cm^{-3}. Certainly the approximation $p_0 \gg n_0$ is valid in this case. Now if 10^{14} EHP/cm^3 are created at $t = 0$, we can calculate the decay of these carriers in time. The approximation $\delta n \ll p_0$ is reasonable, as Fig. 4-7 indicates. This figure shows the decay in time of the excess populations for a carrier recombination lifetime of $\tau_n = \tau_p = 10^{-8}$ s.

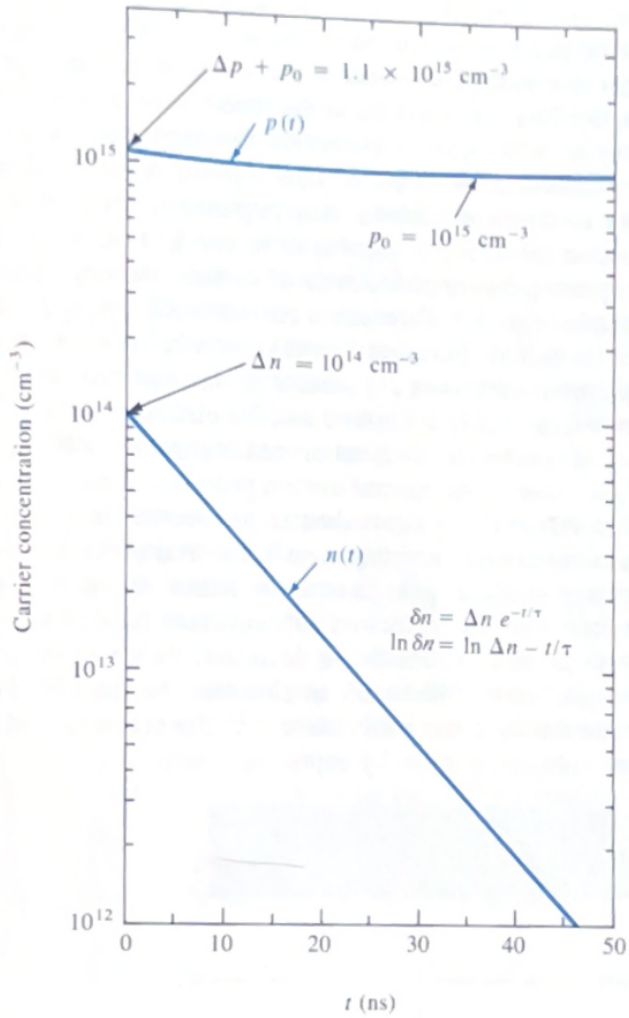


Figure 4-7 Decay of excess electrons and holes by recombination, for $\Delta n = \Delta p = 0.1p_0$, with n_0 negligible, and $\tau = 10$ ns (Example 4-2). The exponential decay of $\delta n(t)$ is linear on this semi-logarithmic graph.

There is a large percentage change in the minority carrier electron concentration in this example and a small percentage change in the majority hole concentration. Basically, the approximations of extrinsic material and low-level injection allow us to represent $n(t)$ in Eq. (4-4) by the excess concentration $\delta n(t)$ and $p(t)$ by the equilibrium value p_0. Figure 4-7 indicates that this is a

good approximation for the example. A more general expression for the carrier lifetime is

$$\tau_n = \frac{1}{\alpha_r(n_0 + p_0)} \tag{4-8}$$

This expression is valid for n- or p-type material if the injection level is low.

4.3.2 Indirect Recombination; Trapping

In column IV semiconductors and in certain compounds, the probability of direct electron-hole recombination is very small (Appendix III). There is some band gap light given off by materials such as Si and Ge during recombination, but this radiation is very weak and may be detected only by sensitive equipment. The vast majority of the recombination events in indirect materials occur via *recombination levels* within the band gap, and the resulting energy loss by recombining electrons is usually given up to the lattice as heat rather than by the emission of photons. Any impurity or lattice defect can serve as a recombination center if it is capable of receiving a carrier of one type and subsequently capturing the opposite type of carrier, thereby annihilating the pair. For example, Fig. 4-8 illustrates a recombination level E_r which is below E_F at equilibrium and therefore is substantially filled with electrons. When excess electrons and holes are created in this material, each EHP recombines at E_r in two steps: (a) hole capture and (b) electron capture.

Since the recombination centers in Fig. 4-8 are filled at equilibrium, the first event in the recombination process is hole capture. It is important to note that this event is equivalent to an electron at E_r falling to the valence band, leaving behind an empty state in the recombination level. Thus in hole capture, energy is *given up* as heat to the lattice. Similarly, energy is given up when a conduction band electron subsequently falls to the empty state in E_r. When both of these events have occurred, the recombination center is back to its original state (filled with an electron), but an EHP is missing. Thus one EHP recombination has taken place, and the center is ready to participate in another recombination event by capturing a hole.

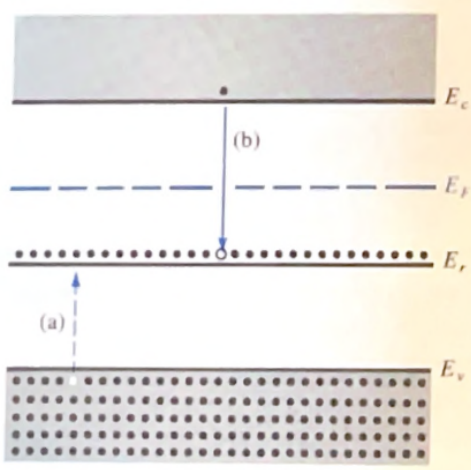


Figure 4-8 Capture processes at a recombination level: (a) hole capture at a filled recombination center; (b) electron capture at an empty center.

The carrier lifetime resulting from indirect recombination is somewhat more complicated than is the case for direct recombination, since it is necessary to account for unequal times required for capturing each type of carrier. In particular, recombination is often delayed by the tendency for a captured carrier to be thermally reexcited to its original band before capture of the opposite type of carrier can occur (Section 4.2.1). For example, if electron capture (b) does not follow immediately after hole capture (a) in Fig. 4-8, the hole may be thermally reexcited to the valence band. Energy is required for this process, which is equivalent to a valence band electron being raised to the empty state in the recombination level. This process delays the recombination, since the hole must be captured again before recombination can be completed.

When a carrier is trapped temporarily at a center and then is reexcited without recombination taking place, the process is often called *temporary trapping*. Although the nomenclature varies somewhat, it is common to refer to an impurity or defect center as a *trapping center* (or simply *trap*) if, after capture of one type of carrier, the most probable next event is reexcitation. If the most probable next event is capture of the opposite type of carrier, the center is predominately a recombination center. The recombination can be slow or fast, depending on the average time the first carrier is held before the second carrier is captured. In general, trapping levels located deep in the band gap are slower in releasing trapped carriers than are the levels located near one of the bands. This results from the fact that more energy is required, for example, to reexcite a trapped electron from a center near the middle of the gap to the conduction band than is required to reexcite an electron from a level closer to the conduction band.

As an example of impurity levels in semiconductors, Fig. 4-9 shows the energy level positions of various impurities in Si. In this diagram a superscript indicates whether the impurity is positive (donor) or negative (acceptor) when ionized. Some impurities introduce multiple levels in the band gap; for example, Zn introduces a level (Zn^-) located 0.31 eV above the valence band and a second level ($Zn^=$) near the middle of the gap. Each Zn impurity atom is capable of accepting two electrons from the semiconductor, one in the lower level and then one in the upper level.

The effects of recombination and trapping can be measured by a *photoconductive decay* experiment. As Fig. 4-7 shows, a population of excess electrons and holes disappears with a decay constant characteristic of the particular recombination process. The conductivity of the sample during the decay is

$$\sigma(t) = q[n(t)\mu_n + p(t)\mu_p] \tag{4-9}$$

Therefore, the time dependence of the carrier concentrations can be monitored by recording the sample resistance as a function of time. A typical experimental arrangement is shown schematically in Fig. 4-10. A source of short pulses of light is required, along with an oscilloscope for displaying the sample voltage as the resistance varies. Microsecond light pulses can be obtained by periodically discharging a capacitor through a flash tube containing a gas such as

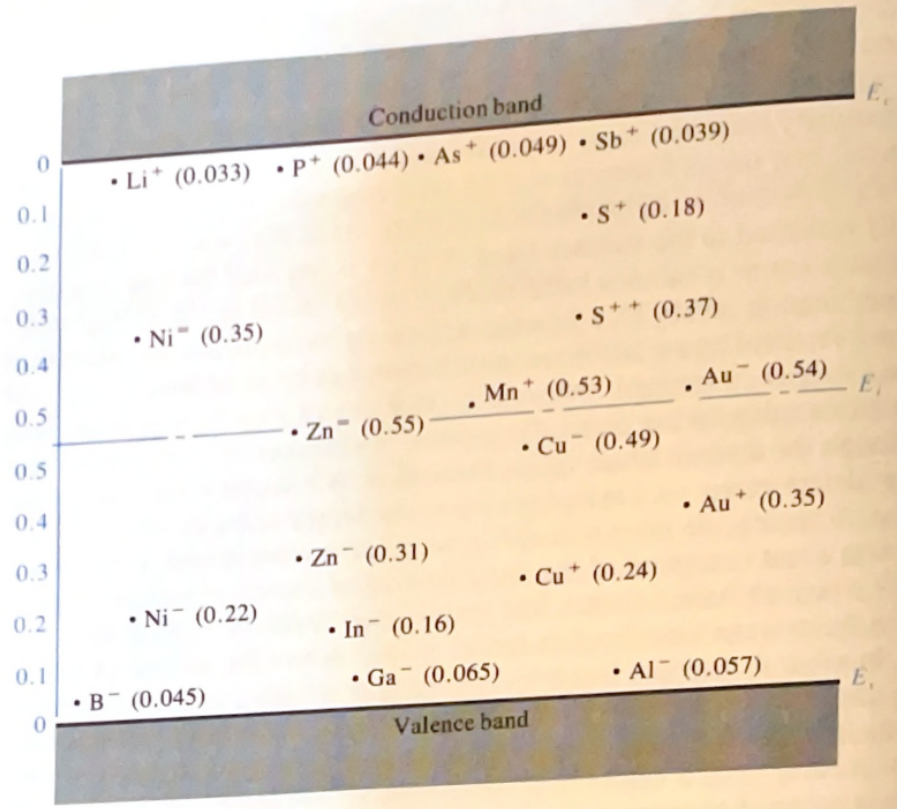


Figure 4-9
Energy levels of impurities in Si. The energies are measured from the nearest band edge (E_v or E_c); donor levels are designated by a plus sign and acceptors by a minus sign.†

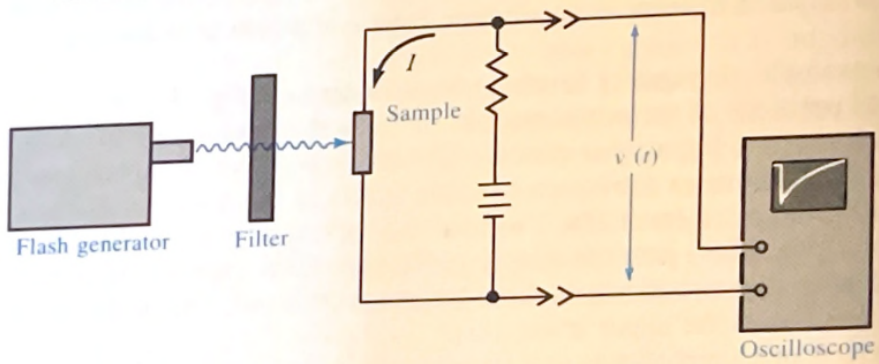


Figure 4-10
Experimental arrangement for photoconductive decay measurements, and a typical oscilloscope trace.

xenon. For shorter pulses, special techniques such as the use of a pulsed laser (Chapter 10) must be used.

4.3.3 Steady State Carrier Generation; Quasi-Fermi Levels

In the discussion above we emphasized the transient decay of an excess EHP population. However, the various recombination mechanisms are also impor-

†References: S.M. Sze and J.C. Irvin, "Resistivity, Mobility, and Impurity Levels in GaAs, Ge, and Si at 300°K," *Solid State Electronics*, vol. 11, pp. 599–602 (June 1968); E. Schibli and A.G. Milnes, "Deep Impurities in Silicon," *Materials Science and Engineering*, vol. 2, pp. 173–180 (1967).

tant in a sample at thermal equilibrium or with a steady state EHP generation–recombination balance.† For example, a semiconductor at equilibrium experiences thermal generation of EHPs at a rate $g(T) = g_i$ described by Eq. (3-7). This generation is balanced by the recombination rate so that the equilibrium concentrations of carriers n_0 and p_0 are maintained:

$$g(T) = \alpha_r n_i^2 = \alpha_r n_0 p_0 \tag{4-10}$$

This equilibrium rate balance can include generation from defect centers as well as band-to-band generation.

If a steady light is shone on the sample, an optical generation rate g_{op} will be added to the thermal generation, and the carrier concentrations n and p will increase to new steady state values. We can write the balance between generation and recombination in terms of the equilibrium carrier concentrations and the departures from equilibrium δn and δp:

$$g(T) + g_{op} = \alpha_r np = \alpha_r(n_0 + \delta n)(p_0 + \delta p) \tag{4-11}$$

For steady state recombination and no trapping, $\delta n = \delta p$; thus Eq. (4-11) becomes

$$g(T) + g_{op} = \alpha_r n_0 p_0 + \alpha_r[(n_0 + p_0)\delta n + \delta n^2] \tag{4-12}$$

The term $\alpha_r n_0 p_0$ is just equal to the thermal generation rate $g(T)$. Thus, neglecting the δn^2 term for low-level excitation, we can rewrite Eq. (4-12) as

$$g_{op} = \alpha_r(n_0 + p_0)\delta n = \frac{\delta n}{\tau_n} \tag{4-13}$$

The excess carrier concentration can be written as

$$\delta n = \delta p = g_{op}\tau_n \tag{4-14}$$

More general expressions are given in Eq. (4-16), which allow for the case $\tau_p \neq \tau_n$, when trapping is present.

EXAMPLE 4-3

As a numerical example, let us assume that 10^{13} EHP/cm^3 are created optically every microsecond in a Si sample with $n_0 = 10^{14}$ cm^{-3} and $\tau_n = \tau_p = 2$ μsec. The steady state excess electron (or hole) concentration is then 2×10^{13} cm^{-3} from Eq. (4-14). While the percentage change in the majority electron concentration is small, the minority carrier concentration changes from

$$p_0 = n_i^2/n_0 = (2.25 \times 10^{20})/10^{14} = 2.25 \times 10^6 \text{ cm}^{-3} \qquad (\textit{equilibrium})$$

†The term *equilibrium* refers to a condition of no external excitation except for temperature, and no net motion of charge (e.g., a sample at a constant temperature, in the dark, with no fields applied). *Steady state* refers to a nonequilibrium condition in which all processes are constant and are balanced by opposing processes (e.g., a sample with a constant current or a constant optical generation of EHPs just balanced by recombination).

to

$$p = 2 \times 10^{13}\ \text{cm}^{-3} \qquad \textit{(steady state)}$$

Note that the equilibrium equation $n_0 p_0 = n_i^2$ cannot be used with the subscripts removed; that is, $np \neq n_i^2$ when excess carriers are present.

It is often desirable to refer to the steady state electron and hole concentrations in terms of Fermi levels, which can be included in band diagrams for various devices. The Fermi level E_F used in Eq. (3-25) is meaningful only when no excess carriers are present. However, we can write expressions for the steady state concentrations in the same *form* as the equilibrium expressions by defining separate *quasi-Fermi levels* F_n and F_p for electrons and holes. The resulting carrier concentration equations

$$\boxed{n = n_i e^{(F_n - E_i)/kT}} \tag{4-15a}$$

$$\boxed{p = n_i e^{(E_i - F_p)/kT}} \tag{4-15b}$$

can be considered as defining relations for the quasi-Fermi levels.†

EXAMPLE 4-4 In Example 4-3, the steady state electron concentration is

$$n = n_0 + \delta n = 1.2 \times 10^{14} = (1.5 \times 10^{10}) e^{(F_n - E_i)/0.0259}$$

where $kT \simeq 0.0259$ eV at room temperature. Thus the electron quasi-Fermi level position $F_n - E_i$ is found from

$$F_n - E_i = 0.0259 \ln(8 \times 10^3) = 0.233 \text{ eV}$$

and F_n lies 0.233 eV above the intrinsic level. By a similar calculation, the hole quasi-Fermi level lies 0.186 eV below E_i (Fig. 4-11). In this example, the equilibrium Fermi level is $0.0259 \ln(6.67 \times 10^3) = 0.228$ eV above the intrinsic level.

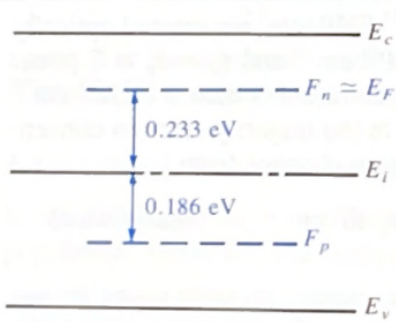


Figure 4-11 Quasi-Fermi levels F_n and F_p for a Si sample with $n_0 = 10^{14}$ cm^{-3}, $\tau_p = 2\ \mu$s, and $g_{op} = 10^{19}$ EHP/cm^3-s (Example 4-4).

†In some texts the quasi-Fermi level is called *IMREF*, which is Fermi spelled backward.

The quasi-Fermi levels of Fig. 4-11 illustrate dramatically the deviation from equilibrium caused by the optical excitation; the steady state F_n is only slightly above the equilibrium E_F, whereas F_p is greatly displaced below E_F. From the figure it is obvious that the excitation causes a large percentage change in minority carrier hole concentration and a relatively small change in the electron concentration.

In summary, the quasi-Fermi levels F_n and F_p are the steady state analogues of the equilibrium Fermi level E_F. When excess carriers are present, the deviations of F_n and F_p from E_F indicate how far the electron and hole populations are from the equilibrium values n_0 and p_0. A given concentration of excess EHPs causes a large shift in the minority carrier quasi-Fermi level compared with that for the majority carriers. The separation of the quasi-Fermi levels $F_n - F_p$ is a direct measure of the deviation from equilibrium (at equilibrium $F_n = F_p = E_F$). The concept of quasi-Fermi levels is very useful in visualizing minority and majority carrier concentrations in devices where these quantities vary with position.

4.3.4 Photoconductive Devices

There are a number of applications for devices which change their resistance when exposed to light. For example, such light detectors can be used in the home to control automatic night lights which turn on at dusk and turn off at dawn. They can also be used to measure illumination levels, as in exposure meters for cameras. Many systems include a light beam aimed at the photoconductor, which signals the presence of an object between the source and detector. Such systems are useful in moving-object counters, burglar alarms, and many other applications. Detectors are used in optical signaling systems in which information is transmitted by a light beam and is received at a photoconductive cell.

Considerations in choosing a photoconductor for a given application include the sensitive wavelength range, time response, and optical sensitivity of the material. In general, semiconductors are most sensitive to photons with energies equal to the band gap or slightly more energetic than band gap. Less energetic photons are not absorbed, and photons with $h\nu \gg E_g$ are absorbed at the surface and contribute little to the bulk conductivity. Therefore, the table of band gaps (Appendix III) indicates the photon energies to which most semiconductor photodetectors respond. For example, CdS ($E_g = 2.42$ eV) is commonly used as a photoconductor in the visible range, and narrow-gap materials such as Ge (0.67 eV) and InSb (0.18 eV) are useful in the infrared portion of the spectrum. Some photoconductors respond to excitations of carriers from impurity levels within the band gap and therefore are sensitive to photons of less than band gap energy.

The optical sensitivity of a photoconductor can be evaluated by examining the steady state excess carrier concentrations generated by an optical generation rate g_{op}. If the mean time each carrier spends in its respective band before capture is τ_n and τ_p, we have

$$\delta n = \tau_n g_{op} \quad \text{and} \quad \delta p = \tau_p g_{op} \tag{4-16}$$

and the photoconductivity change is

$$\Delta\sigma = qg_{op}(\tau_n\mu_n + \tau_p\mu_p) \tag{4-17}$$

For simple recombination, τ_n and τ_p will be equal. If trapping is present, however, one of the carriers may spend little time in its band before being trapped. From Eq. (4-17) it is obvious that for maximum photoconductive response, we want high mobilities and long lifetimes. Some semiconductors are especially good candidates for photoconductive devices because of their high mobility; for example, InSb has an electron mobility of about 10^5 cm^2/V-s and therefore is used as a sensitive infrared detector in many applications.

The time response of a photoconductive cell is limited by the recombination times, the degree of carrier trapping, and the time required for carriers to drift through the device in an electric field. Often these properties can be adjusted by proper choice of material and device geometry, but in some cases improvements in response time are made at the expense of sensitivity. For example, the drift time can be reduced by making the device short, but this substantially reduces the responsive area of the device. In addition, it is often desirable that the device have a large dark resistance, and for this reason, shortening the length may not be practical. There is usually a compromise between sensitivity, response time, dark resistance, and other requirements in choosing a device for a particular application.

4.4 DIFFUSION OF CARRIERS

When excess carriers are created nonuniformly in a semiconductor, the electron and hole concentrations vary with position in the sample. Any such spatial variation (*gradient*) in n and p calls for a net motion of the carriers from regions of high carrier concentration to regions of low carrier concentration. This type of motion is called *diffusion* and represents an important charge transport process in semiconductors. The two basic processes of current conduction are diffusion due to a carrier gradient and drift in an electric field.

4.4.1 Diffusion Processes

When a bottle of perfume is opened in one corner of a closed room, the scent is soon detected throughout the room. If there is no convection or other net motion of the air, the scent spreads by diffusion. The diffusion is the natural result of the *random motion* of the individual molecules. Consider, for example, a volume of arbitrary shape with scented air molecules inside and unscented molecules outside the volume. All the molecules undergo random thermal motion and collisions with other molecules. Thus each molecule moves in an arbitrary direction until it collides with another air molecule, after which it moves in a new direction. If the motion is truly random, a molecule at

the edge of the volume has equal probabilities of moving into or out of the volume on its next step (assuming the curvature of the surface is negligible on the molecular scale). Therefore, after a mean free time $\bar{t}$, half the molecules at the edge will have moved into the volume and half will have moved out of the volume. The net effect is that the volume containing scented molecules has increased. This process will continue until the molecules are uniformly distributed in the room. Only then will a given volume gain as many molecules as it loses in a given time. In other words, net diffusion will continue as long as gradients exist in the distribution of scented molecules.

Carriers in a semiconductor diffuse in a carrier gradient by random thermal motion and scattering from the lattice and impurities. For example, a pulse of excess electrons injected at $x = 0$ at time $t = 0$ will spread out in time as shown in Fig. 4-12. Initially, the excess electrons are concentrated at $x = 0$; as time passes, however, electrons diffuse to regions of low electron concentration until finally $n(x)$ is constant.

We can calculate the rate at which the electrons diffuse in a one-dimensional problem by considering an arbitrary distribution $n(x)$ such as Fig. 4-13a. Since the mean free path $\bar{l}$ between collisions is a small incremental distance, we can divide x into segments $\bar{l}$ wide, with $n(x)$ evaluated at the center of each segment (Fig. 4-13b).

The electrons in segment (1) to the left of x_0 in Fig. 4-13b have equal chances of moving left or right, and in a mean free time $\bar{t}$ one-half of them will move into segment (2). The same is true of electrons within one mean free path of x_0 to the right; one-half of these electrons will move through x_0 from right to left in a mean free time. Therefore, the *net* number of electrons passing x_0 from left to right in one mean free time is $\frac{1}{2}(n_1\bar{l}A) - \frac{1}{2}(n_2\bar{l}A)$, where the area

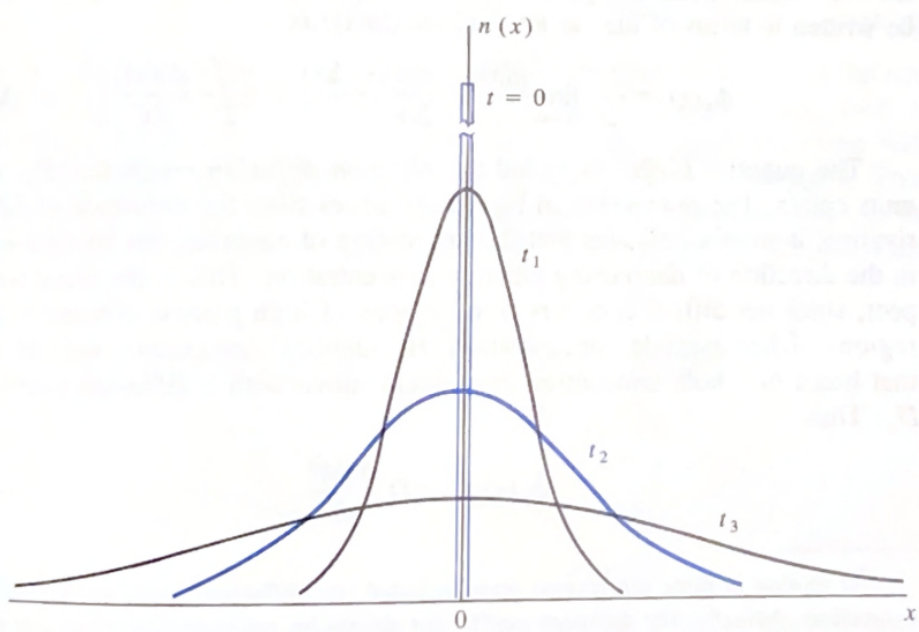


Figure 4-12
Spreading of a pulse of electrons by diffusion.

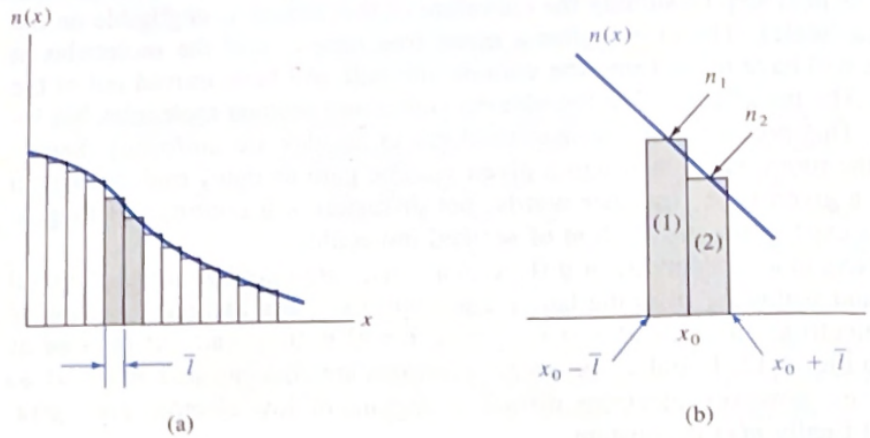


Figure 4-13 An arbitrary electron concentration gradient in one dimension: (a) division of $n(x)$ into segments of length equal to a mean free path for the electrons; (b) expanded view of two of the segments centered at x_0.

perpendicular to x is A. The rate of electron flow in the $+x$-direction per unit area (the electron flux density ϕ_n) is given by

$$\phi_n(x_0) = \frac{\bar{l}}{2\bar{t}}(n_1 - n_2) \tag{4-18}$$

Since the mean free path $\bar{l}$ is a small differential length, the difference in electron concentration $(n_1 - n_2)$ can be written as

$$n_1 - n_2 = \frac{n(x) - n(x + \Delta x)}{\Delta x}\bar{l} \tag{4-19}$$

where x is taken at the center of segment (1) and $\Delta x = \bar{l}$. In the limit of small Δx (i.e., small mean free path $\bar{l}$ between scattering collisions), Eq. (4-18) can be written in terms of the carrier gradient $dn(x)/dx$:

$$\phi_n(x) = \frac{\bar{l}^2}{2\bar{t}} \lim_{\Delta x \to 0} \frac{n(x) - n(x + \Delta x)}{\Delta x} = \frac{-\bar{l}^2}{2\bar{t}} \frac{dn(x)}{dx} \tag{4-20}$$

The quantity $\bar{l}^2/2\bar{t}$ † is called the *electron diffusion coefficient* D_n, with units cm^2/s. The minus sign in Eq. (4-20) arises from the definition of the derivative; it simply indicates that the net motion of electrons due to diffusion is in the direction of *decreasing* electron concentration. This is the result we expect, since net diffusion occurs from regions of high particle concentration to regions of low particle concentration. By identical arguments, we can show that holes in a hole concentration gradient move with a diffusion coefficient D_p. Thus

$$\phi_n(x) = -D_n \frac{dn(x)}{dx} \tag{4-21a}$$

† If motion in three dimensions were included, the diffusion would be smaller in the x-direction. Actually, the diffusion coefficient should be calculated from the true energy distributions and scattering mechanisms. Diffusion coefficients are usually determined experimentally for a particular material, as described in Section 4.4.5.

$$\phi_p(x) = -D_p \frac{dp(x)}{dx} \tag{4-21b}$$

The diffusion current crossing a unit area (the current density) is the particle flux density multiplied by the charge of the carrier:

$$J_n(\text{diff.}) = -(-q)D_n \frac{dn(x)}{dx} = +qD_n \frac{dn(x)}{dx} \tag{4-22a}$$

$$J_p(\text{diff.}) = -(+q)D_p \frac{dp(x)}{dx} = -qD_p \frac{dp(x)}{dx} \tag{4-22b}$$

It is important to note that electrons and holes move together in a carrier gradient [Eqs. (4-21)], but the resulting currents are in opposite directions [Eqs. (4-22)] because of the opposite charge of electrons and holes.

4.4.2 Diffusion and Drift of Carriers; Built-in Fields

If an electric field is present in addition to the carrier gradient, the current densities will each have a drift component and a diffusion component

$$\boxed{\begin{aligned} J_n(x) &= \underbrace{q\mu_n n(x)\mathscr{E}(x)}_{\text{drift}} + \underbrace{qD_n \frac{dn(x)}{dx}}_{\text{diffusion}} \\ J_p(x) &= q\mu_p p(x)\mathscr{E}(x) - qD_p \frac{dp(x)}{dx} \end{aligned}} \quad \begin{aligned} &(4\text{-}23a) \\ &(4\text{-}23b) \end{aligned}$$

and the total current density is the sum of the contributions due to electrons and holes:

$$J(x) = J_n(x) + J_p(x) \tag{4-24}$$

We can best visualize the relation between the particle flow and the currents of Eqs. (4-23) by considering a diagram such as shown in Fig. 4-14. In this figure an electric field is assumed to be in the x-direction, along with carrier distributions $n(x)$ and $p(x)$ which decrease with increasing x. Thus the derivatives in Eqs. (4-21) are negative, and diffusion takes place in the $+x$-direction. The resulting electron and hole diffusion currents [J_n(diff.) and

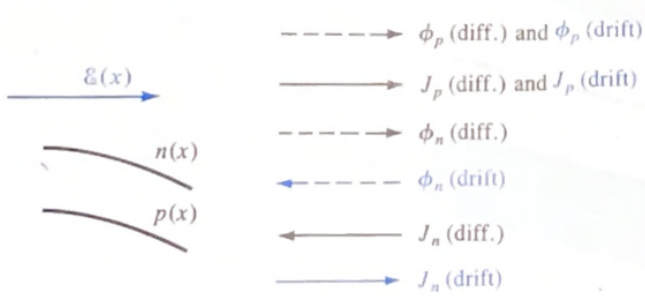


Figure 4-14 Drift and diffusion directions for electrons and holes in a carrier gradient and an electric field. Particle flow directions are indicated by dashed arrows, and the resulting currents are indicated by solid arrows.

J_p(diff.)] are in opposite directions, according to Eqs. (4-22). Holes drift in the direction of the electric field [ϕ_p(drift)], whereas electrons drift in the opposite direction because of their negative charge. The resulting drift current is in the $+x$-direction in each case. Note that the drift and diffusion components of the current are additive for holes when the field is in the direction of decreasing hole concentration, whereas the two components are subtractive for electrons under similar conditions. The total current may be due primarily to the flow of electrons or holes, depending on the relative concentrations and the relative magnitudes and directions of electric field and carrier gradients.

An important result of Eqs. (4-23) is that minority carriers can contribute significantly to the current through diffusion. Since the drift terms are proportional to carrier concentration, minority carriers seldom provide much drift current. On the other hand, diffusion current is proportional to the *gradient* of concentration. For example, in n-type material the minority hole concentration p may be many orders of magnitude smaller than the electron concentration n, but the gradient dp/dx may be significant. As a result, minority carrier currents through diffusion can sometimes be as large as majority carrier currents.

In discussing the motion of carriers in an electric field, we should indicate the influence of the field on the energies of electrons in the band diagrams. Assuming an electric field $\mathscr{E}(x)$ in the x-direction, we can draw the energy bands as in Fig. 4-15, to include the change in potential energy of electrons in the field. Since electrons drift in a direction opposite to the field, we expect the potential energy for electrons to increase in the direction of the field, as in Fig. 4-15. The electrostatic potential $\mathscr{V}(x)$ varies in the opposite direction, since it is defined in terms of positive charges and is therefore related to the electron potential energy $E(x)$ displayed in the figure by $\mathscr{V}(x) = E(x)/(-q)$.

From the definition of electric field,

$$\mathscr{E}(x) = -\frac{d\mathscr{V}(x)}{dx} \tag{4-25}$$

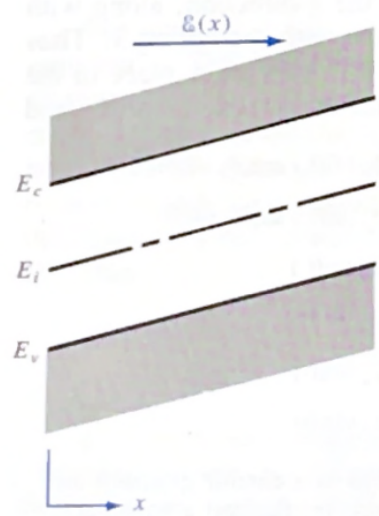


Figure 4-15 Energy band diagram of a semiconductor in an electric field $\mathscr{E}(x)$.

we can relate $\mathscr{E}(x)$ to the electron potential energy in the band diagram by choosing some reference in the band for the electrostatic potential. We are interested only in the spatial variation in $\mathcal{V}(x)$ for Eq. (4-25). Choosing E_i as a convenient reference, we can relate the electric field to this reference by

$$\boxed{\mathscr{E}(x) = -\frac{d\mathcal{V}(x)}{dx} = -\frac{d}{dx}\left[\frac{E_i}{(-q)}\right] = \frac{1}{q}\frac{dE_i}{dx}} \tag{4-26}$$

Therefore, the variation of band energies with $\mathscr{E}(x)$ as drawn in Fig. 4-15 is correct. The direction of the slope in the bands relative to $\mathscr{E}$ is simple to remember: Since the diagram indicates electron energies, we know the slope in the bands must be such that electrons drift "downhill" in the field. Therefore, $\mathscr{E}$ points "uphill" in the band diagram.

At equilibrium, no net current flows in a semiconductor. Thus any fluctuation which would begin a diffusion current also sets up an electric field which redistributes carriers by drift. An examination of the requirements for equilibrium indicates that the diffusion coefficient and mobility must be related. Setting Eq. (4-23b) equal to zero for equilibrium, we have

$$\mathscr{E}(x) = \frac{D_p}{\mu_p}\frac{1}{p(x)}\frac{dp(x)}{dx} \tag{4-27}$$

Using Eq. (3-25b) for $p(x)$,

$$\mathscr{E}(x) = \frac{D_p}{\mu_p}\frac{1}{kT}\left(\frac{dE_i}{dx} - \frac{dE_F}{dx}\right) \tag{4-28}$$

The equilibrium Fermi level does not vary with x, and the derivative of E_i is given by Eq. (4-26). Thus Eq. (4-28) reduces to

$$\boxed{\frac{D}{\mu} = \frac{kT}{q}} \tag{4-29}$$

This result is obtained for either carrier type. This important equation is called the *Einstein relation*. It allows us to calculate either D or μ from a measurement of the other. Table 4-1 lists typical values of D and μ for several semiconductors at room temperature. It is clear from these values that $D/\mu \simeq 0.026$ V.

Table 4-1. Diffusion coefficient and mobility of electrons and holes for intrinsic semiconductors at 300 K. *Note:* Use Fig. 3-23 for doped semiconductors.

	D_n (cm^2/s)	D_p (cm^2/s)	μ_n (cm^2/V-s)	μ_p (cm^2/V-s)
Ge	100	50	3900	1900
Si	35	12.5	1350	480
GaAs	220	10	8500	400

An important result of the balance of drift and diffusion at equilibrium is that *built-in* fields accompany gradients in E_i [see Eq. (4-26)]. Such gradients in the bands at equilibrium (E_F constant) can arise when the band gap varies due to changes in alloy composition. More commonly, built-in fields result from doping gradients. For example, a donor distribution $N_d(x)$ causes a gradient in $n_0(x)$, which must be balanced by a built-in electric field $\mathscr{E}(x)$.

EXAMPLE 4-5 An intrinsic Si sample is doped with donors from one side such that $N_d = N_0 \exp(-ax)$. (a) Find an expression for $\mathscr{E}(x)$ at equilibrium over the range for which $N_d \gg n_i$. (b) Evaluate $\mathscr{E}(x)$ when $a = 1(\mu\text{m})^{-1}$. (c) Sketch a band diagram such as in Fig. 4-15 and indicate the direction of $\mathscr{E}$.

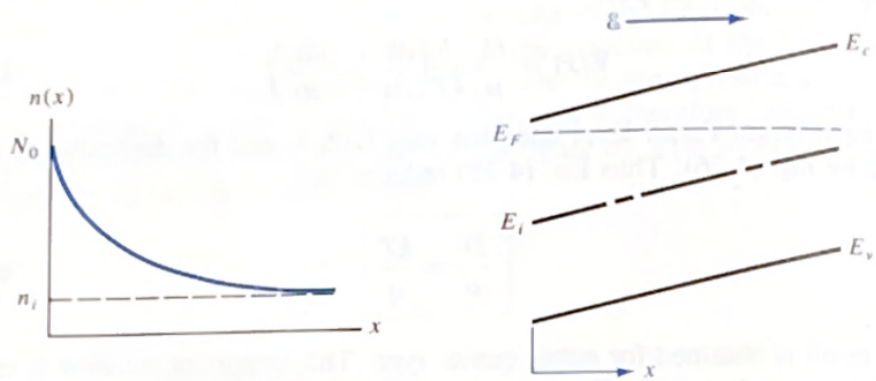
SOLUTION (a) From Eq. (4-23a),

$$\mathscr{E}(x) = -\frac{D_n}{\mu_n}\frac{dn/dx}{n} = -\frac{kT}{q}\frac{N_0(-a)e^{-ax}}{N_0 e^{-ax}} = +\frac{kT}{q}a$$

We notice for this exponential impurity distribution, $\mathscr{E}(x)$ depends on a but not on N_0 or x.

(b) $\mathscr{E}(x) = 0.0259(10^4) = 259$ V/cm

(c)



4.4.3 Diffusion and Recombination; The Continuity Equation

In the discussion of diffusion of excess carriers, we have thus far neglected the important effects of recombination. These effects must be included in a description of conduction processes, however, since recombination can cause a variation in the carrier distribution. For example, consider a differential length Δx of a semiconductor sample with area A in the yz-plane (Fig. 4-16). The hole current density leaving the volume, $J_p(x + \Delta x)$, can be larger or smaller than the current density entering, $J_p(x)$, depending on the generation and recombination of carriers taking place within the volume. The net increase in hole concentration per unit time, $\partial p/\partial t$, is the difference between the hole flux per unit

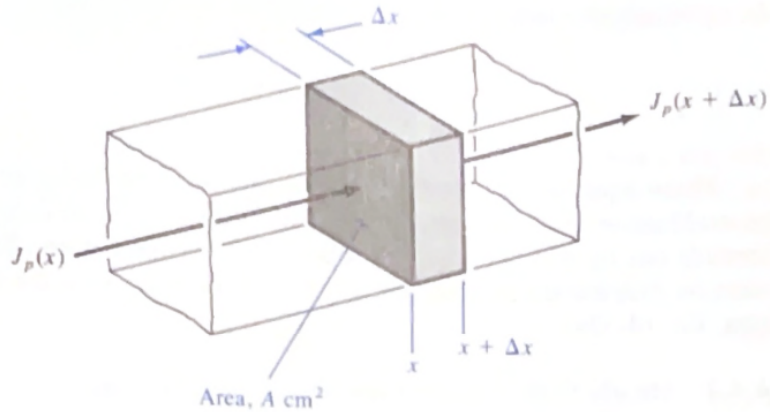


Figure 4-16 Current entering and leaving a volume ΔxA.

volume entering and leaving, minus the recombination rate. We can convert hole current density to hole particle flux density by dividing J_p by q. The current densities are already expressed per unit area; thus dividing $J_p(x)/q$ by Δx gives the number of carriers per unit volume entering ΔxA per unit time, and $(1/q)J_p(x + \Delta x)/\Delta x$ is the number leaving per unit volume and time:

$$\left.\frac{\partial p}{\partial t}\right|_{x \to x+\Delta x} = \frac{1}{q}\frac{J_p(x) - J_p(x + \Delta x)}{\Delta x} - \frac{\delta p}{\tau_p} \tag{4-30}$$

$$\text{Rate of hole buildup} = \text{increase of hole concentration in } \Delta xA \text{ per unit time} - \text{recombination rate}$$

As Δx approaches zero, we can write the current change in derivative form:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial \delta p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{\delta p}{\tau_p} \tag{4-31a}$$

The expression (4-31a) is called the *continuity equation* for holes. For electrons we can write

$$\frac{\partial \delta n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{\delta n}{\tau_n} \tag{4-31b}$$

since the electronic charge is negative.

When the current is carried strictly by diffusion (negligible drift), we can replace the currents in Eqs. (4-31) by the expressions for diffusion current; for example, for electron diffusion we have

$$J_n(\text{diff.}) = qD_n\frac{\partial \delta n}{\partial x} \tag{4-32}$$

Substituting this into Eq. (4-31b) we obtain the *diffusion equation* for electrons,

$$\boxed{\frac{\partial \delta n}{\partial t} = D_n\frac{\partial^2 \delta n}{\partial x^2} - \frac{\delta n}{\tau_n}} \tag{4-33a}$$

and similarly for holes,

$$\boxed{\frac{\partial \delta p}{\partial t} = D_p \frac{\partial^2 \delta p}{\partial x^2} - \frac{\delta p}{\tau_p}} \tag{4-33b}$$

These equations are useful in solving transient problems of diffusion with recombination. For example, a pulse of electrons in a semiconductor (Fig. 4-12) spreads out by diffusion and disappears by recombination. To solve for the electron distribution in time, $n(x, t)$, we would begin with the diffusion equation, Eq. (4-33a).

4.4.4 Steady State Carrier Injection; Diffusion Length

In many problems a steady state distribution of excess carriers is maintained, such that the time derivatives in Eqs. (4-33) are zero. In the steady state case the diffusion equations become

$$\frac{d^2 \delta n}{dx^2} = \frac{\delta n}{D_n \tau_n} \equiv \frac{\delta n}{L_n^2} \tag{4-34a}$$

$$\frac{d^2 \delta p}{dx^2} = \frac{\delta p}{D_p \tau_p} \equiv \frac{\delta p}{L_p^2} \qquad (\textit{steady state}) \tag{4-34b}$$

where $L_n \equiv \sqrt{D_n \tau_n}$ is called the electron *diffusion length* and L_p is the diffusion length for holes. We no longer need partial derivatives, since the time variation is zero for steady state.

The physical significance of the diffusion length can be understood best by an example. Let us assume that excess holes are somehow injected into a semi-infinite semiconductor bar at $x = 0$, and the steady state hole injection maintains a constant excess hole concentration at the injection point $\delta p(x = 0) = \Delta p$. The injected holes diffuse along the bar, recombining with a characteristic lifetime τ_p. In steady state we expect the distribution of excess holes to decay to zero for large values of x, because of the recombination (Fig. 4-17). For this

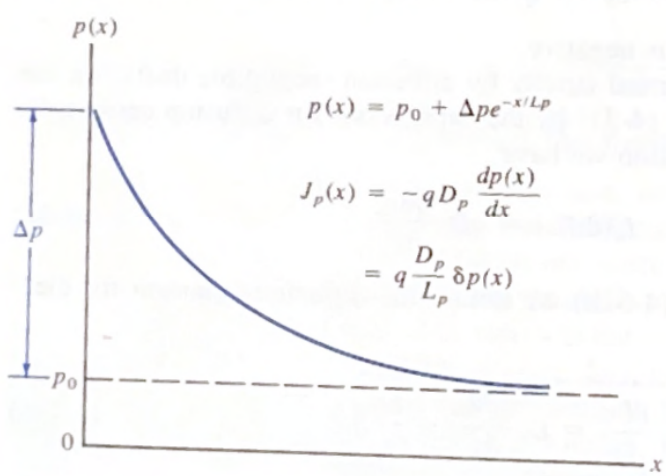


Figure 4-17 Injection of holes at $x = 0$, giving a steady state hole distribution $p(x)$ and a resulting diffusion current density $J_p(x)$.

problem we use the steady state diffusion equation for holes, Eq. (4-34b). The solution to this equation has the form

$$\delta p(x) = C_1 e^{x/L_p} + C_2 e^{-x/L_p} \tag{4-35}$$

We can evaluate C_1 and C_2 from the boundary conditions. Since recombination must reduce $\delta p(x)$ to zero for large values of x, $\delta p = 0$ at $x = \infty$, and therefore $C_1 = 0$. Similarly, the condition $\delta p = \Delta p$ at $x = 0$ gives $C_2 = \Delta p$, and the solution is

$$\boxed{\delta p(x) = \Delta p e^{-x/L_p}} \tag{4-36}$$

The injected excess hole concentration dies out exponentially in x due to recombination, and the diffusion length L_p represents the distance at which the excess hole distribution is reduced to $1/e$ of its value at the point of injection. We can show that L_p *is the average distance a hole diffuses before recombining*. To calculate an average diffusion length, we must obtain an expression for the probability that an injected hole recombines in a particular interval dx. The probability that a hole injected at $x = 0$ *survives* to x without recombination is $\delta p(x)/\Delta p = \exp(-x/L_p)$, the ratio of the steady state concentrations at x and 0. On the other hand, the probability that a hole at x will *recombine* in the subsequent interval dx is

$$\frac{\delta p(x) - \delta p(x + dx)}{\delta p(x)} = \frac{-(d\delta p(x)/dx)\,dx}{\delta p(x)} = \frac{1}{L_p}dx \tag{4-37}$$

Thus the total probability that a hole injected at $x = 0$ will recombine in a given dx is the product of the two probabilities:

$$(e^{-x/L_p})\left(\frac{1}{L_p}dx\right) = \frac{1}{L_p}e^{-x/L_p}\,dx \tag{4-38}$$

Then, using the usual averaging techniques described by Eq. (2-21), the average distance a hole diffuses before recombining is

$$\langle x \rangle = \int_0^\infty x\frac{e^{-x/L_p}}{L_p}dx = L_p \tag{4-39}$$

The steady state distribution of excess holes causes diffusion, and therefore a hole current, in the direction of decreasing concentration. From Eqs. (4-22b) and (4-36) we have

$$J_p(x) = -qD_p\frac{dp}{dx} = -qD_p\frac{d\delta p}{dx} = q\frac{D_p}{L_p}\Delta p e^{-x/L_p} = q\frac{D_p}{L_p}\delta p(x) \tag{4-40}$$

Since $p(x) = p_0 + \delta p(x)$, the space derivative involves only the excess concentration. We notice that since $\delta p(x)$ is proportional to its derivative for an exponential distribution, the diffusion current at any x is just proportional to the excess concentration δp at that position.

Although this example seems rather restricted, its usefulness will become apparent in Chapter 5 in the discussion of p-n junctions. The injection of minority carriers across a junction often leads to exponential distributions as in Eq. (4-36), with the resulting diffusion current of Eq. (4-40).

4.4.5 The Haynes–Shockley Experiment

One of the classic semiconductor experiments is the demonstration of drift and diffusion of minority carriers, first performed by J. R. Haynes and W. Shockley in 1951 at the Bell Telephone Laboratories. The experiment allows independent measurement of the minority carrier mobility μ and diffusion coefficient D. The basic principles of the Haynes–Shockley experiment are as follows: A pulse of holes is created in an n-type bar (for example) that contains an electric field (Fig. 4-18); as the pulse drifts in the field and spreads out by diffusion, the excess hole concentration is monitored at some point down the bar; the time required for the holes to drift a given distance in the field gives a measure of the mobility; and the spreading of the pulse during a given time is used to calculate the diffusion coefficient.

In Fig. 4-18 a pulse of excess carriers is created by a light flash at some point $x = 0$ in an n-type semiconductor ($n_0 \gg p_0$). We assume that the excess carriers have a negligible effect on the electron concentration but change the hole concentration significantly. The excess holes drift in the direction of the electric field and eventually reach the point $x = L$, where they are monitored. By measuring the drift time t_d, we can calculate the drift velocity v_d and, therefore, the hole mobility:

$$\mathsf{v}_d = \frac{L}{t_d} \tag{4-41}$$

$$\mu_p = \frac{\mathsf{v}_d}{\mathscr{E}} \tag{4-42}$$

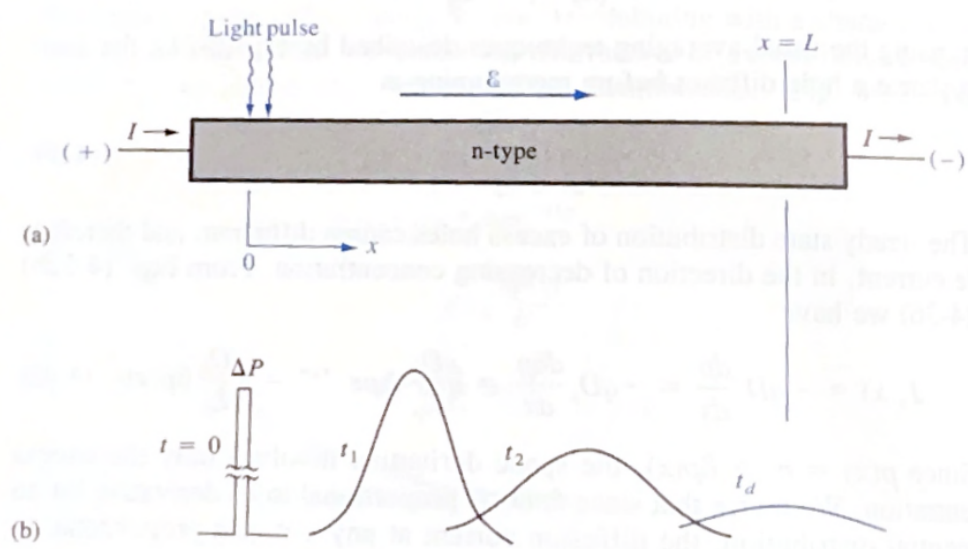


Figure 4-18 Drift and diffusion of a hole pulse in an n-type bar: (a) sample geometry; (b) position and shape of the pulse for several times during its drift down the bar.

Thus the hole mobility can be calculated directly from a measurement of the drift time for the pulse as it moves down the bar. In contrast with the Hall effect (Section 3.4.5), which can be used with resistivity to obtain the *majority* carrier mobility, the Haynes–Shockley experiment is used to measure the *minority* carrier mobility.

As the pulse drifts in the $\mathscr{E}$ field it also spreads out by diffusion. By measuring the spread in the pulse, we can calculate D_p. To predict the distribution of holes in the pulse as a function of time, let us first reexamine the case of diffusion of a pulse *without drift, neglecting recombination* (Fig. 4-12). The equation which the hole distribution must satisfy is the time-dependent diffusion equation, Eq. (4-33b). For the case of negligible recombination (τ_p long compared with the times involved in the diffusion), we can write the diffusion equation as

$$\frac{\partial \delta p(x,t)}{\partial t} = D_p \frac{\partial^2 \delta p(x,t)}{\partial x^2} \tag{4-43}$$

The function which satisfies this equation is called a *gaussian distribution*,

$$\delta p(x,t) = \left[\frac{\Delta P}{2\sqrt{\pi D_p t}}\right] e^{-x^2/4D_p t} \tag{4-44}$$

where ΔP is the number of holes per unit area created over a negligibly small distance at $t = 0$. The factor in brackets indicates that the peak value of the pulse (at $x = 0$) decreases with time, and the exponential factor predicts the spread of the pulse in the positive and negative x-directions (Fig. 4-19). If we designate the peak value of the pulse as $\widehat{\delta p}$ at any time (say t_d), we can use Eq. (4-44) to calculate D_p from the value of δp at some point x. The most convenient choice is the point $\Delta x/2$, at which δp is down by $1/e$ of its peak value $\widehat{\delta p}$. At this point we can write

$$e^{-1}\widehat{\delta p} = \widehat{\delta p}\, e^{-(\Delta x/2)^2/4D_p t_d} \tag{4-45}$$

$$D_p = \frac{(\Delta x)^2}{16 t_d} \tag{4-46}$$

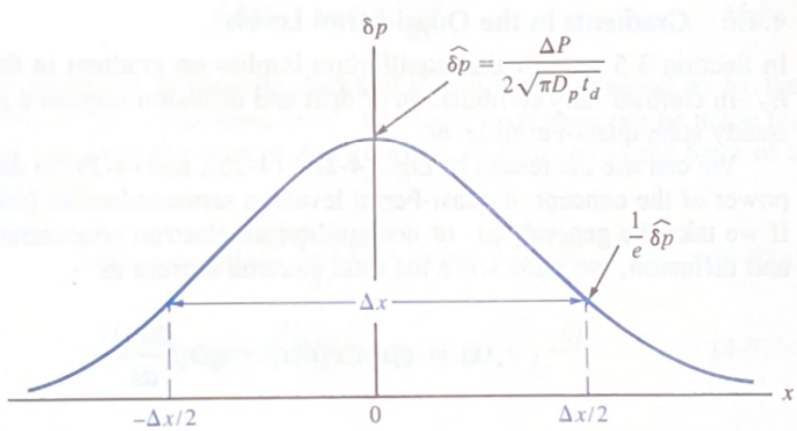


Figure 4-19 Calculation of D_p from the shape of the δp distribution after time t_d. No drift or recombination is included.

Since Δx cannot be measured directly, we use an experimental setup such as Fig. 4-20, which allows us to display the pulse on an oscilloscope as the carriers pass under a detector. As we shall see in Chapter 5, a forward-biased p-n junction serves as an excellent injector of minority carriers, and a reverse-biased junction serves as a detector. The measured quantity in Fig. 4-20 is the pulse width Δt displayed on the oscilloscope in time. It is related to Δx by the drift velocity, as the pulse drifts past the detector point (2).

$$\Delta x = \Delta t \mathrm{v}_d = \Delta t \frac{L}{t_d} \tag{4-47}$$

EXAMPLE 4-6 An n-type Ge sample is used in the Haynes–Shockley experiment shown in Fig. 4-20. The length of the sample is 1 cm, and the probes (1) and (2) are separated by 0.95 cm. The battery voltage E_0 is 2 V. A pulse arrives at point (2) 0.25 ms after injection at (1); the width of the pulse Δt is 117 μs. Calculate the hole mobility and diffusion coefficient, and check the results against the Einstein relation.

SOLUTION

$$\mu_p = \frac{\mathrm{v}_d}{\mathscr{E}} = \frac{0.95/(0.25 \times 10^{-3})}{2/1} = 1900 \text{ cm}^2/(\text{V-s})$$

$$D_p = \frac{(\Delta x)^2}{16 t_d} = \frac{(\Delta t L)^2}{16 t_d^3}$$

$$= \frac{(117 \times 0.95)^2 \times 10^{-12}}{16(0.25)^3 \times 10^{-9}} = 49.4 \text{ cm}^2/\text{s}$$

$$\frac{D_p}{\mu_p} = \frac{49.4}{1900} = 0.026 = \frac{kT}{q}$$

4.4.6 Gradients in the Quasi-Fermi Levels

In Section 3.5 we saw that equilibrium implies no gradient in the Fermi level E_F. In contrast, any combination of drift and diffusion implies a gradient in the steady state quasi-Fermi level.

We can use the results of Eqs. (4-23), (4-26), and (4-29) to demonstrate the power of the concept of quasi-Fermi levels in semiconductors [see Eq. (4-15)]. If we take the general case of nonequilibrium electron concentration with drift and diffusion, we must write the total electron current as

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + qD_n \frac{dn(x)}{dx} \tag{4-48}$$

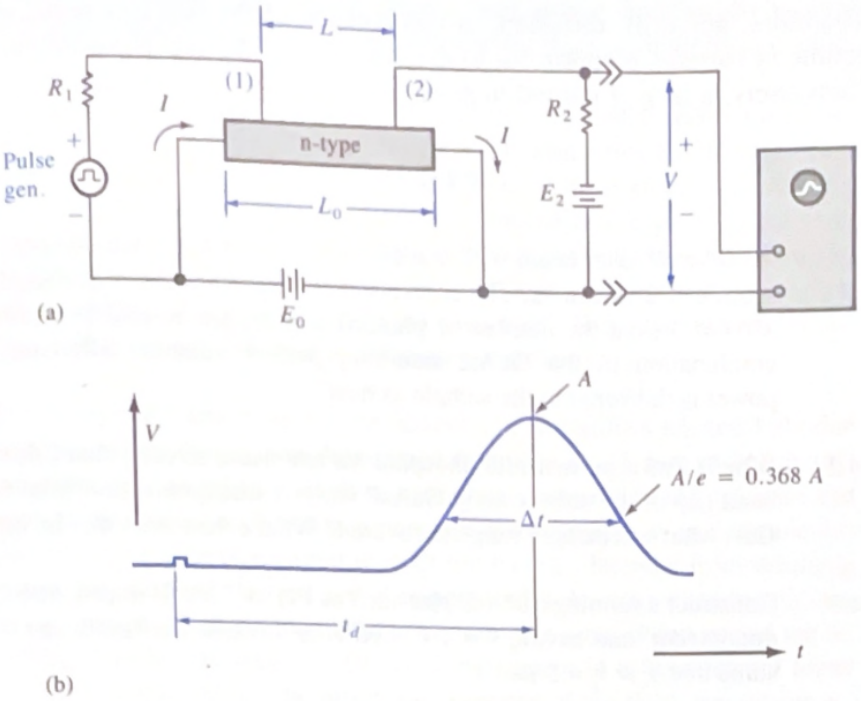


Figure 4-20 The Haynes–Shockley experiment: (a) circuit schematic; (b) typical trace on the oscilloscope screen.

where the gradient in electron concentration is

$$\frac{dn(x)}{dx} = \frac{d}{dx}\left[n_i e^{(F_n - E_i)/kT}\right] = \frac{n(x)}{kT}\left(\frac{dF_n}{dx} - \frac{dE_i}{dx}\right) \tag{4-49}$$

Using the Einstein relation, the total electron current becomes

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + \mu_n n(x)\left[\frac{dF_n}{dx} - \frac{dE_i}{dx}\right] \tag{4-50}$$

But Eq. (4-26) indicates that the subtractive term in the brackets is just $q\mathscr{E}(x)$, giving a direct cancellation of $q\mu_n n(x)\mathscr{E}(x)$ and leaving

$$J_n(x) = \mu_n n(x)\frac{dF_n}{dx} \tag{4-51}$$

Thus, the processes of electron drift and diffusion are summed up by the spatial variation of the quasi-Fermi level. The same derivation can be made for holes, and we can write the current due to drift and diffusion in the form of a *modified Ohm's law*

$$J_n(x) = q\mu_n n(x)\frac{d(F_n/q)}{dx} = \sigma_n(x)\frac{d(F_n/q)}{dx} \tag{4-52a}$$

$$J_p(x) = q\mu_p p(x)\frac{d(F_p/q)}{dx} = \sigma_p(x)\frac{d(F_p/q)}{dx} \tag{4-52b}$$

Therefore, any drift, diffusion, or combination of the two in a semiconductor results in currents proportional to the gradients of the two quasi-Fermi levels. Conversely, a lack of current implies constant quasi-Fermi levels.

PROBLEMS

4.1. A 100-mW laser beam with wavelenth $\lambda = 6328$ Å is focused onto a GaAs sample 100 μm thick. The absorption coefficient at this wavelength is 3×10^4 cm^{-1}. Find the number of photons emitted per second by radiative recombination in the GaAs, assuming perfect quantum efficiency. What power is delivered to the sample as heat?

4.2. Why is ZnS a more useful phosphor for television screens than CdSe? If the band gap of the ternary alloy GaAsP varies with composition from GaAs to GaP, what wavelength range is covered? What colors does this include?

4.3. Construct a semilogarithmic plot such as Fig. 4-7 for Si doped with 2×10^{15} donors/cm^3 and having 4×10^{14} EHP/cm^3 created uniformly at $t = 0$. Assume that $\tau_n = \tau_p = 5$ μs.

4.4. Calculate the recombination coefficient α_r for the low-level excitation described in Prob. 4.3. Assume that this value of α_r applies when the GaAs sample is uniformly exposed to a steady state optical generation rate $g_{op} = 10^{19}$ EHP/cm^3-s. Find the steady state excess carrier concentration $\Delta n = \Delta p$.

4.5. A direct semiconductor with $n_i = 10^{10}$ cm^{-3} is doped with 10^{15} cm^{-3} donors. Its low-level carrier lifetime τ is $\tau_n = \tau_p = 10^{-7}$ s.

(a) If a sample of this material is uniformly exposed to a steady optical generation rate of $g_{op} = 2 \times 10^{22}$ EHP/cm^3-s, calculate the excess carrier concentration $\Delta n = \Delta p$. *Note:* This excitation rate is *not* low level, but you may assume that α_r is the same.

(b) If the carrier lifetime is defined as the excess carrier concentration divided by the recombination rate, what is τ at this excitation level?

4.6. With E_F located 0.3 eV below the conduction band in a Si sample, what charge state would you expect for most Ga atoms in the sample? (See Fig. 4-9.) What would be the predominant charge state of Ni? Au? *Note:* By charge state we mean neutral, singly positive, doubly negative, etc.

4.7. An n-type Si sample with $N_d = 10^{16}$ cm^{-3} is steadily illuminated such that $g_{op} = 10^{21}$ EHP/cm^3-s. If $\tau_n = \tau_p = 10^{-6}$ s for this excitation, calculate the separation in the quasi-Fermi levels $F_n - F_p$. Draw a band diagram such as Fig. 4-11.

of the barrier Φ_B for electron injection from the metal into the semiconductor. This barrier (which is $\Phi_m - \chi$ for the ideal case shown in Fig. 5-33) is unaffected by the bias voltage. We expect the probability of an electron in the metal surmounting this barrier to be given by a Boltzmann factor. Thus

$$I_0 \propto e^{-q\Phi_B/kT} \tag{5-77}$$

The diode equation (5-76) applies also to the metal–p-type semiconductor junction of Fig. 5-32. In this case forward voltage is defined with the semiconductor biased positively with respect to the metal. Forward current increases as this voltage lowers the potential barrier to $V_0 - V$ and holes flow from the semiconductor to the metal. Of course, a reverse voltage increases the barrier for hole flow and the current becomes negligible.

In both of these cases the Schottky barrier diode is rectifying, with easy current flow in the forward direction and little current in the reverse direction. We also note that the forward current in each case is due to the injection of *majority* carriers from the semiconductor into the metal. The absence of minority carrier injection and the associated storage delay time is an important feature of Schottky barrier diodes. Although some minority carrier injection occurs at high current levels, these are essentially majority carrier devices. Their high-frequency properties and switching speed are therefore generally better than typical p-n junctions.

In the early days of semiconductor technology, rectifying contacts were made simply by pressing a wire against the surface of the semiconductor. In modern devices, however, the metal–semiconductor contact is made by depositing an appropriate metal film on a clean semiconductor surface and defining the contact pattern photolithographically. Schottky barrier devices are particularly well suited for use in densely packed integrated circuits, because fewer photolithographic masking steps are required compared to p-n junction devices.

5.7.3 Ohmic Contacts

In many cases we wish to have an *ohmic* metal–semiconductor contact, having a linear I–V characteristic in both biasing directions. For example, the surface of a typical integrated circuit is a maze of p and n regions, which must be contacted and interconnected. It is important that such contacts be ohmic, with minimal resistance and no tendency to rectify signals.

Ideal metal–semiconductor contacts are ohmic when the charge induced in the semiconductor in aligning the Fermi levels is provided by majority carriers (Fig. 5-34). For example, in the $\Phi_m < \Phi_s$ (n-type) case of Fig. 5-34a, the Fermi levels are aligned at equilibrium by transferring electrons from the metal to the semiconductor. This raises the semiconductor electron energies (lowers the electrostatic potential) relative to the metal at equilibrium (Fig. 5-34b). In this case the barrier to electron flow between the metal and the semiconductor is small and easily overcome by a small voltage. Similarly, the case $\Phi_m > \Phi_s$ (p-type) results in easy hole flow across the junction (Fig. 5-34d). Unlike the

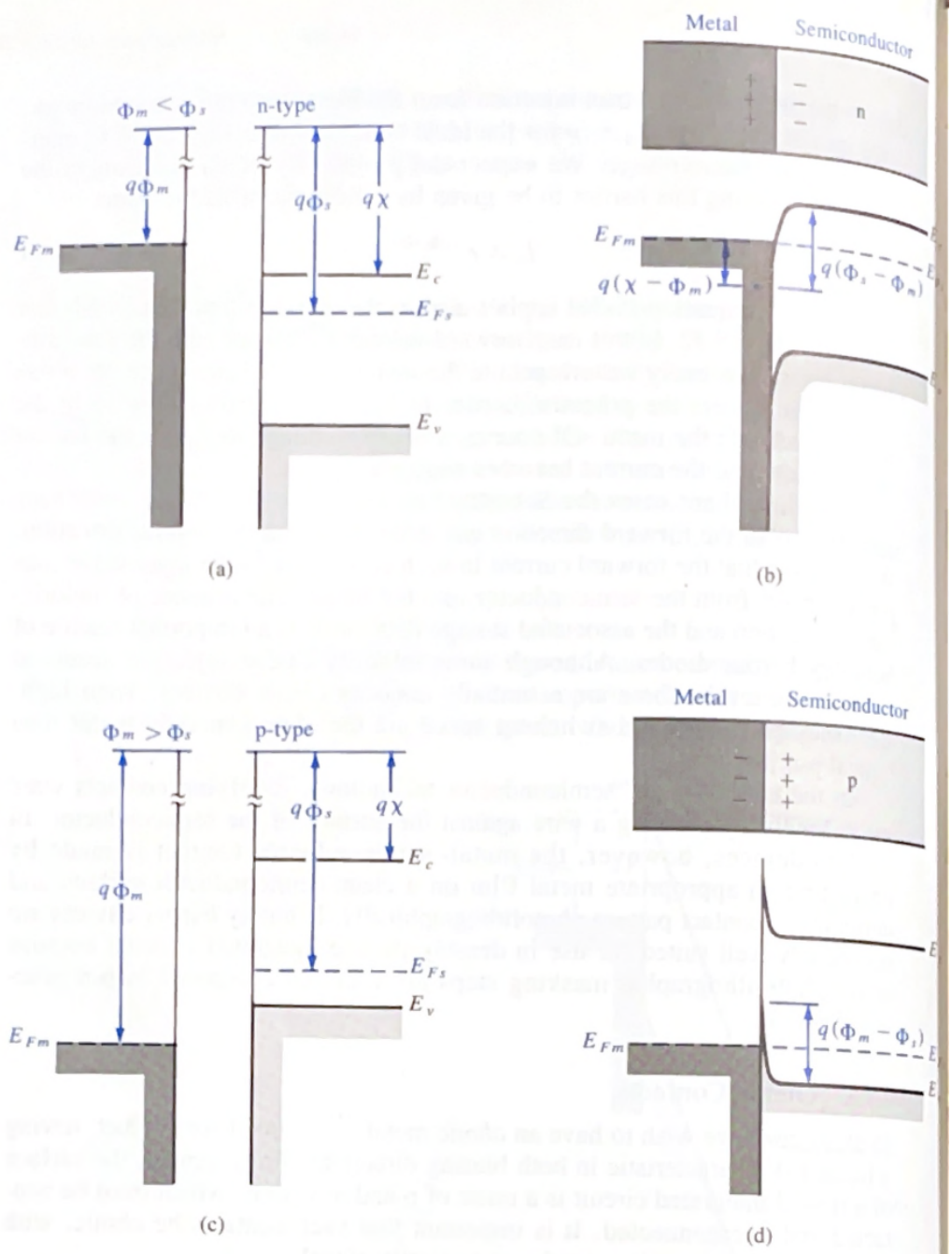


Figure 5-34 Ohmic metal–semiconductor contacts: (a) $\Phi_m < \Phi_s$ for an n-type semiconductor, and (b) the equilibrium band diagram for the junction; (c) $\Phi_m > \Phi_s$ for a p-type semiconductor, and (d) the junction at equilibrium.

rectifying contacts discussed above, no depletion region occurs in the semiconductor in these cases since the electrostatic potential difference required to align the Fermi levels at equilibrium calls for accumulation of majority carriers in the semiconductor.

A practical method for forming ohmic contacts is by doping the semiconductor heavily in the contact region. Thus if a barrier exists at the interface, the depletion width is small enough to allow carriers to tunnel through the barrier. For example, Au containing a small percentage of Sb can be alloyed to n-type

IPR2025-00238 - BOE
Exhibit 1007 - Page 43

Si, forming an n^+ layer at the semiconductor surface and an excellent ohmic contact. Similarly, p-type material requires a p^+ surface layer in contact with the metal. In the case of Al on p-type Si, the metal contact also provides the acceptor dopant. Thus the required p^+ surface layer is formed during a brief heat treatment of the contact after the Al is deposited. As we shall see in the discussion of integrated circuits in Chapter 9, Al makes good ohmic contact to p-type Si, but n-type material requires an n^+ diffusion or implantation before contacting by Al.

5.7.4 Typical Schottky Barriers

The discussion of ideal metal–semiconductor contacts does not include certain effects of the junction between the two dissimilar materials. Unlike a p-n junction, which occurs within a single crystal, a Schottky barrier junction includes a termination of the semiconductor crystal. The semiconductor surface contains *surface states* due to incomplete covalent bonds and other effects, which can lead to charges at the metal–semiconductor interface. Furthermore, the contact is seldom an atomically sharp discontinuity between the semiconductor crystal and the metal. There is typically a thin interfacial layer, which is neither semiconductor nor metal. For example, silicon crystals are covered by a thin (10–20 Å) oxide layer even after etching or cleaving in atmospheric conditions. Therefore, deposition of a metal on such a Si surface leaves a glassy interfacial layer at the junction. Although electrons can tunnel through this thin layer, it does affect the barrier to current transport through the junction.

Because of surface states, the interfacial layer, microscopic clusters of metal–semiconductor phases, and other effects, it is difficult to fabricate junctions with barriers near the ideal values predicted from the work functions of the two isolated materials. Therefore, measured barrier heights are used in device design. In compound semiconductors the interfacial layer introduces states in the semiconductor band gap that pin the Fermi level at a fixed position, regardless of the metal used (Fig. 5-35). For example, a collection of interface

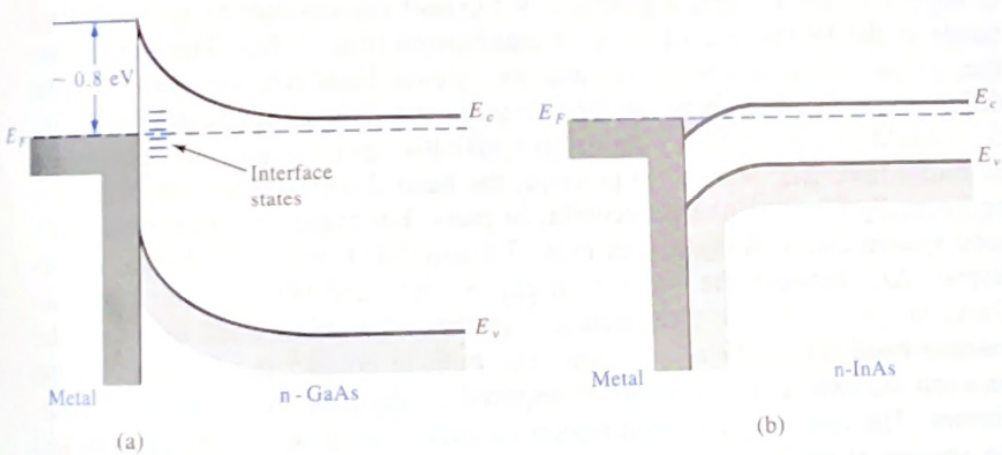


Figure 5-35 Fermi level pinning by interface states in compound semiconductors: (a) E_F is pinned near $E_C - 0.8$ eV in n-type GaAs, regardless of the choice of metal; (b) E_F is pinned above E_C in n-type InAs, providing an excellent ohmic contact.

electrons are confined to the heterojunction, device behavior closely resembles that of a MOSFET. The advantages of the HEMT over the Si MOSFET are the higher mobility and maximum electron velocity in GaAs compared with Si, and the smoother interfaces possible with an AlGaAs/GaAs heterojunction compared with the Si/SiO_2 interface. The high performance of the HEMT translates into an extremely high cutoff frequency, and devices with fast access times.

Although we have discussed the HEMT in terms of the AlGaAs/GaAs heterojunction, other materials are also promising, such as the InGaAsP/InP system. A motivation for avoiding $Al_xGa_{1-x}As$ is the presence of a deep-level defect called the DX center for $x > 0.2$, which traps electrons and impairs the HEMT operation. Since very thin layers are involved, materials with slight lattice mismatch can be grown to form *pseudomorphic* HEMTs. An example of such a system is the use of a thin layer of InGaAs grown pseudomorphically on GaAs, followed by AlGaAs. An advantage of this system is that a useful band discontinuity can be achieved using AlGaAs of low enough Al composition to avoid the DX center problem.

The HEMT, or MODFET, is also referred to as a *two-dimensional electron gas FET (2-DEG FET, or TEGFET)* to emphasize the fact that conduction along the channel occurs in a thin sheet of charge. The device has also been called a *separately doped FET (SEDFET)*, to emphasize the fact that the doping occurs in a separate region from the channel.

8.2.3 Short Channel Effects

As mentioned in Section 8.1.3, a variety of modifications to the simple theory of JFET and MESFET operation must be made when the channel length is small (typically $<1\ \mu m$). Until recently these short-channel effects would be considered unusual, but now it is common to encounter FET devices in which these effects dominate the I–V characteristics. For example, high-field effects occur when 1 V appears across a channel length of 1 μm (10^{-4} cm), giving an electric field of 10 kV/cm.

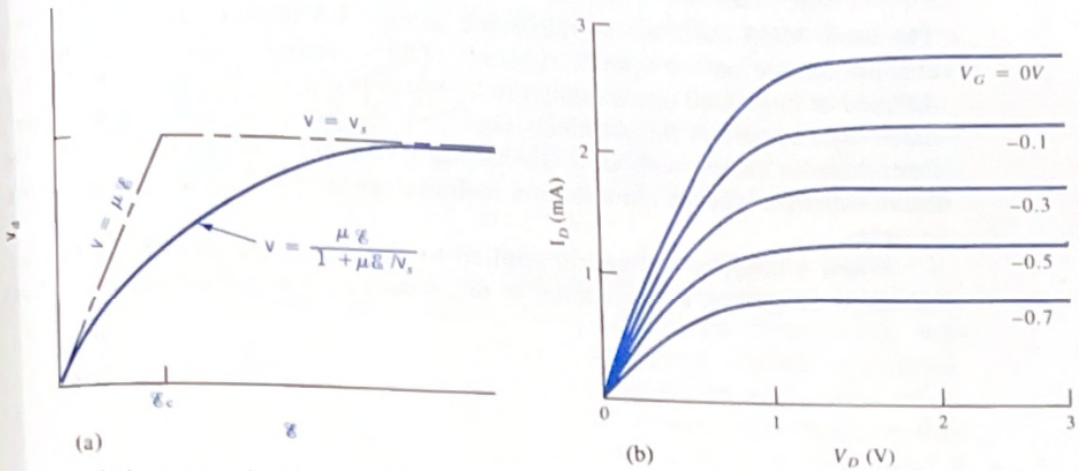
A simple piecewise-linear approximation to the velocity-field curve assumes a constant mobility (linear) dependence up to some critical field $\mathscr{E}_c$ and a constant saturation velocity v_s for higher fields. For Si a better approximation is

$$v_d = \frac{\mu\mathscr{E}}{1 + \mu\mathscr{E}/v_s} \tag{8-12}$$

where μ is the low-field mobility. These two approximations are shown in Fig. 8-7 (a). If we assume that the electrons passing through the channel drift with a constant saturation velocity v_s, the current takes a simple form

$$I_D = qnv_sA = qN_dv_sZh \tag{8-13}$$

where h is a slow function of V_G. In this case the saturated current follows the velocity saturation, and does not require a true pinch-off in the sense of depletion regions meeting at some point in the channel. In the saturated velocity case, the transconductance g_m is essentially constant, in contrast with the constant



Effects of electron velocity saturation at high electric fields: (a) approximations to the saturation of drift velocity with increasing field; (b) drain current–voltage characteristics for the saturated velocity case, showing almost equally spaced curves with increasing gate voltage.

Figure 8-7

mobility case described by Eq. (8-10). As shown in Fig. 8-7(b), the I_d–V_d curves are more evenly spaced if constant saturation velocity dominates, compared with the V_G-dependent spacing shown in Fig. 8-3(b) for the long-channel constant-mobility case.

Most devices operate with characteristics intermediate between the constant mobility and the constant velocity regimes. Depending on the details of the field distribution, it is possible to divide up the channel into regions dominated by the two extreme cases, or to use an approximation such as Eq. (8-12).

Another important short-channel effect, described in Section 8.1.3, is the reduction in effective channel length after pinch-off as the drain voltage is increased. This effect is not significant in long-channel devices, since the change in L due to intrusion of the depletion region is a minor fraction of the total channel length. In short-channel devices, however, the effective channel length can be substantially shortened, leading to a slope in the saturated I–V characteristic that is analogous to the Early (base-width narrowing) effect in bipolar transistors (Fig. 7-22).

8.3 THE METAL–INSULATOR–SEMICONDUCTOR FET

One of the most widely used electronic devices, particularly in digital integrated circuits, is the *metal–insulator–semiconductor (MIS) transistor*. In this device the channel current is controlled by a voltage applied at a gate electrode that is isolated from the channel by an insulator. The resulting device may be referred to generically as an insulated-gate field effect transistor (IGFET). However, since most such devices are made using silicon for the semiconductor, SiO_2 for the insulator, and Al or other metal for the gate electrode, the term *MOS transistor* (MOST) is commonly used.

where V_B will be negative for the n-channel case. As the substrate bias is increased, the threshold voltage becomes more positive. The effect of this bias becomes more dramatic as the substrate doping is increased, since ΔV_T is also proportional to $\sqrt{N_a}$. For a p-channel device the bulk-to-source voltage V_B is positive to achieve a reverse bias, and the approximate change ΔV_T for $V_B \gg 2\phi_F$ is

$$\Delta V_T \simeq -\frac{\sqrt{2\epsilon_s q N_d}}{C_i} V_B^{1/2} \qquad \text{(p channel)} \tag{8-45}$$

Thus the p-channel threshold voltage becomes more negative with substrate bias.

The substrate bias effect (also called the *body effect*) increases V_T for either type of device. This effect can be used to raise the threshold voltage of a marginally enhancement device ($V_T \simeq 0$) to a somewhat larger and more manageable value. This can be an asset for n-channel devices particularly (see Fig. 8-17). The effect can present problems, however, in MOS integrated circuits for which it is impractical to connect each source region to the substrate. In these cases, possible V_T shifts due to the body effect must be taken into account in the circuit design.

8.3.8 Capacitance Effects and Self-Aligned Transistors

The high-frequency operation of MOS transistors depends upon the capacitances associated with various parts of the device. The intrinsic MOS capacitance due to the gate region itself is essentially that of the MOS capacitor of Fig. 8-13, taking into account the effects of threshold voltage. Since the inversion region is contacted by the source and drain in the transistor, the supply of carriers at high frequency is not limited by thermal generation as in the MOS capacitor. Thus the capacitance after inversion is C_i (Fig. 8-24). For a 100-Å gate oxide thickness, this capacitance is 34.5×10^{-8} F/cm^2. Assuming a gate region with $L = 1.0\ \mu$m and $Z = 5\ \mu$m, the gate capacitance is 0.017 pF.

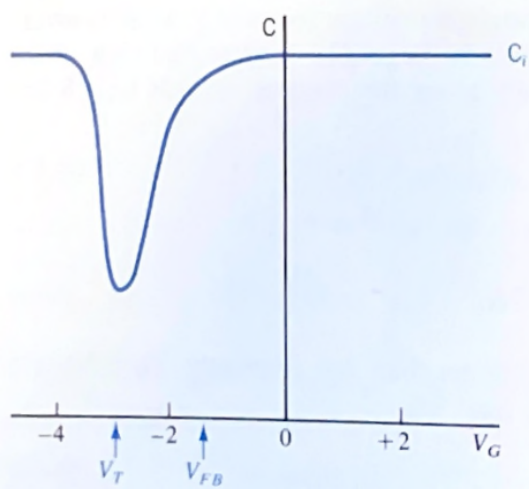


Figure 8-24 Gate capacitance–voltage characteristic of a p-channel MOS transistor with $V_B = 0$.

In addition to the intrinsic MOS capacitance, there are several parasitic capacitances which limit the high-frequency operation of MOS transistors. The junction capacitance associated with the junctions between source and drain and the substrate is described by Eq. (5-61). This capacitance is often comparable to the MOS gate capacitance, and varies with junction voltage. Therefore, the junction capacitances are significant in high-frequency or fast-switching operations. Another parasitic capacitance arises from metal interconnections over the field oxide. Since the field oxide is generally thick, this stray capacitance is most important in circuits with considerable metallization.

A particularly troublesome stray capacitance results from the overlap of the gate with the source and drain regions (Fig. 8-25a). If the gate does not extend to the source and drain diffusions, an incomplete channel will be formed and the device will not operate. To avoid this possibility, standard Al gate processing calls for some overlap of the gate past the source and drain edges. This overlap must be large enough to allow for mask registration tolerances and variation in lateral diffusion in the source and drain diffusion step. Because of this overlap, a parasitic capacitance develops between the gate and source and between gate and drain. Since the gate is usually the input and the drain is the output in typical circuit configurations, this overlap capacitance introduces an undesirable feedback effect between input and output at high frequencies (sometimes called the *Miller effect*). To avoid this problem, several methods of achieving *self-aligned gate* regions have been developed.

One method for avoiding the gate overlap capacitance is to use ion implantation to align the source and drain with the edges of the gate (Fig. 8-25b). In this method the gate metal is made narrower than the distance between the gate and source diffusions. An ion implantation is then performed, using the metal gate as a mask, to extend the source and drain regions to the edges of the gate. This procedure takes advantage of the fact that the ion beam arrives essentially perpendicular to the surface. The primary disadvantage of this self-aligned gate procedure is the requirement of a relatively high dose to achieve sufficient conduction in the implanted source and drain extensions, and the need for subsequent annealing of the radiation damage. Since the Al metal gate cannot be raised to high temperature, annealing must be done in the range of 500°C. Al-

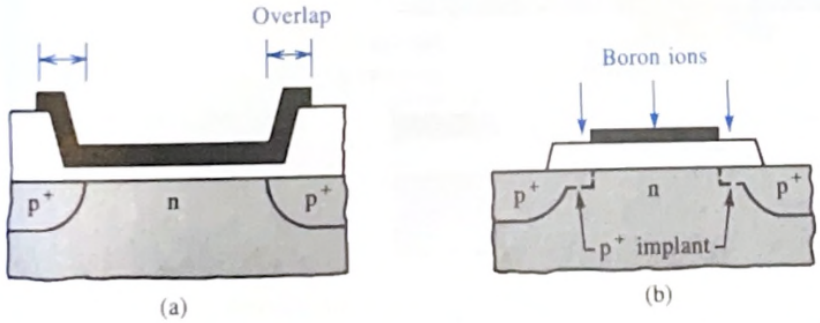


Figure 8-25 Reduction of parasitic capacitance due to gate overlap: (a) standard processing with overlap allowed for mask registration tolerance; (b) extension of source and drain to the edges of the Al gate by boron implantation.

though reasonable annealing can result, the residual damage does reduce the conductivity of the implanted regions.

The silicon gate method described in Section 8.3.6 has the additional advantage of fairly good gate alignment. In this procedure the polycrystalline Si gate material is deposited over the gate oxide, and windows for the source and drain diffusion are opened in the Si film and the underlying oxide (Fig. 8-26). Since the polycrystalline Si can be raised to high temperature, the source and drain diffusion can be done with the gate in place. Two important advantages result from this method—the gate oxide is covered immediately after growth by the deposited Si, thereby protecting it during subsequent processing; in addition the edges of the gate serve to define the windows for the source and drain diffusion. As a result, the overlap is only that which results from lateral diffusion. The mask registration tolerance is eliminated.

An additional advantage of the Si-gate procedure is that the deposited Si can also be used for interconnections to adjacent devices (Fig. 8-26c). After covering the polycrystalline Si with oxide, another layer of metallization (e.g., Al) can be added. Such multilayer interconnection is of great value in the layout of complex integrated circuits. The reduction of gate–substrate work function difference, the self-aligned gate feature, protection of the gate oxide during most processing steps, and the ease of multilayer interconnection all provide attractive advantages for polycrystalline Si gate technology.

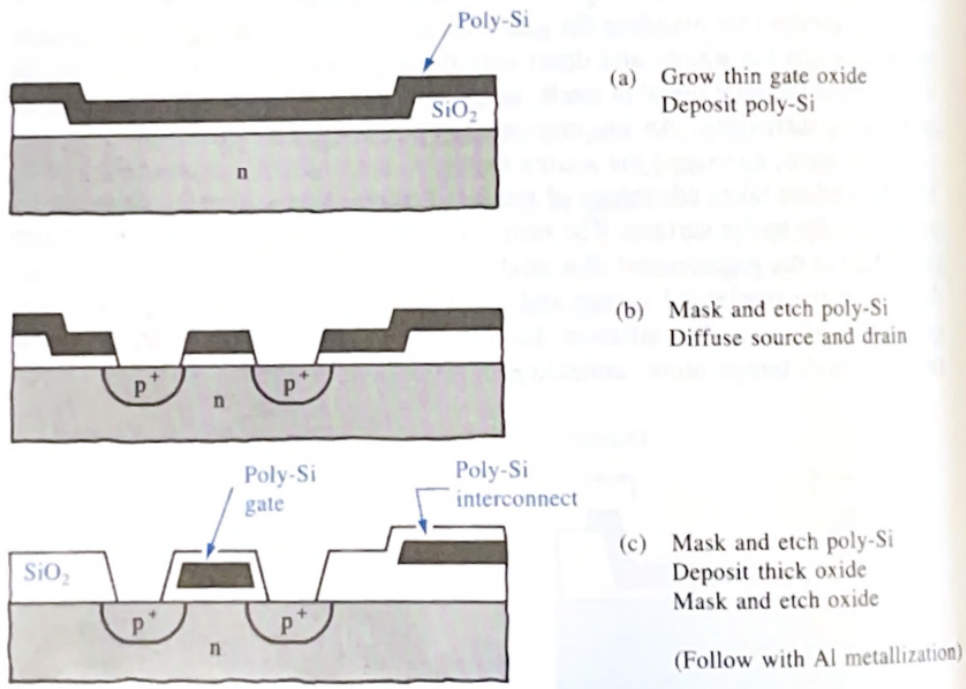


Figure 8-26 Silicon gate technology: (a) polycrystalline Si deposited over SiO_2; (b) diffusion of source and drain using the Si gate as a mask to define the channel length; (c) after further processing, the poly-Si gate and interconnection regions are covered with SiO_2. The source and drain regions can now be contacted by Al metal, and Al interconnections can pass over the buried poly-Si interconnections.

8.3.9 Short-Channel Effects

When MOS transistors are made very small, as required in many high-speed devices and integrated circuits, the size of the depletion region near the drain in saturation can be comparable to the channel length (Fig. 8-27). When this occurs, the saturation current does not remain constant with increased V_D. In deriving Eq. (8-40) we assumed that the channel length L was not changed by the formation of the depletion region at the drain end beyond pinch-off. For short channels the spreading of this depletion region with increased V_D beyond pinch-off causes the effective channel length L' to decrease. Since I_D (sat.) is inversely proportional to channel length in Eq. (8-40), the drain current increases as the drain depletion region spreads into the channel. As a result, I_D (sat.) increases somewhat with V_D, rather than remaining constant. Velocity saturation effects (Section 8.2.3) are also likely to occur in short-channel devices.

A variety of other effects occur as the MOS device dimensions are reduced. For example, as Fig. 8-28 illustrates, the source and drain depletion regions can intrude into the channel even without bias, as these junctions are brought closer together in short-channeled devices. This effect is called *charge*

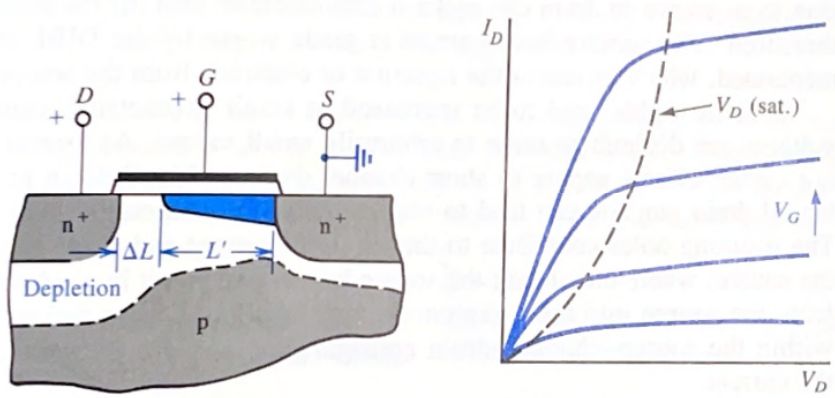


Figure 8-27 Short-channel effects: (a) reduction of the effective channel length L' in saturation; (b) increase of I_D beyond pinch-off due to the reduction of L'.

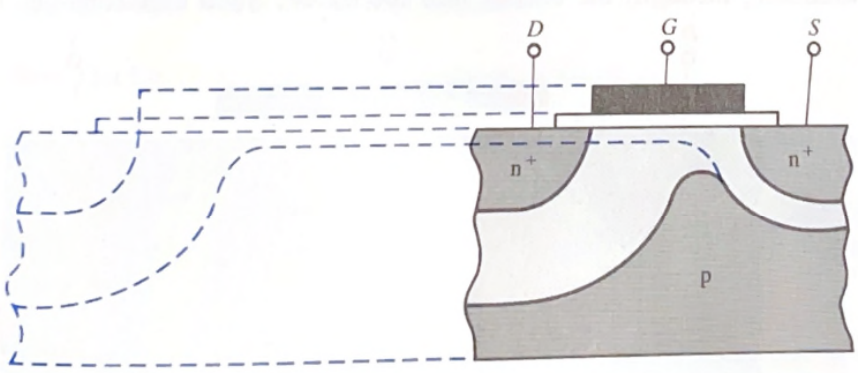


Figure 8-28 Intrusion of the source and drain depletion regions into the channel as the device dimensions are reduced from the long-channel case (dashed lines). The effect of these depletion regions in reducing the usable channel length is called *charge sharing*, and *drain-induced barrier lowering* can result in loss of gate control.

sharing, but it should probably be called charge hogging, because the source and drain in effect take part of the channel charge, which would otherwise be controlled by the gate. As the drain depletion region continues to increase with bias, it can actually interact with the source-to-channel junction and lower the potential barrier. This problem is known as *drain-induced barrier lowering* (DIBL). When the source junction barrier is reduced, electrons are easily injected into the channel and the gate voltage no longer has control of the drain current. Under extreme conditions of encroaching source and drain depletion regions, the two can meet. This punch-through effect results in a continuous depleted region from drain to source. Carriers injected at the source are driven to the drain by the high field between the two electrodes. The current is called *space-charge-limited* current and is proportional to V_D^2/L^3.

An effect that is exacerbated by short-channel designs is the *subthreshold current*, which arises from the fact that some electrons are induced in the channel even before strong inversion is established. For the low electron concentrations typical of the subthreshold regime, we expect diffusion currents (proportional to carrier gradients) to dominate over drift currents (proportional to carrier concentrations). For very short channel lengths, such carrier diffusion from source to drain can make it impossible to turn off the device below threshold. The subthreshold current is made worse by the DIBL effect just mentioned, which increases the injection of electrons from the source.

Electric fields tend to be increased at small geometries, since device voltages are difficult to scale to arbitrarily small values. As a result, various hot carrier effects appear in short-channel devices. The field in the reverse-biased drain junction can lead to impact ionization and carrier multiplication. The resulting holes contribute to the substrate current and some may move to the source, where they lower the source barrier and result in electrons injected from the source into the p-region. In fact, n-p-n transistor action can result within the source–channel–drain configuration and prevent gate control of the current.

Another hot electron effect is the transport of energetic electrons over (or tunneling through) the barrier into the oxide. Such electrons can become

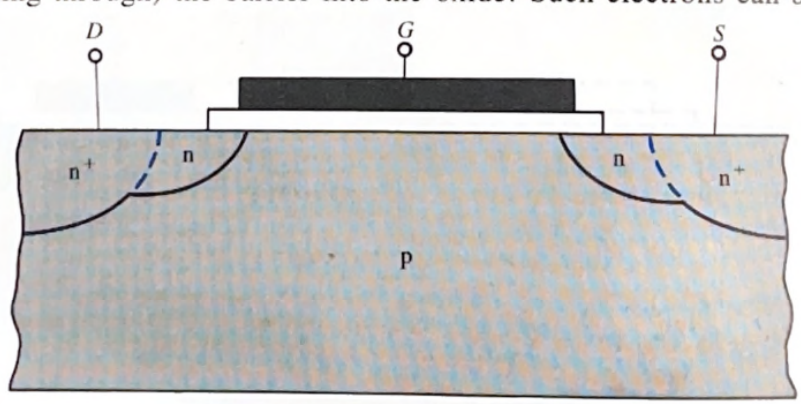


Figure 8-29 The lightly doped drain (LDD) structure, in which the source and drain regions are doped more lightly near the channel to reduce the electric fields and therefore reduce hot electron effects.

trapped in the oxide, where they change the threshold voltage and the I-V characteristics of the device. Hot electron effects can be reduced by reducing the doping in the source and drain regions, so that the junction fields are smaller. However, lightly doped source and drain regions are incompatible with small geometry devices because of contact resistance and other problems. A compromise design called the *lightly doped drain* (LDD) uses two doping levels, with heavy doping over most of the source and drain areas but with light doping in a region adjacent to the channel (Fig. 8-29). The LDD structure decreases the field between the drain and channel regions, thereby reducing injection into the oxide, impact ionization, and other hot electron effects.

PROBLEMS

8.1. Modify Eqs. (8-1) through (8-4) to include effects of the contact potential V_0. Define a true pinch-off voltage V_T to distinguish this case from V_P defined in Eq. (8-3).

8.2. Modify Eqs. (8-6) through (8-9) to include V_0. Let V_P be defined as in Eq. (8-3), and call the true pinch-off voltage V_T.

8.3. Assume the JFET shown in Fig. 8-4 is Si and has p^+ regions doped with 10^{18} acceptors/cm^3 and a channel with 10^{16} donors/cm^3. If the channel half-width a is 1 μm, compare V_P with V_0. What voltage V_{GD} is required to cause pinch-off when V_0 is included? With $V_G = -3$ V, at what value of V_D does the current saturate?

8.4. If the ratio $Z/L = 10$ for the JFET of Prob. 8.3, and $\mu_n = 1000$ cm^2/V-s, calculate $I_D(sat)$ for $V_G = 0, -2, -4$, and -6 V. Plot $I_D(sat)$ vs. $V_D(sat)$.

8.5. For the JFET of Prob. 8.4, plot I_D vs. V_D for the same three values of V_G. Terminate each plot at the point of saturation.

8.6. Show that Eq. (8-8) results from integration of Eq. (8-7).

8.7. Redraw Figs. 8-10, 8-11, and 8-12 for the p-channel (n-type substrate) case.

8.8. Show that the width of the depletion region in Fig. 8-12 is given by Eq. (8-23). Assume the carriers are completely swept out within W, as was done in Section 5.2.3.

8.9. An Al-gate n-channel MOS transistor is made on a p-type Si substrate with $N_a = 10^{17}$ cm^{-3}. The SiO_2 thickness is 100 Å in the gate region, and the effective interface charge Q_i is $5 \times 10^{10} q$ C/cm^2.
(a) Find W_m, V_{FB}, and V_T.
(b) Sketch the C–V curve for this device and give important numbers for the scale.