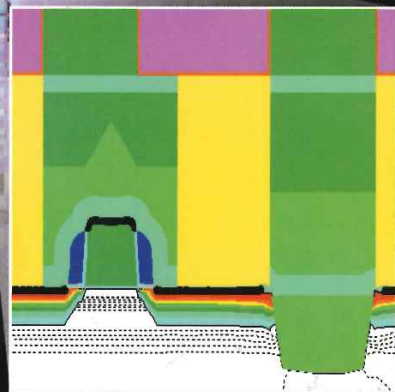
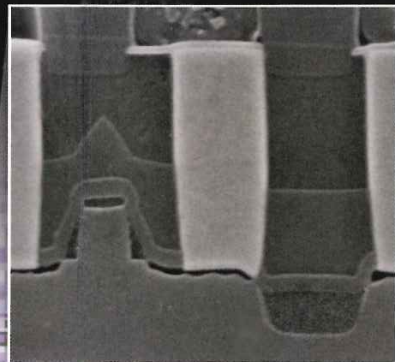


SILICON VLSI TECHNOLOGY

Fundamentals, Practice and Modeling



James D. Plummer • Michael D. Deal • Peter B. Griffin

Prentice Hall Electronics and VLSI Series—Charles Sodini, Series Editor

Silicon VLSI Technology

Fundamentals, Practice and Modeling

James D. Plummer

Michael D. Deal

Peter B. Griffin

*Department of Electrical Engineering
Stanford University*



Prentice Hall
Upper Saddle River, NJ 07458

Library of Congress Cataloging-in-Publication Data

Silicon VLSI technology

p. cm.

ISBN 0-13-085037-3

1. Integrated circuits—Very large scale integration—Design and construction. 2. Silicon. 3. Silicon oxide films. 4. Metal oxide semiconductors. 5. Silicon technology.

TK7874.75.S54 2000

621.39'5—dc21

99-42745

CIP

Publisher: Tom Robbins

Associate Editor: Alice Dworkin

Editorial/Production Supervision: Rose Kernan

Vice President and Editorial Director, ECS: Marcia Horton

Vice President of Production and Manufacturing: David W. Riccardi

Executive Managing Editor: Vince O'Brien

Marketing Manager: Danny Hoyt

Managing Editor: David A. George

Manufacturing Buyer: Pat Brown

Manufacturing Manager: Trudy Piscioti

Art Director: Jayne Conte

Cover Design: Bruce Kenselaar

Editorial Assistant: Jesse Power

Copy Editor: Martha Williams

Composition: D&G Limited, LLC



©2000 by Prentice Hall, Inc.

Upper Saddle River, New Jersey 07458

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.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Printed in the United States of America

10 9 8 7 6 5 4 3 2

ISBN 0-13-085037-3

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*

Pearson Education Pte., Ltd., *Singapore*

Editora Prentice Hall do Brasil, Ltda., *Rio de Janeiro*

Contents

Preface.....	xi
Chapter 1 Introduction and Historical Perspective	1
1.1 Introduction.....	1
1.2 Integrated Circuits and the Planar Process—Key Inventions That Made It All Possible.....	7
1.3 Semiconductors.....	13
1.4 Semiconductor Devices.....	33
1.4.1 PN Diodes.....	33
1.4.2 MOS Transistors.....	36
1.4.3 Bipolar Junction Transistors.....	39
1.5 Semiconductor Technology Families.....	41
1.6 Modern Scientific Discovery—Experiments, Theory, and Computer Simulation.....	43
1.7 The Plan For This Book.....	45
1.8 Summary of Key Ideas.....	46
1.9 References.....	46
1.10 Problems.....	47
Chapter 2 Modern CMOS Technology	49
2.1 Introduction.....	49
2.2 CMOS Process Flow.....	50
2.2.1 The Beginning—Choosing a Substrate.....	51
2.2.2 Active Region Formation.....	52
2.2.3 Process Option for Device Isolation—Shallow Trench Isolation.....	57
2.2.4 N and P Well Formation.....	60
2.2.5 Process Options for Active Region and Well Formation.....	63
2.2.6 Gate Formation.....	71
2.2.7 Tip or Extension (LDD) Formation.....	76
2.2.8 Source/Drain Formation.....	80
2.2.9 Contact and Local Interconnect Formation.....	82
2.2.10 Multilevel Metal Formation.....	84
2.3 Summary of Key Ideas.....	90
2.4 Problems.....	91

3.1	Introduction	93
3.2	Historical Development and Basic Concepts	93
3.2.1	Crystal Structure	94
3.2.2	Defects in Crystals	97
3.2.3	Raw Materials and Purification	101
3.2.4	Czochralski and Float-Zone Crystal Growth Methods	102
3.2.5	Wafer Preparation and Specification	105
3.3	Manufacturing Methods and Equipment	109
3.4	Measurement Methods	111
3.4.1	Electrical Measurements	111
3.4.1.1	Hot Point Probe	112
3.4.1.2	Sheet Resistance	113
3.4.1.3	Hall Effect Measurements	115
3.4.2	Physical Measurements	117
3.4.2.1	Defect Etches	117
3.4.2.2	Fourier Transform Infrared Spectroscopy (FTIR)	118
3.4.2.3	Electron Microscopy	119
3.5	Models and Simulation	121
3.5.1	Czochralski Crystal Growth	122
3.5.2	Dopant Incorporation during CZ Crystal Growth	125
3.5.3	Zone Refining and FZ Growth	128
3.5.4	Point Defects	131
3.5.5	Oxygen in Silicon	138
3.5.6	Carbon in Silicon	142
3.5.7	Simulation	143
3.6	Limits and Future Trends in Technologies and Models	144
3.7	Summary of Key Ideas	146
3.8	References	147
3.9	Problems	148

Chapter 4	Semiconductor Manufacturing—Clean Rooms, Wafer Cleaning, and Gettering	151
4.1	Introduction	151
4.2	Historical Development and Basic Concepts	154
4.2.1	Level 1 Contamination Reduction: Clean Factories	157
4.2.2	Level 2 Contamination Reduction: Wafer Cleaning	159
4.2.3	Level 3 Contamination Reduction: Gettering	161
4.3	Manufacturing Methods and Equipment	165
4.3.1	Level 1 Contamination Reduction: Clean Factories	165
4.3.2	Level 2 Contamination Reduction: Wafer Cleaning	166
4.3.3	Level 3 Contamination Reduction: Gettering	167
4.4	Measurement Methods	169
4.4.1	Level 1 Contamination Reduction: Clean Factories	169

4.4.2	Level 2 Contamination Reduction: Wafer Cleaning.....	173
4.4.3	Level 3 Contamination Reduction: Gettering.....	176
4.5	Models and Simulation.....	180
4.5.1	Level 1 Contamination Reduction: Clean Factories.....	181
4.5.2	Level 2 Contamination Reduction: Wafer Cleaning.....	184
4.5.3	Level 3 Contamination Reduction: Gettering.....	186
4.5.3.1	Step 1: Making the Metal Atoms Mobile.....	186
4.5.3.2	Step 2: Metal Diffusion to the Gettering Site.....	187
4.5.3.3	Step 3: Trapping the Metal Atoms at the Gettering Site.....	190
4.6	Limits and Future Trends in Technologies and Models.....	193
4.7	Summary of Key Ideas.....	196
4.7	References.....	196
4.9	Problems.....	198

Chapter 5 Lithography 201

5.1	Introduction.....	201
5.2	Historical Development and Basic Concepts.....	203
5.2.1	Light Sources.....	206
5.2.2	Wafer Exposure Systems.....	208
5.2.2.1	Optics Basics—Ray Tracing and Diffraction.....	209
5.2.2.2	Projection Systems (Fraunhofer Diffraction).....	212
5.2.2.3	Contact and Proximity Systems (Fresnel Diffraction).....	219
5.2.3	Photoresists.....	221
5.2.3.1	g-line and i-line Resists.....	223
5.2.3.2	Deep Ultraviolet (DUV) Resists.....	225
5.2.3.3	Basic Properties and Characterization of Resists.....	227
5.2.4	Mask Engineering—Optical Proximity Correction and Phase Shifting..	230
5.3	Manufacturing Methods and Equipment.....	234
5.3.1	Wafer Exposure Systems.....	234
5.3.2	Photoresists.....	238
5.4	Measurement Methods.....	241
5.4.1	Measurement of Mask Features and Defects.....	242
5.4.2	Measurement of Resist Patterns.....	244
5.4.3	Measurement of Etched Features.....	244
5.5	Models and Simulation.....	246
5.5.1	Wafer Exposure Systems.....	247
5.5.2	Optical Intensity Pattern in the Photoresist.....	253
5.5.3	Photoresist Exposure.....	259
5.5.3.1	g-line and i-line DNQ Resists.....	259
5.5.3.2	DUV Resists.....	263
5.5.4	Postexposure Bake (PEB).....	264
5.5.4.1	g-line and i-line DNQ Resists.....	264
5.5.4.2	DUV Resists.....	266
5.5.5	Photoresist Developing.....	267
5.5.6	Photoresist Postbake.....	270

5.6.2	X-ray Lithography	275
5.6.3	Advanced Mask Engineering	277
5.6.4	New Resists	278
5.7	Summary of Key Ideas	281
5.8	References	281
5.9	Problems	283

Chapter 6 Thermal Oxidation and the Si/SiO₂ Interface 287

6.1	Introduction	287
6.2	Historical Development and Basic Concepts	290
6.3	Manufacturing Methods and Equipment	296
6.4	Measurement Methods	298
6.4.1	Physical Measurements	299
6.4.2	Optical Measurements	299
6.4.3	Electrical Measurements—The MOS Capacitor	301
6.5	Models and Simulation	312
6.5.1	First-Order Planar Growth Kinetic —The Linear Parabolic Model	313
6.5.2	Other Models for Planar Oxidation Kinetics	322
6.5.3	Thin Oxide SiO ₂ Growth Kinetics	326
6.5.4	Dependence of Growth Kinetics on Pressure	328
6.5.5	Dependence of Growth Kinetics on Crystal Orientation	329
6.5.6	Mixed Ambient Growth Kinetics	332
6.5.7	2D SiO ₂ Growth Kinetics	333
6.5.8	Advanced Point Defect Based Models for Oxidation	339
6.5.9	Substrate Doping Effects	343
6.5.10	Polysilicon Oxidation	345
6.5.11	Si ₃ N ₄ Growth and Oxidation Kinetics	347
6.5.12	Silicide Oxidation	350
6.5.13	Si/SiO ₂ Interface Charges	352
6.5.14	Complete Oxidation Module Simulation	357
6.6	Limits and Future Trends in Technologies and Models	359
6.7	Summary of Key Ideas	361
6.8	References	361
6.9	Problems	364

Chapter 7 Dopant Diffusion 371

7.1	Introduction	371
7.2	Historical Development and Basic Concepts	374
7.2.1	Dopant Solid Solubility	375
7.2.2	Diffusion from a Macroscopic Viewpoint	377
7.2.3	Analytic Solutions of the Diffusion Equation	379
7.2.4	Gaussian Solution in an Infinite Medium	380

7.2.5	Gaussian Solution Near a Surface	381
7.2.6	Error-Function Solution in an Infinite Medium	382
7.2.7	Error-Function Solution Near a Surface	384
7.2.8	Intrinsic Diffusion Coefficients of Dopants in Silicon	386
7.2.9	Effect of Successive Diffusion Steps	388
7.2.10	Design and Evaluation of Diffused Layers	389
7.2.11	Summary of Basic Diffusion Concepts	392
7.3	Manufacturing Methods and Equipment	392
7.4	Measurement Methods	395
7.4.1	SIMS	396
7.4.2	Spreading Resistance	397
7.4.3	Sheet Resistance	398
7.4.4	Capacitance Voltage	399
7.4.5	TEM Cross Section	399
7.4.6	2D Electrical Measurements Using Scanning Probe Microscopy	400
7.4.7	Inverse Electrical Measurements	402
7.5	Models and Simulation	403
7.5.1	Numerical Solutions of the Diffusion Equation	403
7.5.2	Modifications to Fick's Laws to Account for Electric Field Effects	406
7.5.3	Modifications to Fick's Laws to Account for Concentration-Dependent Diffusion	409
7.5.4	Segregation	413
7.5.5	Interfacial Dopant Pileup	415
7.5.6	Summary of the Macroscopic Diffusion Approach	417
7.5.7	The Physical Basis for Diffusion at an Atomic Scale	417
7.5.8	Oxidation-Enhanced or -Retarded Diffusion	419
7.5.9	Dopant Diffusion Occurs by Both I and V	422
7.5.10	Activation Energy for Self-Diffusion and Dopant Diffusion	426
7.5.11	Dopant-Defect Interactions	426
7.5.12	Chemical Equilibrium Formulation for Dopant-Defect Interactions	432
7.5.13	Simplified Expression for Modeling	434
7.5.14	Charge State Effects	436
7.6	Limits and Future Trends in Technologies and Models	439
7.6.1	Doping Methods	440
7.6.2	Advanced Dopant Profile Modeling—Fully Kinetic Description of Dopant-Defect Interactions	440
7.7	Summary of Key Ideas	442
7.8	References	443
7.9	Problems	445

Chapter 8 Ion Implantation 451

8.1	Introduction	451
8.2	Historical Development and Basic Concepts	451
8.2.1	Implants in Real Silicon—The Role of the Crystal Structure	461
8.3	Manufacturing Methods and Equipment	463

8.5	Models and Simulations	470
8.5.1	Nuclear Stopping	471
8.5.2	Nonlocal Electronic Stopping	473
8.5.3	Local Electronic Stopping	474
8.5.4	Total Stopping Powers	475
8.5.5	Damage Production	476
8.5.6	Damage Annealing	479
8.5.7	Solid-Phase Epitaxy	482
8.5.8	Dopant Activation	484
8.5.9	Transient-Enhanced Diffusion	486
8.5.10	Atomic-Level Understanding of TED	488
8.5.11	Effects on Devices	497
8.6	Limits and Future Trends in Technologies and Models	499
8.7	Summary of Key Ideas	500
8.8	References	500
8.9	Problems	502

Chapter 9 Thin Film Deposition 509

9.1	Introduction	509
9.2	Historical Development and Basic Concepts	511
9.2.1	Chemical Vapor Deposition (CVD)	512
9.2.1.1	Atmospheric Pressure Chemical Vapor Deposition (APCVD)	513
9.2.1.2	Low-Pressure Chemical Vapor Deposition (LPCVD)	525
9.2.1.3	Plasma-Enhanced Chemical Vapor Deposition (PECVD)	527
9.2.1.4	High-Density Plasma Chemical Vapor Deposition (HDPCVD)	530
9.2.2	Physical Vapor Deposition (PVD)	530
9.2.2.1	Evaporation	531
9.2.2.2	Sputter Deposition	539
9.3	Manufacturing Methods	554
9.3.1	Epitaxial Silicon Deposition	556
9.3.2	Polycrystalline Silicon Deposition	558
9.3.3	Silicon Nitride Deposition	561
9.3.4	Silicon Dioxide Deposition	563
9.3.5	Al Deposition	565
9.3.6	Ti and Ti-W Deposition	566
9.3.7	W Deposition	567
9.3.8	TiSi ₂ and WSi ₂ Deposition	567
9.3.9	TiN Deposition	568
9.3.10	Cu Deposition	570
9.4	Measurement Methods	572

9.5	Models and Simulation.....	573
9.5.1	Models for Deposition Simulations	573
9.5.1.1	Models in Physically Based Simulators Such as SPEEDIE	574
9.5.1.2	Models for Different Types of Deposition Systems.....	582
9.5.1.3	Comparing CVD and PVD and Typical Parameter Values	587
9.5.2	Simulations of Deposition Using a Physically Based Simulator, SPEEDIE	590
9.5.3	Other Deposition Simulations	598
9.6	Limits and Future Trends in Technologies and Models	601
9.7	Summary of Key Ideas	602
9.8	References	603
9.9	Problems	605

Chapter 10 Etching 609

10.1	Introduction.....	609
10.2	Historical Development and Basic Concepts	612
10.2.1	Wet Etching.....	612
10.2.2	Plasma Etching	619
10.2.2.1	Plasma Etching Mechanisms	621
10.2.2.2	Types of Plasma Etch Systems	628
10.2.2.3	Summary of Plasma Systems and Mechanisms	636
10.3	Manufacturing Methods.....	637
10.3.1	Plasma Etching Conditions and Issues	638
10.3.2	Plasma Etch Methods for Various Films	643
10.3.2.1	Plasma Etching Silicon Dioxide.....	644
10.3.2.2	Plasma Etching Polysilicon.....	647
10.3.2.3	Plasma Etching Aluminum.....	649
10.4	Measurement Methods.....	650
10.5	Models and Simulation.....	653
10.5.1	Models for Etching Simulation.....	653
10.5.2	Etching Models—Linear Etch Model	656
10.5.3	Etching Models—Saturation/Adsorption Model for Ion-Enhanced Etching.....	663
10.5.4	Etching Models—More Advanced Models.....	669
10.5.5	Other Etching Simulations	671
10.6	Limits and Future Trends in Technologies and Models	675
10.7	Summary of Key Ideas	676
10.8	References	677
10.9	Problems	679

Chapter 11 Back-End Technology 681

11.1	Introduction.....	681
11.2	Historical Development and Basic Concepts	687
11.2.1	Contacts.....	688
11.2.2	Interconnects and Vias.....	695

11.3.3	Contact Formation	719
11.3.4	Global Interconnects	721
11.3.5	IMD Deposition and Planarization	723
11.3.6	Via Formation	724
11.3.7	Final Steps	725
11.4	Measurement Methods.	725
11.4.1	Morphological Measurements	726
11.4.2	Electrical Measurements	726
11.4.3	Chemical and Structural Measurements	732
11.4.4	Mechanical Measurements	734
11.5	Models and Simulation.	737
11.5.1	Silicide Formation.	738
11.5.2	Chemical-Mechanical Polishing	744
11.5.3	Reflow	746
11.5.4	Grain Growth	753
11.5.5	Diffusion in Polycrystalline Materials	762
11.5.6	Electromigration	765
11.6	Limits and Future Trends in Technologies and Models	776
11.7	Summary of Key Ideas	780
11.8	References	781
11.9	Problems	784

Appendices 787

A.1	Standard Prefixes	787
A.2	Useful Conversions.	787
A.3	Physical Constants	788
A.4	Physical Properties of Silicon	788
A.5	Properties of Insulators Used in Silicon Technology	789
A.6	Color Chart for Deposited Si_3N_4 Films Observed Perpendicularly under Daylight Fluorescent Lighting.	789
A.7	Color Chart for Thermally Grown SiO_2 Films Observed Perpendicularly under Daylight Fluorescent Lighting.	790
A.8	Irwin Curves	791
A.9	Error Function.	793
A.10	List of Important Symbols	797
A.11	List of Common Acronyms	798
A.12	Tables in Text	801
A.13	Answers to Selected Problems	802

Index 805

provided substantial inputs and comments. Our own work in this field has been supported over many years by DARPA (the Defense Advanced Research Projects Agency) and by the Semiconductor Research Corporation (SRC). We owe them a considerable debt of gratitude for making our work possible. We welcome comments or suggestions on this text by email at plummer@ee.stanford.edu, deal@ee.stanford.edu, or griffin@stanford.edu.

Introduction and Historical Perspective

1

1.1 Introduction

Since its invention about 40 years ago, the Integrated Circuit (IC) has literally changed our world. From a modest beginning, which permitted the integration of a few semiconductor devices and passive components on a common substrate, has come a technology which can integrate tens of millions of components in a square cm of silicon. From a beginning in which integrated circuits were applied only to a very small set of specialized applications has come a technology which is pervasive in today's world. Many have referred to these events as a true revolution, the beginning of the information age. Most industrialized nations today believe that integrated circuits will be one of the basic building blocks for their future economies. Global competition for a share of these market opportunities is intense today and will remain so for the foreseeable future.

What are the factors that have contributed to this remarkable growth in integrated circuit capability? Fundamentally, integrated circuits (chips) consist of a large number of individual components (transistors, resistors, capacitors, etc.), fabricated side by side in a common substrate and wired together to perform a particular circuit function. Over time, more complex circuits are built simply by making the individual components smaller which allows more of them to be integrated in a given area, and by making the overall size of chips larger. The higher levels of integration permitted by these smaller devices and bigger chips have made it possible to build progressively more complex, higher-performance, and more economical integrated systems as time has passed.

Figure 1-1 illustrates the increasing complexity of integrated circuits over the past 35 years. Technology improvements over that period have resulted in a doubling of the number of transistors on a chip roughly every two years, a "law" that has become known as "Moore's law" [1.1]. Decreasing component size is often characterized in IC technology by the minimum line width. This is usually defined as the smallest lateral feature size that is printed on the wafer surface during the fabrication process. This dimension is plotted in Figure 1-2 over the past 35 years and extrapolated into the future. On average, this dimension has decreased by 13% each year. In recent years, this reduction has been closer to 10% per year. The other principal contributor to increasing chip complexity, overall chip size, has risen by about 16% each year over the past 35 years. In recent years this number has also slightly decreased, ranging from about 6.3% per year for microprocessors, to 12% per year for dynamic memory or DRAMs.

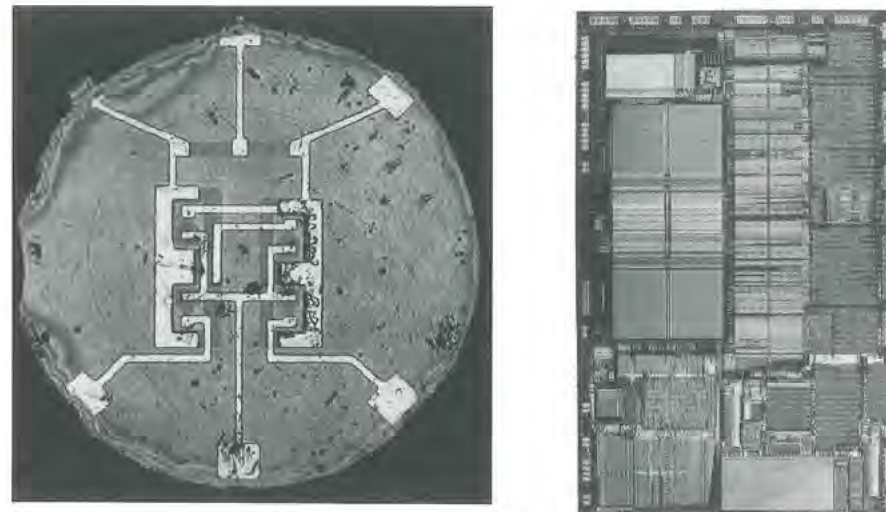


Figure 1-1 Photomicrographs of state-of-the-art ICs manufactured in the early 1960s (left) and in the early 1990s (right). The 1960s IC contains four bipolar transistors and several resistors. The 1990s chip contains over a million MOS transistors.

In 1994 the Semiconductor Industry Association (SIA) organized a workshop in Boulder, Colorado which was chartered to produce a roadmap for the semiconductor industry. The result was the National Technology Roadmap for Semiconductors or NTRS [1.2]. This roadmap was updated in 1997 through a series of industrywide meetings [1.3] and is being updated again in 1999. The NTRS is currently the most widely cited “strategic plan” for the semiconductor industry. The fundamental premise of the NTRS is that industry history as described by Moore’s law should be maintained through 2012. In other words, a new generation of technology should appear on average every three years, starting with 0.25- μm technology today and extending to 0.05- μm technology in 2012. The NTRS is a needs-driven document which assumes that silicon CMOS technology will dominate the semiconductor industry for this entire period. The starting point is a standard set of industry products—DRAMs, static memory or SRAMs, microprocessors, and Application-Specific Integrated Circuits or ASICs—which result in the technology requirements in Table 1-1, abstracted from the 1997 roadmap.

Several trends can be observed from this table. First, of course, are the steady progressions to smaller feature sizes and larger chips. Together these make possible the increases in transistors per chip. Note that a factor greater than 100X is expected in the complexity of microprocessor chips over the next 15 years. The number of wiring levels is projected to steadily increase over this time period as well. This is because of the great difficulty in connecting all the components in these very complex chips. This is the same trend that has occurred in the number of wiring levels in printed circuit boards over the

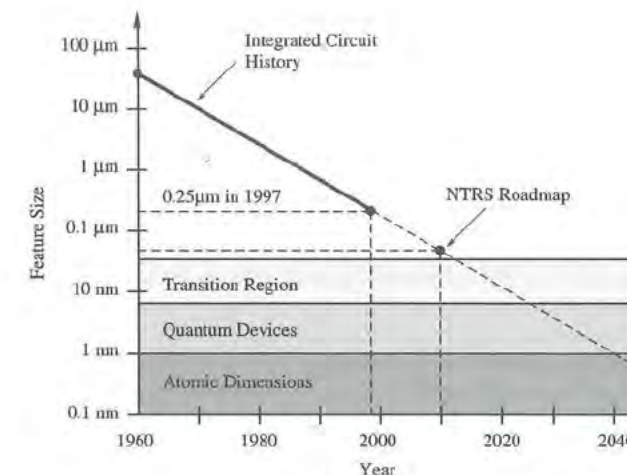


Figure 1-2 Historical trends and future projections for the minimum feature size used in integrated circuits in manufacturing.

years, for the same reasons. The mask count is a measure of the number of lithography or printing operations required to manufacture the chips. It is really a measure of the complexity of the manufacturing process. Here the trend is obviously upward as well, primarily to accommodate the increased wiring levels needed in the chips. Finally, the operating voltage of future chips is expected to decrease significantly. This is driven by the fact that as transistors get smaller, the electric fields in these devices will increase to unacceptable levels unless operating voltages are reduced. Just as important, the power dissipation of chips is proportional to the square of the operating voltage and this is also becoming a major issue as more and more complexity is packed onto a single chip.

Table 1-1 Future projections for silicon technology taken from the SIA NTRS [1.3]

Year of first DRAM shipment	1997	1999	2003	2006	2009	2012
Minimum Feature Size	250 nm	180 nm	130 nm	100 nm	70 nm	50 nm
DRAM Bits/Chip	256M	1G	4G	16G	64G	256G
DRAM Chip Size (mm^2)	280	400	560	790	1120	1580
Microprocessor Transistors/chip	11M	21M	76M	200M	520M	1.40B
Maximum Wiring Levels	6	6-7	7	7-8	8-9	9
Minimum Mask Count	22	22-24	24	24-26	26-28	28
Minimum Supply Voltage (volts)	1.8-2.5	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6

One final point about the NTRS is worth making. This is the assumption that in the future, silicon will remain the basic material used in the fabrication of ICs. There are

many other material possibilities that have been investigated over the years, ranging from elemental semiconductors like Ge to compound materials like GaAs. None of these, however, has the combination of properties that together provide a viable alternative to silicon. We will see why this is so throughout this text as we discuss the various processing technologies used to make silicon ICs today. Simply stated, silicon will be the dominant material used for chips for as far into the future as we can reliably see.

The primary devices used in today's integrated circuits are Metal Oxide Semiconductor (MOS) transistors and Bipolar Junction Transistors (BJTs). The basic physical principles underlying these devices were understood before the invention of the integrated circuit. The field effect mechanism on which the MOS device is based was first investigated in the 1930s. The operating principles of the BJT device were first described in the late 1940s. While the first ICs were built with BJT devices, the MOS transistor has become the dominant device in the past two decades. MOS devices offer higher density, lower power operation, and considerably greater design flexibility than do BJT devices. BJT devices are still used today in analog and a few other applications, but their market share will likely continue to shrink. As a result, most efforts to develop new IC technologies today focus on MOS technologies and that is the emphasis we will take in this book.

Many studies have probed the question of how small MOS devices can be made before their basic operating principles change. Stated another way, how long can the historical trends of shrinking transistors and increasing chip complexity continue? While there is still disagreement about the exact answers to these questions, it is clear that quantum mechanical effects will become important when device dimensions reach 10 or 20 nanometers. It is also clear that toward the end of the NTRS or perhaps just beyond it, today's MOS device structures will begin to encounter significant problems. Lateral device dimensions between $0.05\text{ }\mu\text{m}$ (50 nm) and about $0.01\text{ }\mu\text{m}$ (10 nm) fall into this category, labeled the "transition region" in Figure 1-2. An example of an issue that is important at these dimensions is that gate oxides in MOS devices likely cannot be shrunk below about 1–1.5 nm because electrons can easily tunnel through such thin insulators. (A 1-nm oxide thickness would be required in MOS devices with lateral dimensions on the order of $0.05\text{ }\mu\text{m}$ or 50 nm according to the NTRS.) Also at these dimensions, it is very difficult to keep the depletion regions associated with PN junctions from interacting with each other. In any case, device innovations will be required in order to continue shrinking dimensions in integrated circuits much beyond the NTRS. If history is any guide, these innovations will occur, resulting in the ability to continue shrinking device dimensions and increasing chip capability well into the next century. In fact the 1999 version of the NTRS has added another technology generation at 35 nm.

Many researchers today believe that the ultimate limits of silicon CMOS technology lie well beyond the above roadmap. Theoretical studies have indicated that fairly standard silicon transistor structures can work properly down to dimensions of 20–30 nm. If such devices can be manufactured economically, they will become the basis of chips with 10^{10} to 10^{12} devices around 2020. Such extraordinary "systems on a chip" will allow products that we cannot even envision clearly today.

It is interesting to note that even though there are device problems related to continued scaling, there are no known fundamental fabrication limitations until far into the



Figure 1-3 Xenon atoms individually placed on a nickel single crystal surface to spell IBM. The spacing between xenon atoms is about 1 nm. Reprinted with permission of IBM.

future. Integrated circuits today are built using optical techniques to produce the fine line patterns required for device structures. These techniques will likely have to be replaced at dimensions below $0.1\text{ }\mu\text{m}$ (although continued innovation may reduce this limit). There are a number of possibilities that have been shown to work at smaller dimensions, such as electron-beam and X-ray lithography. Other possibilities also exist, including some very recent and quite remarkable work (Figure 1-3) in which a scanning tunneling microscope was used to control the placement of individual atoms on a single-crystal Ni surface [1.4]. It is clear that technology options will allow smaller structures and continued device innovation far into the future.

In addition to shrinking individual device sizes, overall chip sizes have increased dramatically over the past three decades. Larger chips permit more complex functions to be integrated, improving system reliability and generally also producing higher system performance. This is because it is often the case that driving signals over the relatively large distances between chips is the limiting factor in overall system speed. Eliminating these long signal paths by combining functions on a single chip can therefore improve performance. It is also the case that the smaller devices characteristic of improved technologies are also intrinsically faster than their larger counterparts and hence also contribute to higher-performance systems.

At any point in time, there is an upper bound on the size of a chip which is economically manufacturable. Defects during the manufacturing process can arise from a number of sources including airborne particulates, machine introduced particles, or the printing or lithographic process. Such defects tend to be randomly distributed over a wafer and are often characterized with an average defect density per square centimeter. Since in most ICs redundancy is not used, a single defect larger than some critical size usually means that the chip will not function correctly. Manufacturing ICs with high yield therefore requires low defect densities. This has been the driving force for the ultra clean facilities in which ICs are manufactured today. When new chips are introduced to the marketplace, their manufacturing yield may be less than 20%. However, as experience is gained with manufacturing the part and as improvements are made in

manufacturing processes, yields often approach 80% or more in mature products. There are no fundamental reasons why chips much larger than those built today cannot be manufactured. We should expect the historical trends in this regard to continue for the foreseeable future.

There is, however, one byproduct of this drive for larger chips and higher manufacturing yields that should be mentioned. The cost of the factories in which such chips are built has also risen astronomically in the past three decades. In 1970, a state-of-the-art fabrication facility could be built for a few million dollars. The costs today are over \$1 billion. Concerns have been expressed that these very high costs will necessarily drive many companies out of high volume commodity IC markets (memory chips, for example) in the future. This has not happened as yet. Innovations in manufacturing such as designing and building plants so that capacity can be added in phases as the market grows, or sharing the costs of large plants across a broader product spectrum have helped to mitigate the high plant costs. The recent surge in foundries, which manufacture chips for a wide spectrum of customers, is an interesting development in this regard. In the foundry scenario, an independent company builds an expensive chip fabrication plant and then fabricates products using standard processes for many other companies which often do not have their own chip manufacturing plants. Also, of course, because of the economic importance of semiconductors, governments around the world have chosen to invest substantial resources in establishing chip manufacturing in their countries.

As the complexity of integrated circuits has increased, so have the costs associated with designing them. Gordon Moore pointed out in 1979 [1.5] that the cost required to design an IC was increasing at about the same rate as the complexity of the technology. Transistor count was doubling in ICs every two years; design costs were not far behind. Had this continued, there would be very few ICs being designed today at the capability of the technology (tens of millions of transistors), simply because it would be difficult to define products that could return the huge investment needed to design such chips. Beginning in the late 1970s however, a revolution took place, which dramatically reduced design costs. That revolution was the introduction of very powerful Computer-Aided Design (CAD) tools, which take care of many of the design details. CAD tools continue to be refined today and they have opened up the opportunity to design ICs to a wide range of people who do not necessarily have to know about the details of transistor physics or IC technology. This CAD revolution has included the introduction of ideas like gate arrays, programmable logic arrays, standard cells, and powerful tools to design fully custom ICs. The net result has been that IC design costs have remained low enough that thousands of new chips are introduced each year. In fact the number of new chip designs introduced each year continues to increase.

In the remainder of this chapter we will discuss some of the important inventions, particularly in the 1940s and 50s, that made integrated circuits possible. We will also review basic ideas about semiconductors, materials, devices and circuit families. This review is not intended to be comprehensive since there are many complete references, which cover these topics in detail. However, since we will make use of many of these concepts in later chapters, the background provided here will be important to those discussions.

1.2 Integrated Circuits and the Planar Process—Key Inventions That Made It All Possible

The devices which are used in today's integrated circuits, primarily MOS and bipolar transistors, were invented long before the technologies were available to manufacture them in high volume. Many of the basic properties of semiconductors such as rectification and photoconductivity were discovered before 1900 although they were not understood at that time. Simple devices based on these properties were available in the form of selenium rectifiers and photodetectors by the mid 1930s. It was also during this time period when the physical principles underlying the behavior of metal/semiconductor contacts began to be understood, particularly through the work of Schottky and Mott.

The behavior of semiconductor surfaces plays a key role in all modern devices. Lian-drat first proposed in 1935 that the conductivity of a surface region in a semiconductor could be modulated by the application of a perpendicular electric field. This "field effect" concept led to many studies of semiconductor surfaces over the next three decades.

The Second World War interrupted much of the basic work on semiconductors that was ongoing at the time, particularly at Bell Telephone Laboratories where an effort was underway to find a solid-state device for switching telephone signals. Shortly after the end of the war, this effort resumed, and it was not long until a major breakthrough was made with the demonstration of the point contact transistor in December 1947 (Figure 1-4). This and subsequent work, which resulted in the invention of the bipolar transistor, resulted in the Nobel Prize in physics for John Bardeen, Walter Brattain, and William Shockley in 1956. An interesting historical account of many of these events can be found in [1.6].



Figure 1-4 The point contact transistor invented at Bell Telephone Laboratories in 1947 by Bardeen, Brattain, and Shockley. Reprinted with permission of Lucent Technologies, Bell Labs Innovations.

In the 1950s, interest in semiconductor surfaces was renewed when it became apparent that reliability problems associated with the new transistor structures were related to surface effects. In a classic experiment in 1953, Brattain and Bardeen found that the surface properties of semiconductors could be controlled by exposing them to oxygen, water, or ozone ambients. Many experiments over the next few years led to the first high-quality SiO_2 layers grown on Si substrates around 1960. At about this same time, the stable properties of SiO_2 layers on Si began to be understood theoretically.

Although the first point contact transistors in 1947 were built in polycrystalline germanium, within a year or two after that the device was also demonstrated in silicon and in single-crystal material. These two changes would also have a major impact on future integrated circuits. Single crystals provide uniform and reproducible device characteristics and thus provide the ability to integrate millions of identical components side by side on a chip. Silicon provides a controllable, stable, and reproducible surface passivation layer (SiO_2) which has made possible modern IC technology. Much of the credit for developing single crystal source material (Si and Ge) belongs to Gordon Teal of Bell Labs. Chapter 3 describes these developments in more detail.

By the mid 1950s, both grown junction and alloy junction bipolar transistors were commercially available. Ge was still the dominant material used at this time. Figures 1-5 and 1-6 illustrate the methods used to fabricate these devices. In the grown junction process, a single crystal of silicon or germanium was grown which contained a thin region (P type in Figure 1-5) of opposite doping to the main part of the crystal. Sawing removed most of the upper and lower N-type material, leaving a thin wafer with an NPN structure. Further sawing of this wafer resulted in a number of individual transistor structures to which leads were attached to provide external connections. In silicon-based devices, it was common to use Al wires to connect to the middle base (P) region since Al forms an ohmic (conducting) contact to P-type material and a rectifying

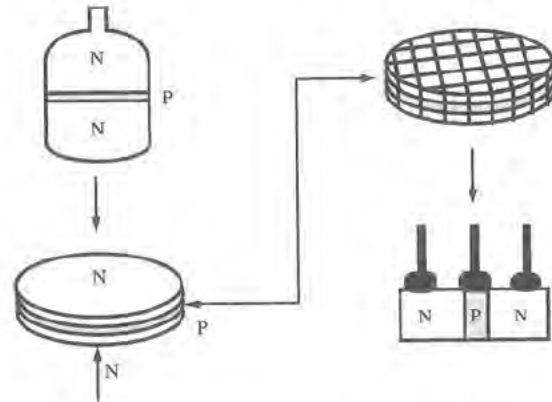


Figure 1-5 Grown junction transistor technology of the 1950s.

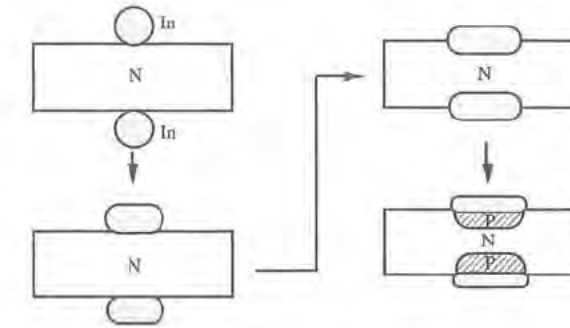


Figure 1-6 Alloy junction technology of the 1950s.

(Schottky) contact to lightly doped N regions. As a result it was not critical that the middle connection be placed exactly on the P region. In Ge based structures of this type, PNP type devices were usually made, with Au contacts to the middle N region.

The alloy junction technology illustrated in Figure 1-6 was even simpler. A metal such as indium was placed on the semiconductor (usually Ge). The structure was then heated, melting the In and allowing it to dissolve into the Ge. Indium is a P-type dopant so P regions were formed creating a PNP structure along with contacts to the P regions.

While such devices were very useful components, the technologies used to build them were not extendible to multitransistor integrated circuits. Exposed junctions were present on the semiconductor surface and no means to interconnect multiple devices was provided. New inventions were needed to overcome these problems.

Part of the solution was provided by the invention of gas phase diffusion processes, again at Bell Laboratories. That led to the commercial availability of diffused mesa bipolar transistors by 1957. Figure 1-7 illustrates this device structure. Beginning with an N-type crystal, the wafer was exposed in a high temperature furnace to a gaseous source of a P-type dopant such as boron. The boron diffused into the crystal by solid-state diffusion, resulting in the P-type layer. The process was then repeated with an N-type source, producing the final NPN structure. After contacts were alloyed to the surface N and P regions, a Si etch created the mesa structure to localize the N and P regions on the surface. This double diffused process had the great advantage that multiple devices could be produced from a single substrate and multiple substrates (wafers) could be produced from one grown crystal. However, all of the processes described to this point still had the great disadvantage that exposed junctions were present on the wafer surface or at the wafer edges.

The next breakthrough came with the invention of the planar process, illustrated in Figure 1-8, by Jean Hoerni of Fairchild Semiconductor. This process relied on the gas phase diffusion of dopants to produce N- and P-type regions, but also on the ability of SiO_2 to mask these diffusions. This was a major advance and it was largely responsible for the switch from Ge to Si. Silicon is unique among semiconductor materials in its

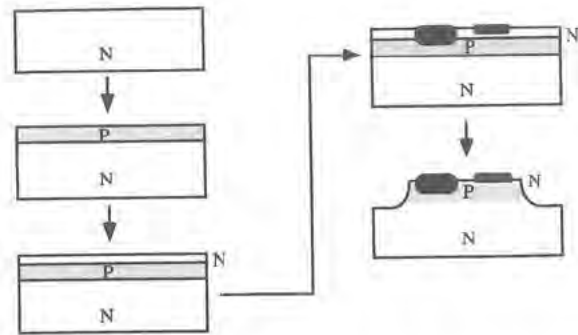


Figure 1-7 Double diffused mesa transistor technology of the late 1950s.

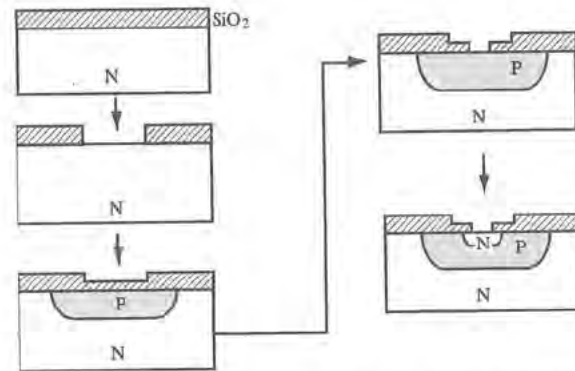


Figure 1-8 The planar process invented by Jean Hoerni of Fairchild in the late 1950s.

ability to be oxidized to produce a stable insulating coating. SiO_2 has many properties that make it almost an ideal surface layer for semiconductor devices. We will discuss oxidation and SiO_2 in much more detail in Chapter 6, but one of those properties is the ability to mask most of the common dopants that are used in semiconductor fabrication. In other words, most dopants like As, B, and P have much smaller rates of diffusion in SiO_2 than they do in silicon. The planar process makes use of this property by using SiO_2 to selectively block the dopants from diffusing into the silicon substrate. This produces junctions that terminate under an SiO_2 surface and which as a result are passivated. (We will define passivation more carefully in Chapter 6, but we basically mean by this term that the junction electrical properties are not significantly degraded at the surface, compared to their bulk properties.)

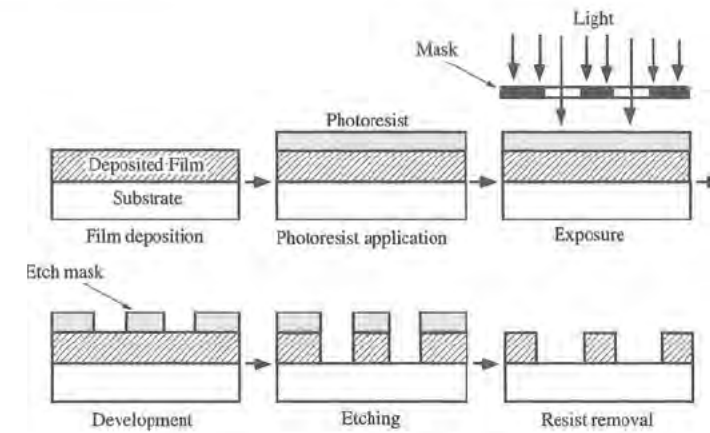


Figure 1-9 Basic photolithography process. The IC pattern is transferred from a mask to the silicon by printing it on the wafer using a light sensitive resist material.

A second essential part of the planar process, illustrated in Figure 1-9, is the ability to pattern the SiO_2 . This is normally accomplished by a process called photolithography that we will discuss in detail in Chapter 5. The essence of this process involves the use of a light sensitive resist (or "film"), which is deposited onto the SiO_2 surface. The resist is then exposed through a mask and developed to produce a pattern in the resist. During the developing, the resist washes away in the regions where diffusions are desired. The resist is then used as a mask to selectively etch the SiO_2 layer.

One final key invention was needed to make modern IC technology possible. That was the ability to integrate multiple components on the same chip and to interconnect them to form a circuit. The invention of the integrated circuit occurred in 1959 and was due to Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Semiconductor. An interesting historical account of these events is given in [1.7]. The key ideas were the extensions of the ideas in Figures 1-8 and 1-9 to using masking and lithography to provide multiple devices and interconnects on the same chip. Figure 1-10 illustrates these extensions for a simple circuit example. Kilby and Noyce have been widely recognized for their work leading to the integrated circuit. By combining P- and N-type diffusions and SiO_2 passivation layers, many types of devices including transistors, resistors, and capacitors are possible in modern IC structures.

Although it is perhaps obvious from Figure 1-10, it should be noted that when multiple diffusions and thin film layers are used to construct an integrated circuit, the masks associated with each of the layers must be precisely aligned with respect to each other. That is, the N-type diffusion in the center NPN transistor must be placed inside the P-type region and the contact holes which allow the Al metal to contact the device regions must be placed inside those regions. In general the placement accuracy during

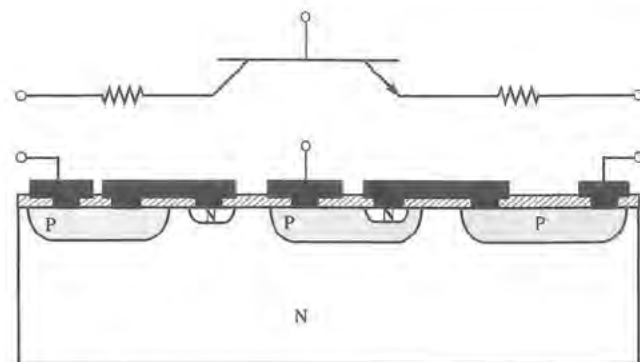


Figure 1-10 Integrated circuits use photolithography and masking to fabricate multiple components in a common substrate.

photolithography must be on the order of $1/4$ to $1/3$ of the linewidth being printed. In today's modern factories, this alignment process is carried out automatically in machines called steppers which we will discuss in more detail in Chapter 5.

Since 1960, the basic technologies used to manufacture integrated circuits have not changed. They have, of course, been significantly improved and many new ways of depositing, etching, diffusing, and patterning have been developed. These improvements have been more evolutionary than revolutionary however. In spite of this fact, the cumulative impact of rapid evolutionary developments over the past 35 years in IC technology has been enormous and we should expect many more years of the same kinds of developments.

Figure 1-11 shows a schematic cross section of a modern silicon integrated circuit. The silicon wafer is typically $> 500 \mu\text{m}$ thick; the layers on the top surface are typically $1 \mu\text{m}$ or thinner. Thus the full wafer thickness is not normally shown in these drawings. The N and P regions have been selectively doped to change their conductivity and they form the electrically active parts of the devices. The layers above the surface may be insulators like SiO_2 or Si_3N_4 , conductors like Al, or semiconductors like polycrystalline silicon. All of these layers are either selectively deposited only in certain areas or are patterned after a blanket deposition using photolithography and etching so that they remain only in local areas. The lateral scale of the various features is determined by the patterning capability of the technology. In high-volume manufacturing today, such features might have a minimum size of perhaps $0.25 \mu\text{m}$. The particular technology illustrated in Figure 1-11 is called CMOS (Complementary MOS) since it integrates both N- and P-type MOS devices side by side on the same chip. This is the technology that forms the basis of the NTRS discussed earlier and outlined in Table 1-1. Figure 1-12 shows an actual cross section of a modern CMOS IC. Note that the interconnect layers dominate the overall chip cross section. The devices themselves are all located in a thin layer near the bottom of the photomicrograph. Most of the silicon wafer itself is not shown and is off the bottom of the photograph.

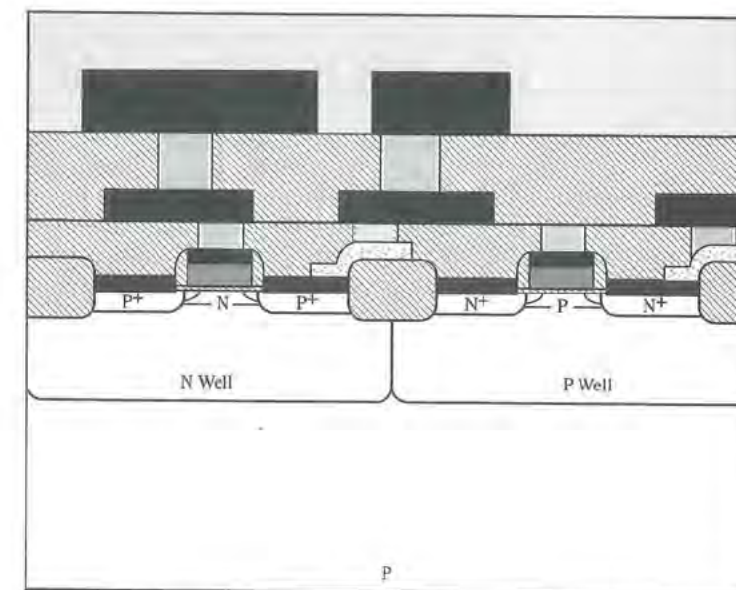


Figure 1-11 Schematic cross section of a modern silicon integrated circuit. The particular technology shown is CMOS with an NMOS device on the right and a PMOS device on the left. There are two levels of wiring shown. We will discuss this technology in detail in Chapter 2.

1.3 Semiconductors

This section provides a brief review of the basic properties of semiconductor materials. Many of the ideas discussed here will be used throughout this book. For readers already familiar with this material, this section can be skimmed or skipped.

Semiconductors are a class of materials which have the unique property that their electrical conductivity can be controlled over a very wide range by the introduction of dopants. While this property can easily be observed in crystalline, polycrystalline, or amorphous semiconductor materials, crystalline materials provide the most reproducible properties and the highest performance devices and are almost always used in integrated circuits. Dopants are atoms that generally contain either one more or one fewer electrons in their outermost shell than the host semiconductor. They provide one extra electron or one missing electron (a "hole") compared to the host atoms. These excess electrons and holes are the carriers, which carry current in semiconductor devices. The key to building semiconductor devices and integrated circuits lies in the ability to control the local doping and hence the local electronic properties of a semiconductor crystal.

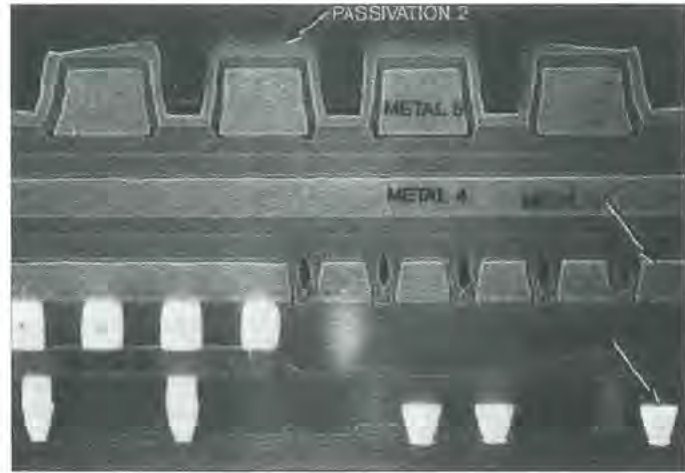


Figure 1-12 Actual cross section of a modern IC (IBM's PowerPC chip). Note the multiple layers of metal for wiring above the silicon surface. The active parts of the transistors are barely visible at the bottom of the photograph. Reprinted with permission of Integrated Circuit Engineering Corporation.

Consider silicon as a representative semiconductor. As illustrated in Figure 1-13, silicon has four electrons in its outermost shell. These are known as the valence electrons since they are the participants in chemical reactions and chemical bonding. When silicon atoms combine to form a solid crystal, a particularly stable electronic arrangement can be formed if the silicon atoms form covalent or shared electron bonds with their four nearest neighbors as illustrated in Figure 1-13. This arrangement is favored because each silicon atom can then “fill up” its outermost shell to a total of eight shared electrons. (In the periodic table, elements with full outer electron shells are the inert gases like Ar and Ne which are chemically rather inert and very stable.) The covalent bonding or sharing arrangement in essence populates the entire outer shell for each Si atom, resulting in a stable structure in which all electrons are bound to atoms, at least at very low temperatures.

Elemental semiconductors, all of which have this same bonding arrangement, lie in Column IV of the periodic table as shown in Figure 1-14. This same type of bonding arrangement can be produced using mixtures of elements from other columns of the periodic table. For example GaAs consists of alternating Ga (column III) and As (column V) atoms which have an average of four electrons per atom and so the same covalent bonding arrangement works. More complex examples like $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ are also possible. Thus nature provides many possible materials which can act as semiconductors.

At temperatures above absolute zero, thermal energy can break some of the Si-Si bonds as illustrated in Figure 1-15. This creates both a free or mobile electron and a mobile hole (or missing electron). The concentrations of electrons and holes are exactly equal in pure semiconductors and are referred to as the intrinsic carrier concentration

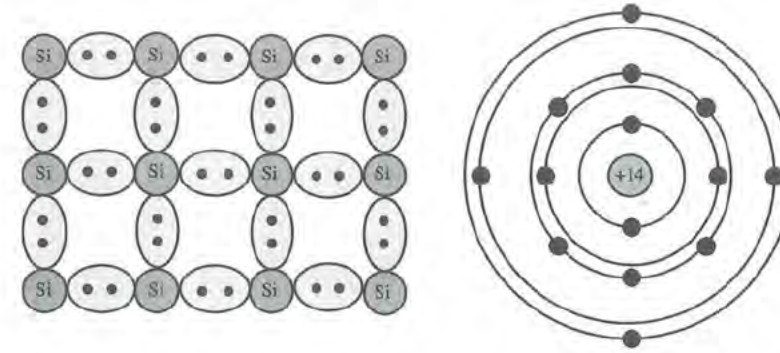


Figure 1-13 Simple representation of silicon atoms bonded in a crystal (left). The dotted areas are covalent or shared electron bonds. The electronic structure of a single Si atom is shown conceptually on the right. The four outermost electrons are the valence electrons that participate in the covalent bonds.

		III ^A	IV ^A	V ^A	VI ^A
		5	6	7	8
		B 10.81	C 12.01	N 14.01	O 16.00
		13	14	15	16
		Al 26.98	Si 28.09	P 30.97	S 32.06
II ^B		30	31	32	33
		Zn 65.39	Ga 69.72	Ge 72.59	As 74.92
		48	49	50	51
		Cd 112.4	In 114.8	Sn 118.7	Sb 121.8
		80	81	82	83
		Hg 200.6	Tl 204.4	Pb 207.2	Bi 209.0
					84
					Po 209

Figure 1-14 Portion of the periodic table relevant to semiconductor materials and doping. Elemental semiconductors are in column IV. Compound semiconductors are combinations of elements from columns III and V (or II and VI).

n_i . Figure 1-16 plots this concentration as a function of temperature. At room temperature, n_i has a value of about $1.4 \times 10^{10} \text{ cm}^{-3}$ in silicon. Since there are about $5 \times 10^{22} \text{ cm}^{-3}$ atoms in silicon, fewer than 1 in 10^{12} bonds are broken at room temperature. As a result, pure silicon is a very poor conductor. The mobile carriers carry electrical currents in devices. With so few of these present in pure silicon, the currents would be far too small to be useful in devices.

Fortunately, semiconductors have the property that they can be doped. This process is illustrated in Figure 1-17. Doping could be accomplished by the gas phase diffusion process we discussed in connection with Figures 1-7 and 1-8. However, today it is generally accomplished by a process called ion implantation that we will discuss in detail in

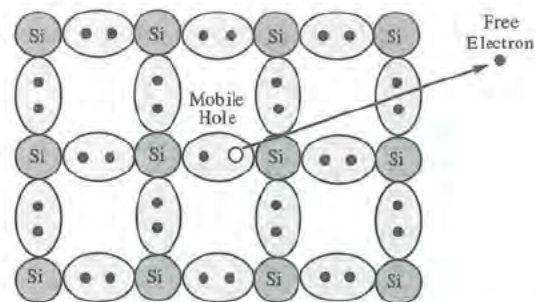


Figure 1-15 Electron (-) and hole (+) pair generation represented by a broken bond in the crystal. Both carriers are mobile and can carry currents in devices.

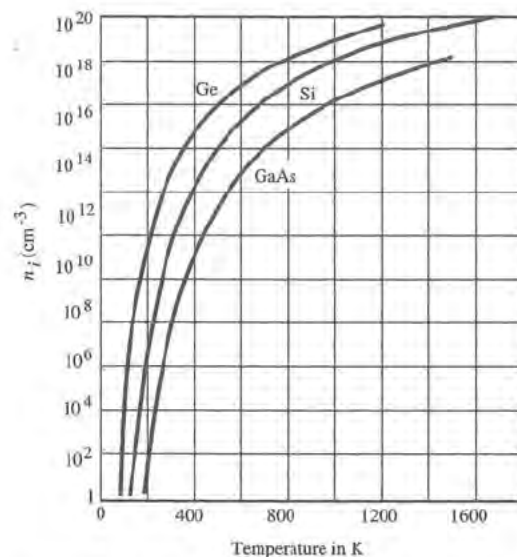


Figure 1-16 Intrinsic carrier concentration versus temperature in common semiconductors. After [1.9].

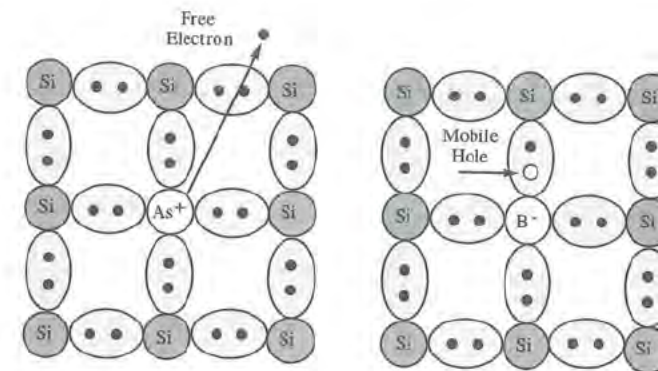


Figure 1-17 Doping of group IV semiconductors using elements from column V (arsenic) or III (boron) of the periodic table.

Chapter 8. Doping results in a column V (As, for example) or a column III (B, for example) atom replacing a silicon atom in the crystal structure. Such dopants either contribute an extra electron (column V) to the crystal and become N-type dopants, or they contribute a hole (column III) and become P-type dopants. The electrons or holes are introduced on a one for one basis by the dopants. As a result, to the extent that we can control the doping concentration accurately, we can precisely control the free electron and hole concentrations and therefore the conductivity of the silicon. N_D and N_A are used to refer to the N-type (donor) and P-type (acceptor) concentrations, respectively. In semiconductor devices, this doping is done locally as illustrated in Figure 1-11 using photolithography and masking.

Modern integrated circuit technology generally uses ion implantation to dope the semiconductor. This permits the controlled introduction of parts per million to parts per hundred of dopant atoms. As a result, the conductivity of silicon can be controlled over a very wide range, permitting many types of semiconductor devices to be fabricated. Figure 1-18 illustrates the range of conductivity it is possible to achieve in silicon by doping. The terms N^+ , N^- , P^+ , and so on are often used in semiconductor devices to describe relative levels of doping in a particular region. While there are no formal definitions of the ranges of doping each of these terms represents, the following approximate definitions are often used:

$$\begin{aligned} N^{--} \text{ or } P^{--} &: N_D \text{ or } N_A < 10^{14} \text{ cm}^{-3} \\ N^- \text{ or } P^- &: 10^{14} \text{ cm}^{-3} < N_D \text{ or } N_A < 10^{16} \text{ cm}^{-3} \\ N \text{ or } P &: 10^{16} \text{ cm}^{-3} < N_D \text{ or } N_A < 10^{18} \text{ cm}^{-3} \\ N^+ \text{ or } P^+ &: 10^{18} \text{ cm}^{-3} < N_D \text{ or } N_A < 10^{20} \text{ cm}^{-3} \\ N^{++} \text{ or } P^{++} &: N_D \text{ or } N_A > 10^{20} \text{ cm}^{-3} \end{aligned}$$

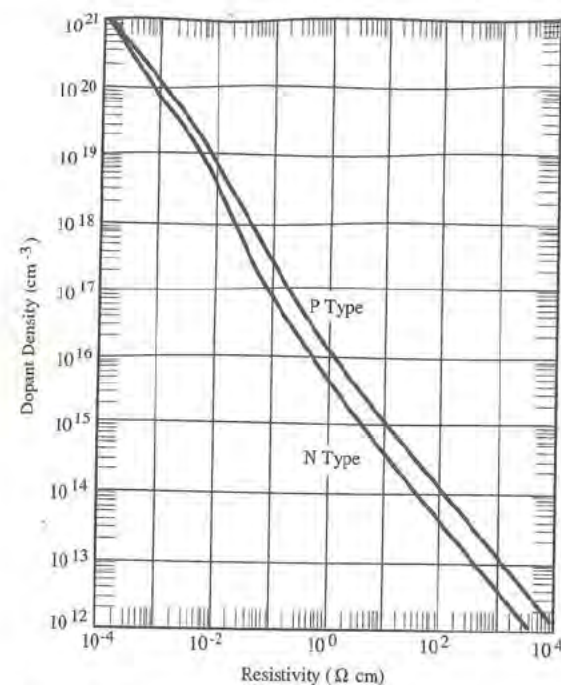


Figure 1-18 Resistivity versus doping for N- and P-type silicon. After [1.10].

Since the lattice density in silicon is $5 \times 10^{22} \text{ cm}^{-3}$, even the heaviest doped regions, N^{++} and P^{++} , normally use doping concentrations less than 1%.

The resistivity plotted in Figure 1-18 is inversely proportional to the carrier concentrations and is defined as

$$\rho = \frac{1}{q\mu_n n + q\mu_p p} \quad (1.1)$$

where n and p are the electron and hole concentrations and μ_n and μ_p are the electron and hole mobilities, respectively. μ_n and μ_p vary with temperature, doping, and electric field in semiconductors. In silicon at room temperature, with low doping concentration and small fields, maximum values of $\mu_n \approx 1500 \text{ cm}^2 \text{ volt}^{-1} \text{ sec}^{-1}$ and $\mu_p \approx 500 \text{ cm}^2 \text{ volt}^{-1} \text{ sec}^{-1}$ are observed. n and p are determined by the dopant concentrations N_D and N_A through relationships we will describe shortly. In general only one of the two terms in Eq. (1.1) is significant since $n \gg p$ in N-doped material and $p \gg n$ in P-doped material.

The electrical properties of semiconductors are often described through the use of two types of models. The first of these is the "bond" model that we used beginning with Figure 1-13. The second model is the "band" model. Both of these approaches model the

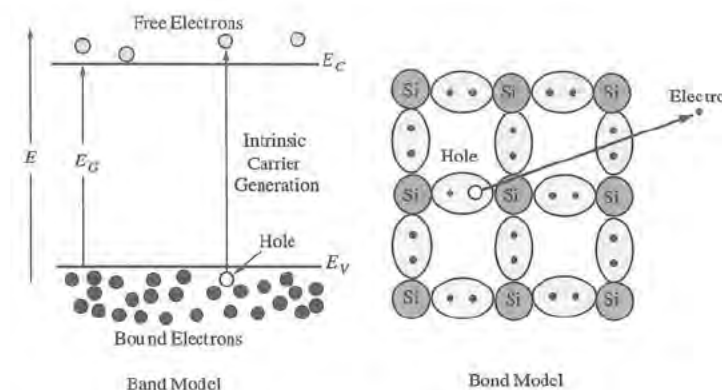


Figure 1-19 Simple band and bond diagram representations of pure silicon. Bonded electrons lie at energy levels below E_V ; free electrons are above E_C . The process of intrinsic carrier generation is illustrated in each model.

same physical phenomena. They are simply different ways of explaining the same underlying mechanisms. We will make use of both approaches in this book because some concepts are easier to understand with the bond model, others with the band model.

Figure 1-19 illustrates both models for the case of pure silicon. In the band model representation, the vertical direction is the potential energy of electrons in the crystal. From the bond model, we can see that in a perfect crystal, electrons are either bound to host silicon atoms or they are free. A finite amount of energy is required to free the electron (that is, to break a Si-Si bond). In the band diagram, we represent this with the bound electron energy levels below the valence band (E_V) and the free electron energy levels above the conduction band (E_C). The energy gap between E_V and E_C represents the energy needed to free the electron. In a perfect crystal, there are no allowed energy levels between E_V and E_C (an electron is either free or it is bound). Thus the breaking of a Si-Si bond, which creates one free electron and one free hole, is illustrated in the band model by elevating an electron up to the conduction band.

In Figure 1-20, the introduction of dopants is represented in the bond model and in the band diagram by the E_D and E_A energy levels. Consider the E_D level first, which represents a donor atom (As for example). When an As atom is introduced into the crystal, it brings with it an extra electron in its outermost shell, beyond the four needed for covalent bonding in the silicon lattice. This electron will be bound weakly to the As atom because the As atom also brings along with it a positive charge in its nucleus (proton) to balance each electron. The binding energy of this fifth valence electron is usually rather small since it is not needed for the four covalent bonds the As atom forms in the silicon lattice. As a result, the fifth electron can be quite easily freed. We represent this in the band diagram with an energy level E_D that is close to E_C . The small difference in energy between E_C and E_D represents the energy, which must be supplied to free the fifth electron. When $E_C - E_D$ is small, we refer to the dopant as a "shallow" donor. For

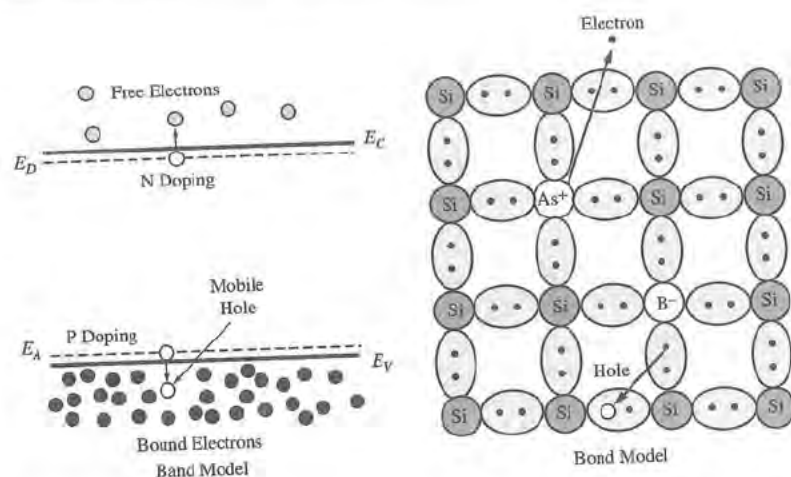
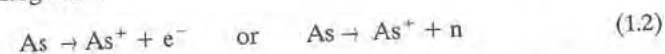


Figure 1-20 Simple band and bond diagram representations of doped silicon. E_A and E_D represent acceptor and donor energy levels respectively. P- and N-type doping are illustrated in each model, using As as the donor and B as the acceptor.

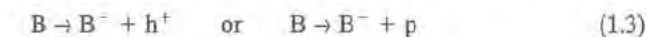
such donors, except at very low temperatures, this energy is always available due to the thermal energy of the crystal and so each donor provides one free electron in the crystal. Note that once the As atom has donated its extra electron to the crystal, the As will have a net positive charge As^+ .



P-type dopants are represented in an analogous manner as also shown in Figure 1-20 for both the bond and band models. E_A represents the energy level introduced into the crystal by the P-type dopant (B, for example). Such dopants have one fewer valence electrons per atom than the host crystal. The hole or missing electron is weakly bound to the B atom. An electron from a neighboring Si atom can exchange positions with the hole, effectively allowing the hole to move through the crystal as a positively charged carrier. The energy difference between E_V and E_A represents this binding energy of the hole to the B atom and is small in the case of a "shallow" acceptor. As is the case with N-type dopants, thermal energy supplies enough energy to free the hole at all but the very lowest temperatures.

An analogy to bubbles in a liquid is often used to describe hole behavior in the band model. Bubbles tend to rise to the surface. Thus energy needs to be provided to these bubbles or holes to push them down from the E_A level into the valence band where they become mobile positively charged carriers.

Once the boron gives up its hole to the crystal, it has a net negative charge B^- .



We can quantify these ideas a little more carefully. The intrinsic carrier concentration in silicon is given by [1.8]

$$n_i = 3.1 \times 10^{16} T^{3/2} \exp\left(-\frac{0.603 \text{ eV}}{kT}\right) \text{ cm}^{-3} \quad (1.4)$$

This equation is an empirical fit to experimental data. It describes the process by which an electron can be excited from the valence band (bound states) to the conduction band (free states). As we saw in Figure 1-16, the number of such carriers in materials like silicon is relatively small at room temperature. Since the process responsible for generating n_i is bond breaking, we would expect that semiconductors that have weaker bonds also have higher n_i values. This is indeed the case. Weaker bonds are represented in the band model by smaller bandgaps. Table 1-2 summarizes these key numbers for three typical semiconductors. n_i depends exponentially on the bandgap with wider bandgap materials having smaller values of n_i .

Table 1-2 Bandgap and intrinsic carrier concentration of three common semiconductors.

Material	E_G at Room Temp	n_i at Room Temp
Germanium	0.66 eV	$2.5 \times 10^{13} \text{ cm}^{-3}$
Silicon	1.12 eV	$1.45 \times 10^{10} \text{ cm}^{-3}$
GaAs	1.42 eV	$2 \times 10^6 \text{ cm}^{-3}$

In an undoped material, it is also the case that $n = p$, so it must be true that

$$np = n_i^2 \quad (1.5)$$

Although we have not proven it, Eq. (1.5) holds in doped as well as in undoped semiconductors. This is often called the law of mass action and is an important relationship in describing carrier concentrations in semiconductors. An important implication of this equation is that if we increase the electron population in the crystal by introducing N-type dopants, n will increase and p will decrease. Physically, the reason for this is that the processes of generation of hole-electron pairs by Si-Si bond breaking and the reverse recombination process are constantly occurring in a crystal to produce an equilibrium number of carriers. If we introduce more electrons by doping, we will increase the probability of electrons and holes encountering each other in the crystal and hence the probability of recombination. An increase in n will therefore drive the hole population down as described by Eq. (1.5).

At very high doping concentrations, the n_i value used in Eq. (1.5) must be changed to account for what are called "heavy doping effects." Fundamentally such effects occur

because the simple picture of a Si crystal with widely separated doping atoms is not a correct picture at high doping concentrations. At n or p values greater than about 10^{19} cm^{-3} , the doping atoms are on average close enough to each other that they begin to interact and to affect the Si crystal structure. (See problem 1.2.) Generally this results in a reduction in the Si bandgap and an increase in n_i .

While shallow donors or acceptors are always ionized at room temperature (that is they contribute free electrons or holes to the crystal), this is not always the case at low temperatures. Also, for atoms that introduce energy levels deeper in the bandgap, full ionization often does not occur even at room temperature because thermal energy is not large enough to excite the extra electron or hole up to E_C (or down to E_V).

Figure 1-21 illustrates the general behavior of the free carrier concentration versus temperature, for a specific case of a shallow donor in silicon. At temperatures below about 100K, freezeout of the As donor is observed and the free electron concentration drops. At temperatures above about 600K, n_i begins to dominate the doping and n increases. We refer to this n_i dominated regime as the intrinsic region. For lower temperatures, the material is extrinsic or dopant dominated. Devices are almost always operated in the temperature regime in which the material is extrinsic, that is, the doping controls the behavior. In the processing world, however, much higher temperatures are involved and the material is often intrinsic. This difference will have significant impact throughout this book. Note also that the value of n_i that is appropriate for device operation is the room temperature value of $1.45 \times 10^{10} \text{ cm}^{-3}$ in silicon, whereas much higher values ($10^{16} - 10^{19} \text{ cm}^{-3}$) are appropriate in the processing world.

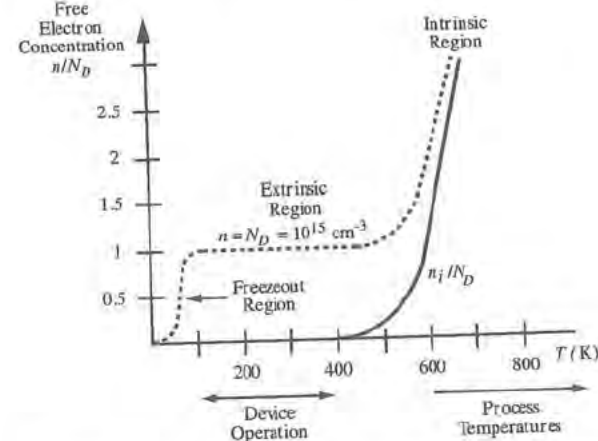


Figure 1-21 Behavior of free carrier concentration versus temperature. Arsenic in silicon is qualitatively illustrated as a specific example ($N_D = 10^{15} \text{ cm}^{-3}$). Note that at high temperatures n_i becomes larger than the 10^{15} doping and $n \approx n_i$. Devices are normally operated where $n = N_D$. Fabrication occurs at much higher temperatures where n is often $\approx n_i$.

The probability of an electron occupying any particular energy level E in a system of discrete energy levels, when the electrons are subject to the Pauli exclusion principle is given by the Fermi-Dirac probability function:

$$F(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{kT}\right)} \quad (1.6)$$

This is the appropriate expression for electrons and holes in a semiconductor crystal. The quantity E_F in this equation is called the Fermi level and is defined by this equation to be simply the energy level at which the probability of finding an electron is exactly 0.5. $1 - F(E)$ is the probability of not finding an electron at energy E , or in other words, the probability of finding a hole there.

Figure 1-22 illustrates the concept of the Fermi level. The left-hand case represents an undoped semiconductor. In that case, $n = p$ and typically there will be a very small number of electrons in the conduction band and holes in the valence band. Thus the probability of a valence band energy level being occupied by an electron is essentially 1; the probability of finding an electron in the conduction band is essentially 0. Therefore, E_F must be located somewhere in the forbidden band between E_V and E_C . In fact, it is almost exactly in the middle of the bandgap. E_F is usually called the intrinsic Fermi level E_{Fi} or E_i in this case. If we add N-type doping (middle case in Figure 1-22), we know that the number of electrons in the conduction band increases, so the probability of finding an electron above E_C also increases. The Fermi level will move up in the bandgap to reflect this. In fact, a little experience with these sorts of band diagrams will allow the reader to quickly estimate the electron or hole populations in a given situation from a picture, which shows E_F . The higher E_F is in the bandgap, the higher n is.

Note also in Figure 1-22 that we are now showing (schematically) only the holes in the valence band rather than all the bound electrons as we have in earlier figures. The holes are the mobile charge carriers in the valence band and they are what we care about when we describe device operation or process physics issues. From this point forward we will show only the mobile carriers in these diagrams—holes in the valence band and electrons in the conduction band.

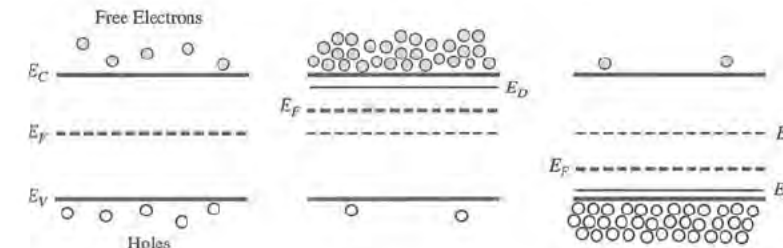


Figure 1-22 Fermi level position in an undoped (left), N-type (center), and P-type (right) semiconductor. The dots represent free electrons; the open circles represent mobile holes.

P-type doping is analogous (right side of Figure 1–22). Introducing P-type dopants increases the hole population in the valence band and decreases the free electron population in the conduction band. E_F moves down in the bandgap to reflect the lower free electron population. The closer E_F is to E_V , the higher the hole population is. Thus, while E_F is a well-defined mathematical quantity, perhaps its greatest use in semiconductors is in visualizing the electronic properties (carrier concentrations and type) through the band diagram concept.

To actually calculate n and p in a given situation, we need to know not only the probability of finding them at an energy level E [Eq. (1.6)], but also the number and position of allowed energy levels. Quantum mechanics tells us how the electrons fill the energy levels in an atom or in a system of atoms like a crystal. In such a system the energy levels are not continuous (no two electrons can have exactly the same set of quantum numbers). From these concepts comes an approximate expression for the allowed electron energy levels at an energy E :

$$N(E) = \frac{4\pi}{h^3} (m_e^*)^{\frac{3}{2}} (E - E_C)^{\frac{1}{2}} \text{ for } E > E_C \quad (1.7)$$

$$N(E) = \frac{4\pi}{h^3} (m_h^*)^{\frac{3}{2}} (E_V - E)^{\frac{1}{2}} \text{ for } E < E_V \quad (1.8)$$

where the first equation describes the allowed energy levels for electrons in the conduction band and the second, the allowed levels for holes in the valence band. In this equation m_e^* and m_h^* are the density of states effective masses of the electron and hole, respectively, in the crystal and are different from the electron rest mass. m_e^* and m_h^* account for the fact that the electrons and holes are located in a crystal rather than in free space. $N(E)$ is of course zero within the bandgap of a pure semiconductor.

Given these expressions for the allowed energy levels and Eq. (1.6) which describes the probability that an allowed level will actually be occupied, we can now determine how many free electrons and holes are actually present in the crystal. Figure 1–23 illustrates graphically how we do this. $F(E)$ is the probability distribution for electrons, which, for an undoped crystal, has a value of 0.5 at the middle of the forbidden band. $N(E)$, the density of allowed states, is plotted from Eqs. (1.7) and (1.8). Note that there are no allowed energy levels in the forbidden band. Above E_C or below E_V , the densities increase as the square root of the energy. The multiple lines in this part of the figure are meant to schematically represent the discrete allowed states that exist at any energy E . For a crystal with an appreciable number of atoms, the discrete energy levels are so close together and so numerous they appear to be a continuous distribution as shown in Eqs. (1.7) and (1.8). The product of the $F(E)$ and $N(E)$ curves on the right represents the electron population at any energy. Since we are interested only in the free electrons, we have shown this population only in the conduction band. In an analogous way, the product of $N(E)$ and $1 - F(E)$ represents the hole population, which we have shown only in the valence band since this is where holes are mobile. Note that most of the free electrons are located fairly close to E_C and most of the free holes close to E_V , because the respective probability functions rapidly fall towards zero further away from

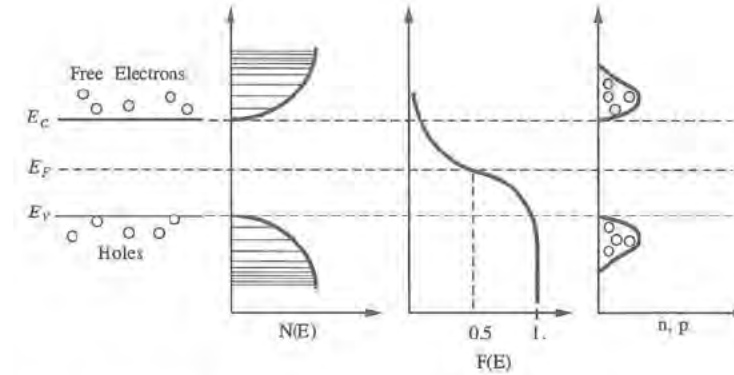


Figure 1–23 Density of allowed states, probability function, and resulting electron and hole populations in a semiconductor crystal.

the band edges. This is in spite of the fact that the number of allowed energy levels increases away from E_C or E_V .

Usually in device physics or in process physics, we are interested in the total number of electrons in the conduction band and the total number of holes in the valence band. These quantities are the areas under the right-hand curves in Figure 1–23 and can be calculated as follows:

$$n = \int_{E_C}^{\infty} F(E)N(E)dE \cong N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad (1.9)$$

$$p = \int_{-\infty}^{E_V} [1 - F(E)]N(E)dE \cong N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \quad (1.10)$$

$$\text{where } N_C = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{\frac{3}{2}}, N_V = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{\frac{3}{2}} \quad (1.11)$$

N_C and N_V are often called the effective densities of states in the conduction and valence bands. They have values of $2.8 \times 10^{19} \text{ cm}^{-3}$ and $1.04 \times 10^{19} \text{ cm}^{-3}$, respectively, in silicon at room temperature.

In integrating Eqs. (1.9) and (1.10), we have made use of the fact that the Fermi-Dirac probability function in Eq. (1.6) can be approximated by the Boltzmann distribution when the Fermi level is at least a few kT away from the conduction and valence bands and from any other allowed energy levels in the bandgap. When this is true, the $E - E_F$ term in Eq. (1.6) is much larger than kT and the 1 in the denominator may be dropped, resulting in the Boltzmann distribution function. Eqs. (1.9) and (1.10) apply only as long as this is true. When E_F approaches E_C , E_V , or other allowed energy levels

in the bandgap, the full Fermi-Dirac distribution function must be used to describe the electrons populating the various energy levels. This may be required at low temperatures when not all donors or acceptors are ionized and E_F can approach E_D or E_A . Fermi-Dirac statistics are also generally required when doping levels exceed 10^{19} cm^{-3} in silicon, since then E_F moves up into the conduction band or down into the valence band, and allowed energy levels exist near E_F . Such heavily doped semiconductors are often called degenerate and act more like metals than semiconductors. We will in general use the simple results in Eqs. (1.9) and (1.10) in this text. However there will be cases where this is not valid. Generally when this occurs, computer techniques are used to calculate n , p , and E_F . Such programs are widely available.

By combining Eq. (1.5) with Eqs. (1.9) and (1.10), we arrive at the result that

$$np = n_i^2 = N_C N_V \exp\left(-\frac{E_G}{kT}\right) = KT^3 \exp\left(-\frac{E_G}{kT}\right) \quad (1.12)$$

where $E_G = E_C - E_V$. Note the similarity of this result to Eq. (1.4). The exponential behavior of n_i with temperature is also apparent in Figure 1-16.

By combining Eqs. (1.9), (1.10) and (1.12), we obtain the following expressions

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad (1.13)$$

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad (1.14)$$

These forms of the equations are often convenient to use instead of Eqs. (1.9) and (1.10).

Eqs. (1.13) and (1.14) reflect the idea that as E_F moves above E_i , the electron population increases exponentially and the hole population decreases. Similarly, as E_F moves below E_i , n decreases and p increases, both exponentially with $|E_F - E_i|$. Eqs. (1.9) and (1.10) write this equivalently in terms of N_C or N_V . Figure 1-22 illustrates these ideas.

It is sometimes the case in semiconductors that N- and P-type doping concentrations are of the same order. Other situations may arise in which n and p are on the order of n_i . This latter case often occurs at high temperatures (such as those occurring during processing), as illustrated in Figure 1-21. In these cases a more general statement of overall charge neutrality can be used to calculate n and p . Overall charge neutrality requires that

$$N_D^+ + p = N_A^- + n \quad (1.15)$$

[Recall from Eqs. (1.2) and (1.3) that the dopants are charged, for example, As^+ or B^- , after they donate electrons or holes.]

Combining this equation with Eq. (1.5), results in the following general equations for electron and hole concentrations. Unless the donors or acceptors have energy levels deep in the forbidden band, or the temperature is very low, N_D and N_A can generally be replaced in these expressions by N_D and N_A , respectively.

$$n = \frac{1}{2}[(N_D^- - N_A^-) + \sqrt{(N_D^- - N_A^-)^2 + 4n_i^2}] \quad (1.16)$$

$$p = \frac{1}{2}[(N_A^- - N_D^+) + \sqrt{(N_A^- - N_D^+)^2 + 4n_i^2}] \quad (1.17)$$

One additional piece of information is necessary in order to apply these concepts in the broadest sense. The bandgap in semiconductors changes as a function of temperature. Physically this occurs because the lattice constant (spacing between the atoms) changes with temperature and hence the energy needed to break bonds changes with temperature. Generally the lattice constant increases with temperature (the atoms move further apart). This weakens the bonds and the bandgap decreases. For silicon, this effect can be approximated with the following equation [1.9] where T is in $^\circ\text{K}$. This effect is particularly important when we consider semiconductors at processing temperatures.

$$E_G(\text{eV}) = 1.17 - \frac{4.73 \times 10^{-4} T^2}{T + 636} \approx 1.16 - (3 \times 10^{-4})T \quad (1.18)$$

Figure 1-24 illustrates some of these concepts. The gray bands at the top and bottom represent the conduction and valence band edges. At higher temperatures, the bandgap narrows, as described by Eq. (1.18). The series of solid curves plot the position of the Fermi level for a specific doping concentration as temperature changes. Each of these

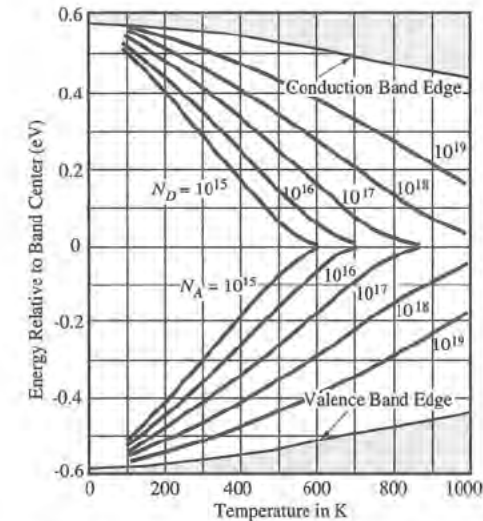


Figure 1-24 Fermi level position in the forbidden band for a given doping level, as a function of temperature. After [1.10].

curves moves closer to E_i as T increases, reflecting the fact that the material becomes more intrinsic at higher temperatures. Normally in studying semiconductor device physics, we are concerned with temperatures around 300K. In the processing world, however, temperatures are usually much higher. This means that in modeling process phenomena, we will use values of n_i which are much larger, values for E_g which are significantly smaller, and analyze materials which are more often intrinsic. In later chapters we will see the implications of these observations.

Example A sample of silicon which is doped with both N- and P-type dopants is shown in Figure 1-25. Calculate n , p , and E_F for (a) room temperature—a typical device operating condition; and (b) 1000°C—a typical processing temperature, which might be used during device fabrication.

Answer In general, in problems of this type, there are three unknowns, n , p , and E_F , assuming the dopants are fully ionized and $N_D = N_D^+$ and $N_A = N_A^-$. There are three independent equations taken from those presented above, which must be solved simultaneously. (There are actually four unknowns including n_i but it may be calculated directly from Eq. (1.4) since the temperature is known.) Generally we know N_D , N_A , and T , as in this problem. The three equations with three unknowns are as follows. Note that Eqs. (1.16) and (1.17) and (1.13) and (1.14) are pairs of equations, which are not independent but are related through Eq. (1.12).

$$\Rightarrow \left\{ \begin{aligned} n &= \frac{1}{2} \left[(N_D^+ - N_A^-) + \sqrt{(N_D^+ - N_A^-)^2 + 4n_i^2} \right] \\ p &= \frac{1}{2} \left[(N_A^- - N_D^+) + \sqrt{(N_A^- - N_D^+)^2 + 4n_i^2} \right] \end{aligned} \right\} \quad (1.16, 1.17)$$

$$\Rightarrow \left\{ \begin{aligned} n &= n_i \exp\left(\frac{E_F - E_i}{kT}\right) \\ p &= n_i \exp\left(\frac{E_i - E_F}{kT}\right) \end{aligned} \right\} \quad (1.13, 1.14)$$

$$\Rightarrow np = n_i^2 \quad (1.12)$$

In virtually every case of interest, physical insight can greatly simplify this situation, usually to the point where simultaneous equations do not have to be solved. In those cases in which simplifications cannot be made, there are computer programs available which efficiently calculate a solution. In solving this example, we will take the approach of first using physical insight to simplify the problem.

a. Room Temperature Solution At room temperature, it is reasonable to assume that all of the dopants are ionized. Both As ($E_D = 0.049$ eV) and B ($E_A = 0.045$ eV) are shal-

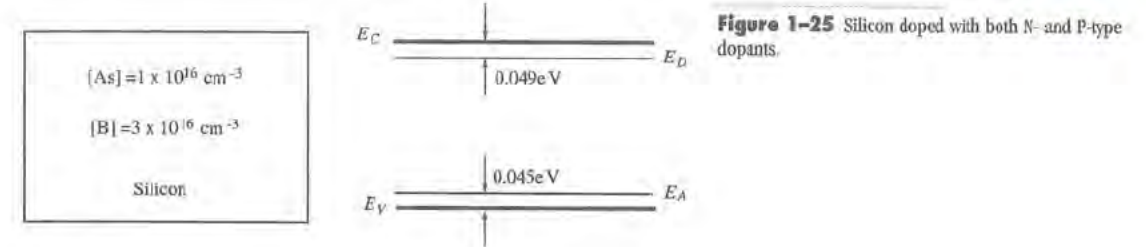


Figure 1-25 Silicon doped with both N- and P-type dopants.

low dopants. Furthermore, the net doping, $N_A - N_D$ is much greater than n_i so we have directly from Eqs. (1.16) and (1.12) that

$$p = N_A - N_D = 2 \times 10^{16} \text{ cm}^{-3}$$

$$n = \frac{n_i^2}{p} = \frac{2.1 \times 10^{20}}{2 \times 10^{16}} = 1.05 \times 10^4 \text{ cm}^{-3}$$

The Fermi level position can now be calculated directly from Eq. (1.14):

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right)$$

$$\therefore E_i - E_F = kT \ln \frac{p}{n_i} = (8.62 \times 10^{-5} \text{ eV K}^{-1})(298 \text{ K}) \ln \frac{2 \times 10^{16}}{1.45 \times 10^{10}} = 0.36 \text{ eV}$$

Another observation at this point may be helpful. To first order, it is the case that energy levels above E_F are not occupied by electrons, while levels below E_F are occupied. (Recall that E_F by definition is the level at which the probability of finding an electron is exactly 0.5. The probability drops rapidly toward 0 above E_F and approaches 1 below E_F as shown in Figure 1-23.) The result in Figure 1-26(a) shows E_F to be well below E_D which means that the donor levels are not occupied or have donated their electrons. E_F is also well above E_A which means that the acceptor levels are occupied or have accepted electrons. This is consistent with our assumption of complete ionization.

b. High Temperature Solution At 1000°C, we can calculate directly from Eq. (1.4) that $n_i = 7.14 \times 10^{18} \text{ cm}^{-3}$. This is much larger than the donor or acceptor concentrations so that the material will be intrinsic. Therefore, $n = p = n_i$ and the Fermi level will be essentially in the middle of the bandgap. Note that E_g is significantly smaller at this temperature than it is at room temperature [0.79 eV from the second expression in Eq. (1.18)]. This solution is shown in Figure 1-26(b). The solution is consistent with all the donors and acceptors being ionized since E_F is well below E_D and well above E_A .

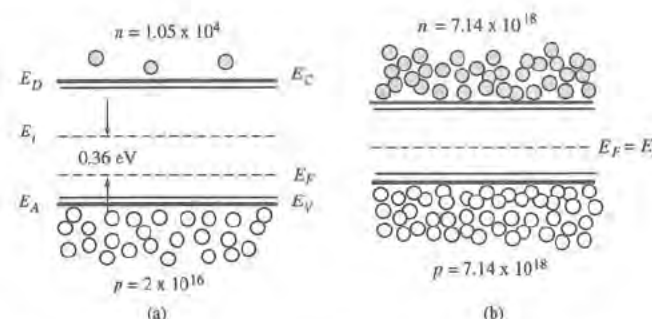


Figure 1-26 Solutions for the example at (a) room temperature, and (b) 1000°C.

The concepts we have been using to describe semiconductors and their band structure are not sufficient to describe many advanced aspects of device performance. For example, in all of our discussion of band diagrams, we have used one-dimensional pictures to represent the material properties. In actual fact, such pictures should be two-dimensional with the vertical axis being electron potential energy (as in our figures) and the horizontal axis being the crystal momentum (k). These more complete representations show that the maximum of the valence band and the minimum of the conduction band in general do not occur at the same value of momentum and are discussed in most device texts [1.11]. This means that in any electron transition from the valence to conduction band we must account for not only the energy change of the electron, but also its momentum change. This becomes important in some aspects of device operation. For example, efficient light emission such as in Light Emitting Diodes (LEDs) is generally only possible in semiconductors in which the conduction band minimum is at the same value of momentum as the valence band maximum. This is referred to as a direct bandgap semiconductor. In this case when an electron falls from the conduction band back to the valence band and recombines there with a hole, the resulting energy can be transferred to a photon (which has essentially no momentum). This occurs in many compound semiconductors such as GaAs.

In silicon, the same recombination process generally occurs through an intermediate energy level in the bandgap (such as a gold atom) which captures the electron and hole necessary for recombination and transfers the excess momentum to a phonon (lattice vibration). This sequence of events conserves both energy and momentum and is referred to as Shockley, Read, Hall or SRH recombination after the scientists who first developed a theoretical description of it.

We will need to understand the indirect recombination process through SRH traps a little more carefully in order to understand the role that impurities like Au, Cu, Fe, and other heavy metals play in silicon technology. These elements act as catalysts for recombination or “traps” in silicon. Consider the deep energy level in Figure 1-27 and

assume that it represents an allowed energy level in the semiconductor bandgap. Elements like Au, Cu, and Fe introduce such levels in silicon as do lattice defects like vacancies and interstitials which we will discuss in more detail in Chapter 3. In indirect bandgap materials like silicon, the direct recombination of an electron and a hole is not a very probable process because of the need to conserve momentum. What is much more likely, and what generally occurs in silicon, is recombination through an intermediate trap level. The trap level first captures either a hole or an electron and then captures the other carrier, completing the recombination event. The excess momentum is given to the crystal through the trapping atom which generally sits on a lattice site.

These steps are illustrated in Figure 1-27 for the specific case of an N-type semiconductor. In N-type material, there are many electrons present, so the likelihood of one of them encountering the trap is much higher than for holes, which are few in number. The typical condition for the trap is therefore to have captured an electron and be waiting for a hole to come into close enough proximity to be captured. It is thus the minority carrier capture process that limits the overall recombination rate and we usually speak of the minority carrier lifetime as representing the time required for recombination to occur. This bulk lifetime to first order is given by

$$\tau_R = \frac{1}{\sigma v_{th} N_t} \quad (1.19)$$

where τ_R is the recombination lifetime, σ is the capture cross section of the trap (on the order of the atom cross sectional area or about 10^{-16} cm^2), v_{th} is the minority carrier thermal velocity (about 10^7 cm sec^{-1}), and N_t is the density of traps per cm^3 . We will use

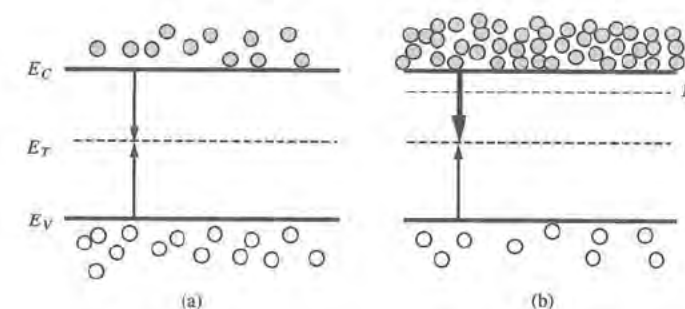


Figure 1-27 Shockley-Read-Hall recombination process. In (a) the trap level captures each carrier separately to accomplish recombination. In (b), the case of an N-type semiconductor is illustrated. The relative sizes of the arrows represent the easy capture of the majority carrier (electron) and the more difficult capture of the minority carrier (hole).

this equation in Chapter 4 to estimate the allowable contamination levels in silicon crystals (Section 4.2).

The band diagram on the right of Figure 1-27 illustrates another way of visualizing the recombination process. In N-type material, E_F will be in the upper half of the bandgap, above the trap level. Energy levels below E_F are normally occupied by electrons since the probability of finding an electron at a level below E_F is close to one. Thus we again conclude that the trap level will normally be occupied by an electron and the hole (minority carrier) capture will be the rate-limiting step. A similar argument in P-type material would show that the capture of the minority carrier electron is the rate-limiting step. This process of recombination is important in many types of devices, but especially in devices that depend on the behavior of minority carriers. Bipolar transistors are one example.

The inverse process—generation—through SRH traps is also possible and is often the dominant carrier generation process in indirect bandgap semiconductors like silicon. In this process, the trap level first generates an electron or hole and then generates the other species. These events can then be repeated to supply carriers to the semiconductor when they are needed. In this case, the material is characterized by a generation rather than a recombination lifetime. Depletion regions are a very common example in which such generation takes place. In general, the two lifetimes τ_R (recombination) and τ_G (generation) are not the same since the processes involved in each are somewhat different (capture of a free carrier versus emission of a carrier from a trap). In silicon τ_G is normally longer than τ_R . Eq. (1.19) provides a simple estimate of carrier lifetime and we will often not have to distinguish between τ_G and τ_R in this text.

SRH theory gives the rate of recombination (or generation) as

$$U = \frac{np - n_i^2}{\tau \left(p + n + 2n_i \cosh\left(\frac{E_T - E_i}{kT}\right) \right)} \quad (1.20)$$

where τ is the carrier lifetime. In equilibrium $np = n_i^2$ and U goes to zero. In this case, recombination and generation exactly balance each other. When excess carriers are present, $np > n_i^2$ and U is positive (net recombination). When $np < n_i^2$, U is negative (net generation). Note that the rate of recombination (or generation) is maximized when the trap level is near the middle of the bandgap.

At the surface of the semiconductor, the crystal structure terminates, usually at a thermally grown SiO_2 layer. Traps at such an interface can also act as SRH centers and can both generate and recombine carriers. The surface equivalent of Eq. (1.19) is

$$s = \sigma_s v_{th} N_{it} \quad (1.21)$$

where σ_s is the capture cross section of the surface trap (cm^2) and N_{it} is the density of traps per cm^2 at the surface. s has units of cm sec^{-1} and is called the surface recombination velocity. s is thus related to the rate at which carriers recombine at the surface. $1/\tau$ is the equivalent bulk recombination rate. When surface effects dominate, an analogous expression to Eq. (1.20) gives the net rate of recombination or generation.

$$U = \frac{s(np - n_i^2)}{p + n + 2n_i \cosh\left(\frac{E_T - E_i}{kT}\right)} \quad (1.22)$$

A final point is important to make. In direct bandgap materials like GaAs in which carriers can recombine without an intermediate trap level, minority carrier lifetimes are generally very short (on the order of nsec). In addition, these lifetimes are fundamental properties of the material and are not generally available to the process or device designer as design variables. In indirect bandgap materials like Si, the minority carrier lifetime is inversely proportional to the density of trapping centers. There is thus the possibility of designing a particular lifetime by controlling the concentration of these traps. In silicon technology, this generally means controlling the concentrations of impurities like Au, Cu, and Fe. Lattice defects created by ion implantation or electron beam radiation can also act as SRH centers. In many cases, it is simply important to minimize SRH centers, that is, to maximize the carrier lifetime. In other cases, it may be desirable to reduce the lifetime by the careful introduction of controlled concentrations of deep level traps.

1.4 Semiconductor Devices

This is, of course, a book on semiconductor technology and not semiconductor devices. However, a first-order understanding of how the basic devices used in ICs operate is useful, because it provides some understanding of the objectives we have for the technology. In this section we review the basic operation of PN junctions and MOS and bipolar transistors. Many excellent texts on these topics exist and the interested reader is referred to them for a more detailed description [1.11].

1.4.1 PN Diodes

The simplest semiconductor device is the PN junction, illustrated in Figure 1-28. Conceptually, simply bringing two pieces of semiconductor material, one N-type and the other P-type, into contact forms the device. In practice, such junctions are formed by starting with a wafer of one type and then doping the surface layer the opposite type. This results in a doping profile as shown on the right side of Figure 1-28.

With no applied bias, a PN junction will form a depletion region or a region in which there are no mobile carriers across the junction. It is easy to see why this happens by considering the fact that across the junction there are very high gradients of holes and electrons. If the doping on the N side is 10^{15} cm^{-3} , then the N side contains 10^{15} cm^{-3} electrons and 10^5 cm^{-3} holes. The P side contains carrier concentrations approximately the opposite of these. Across the junction then, an enormous gradient in each carrier type exists. If something did not happen, these enormous gradients would result in large currents flowing across the junction because the carriers would diffuse down the gradients. However, something does happen.

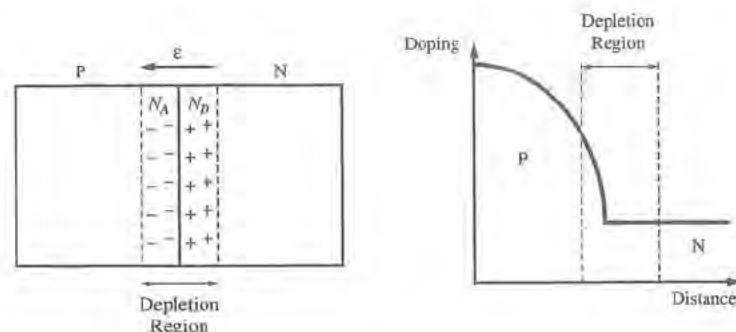


Figure 1-28 Schematic representation of a PN diode on the left. Typical doping profile on the right.

The free carriers in the vicinity of the junction do initially flow to the opposite side. This creates the depletion region and sets up an electric field. This field arises simply because once the free carriers move away, they leave behind the charged donor and acceptor ions N_D^+ and N_A^- bonded in substitutional sites in the silicon lattice. The dopant ions and the free carriers normally balance each other, resulting in a charge neutral material. Without the carriers, the field must exist as illustrated in Figure 1-28.

In the PN junction in equilibrium, drift and diffusion exactly balance, so that no net current flows. This actually results in rather large fields in the depletion region in order to “hold back” the large carrier gradients. Typical fields are on the order of a few volts per micron.

If we apply a positive voltage to the N side of the junction, the electric field across the junction will increase. This polarity adds to the built-in field that is already there. This added field creates additional positive charge on the N side and negative charge on the P side, widening the depletion region in Figure 1-28. We should not expect much current to flow under this bias condition since the internal field by itself is enough to stop carrier diffusion and we have simply made this field larger with this polarity of applied voltage. In fact, only very small leakage currents flow in a PN junction under such reverse bias.

If we now consider making the P side positive, exactly the opposite happens. The applied field partially cancels the built-in field, allowing majority carriers from both sides to diffuse across the junction. Under forward bias in a PN junction, large currents flow since the internal field holding back the diffusion of carriers is removed by this polarity of applied bias. The overall current-voltage relationship for a PN junction is expressed simply as

$$I = I_0 \left(\exp \frac{qV}{kT} - 1 \right) \quad (1.23)$$

where I_0 is the reverse leakage current. For positive voltages on the P side (forward bias), the current increases exponentially with applied bias. For reverse voltages, the current saturates at I_0 which is very small. Typically currents in the forward direction are on the order of mA; in the reverse direction, they are pA or smaller.

The built-in electric field across the junction produces a built-in voltage across the junction given by

$$\phi_i = \frac{kT}{q} \ln \left(\frac{N_D N_A}{n_i^2} \right) \quad (1.24)$$

For typical doping levels in silicon diodes this built-in voltage is about 0.7 eV, so an external forward bias on the order of 0.7 volts will therefore essentially eliminate this barrier, allowing large currents to flow. This is often referred to as the “turn-on” voltage of a silicon diode, although it is clear from Eq. (1.23) that there is no specific voltage at which the current turns on.

It is also easy to see physically why the I - V relationship should be exponential. Recall from Eq. (1.6) that the population of electrons falls off exponentially above the conduction band minimum. As forward bias is applied to the diode and the band bending is decreased, an exponentially increasing number of electrons are able to surmount the barrier and therefore I increases exponentially with voltage. The PN diode then really acts like a two-terminal switch. It is turned on (conducts) under forward bias and is turned off (blocks) under reverse bias.

One final property of PN diodes will be useful to us in later discussions. The depletion region associated with the diode may be thought of as a parallel plate capacitor. The insulating dielectric in the capacitor is of course the depleted silicon; the plates of the capacitor are formed by the conducting charge-neutral P and N regions adjacent to the depletion region. The capacitance of such a structure is given simply by

$$\frac{C}{A} = \frac{\epsilon_s}{x_d} = \left[\sqrt{\frac{q\epsilon_s}{2} \left(\frac{N_A N_D}{N_A + N_D} \right)} \frac{1}{(\phi_i \pm V)} \right] \quad (1.25)$$

Here ϵ_s is the permittivity of silicon and x_d is the width of the depletion region. ϕ_i and the applied voltage V add under reverse bias and subtract under forward bias. The capacitance decreases as reverse bias is applied to the junction because the depletion layer widens; it increases under forward bias as the width decreases. Eq. (1.25) is valid for a junction with uniform doping on the two sides of the junction and only as long as the forward current in the junction is small. Quite often junctions in semiconductor devices are asymmetrically doped (as in Figure 1-28). In this case only the doping on the more lightly doped side matters. Real diodes may also have both N_D and N_A as functions of position. In this case all of the above equations become more complex and generally computer techniques must be used to calculate ϕ_i , $C(V)$, and $I(V)$. Even in these more complex cases, however, the simple equations above usually provide good estimates.

1.4.2 MOS Transistors

All practical circuits today require three terminal switching devices as shown schematically on the left side of Figure 1-29. Circuits are much easier to design if the control terminal is electrically isolated from the output. In digital circuits, the simplest representation of the switching element is simply a switch which is closed and opened by an isolated control terminal. The two states, closed and open, represent digital 0 and 1 states. In analog circuits generally an input signal needs to be amplified. Again in simplified form, this is accomplished by a controlled current source whose current is proportional to an applied input or control signal.

There are two dominant types of such devices used in silicon ICs today, the MOS transistor and the bipolar junction or BJT device. Both of them operate to first order like the simple equivalent circuits in Figure 1-29. Historically, as we saw in Section 1.2, the MOS device was invented first, although it was not used in large-scale production until much later than the BJT because of problems in controlling its properties and in producing stable devices. The basic MOS device is shown in Figure 1-30, in this case an NMOS or N-type device. The device is a three-terminal structure with an input (source), output (drain), and a control terminal (gate).

The N^+ regions at the source and drain of the device provide a means of contacting the region under the gate, which is the active part of the device. The N^+ regions also provide a source of electrons in normal device operation, as we will see. Consider the central part of the device, under the gate electrode. The structure here is metal on top of an insulator (generally SiO_2) on top of silicon and thus provides the name for the device (MOS or Metal Oxide Semiconductor). The silicon is P-type and hence there are very few electrons normally present. Both PN junctions in the device are normally at zero bias or reversed biased, which means that very little current flows across these junctions. Thus with a negative or zero voltage on the gate, there is no connection between the input and output N^+ diffusions and the device operates as an open circuit. This condition is illustrated on the left of Figure 1-30. With negative bias on the gate, majority carrier holes are attracted to the surface, which makes the surface more strongly P-type. With two zero-biased or reverse-biased junctions in series, no current flows.

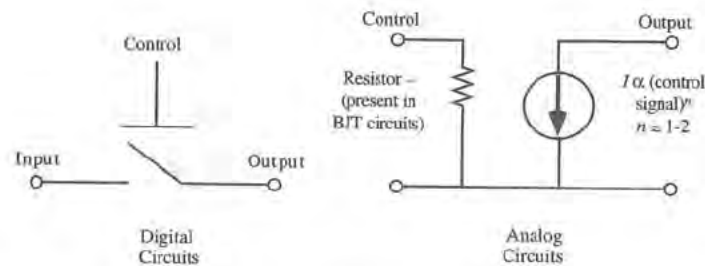


Figure 1-29 Simplified models for a three-terminal device used in digital and analog applications.

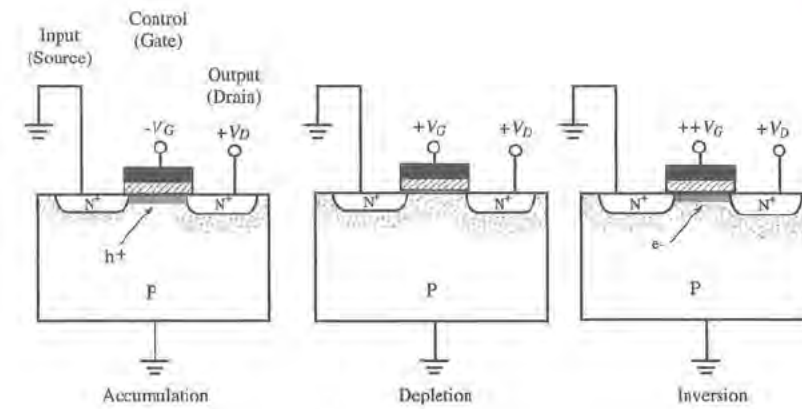


Figure 1-30 Simplified cross section of an MOS transistor biased in its OFF (accumulation), intermediate (depletion), and ON (inversion) states. The dotted areas are the depletion regions.

If we now apply a positive voltage to the gate electrode, the vertical electric field produced across the insulating gate oxide will attract electrons to the silicon surface and repel positively charged holes. Initially, this creates a depletion region in the central part of the device as shown in the middle of Figure 1-30. If a sufficiently positive voltage is applied to the gate, the electrons can actually become the dominant carrier in a narrow layer at the surface, effectively making the surface N-type as shown on the right of Figure 1-30. This process is called surface inversion and results in an inversion layer of electrons about 5 nm thick at the silicon surface. With this inversion layer present, the input and output N^+ diffusions are now effectively connected, resulting in a closed switch. The gate electrode thus allows the device to be turned on ($+V_G$) or turned off ($-$ or zero V_G). When the MOS transistor is turned on, the number of electrons in the inversion layer is proportional to the gate voltage. Thus the resistance of the inversion layer depends on the gate voltage and therefore the current that flows in the device depends on the gate voltage. The result is an equivalent circuit like that shown on the right of Figure 1-29, which allows the device to be used in analog as well as digital applications.

The process of turning the MOS device on or off would actually be very slow if it were not for the presence of the N^+ regions. Because the substrate is P-type, there are very few electrons present there and the only way for the substrate to provide the electrons needed for the inversion layer would be by actually generating them through breaking Si-Si bonds or through the Shockley-Read-Hall generation process described earlier. The only way for the substrate to get rid of the electrons in the inversion layer when the device is turned off is by electron hole recombination. These processes would limit the turn on and turn off times to msec or seconds, depending on the rates of generation and recombination and would certainly not allow the nsec or psec switching times that are common in today's ICs. The solution to this problem resides in the N^+

regions. These regions are rich in electrons and can provide a source or a sink for these carriers in the times required in modern circuits (psec). The inversion layer electrons thus flow out of the source and drain when the device is switched on, and back into these regions when it is switched off.

The depletion regions indicated in Figure 1-30 (the shaded areas) are the regions of the substrate in which there are essentially no mobile carriers. Beneath the gate, the holes in the P-type substrate are driven away by positive gate voltages. In the regions around the source and drain, the electric fields, due to reverse bias on the junctions or due to the built-in junction fields which are present even with zero applied external bias, also drive the mobile carriers away from the immediate vicinity of the junctions.

To fabricate the MOS transistor, we need techniques to dope the silicon N-type (for the NMOS transistor shown here) or P-type for PMOS devices. We will see that ion implantation followed by solid state diffusion provides this capability. We also need the ability to deposit or grow very thin insulating layers to produce the gate dielectric. Thermal growth or deposition of SiO_2 provide this capability and can also provide an Si/SiO₂ interface, which is almost perfect electrically. By this we mean that there are very few charges, electron or hole traps or other undesirable features of the interface that would degrade the operation of MOS devices. We also need to deposit and define thin films of polysilicon (and other materials) to build the gate electrode and thin metal films to provide a wiring capability on the chip. IC technology provides all these capabilities.

While the field effect mechanism to control semiconductor surfaces and convert them from P- to N-type, or vice versa, was known long before the invention of the IC, practical MOS devices were not manufactured until the mid 1960s. The principal reason for this was simply that attempts to build the devices earlier than this usually resulted in structures whose electrical characteristics were unstable or nonreproducible. Many laboratories investigated these problems during the 1960s and the problem was finally traced to charges (defects) present at the Si/SiO₂ interface and to minute (parts per million) concentrations of alkali ions in the MOS gate dielectrics. Charges at the Si/SiO₂ interface are sensitive to fabrication conditions and were gradually brought under control and minimized by better fabrication methods. Ions like Na^+ and K^+ are normally charged and mobile in SiO₂ even at room temperature and can drift around in the gate dielectric as the gate voltage is switched under normal device operating conditions. Charges moving inside the gate dielectric act exactly like a changing gate voltage because they change the vertical electric field in the active part of the MOS device. The result is an MOS device whose I - V characteristics change as the alkali ions drift around.

Many readers may have already drawn the conclusion that since the human body has large concentrations of Na^+ and K^+ , the origin of the contamination problem may well have been handling of the IC wafers during device manufacturing. This is in fact correct, although other sources (chemicals and manufacturing equipment) were also identified. Strict attention to cleanliness in handling and processing the wafers largely solved the MOS stability problem by the end of the 1960s and made possible their large-scale manufacturing. Today's IC plants employ clean rooms with highly controlled environments to prevent a recurrence of these MOS stability problems, but even today, occasional MOS device instabilities arise because of improperly trained technicians or a batch of contaminated chemicals. We will discuss these issues in more detail in Chapter 4.

Charges present in the insulator layer, or at the Si/SiO₂ interface, affect the underlying semiconductor surface properties simply because these charges provide termination points for the electric field lines originating on the gate in MOS devices. If no such charges are present, then applying a voltage on the gate means that the electric field lines must penetrate through the insulator and into the underlying semiconductor where they can then terminate on mobile carriers or on ionized impurity atoms (n , p , N_D^+ , and N_A^-). Thus depletion, inversion, and accumulation are possible. If large numbers of charges are present in the insulator or at the insulator/semiconductor interface, the field lines can terminate before they reach the underlying semiconductor and thus the field effect mechanism does not operate. Nature has provided us with only one superb insulator/semiconductor structure that we know of today, in which these charges are practically zero, the Si/SiO₂ interface. This is one of the main reasons why silicon is the material of choice for today's ICs.

1.4.3 Bipolar Junction Transistors

The other principal type of semiconductor device used in modern integrated circuits is the bipolar transistor, shown in Figure 1-31. In this device, the PN junctions play a more active role in the operation. There are three terminals just as there were in the case of the MOS device. The emitter forms the input, the collector is the output and the base is the control terminal. The simplest equivalent circuits for the device are similar to the MOS transistor (Figure 1-29).

In operation, the emitter is grounded, a small positive voltage is applied to the base, and a larger positive voltage is applied to the collector. The built-in electric field across the base-emitter junction is reduced, allowing the majority carriers on each side to diffuse across the junction. Electrons from the emitter flow into the base and holes from

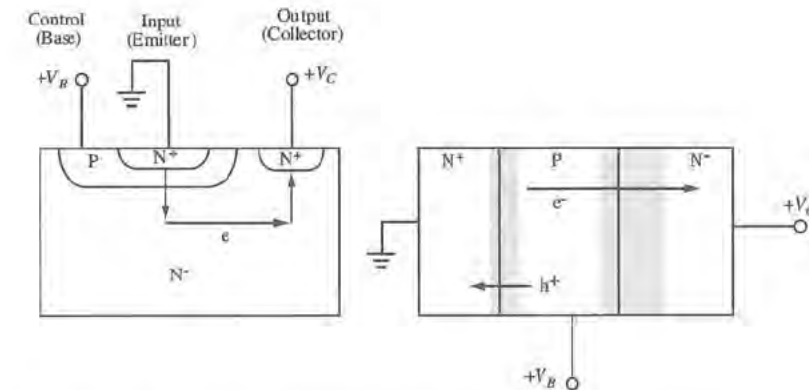


Figure 1-31 Simplified cross section (left) and 1D representation (right) of a bipolar transistor. The shaded areas are the depletion regions. The arrows indicate the path of carrier flow through the device.

the base flow into the emitter. In a well-designed BJT, the emitter doping is several orders of magnitude higher than the base doping, so that the electron population is much higher in the emitter than is the hole population in the base. As a result, when the electric field across the junction is reduced by applying an external V_{BE} , more electrons flow from the emitter into the base than do holes from the base into the emitter. (In other words, the injection of carriers across the junction is asymmetric because the doping is asymmetric.)

If there were no second PN junction present, then the structure would simply be the PN diode we discussed in Figure 1–28. In this case, the injected electrons and holes would diffuse away from the junction on both sides. Since they are minority carriers once they are injected across the junction (the electrons are injected into P-type material and the holes into N-type material), the injected carriers will tend to recombine with the majority carriers in each region. Each time such a recombination event takes place, a majority carrier must be supplied by the contacts to the P and N regions and hence a current flows into the device. From an external circuit point of view, then, we measure a current flowing into the device which is exactly equal to the current (electrons and holes) being injected across the junction. If the carriers injected across the junction do not fully recombine before they reach the ends of the structure, then they quickly recombine at the metal contacts where the rate of recombination is much higher than inside the device. Externally, it makes little difference where the recombination occurs. We simply measure a current going into the device.

The key additional idea that is needed to make the bipolar transistor is the second junction—the base collector junction. This junction is normally reverse biased so that it does not inject carriers itself. However, if carriers from the base emitter junction, particularly electrons from the emitter, diffuse to the collector junction, they will be swept across it by the electric field and “collected” by the collector. In fact, if the base is made narrow enough, virtually all of the electrons injected by the emitter will successfully diffuse through the base without recombining and be collected by the collector. From an external point of view then, we see a large number of electrons flowing in the emitter contact and collect almost all of them at the collector contact. This is the main current path through the device. A small current must be supplied at the control terminal (base) to account for holes that are injected from the base into the emitter and to provide holes for the few recombination events that take place in the base as the electrons flow through there on their way to the collector. The main current path from emitter to collector can easily and quickly be turned on and off by applying or removing V_{BE} . Conceptually, then, the device looks like a switch, which can be turned on and off, just like the MOS device, although the internal physics are quite different in the two cases. Since there is a nonzero input current from the control terminal in the BJT, a better equivalent circuit (Figure 1–29) would include a resistor to allow this current to flow.

Since most of the current in a BJT flows below the silicon surface, the device is much less sensitive to passivation problems than is the MOS transistor. For this reason, the BJT was used in the earliest ICs in the 1960s while researchers were trying to understand the stability problems of the Si/SiO₂ interface. Once these issues were resolved and stable manufacturing of MOS devices became routine, MOS devices became the switching devices of choice in most IC applications. Bipolar devices still find some uses

in ICs today, but more than 90% of the ICs manufactured today rely on the MOS transistor for the basic switching element.

The fabrication of the bipolar device requires essentially the same technologies as in the MOS case. We need methods to locally dope the semiconductor N- and P-type. We need to be able to deposit or grow insulating layers and to deposit and define thin film layers on the semiconductor surface. All of these capabilities are provided in modern IC fabrication processes.

1.5 Semiconductor Technology Families

Active components are of course combined in circuits to implement complete systems on an IC chip. In digital systems, the basic circuit is an inverter, which simply converts a digital 0 to a 1, or vice versa. The implementation of this basic circuit has gone through a number of changes over the past 40 years as IC technology has changed.

In the 1960s, MOS transistors were not manufacturable, for the reasons briefly discussed earlier. As a result, the BJT was used as the basic switching device. Figure 1–32 illustrates the basic technology common in that decade. The NPN structure of the bipolar device is on the left side. The right side shows one implementation of a resistor, which was also needed to construct logic circuits. The basic technology used N⁺ layers grown epitaxially on P-type substrates. Deep P⁺ diffusions were used to provide reverse biased PN junctions which laterally isolated between devices. Six to eight photolithography steps were commonly used. Figure 1–1 (left) illustrates an early example of such an IC.

By the 1970s, MOS technology became manufacturable because the contamination problems described earlier were brought under control. The technology typical of that decade is shown in Figure 1–33. N-type or NMOS devices were generally used because in silicon the electron mobility is about three times higher than the hole mobility,

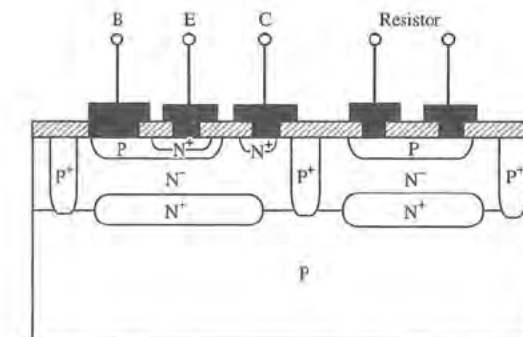


Figure 1–32 Technology typical of the 1960s. Bipolar transistors and resistors were the dominant components.

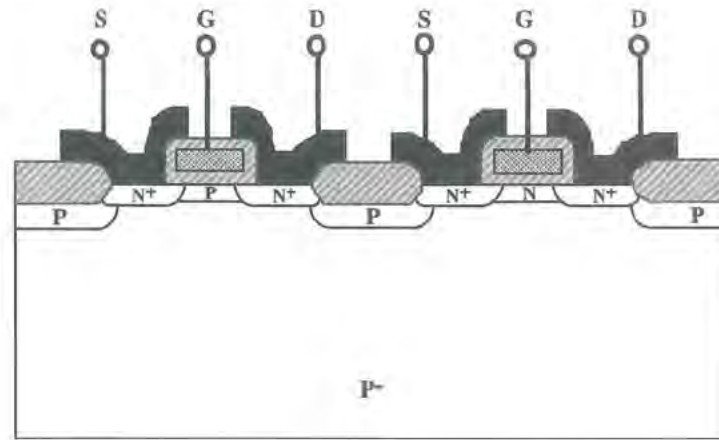


Figure 1-33 Technology typical of the 1970s. Enhancement mode and depletion mode NMOS transistors were the dominant components.

resulting in a 3X advantage in current capability for NMOS versus PMOS devices. The most common technology was Enhancement/Depletion or E/D illustrated in the figure. The device on the left has a P-type channel which requires positive gate voltage to invert or turn on. This is the enhancement device. The device on the right has an N-type channel and a negative threshold voltage. This is the depletion mode device. Although it may not be obvious, a depletion transistor can actually be implemented in a significantly smaller area than a resistor, resulting in denser circuits. NMOS E/D technology was built with about the same number of masks as the 1960s bipolar process.

The 1980s brought a further change in the basic technology. Complementary MOS or CMOS, illustrated in Figure 1-34, became the dominant technology. The basic change was to replace the depletion mode NMOS device in Figure 1-33, with a PMOS transistor. The basic CMOS inverter circuit is also shown in Figure 1-34. This circuit has a major advantage over earlier implementations. For positive input voltages, the NMOS transistor is turned on, pulling the output low. However, the PMOS device is turned off by positive inputs since its threshold voltage is negative. Thus no current flows through the NMOS/PMOS stack. If the input is zero, the PMOS turns on, pulling the output high, but the NMOS is off, so again no current flows in the inverter. Thus the CMOS inverter consumes no current in either state (no DC power). Only when the inverter is actually switching does a transient current flow. Other types of CMOS circuits have the same property. In large chips, with millions of components, this is a significant advantage in terms of overall power consumption and cooling requirements. CMOS technology increased the mask count to about 12–14 masks, depending on the particular technology.

The early 1990s brought a further change with the introduction of BICMOS technology. The structure basically combines bipolar and CMOS devices on the same chip.

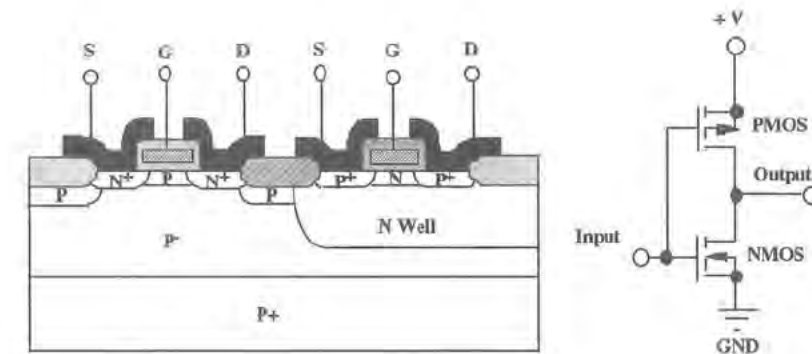


Figure 1-34 Technology typical of the 1980s. CMOS circuits with both NMOS and PMOS devices were dominant.

The penalty for doing this is, of course, additional masks. A BICMOS process typically requires more than 20 masking levels. This technology was used to implement many circuits in the early 1990s. Often much of the internal circuitry on a chip was implemented in CMOS, since this is a denser way to implement functions, with the output devices using bipolar drivers.

Today, technology is driven by the need for high performance, but also by the need for low power. Portable systems obviously need low power, but even in desktop systems, increasing integration levels have made chip and system cooling issues important. The simplest solution to these problems is to simply scale down the operating supply voltage of ICs, since power consumption generally reduces as the square of the supply voltage. These trends are apparent in the NTRS described earlier in this chapter. These trends also put a premium on circuit technologies, which work well at reduced supply voltages. CMOS technology excels at this and for this and other reasons, most workers today believe that it will be the dominant technology well into the next century. This is, in fact, what the NTRS assumes in its projections for the future. We will consider CMOS technology in detail in Chapter 2.

1.6 Modern Scientific Discovery—Experiments, Theory, and Computer Simulation

The history of scientific discovery has until quite recently usually been characterized by laboratory experiments followed by the development of physical models or theory. The invention of the integrated circuit described briefly in Section 1.2, is a good example. The basic contributions of Kilby, Noyce, Hoerni, and many others were largely based on

experiments done in laboratories, and on insight gained from those experiments. In fact, as we shall see throughout this book, the basic physical mechanisms controlling IC technology are in many cases still not fully understood. This lack of full scientific understanding has obviously not prevented widespread industrial application of the technology, however! The invention of the transistor described by Shockley [1.5] is another classic example of experimental science preceding the development of a complete theoretical understanding. In his historical description of those events in the 1940s and 50s, Shockley wrote of the “creative failure methodology” by which he meant that experiments which produce unexpected or undesired results, often lead scientists down new paths to invention. There certainly are examples in science of theoretical predictions preceding and suggesting experiments. However, the reverse has more often been the case.

With the remarkable developments in computer capability in the past decade (due largely to IC technology) has come an opportunity to do experiments in a new way. In fields of science in which theories based on first principles are known, or in which models have been developed, it is quite common today to program those theories or those models in a computer. To the extent that such models are correct, it is then possible to do “computer experiments” or computer simulations. Chemicals can be mixed in a simulated beaker and the resulting reactions observed on a computer screen. Bridges can be constructed in a computer with stresses placed on them to determine failure points and failure mechanisms.

In the integrated circuit world, these tools have become useful adjuncts to real laboratory experiments in recent years. The basic equations, which describe electron and hole transport in semiconductors due to drift and diffusion, have been incorporated into powerful simulators which can today accurately predict the electrical characteristics of proposed device structures. In a similar way, models describing the steps used in fabricating ICs have also been incorporated into process simulators. It is therefore quite possible today to “build” new semiconductor structures and predict their performance using these computer tools.

The state of the art in such simulators today is that they are very useful, but they cannot completely replace real laboratory experiments. This is simply because the models used in the simulators are not complete in some cases, or are purely empirical in other cases. As the models are improved with additional research, the simulators will become more robust and therefore more generally useful. There is great motivation to do this because real laboratory experiments are very expensive and very time consuming, especially as chip technology continues to advance. In an industry as competitive as the semiconductor industry, the time required to get new products to market is a key indicator of a company’s success. To the extent that simulators can decrease this time and the development costs, they can improve competitiveness.

We will make extensive use of process or Technology Computer-Aided Design (TCAD) simulators in this book. This is not only because they are becoming increasingly important as industrial tools, but also because we believe they can greatly enhance the learning process. Our experience with them is that they provide remarkable physical insight and intuition and we will use them throughout the book to illustrate physical principles and to provide accurate illustrations of what real integrated circuit

structures look like. These tools also provide the ability to “look inside” structures or devices and to observe physical phenomena in action that often cannot be easily observed in the laboratory.

1.7 The Plan For This Book

Modern silicon integrated circuit processes are built from a number of what are often called unit process steps. Examples include oxidation, ion implantation, and lithography. Complete process flows are combinations of such steps, often with many hundreds of individual steps. Understanding such a complex fabrication sequence begins of course with understanding the individual steps.

There are generally four things that are required (or are at least highly desirable) in order to apply one of the unit process steps to integrated circuit fabrication:

- Reproducible fabrication techniques
- Measurement methods
- Models of the unit process
- Simulation capability

We will use these four criteria in describing the various unit processes throughout this book. We can take oxidation as a simple example to illustrate this approach. Oxidation simply involves placing a silicon wafer in a high temperature ambient (usually O_2 or H_2O). This results in a chemical reaction, which converts the silicon surface to SiO_2 . To use such a process in building integrated circuits, we must first be able to accurately and reproducibly grow such layers (i.e., fabricate them). This will involve machines (in this case oxidation furnaces), controllers, and suitable recipes.

Once the SiO_2 layer is grown, we will need to measure its thickness and perhaps other properties like its dielectric constant or index of refraction. We need to be certain that the recipe worked and that the properties of the film are close to the desired values. Measuring generally involves instruments which are used after the process step is complete (ellipsometers or spectrophotometers in this case, which can measure the SiO_2 thickness).

The third requirement for a unit process to be incorporated into an overall IC manufacturing environment is that we understand it and be able to model it. We need to be able to predict the results of a given process sequence. In the case of oxidation, we need to know how the oxide properties (thickness, index of refraction, and perhaps defect density), depend on the process control parameters (time, temperature, etc.). There are many reasons why we need such a model. Clearly it is needed in designing new process conditions. If we desire a $1\ \mu m$ oxide in a particular structure, we should be able to calculate a set of parameters that will give that thickness. We also often need to be able to estimate process sensitivities through our models. For example, we might like to know how sensitive our oxide thickness is to small changes in the furnace temperature. Such information would provide some indication of how robust our manufacturing process is. Finally, the existence of a model, particularly a physically based as opposed to an

empirically based model, implies that we really do understand the unit process step. Such understanding is often crucial in solving problems that arise during manufacturing.

The last requirement is an ability to simulate the unit process. This is a fairly recent addition to the list and in fact is not always satisfied in silicon technology. However, if we have a model and we have incorporated that model in a computer simulation program, then we will have the ability to accurately and rapidly answer the kinds of questions posed in the preceding paragraph. Such simulators exist today for many, but not all, unit process steps and we will describe and use them throughout this book.

Many texts on silicon technology begin with a thorough discussion of each of the unit process steps. Only after this is done is the integration of these steps into an overall manufacturing sequence considered. Our experience is that considering the integration first has a significant advantage. That advantage is that the context for each of the unit processes is established first. The disadvantage, of course, is that a reader unfamiliar with silicon technology may not fully appreciate a description of a complete process flow before the individual steps have been discussed in detail. We devote Chapter 2 to such a process flow (a CMOS process) and recommend that it be read before the other chapters. Readers relatively new to silicon technology will likely not completely understand all the details of the CMOS process at first. A second reading of Chapter 2 after the others have been covered may provide additional insight.

1.8 Summary of Key Ideas

Integrated circuits are widely regarded as one of the key components of today's world. They are pervasive in their application. Modern CAD tools have enabled a wide variety of people from many different disciplines to design and use ICs in many different kinds of applications. Approximately every two years the complexity of ICs doubles; that is, the technology used to make them allows twice as many components to be integrated on a single chip, as was possible two years earlier. In addition, the performance of the chips improves every year through scaling down of device sizes. These trends are likely to continue through at least the first decade of the 21st century. Innovations and inventions are likely to continue to occur beyond that time, with the result that the complexity and performance of chips will continue to improve. Ultimately, the ability to turn ideas into useful products depends on the ability to actually build these chips. We will explore the technology that has enabled us to do this in detail in the remainder of this book.

1.9 References

- [1.1]. Gordon Moore was one of the founders of Intel and first pointed out this "law" around 1970.
- [1.2]. "National Technology Roadmap for Semiconductors," SIA, 1994.
- [1.3]. "National Technology Roadmap for Semiconductors," SIA, 1997.
- [1.4]. D. M. Eigler and E. K. Schweizer, "Positioning Single Atoms with a Scanning Tunneling Microscope," *Nature*, vol. 344, p. 524, 1990.

- [1.5]. G. E. Moore, "Are We Really Ready for VLSI?" Keynote Address, 1979 ISSCC.
- [1.6]. W. Shockley, "The Path to the Conception of the Junction Transistor," *IEEE Trans. Elec. Dev.*, vol. ED-31, p. 1523, 1984.
- [1.7]. J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. Elec. Dev.*, vol. ED-23, p. 648, 1976.
- [1.8]. E. H. Putley and W. H. Mitchell, "The Electrical Conductivity and Hall Effect of Silicon," *Proc. Phys. Soc.*, vol. 72, p. 193, 1958.
- [1.9]. C. D. Thurmond, "The Standard Thermodynamic Function for the Formation of Electrons and Holes in Ge, Si, GaAs and GaP," *J. Electrochem. Soc.*, vol. 122, p. 1133, 1975.
- [1.10]. W. E. Beadle, J. C. Tsai and R. D. Plummer, *Quick Reference Manual for Silicon Integrated Circuit Technology*, John Wiley & Sons, 1985.
- [1.11]. See, for example, B. G. Streetman, *Solid State Electronic Devices*, Prentice Hall 1995, or R. F. Pierret, *Semiconductor Device Fundamentals*, Addison-Wesley, 1996.

1.10 Problems

- 1.1. Plot the NRTS roadmap data from Table 1-1 (feature size versus time) on an expanded scale version of Figure 1-2. Do all the points lie exactly on a straight line? If not, what reasons can you suggest for any deviations you observe?
- 1.2. Assuming dopant atoms are uniformly distributed in a silicon crystal, how far apart are these atoms when the doping concentration is (a) 10^{15} cm^{-3} , (b) 10^{18} cm^{-3} , (c) $5 \times 10^{20} \text{ cm}^{-3}$?
- 1.3. Consider a piece of pure silicon 100 μm long with a cross sectional area of 1 μm^2 . How much current would flow through this "resistor" at room temperature in response to an applied voltage of 1 volt?
- 1.4. Estimate the resistivity of pure silicon in $\Omega \text{ cm}$ at (a) room temperature, (b) 77K, and (c) 1000°C. You may neglect the temperature dependence of the carrier mobility in making this estimate.
- 1.5. a. Show that the minimum conductivity of a semiconductor sample occurs when

$$n = n_i \sqrt{\frac{\mu_p}{\mu_n}}$$

- b. What is the expression for the minimum conductivity?
- c. Is this value greatly different than the value calculated in problem 1.4 for the intrinsic conductivity?
- 1.6. When a Au atom sits on a lattice site in a silicon crystal, it can act as either a donor or an acceptor. E_D and E_A levels both exist for the Au and both are close to the

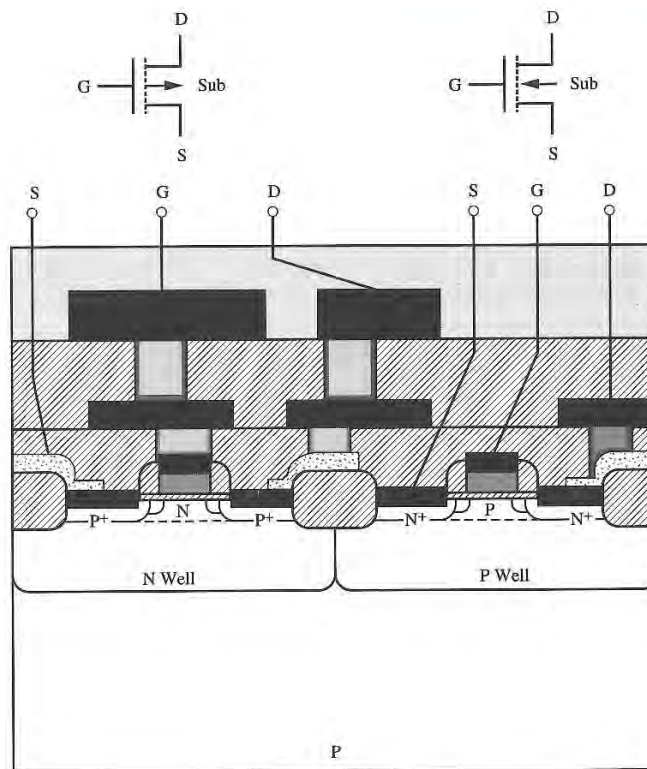


Figure 2-2 Cross section of the final CMOS integrated circuit. A PMOS transistor is shown on the left, an NMOS device on the right.

parts of the structure provide multiple wiring levels above the active devices to interconnect them to perform particular circuit functions. Finally, some regions are included simply to improve the performance of the individual devices by decreasing parasitic resistances or improving voltage ratings. As we proceed through the steps required to build this chip, we will discuss each of these points in greater detail.

2.2.1 The Beginning—Choosing a Substrate

Before we begin actual wafer fabrication, we must of course choose the starting wafers. In general this means specifying type (N or P), resistivity (doping level), crystal orientation, wafer size, and a number of other parameters having to do with wafer flatness, trace impurity levels, and so on. The major choices are the type, resistivity, and orientation.

Figure 2-2 indicates that the final structure has a P-type substrate. In most CMOS integrated circuits, the substrate has a moderately high resistivity (25–50 Ωcm) which corresponds to a doping level on the order of 10^{15} cm^{-3} (Figure 1-18). As is apparent from Figure 2-2, the active devices are actually built in wells diffused into the surface of the wafer. The doping levels in these wells are chosen to optimize the electrical properties of the active devices, as we will see later in this chapter. Typically the well doping levels are on the order of $10^{16} - 10^{17}\text{ cm}^{-3}$ near the wafer surface. In order to reproducibly manufacture such wells, the background doping (the substrate doping in this case) needs to be significantly less than the well doping. Thus the substrate doping is normally chosen to be on the order of 10^{15} cm^{-3} .

The observant reader might notice that the NMOS device could actually be built directly in the P substrate without adding the P well near the surface. In fact this is exactly the way the structure was sketched in Figure 1-34 for simplicity. While some CMOS circuits are actually built this way today, the twin well process illustrated in Figure 2-2 is much more common because the doping process used to produce the P well (ion implantation) is much better controlled in manufacturing than is the substrate doping. Also, since the P well and N well doping concentrations are on the same order, it is easier to start with a much more lightly doped substrate and tailor the wells for the NMOS and PMOS devices individually.

The observant reader might also note in Figure 1-34 that a substrate consisting of a P layer on a P^+ substrate was illustrated. This is one of the technology options we will consider later in Section 2.2.5.

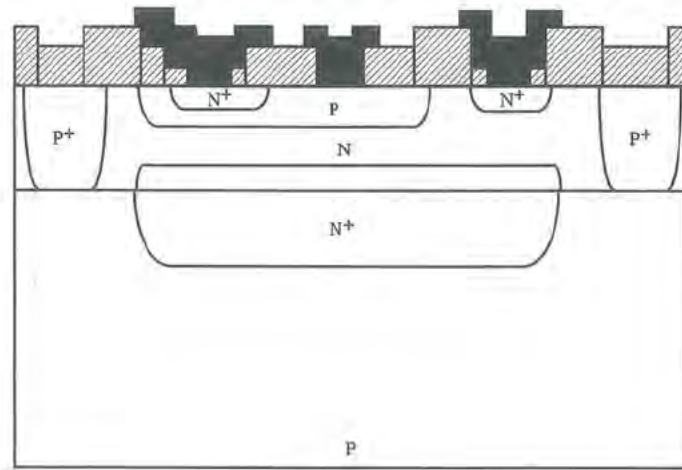
The only other major parameter we need to specify in the starting substrate is the crystal orientation. We will discuss crystal structure in more detail in Chapter 3. However, virtually all modern silicon integrated circuits are manufactured today from wafers with a (100) surface orientation. The principal reason for this is that the properties of the Si/SiO_2 interface are significantly better when a (100) crystal is used. We will discuss the reasons for this in detail in Chapter 6, but the key idea is that the electrical properties of this interface are intimately connected with the atomic bonding between Si and O that takes place when an SiO_2 layer is thermally grown on Si. It is found experimentally that there are fewer imperfections (unsatisfied bonds) on a (100) surface than is the case on other silicon surfaces. Primarily for this reason, we will choose a (100) surface orientation for our starting wafers.

We will also discuss in Chapter 4 some processing which is often done on the starting substrates before any actual device fabrication is begun. This processing is aimed at minimizing the sensitivity of the wafers to trace contaminants that can be introduced to the wafers during the many manufacturing steps they go through to build circuits. These preliminary processing steps are called *gettering* and the most common process today is known as “intrinsic gettering.” Since these steps are not essential to the device fabrication process, we will defer discussion of them to Chapter 4.

2.2.2 Active Region Formation

Modern CMOS chips integrate millions of active devices (NMOS and PMOS) side by side in a common silicon substrate. Circuits are designed with these devices to imple-

- 2.3. The cross section below illustrates a simple bipolar transistor fabricated as part of a silicon IC. (See also Figure 1–32.) Design a plausible process flow to fabricate such a structure, following the ideas of the CMOS process flow in this chapter. You do not have to include any quantitative process parameters (times, temperatures, doses, etc.) Your answer should be given in terms of a series of sketches of the structure after each major process step, like the figures in this chapter. Briefly explain your reasoning for each step and the order you choose to do things.



Crystal Growth, Wafer Fabrication and Basic Properties of Silicon Wafers

3

3.1 Introduction

In Chapter 1 we traced some of the important historical developments that led to the silicon IC industry as we know it today. Almost from the very beginning, it was clear that silicon was the best choice for the material on which to base this industry. The abundance of silicon, the availability of simple techniques for refining it and growing single crystals, the essentially ideal properties of the Si/SiO₂ interface, and the invention of manufacturing techniques based on the planar process, all led to the dominance of silicon based devices by the early 1960s.

The silicon industry depends on a ready supply of inexpensive, high-quality single-crystal wafers. In this chapter we will discuss the methods by which such wafers are prepared and some of the basic properties of these wafers. Integrated circuit manufacturers typically specify physical parameters (diameter, thickness, flatness, mechanical defects like scratches, crystallographic defects like dislocation density, etc.), electrical parameters (N or P type, dopant, resistivity, etc.), and finally, impurity levels (oxygen and carbon in particular) when purchasing wafers. All these parameters must be tightly controlled in order to use the wafers in the high-volume manufacturing of complex chips. We saw a specific example of this type of specification in Chapter 2 in the CMOS process. In that example, we specified a starting substrate which was P type (boron doped), (100) orientation, with a resistivity between 25 and 50 Ωcm. This resistivity corresponds to a doping concentration of about 10^{15} cm^{-3} or about 20 parts per billion boron. Although these are the principal parameters that need to be specified, a more complete wafer description would actually include many additional parameters as outlined above.

3.2 Historical Development and Basic Concepts

Although the first transistors were built in polycrystalline germanium, all modern integrated circuits depend on single-crystal material for optimum device operation. Polycrystalline or even amorphous materials can be used to build semiconductor devices,

but the defects present in these materials act as generation and recombination centers and have significant effects on basic device properties like mobility and minority carrier lifetime. The result is poor performance in the devices—lower output currents for a given applied voltage, higher leakage currents in junctions, and poor reproducibility from one device to another. We begin this chapter with a discussion of some of the basic properties of crystals.

3.2.1 Crystal Structure

In crystalline materials, the atoms are arranged spatially in a periodic fashion. A basic unit cell can be defined which repeats in all three dimensions. A standard terminology is used to describe crystal directions and planes, which we will briefly describe here. In some cases, the electrical properties of devices fabricated in silicon depend on the particular crystal orientation used for the starting wafer. [Recall that in Chapter 2 in our discussion of the CMOS process, we chose a (100) crystal.] Quite often, these orientation-dependent effects are associated with wafer surfaces or material interfaces, where the crystal terminates. Here, atoms may be incompletely bonded and the exact atomic arrangement can make a difference in the electrical properties. Bulk properties in silicon are generally isotropic (independent of direction) because of the cubic symmetry of the silicon crystal.

Figure 3-1 illustrates three simple crystal unit cells. All are based on a cubic structure. The Body-Centered Cubic or BCC cell has an extra atom in the center of the cube; the Face-Centered Cubic or FCC cell has an extra atom in the center of each face of the cube. Silicon and other semiconductors have more complex unit cells than these simple examples, but we can illustrate all of the basic ideas about crystal directions and planes with these simpler structures. An xyz coordinate system defines the directions in the crystal. The dimension a is called the lattice constant and is the basic distance over which the unit cell repeats in the cubic crystal. Directions in crystals are expressed in terms of three integers, which are the same as the components of a vector in that direction. For example, to get from atom A in the cubic crystal to atom B, we move 1 unit in each (x , y , and z) direction. This is then the $[111]$ direction in this crystal. To get from

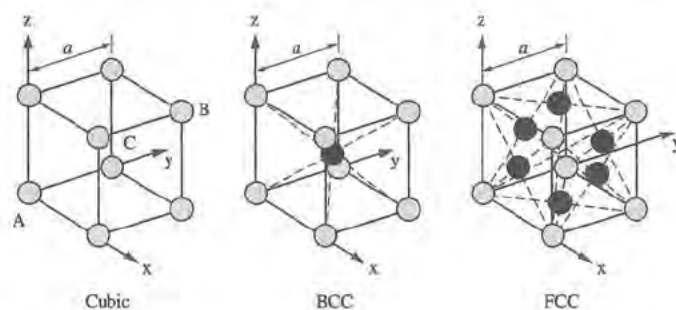


Figure 3-1 Unit cells in simple cubic crystals.

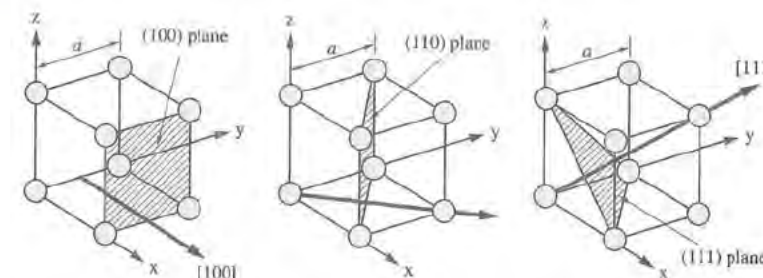


Figure 3-2 Crystal planes and major directions for a cubic lattice. The heavy arrows in each case illustrate crystal directions, designated $[hkl]$. The cross-hatched areas are the corresponding crystal planes, designated (hkl) .

atom A to atom C, we would move in the $[010]$ direction. By symmetry, many directions in a crystal are often equivalent, simply depending on our choice of reference. For example, the $[100]$, $[010]$, and $[001]$ directions are all equivalent in this example and by convention, are referred to as $\langle 100 \rangle$ directions.

It is also useful to describe planes in a crystal and a convention has been adopted for doing this. Figure 3-2 illustrates three simple planes in a cubic crystal. Such planes are described by Miller indices, which are a set of three integers calculated as follows. For a particular plane, the intercepts of that plane with the three crystal axes are determined. In the middle example in Figure 3-2, the intercepts of the (110) plane with the x -, y -, and z -axes are 1, 1, and ∞ , respectively. The reciprocals of these three intercepts are then taken, which gives $1/1$, $1/1$, and $1/\infty$ or simply the (110) plane. (The reciprocals eliminate infinities in the notation.) A minor complication can arise in this procedure if we pick a plane with intercepts outside the basic unit cell. For example, if we picked a plane with xyz intercepts of 3, 4, and 2, respectively, the reciprocals would be $1/3$, $1/4$, and $1/2$. In this case the smallest set of integers that have the same relative values are chosen for the Miller indices, which would result in the (436) plane in this case. There is also no reason why we could not pick a plane with a negative intercept, for example 1, -1 , 2. In this case the reciprocals are $1/1$, $1/-1$, and $1/2$, respectively. The nomenclature in this case places a bar over the corresponding Miller index, so that this plane becomes the $(\bar{2}21)$ plane. Finally, as was the case for crystal directions, many crystal planes are also equivalent. The (100) , (010) , and $(0\bar{1}0)$ planes are all equivalent and are designated as $\{100\}$ planes. Note that there are conventions which are used here for the style of bracket used to describe crystal directions and planes.

It should be noted that in cubic lattices, the direction $[hkl]$ is perpendicular to a plane with the identical three integers (hkl) , as illustrated in Figure 3-2. This is sometimes helpful in visualizing directions and planes in these crystals.

Silicon has a diamond cubic lattice structure, shown in Figure 3-3. This structure is most easily visualized as two merged FCC lattices with the origin of the second lattice offset from the first by $a/4$ in all three directions. Equivalently, one could think of the structure as an FCC lattice with two atoms at each lattice point, with the two offset from each other by $a/4$ in all three directions. Each atom is individually covalently bonded to four nearest neighbors, providing the electrical properties that we described in Chapter

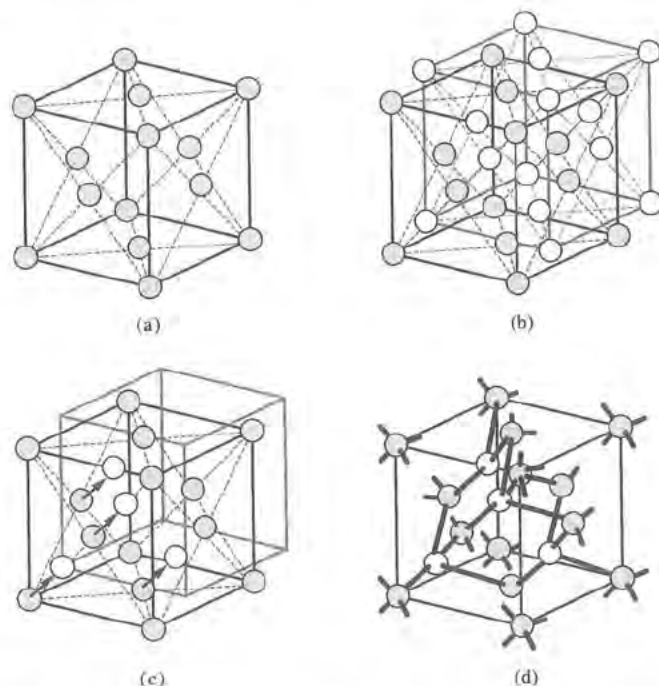


Figure 3-3 Diamond crystal structure of silicon. The unit cell may be visualized as two merged FCC cells, offset from each other by $a/4$ in all three directions. (a) shows the basic FCC unit cell. (b) shows two merged cells, offset by $a/4$. In (c), only the four atoms in the second FCC cell that lie within the first FCC cell are shown. The arrows show the corresponding $a/4$ displacement of these four atoms. In (d), the actual tetragonal bonding between the atoms in the diamond cell is shown.

1. Figure 3-3 illustrates the construction of the diamond lattice from the two merged FCC lattices.

There are two principal silicon crystal orientations that are used in manufacturing integrated circuits, (111) and (100) (meaning that the crystal terminates at the wafer surface on {111} or {100} planes respectively). As mentioned earlier, bulk properties of silicon are generally isotropic since the cubic crystal is symmetric. Thus we find, for example, that dopant diffusion coefficients are independent of the direction in which the dopant is diffusing as long as surfaces play no role in the process (Chapter 7). However, real devices are always built near surfaces and which crystal plane the surface terminates on can make a difference in the surface electrical and physical properties. {111} planes in silicon have the largest number of silicon atoms per cm^2 , {100} planes the lowest. This results in a number of differences in properties. {111} planes oxidize faster than {100} because the oxidation rate is proportional to the number of silicon atoms avail-

able for reaction (Chapter 6). {111} surfaces have higher densities of electrical defects (interface states) because at least some of these defects are believed to be associated with dangling silicon bonds (Chapter 6). Finally, dopant diffusion coefficients and other properties in bulk silicon usually do depend on the surface orientation of the crystal because surfaces are often active interfaces and can perturb bulk properties large distances away from the surface (Chapter 6). Primarily because of the superior electrical properties of the (100) Si/SiO₂ interface, (100) silicon is dominant in manufacturing today. Virtually all MOS-based technologies use crystals with this orientation (like the CMOS process in Chapter 2). Historically many bipolar technologies used (111) crystals because this orientation is somewhat easier to grow by the Czochralski method described later in this chapter; however as bipolar devices have shrunk in size in recent years, they have become more sensitive to surface properties and increasingly also use (100) wafers.

A final point should be made regarding surface effects on device and process parameters. The CMOS process flow described in Chapter 2 illustrated the fact that many etching, oxidation, and deposition steps are used in modern processes. Such steps can end up producing a very nonplanar surface. In other words, the shaping and patterning of the silicon inherent to device fabrication will produce a device surface, which contains {100} surface regions along with almost every other possible crystal orientation. Physical or electrical parameters, which vary with surface orientation, will therefore vary with position in modern ICs. Modeling such effects is only possible with modern computer simulation tools which need to first predict the structure accurately enough to know the surface orientation as a function of position, and then predict the resulting orientation effects on device parameters.

3.2.2 Defects in Crystals

There are several types of defects which are commonly found in crystals. It is useful to classify these into four categories, depending on their dimensionality. The usual categories are point defects, line defects, area defects, and volume defects. Figure 3-4 illustrates some of the more common defects in a simple two-dimensional representation. We will find in later chapters that a number of these defects play important roles in various aspects of silicon technology.

Point defects are the simplest to visualize and they play crucial roles in impurity diffusion (Chapter 7), in ion implantation damage (Chapter 8), and lesser roles in oxidation kinetics (Chapter 6) and other phenomena. In general, anything other than a silicon atom on a lattice site constitutes a point defect. By this definition, for example, a substitutional doping atom is a point defect and might be referred to as an impurity-related defect. However, there are two principal types of point defects we shall be primarily concerned with which are often referred to as native point defects. The first is simply a missing silicon lattice atom or vacancy, which we shall designate simply as V . The other is an extra silicon atom which we shall designate simply as I . Figure 3-4 illustrates two possible arrangements of such an extra atom in the lattice. The upper case illustrates a purely interstitial atom, that is, an atom sitting unbonded in one of the available sites between silicon atoms. The second illustrates two atoms sharing one lattice

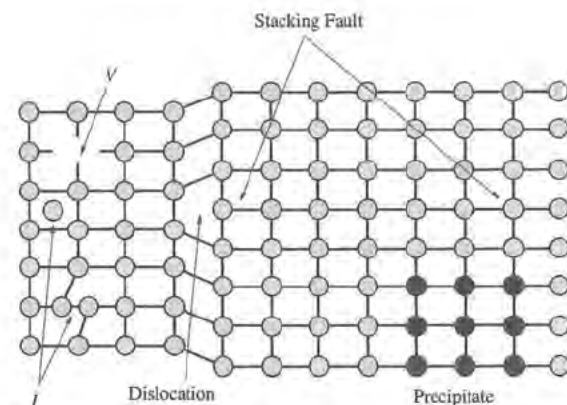


Figure 3-4 Simple 2D representation of some of the common defects found in crystals. *V* and *I* are point defects, the edge dislocation represents a typical line defect, the stacking fault is an area defect, and the precipitate is a volume defect.

site, a defect usually referred to as an interstitialcy. Most of the literature on silicon technology does not distinguish between these two defects, designating both of them as simply a silicon interstitial or *I*. Since the distinction is not usually important in modeling processes in silicon, we shall adopt this strategy in this book and simply use the symbol *I* to refer to an excess silicon atom. Most workers in the field today believe that the interstitialcy is likely the defect that is important in silicon processes, because of the lower energies required to create this defect compared to a pure interstitial. In general, the concentrations of both *V* and *I* increase as temperature increases. The equilibrium concentrations of these defects are essentially zero at room temperature, but they are much higher at process temperatures (10^{12} – 10^{15} cm $^{-3}$ at 1000°C). Quantitative models have been developed which describe the equilibrium concentrations of these defects in silicon. We will discuss these models later in this chapter.

One-dimensional defects in crystals are known as dislocations. An example is illustrated in Figure 3-4. The right-hand part of the crystal contains an extra plane of atoms, which terminates at a dislocation. The dislocation itself, then, is a linear defect in the direction into the paper. Dislocations either terminate at the edge of the crystal (edge dislocation), or they form a closed loop within the crystal (dislocation loops). Macroscopic edge dislocations are never present in silicon wafers at the start of IC manufacturing; such wafers are normally specified as “dislocation free.” However, such defects can be generated during the high-temperature steps normally used in wafer fabrication, particularly if thin films present on the wafer surface generate high stresses. We saw one example of such a process in Chapter 2 when we considered the LOCOS method of forming SiO $_2$ locally on the wafer surface. The Si $_3$ N $_4$ layer used in Figures 2-3 to 2-5 as a mask against oxidation is usually under large tensile stress, resulting in corresponding

compressive stress in the silicon substrate. If these stresses are large enough, they can result in dislocation formation during the high-temperature LOCOS oxidation.

Another way in which such stresses can be generated is through temperature gradients. These can occur during crystal growth, when the seed is first inserted into the melt and the outside of the seed heats up more rapidly than the inside, or during wafer processing if a heating system is used which is nonuniform. In the crystal growth case, dislocation formation is inevitable when the seed is first inserted into the liquid silicon, but such dislocations can be grown out to the edge of the crystal during the initial stages of growth as discussed in Section 3.3. In the wafer heating case, large temperature gradients are not normally a problem during gradual heating or cooling cycles such as those encountered in furnaces. However, they can be a problem in newer rapid thermal processing systems, which use high-intensity lamps to heat wafers at a rate of hundreds of °C per second. Such systems are finding increased use for Rapid Thermal Annealing (RTA) of implant damage (Chapter 7), Rapid Thermal Oxidation (RTO) described in Chapter 6 and for other types of Rapid Thermal Processing (RTP). If the lamps in such systems are not configured to maintain a uniform temperature across the wafer, temperature gradients can result. Note that maintaining a uniform temperature is not necessarily the same thing as maintaining uniform energy input into the wafer because heat loss due to radiation or conduction may vary with position across a wafer, especially at the edges. If temperature gradients exist, the resulting thermal stresses are given approximately by

$$\sigma = \alpha Y \Delta T \quad (3.1)$$

where α is the thermal expansion coefficient of silicon, Y is Young's modulus and ΔT is the temperature difference across the wafer. The yield strength of silicon is on the order of 0.5×10^9 dyne cm $^{-2}$ at process temperatures. Such stress levels can be reached, and dislocations therefore generated, with ΔT values on the order of 100°C.

Microscopic dislocation loops usually are present in silicon starting wafers. At the high temperatures associated with crystal growth, point defect concentrations are quite high. As the crystal is pulled from the melt and cools, there may not be sufficient time for the populations of these defects to reduce to the much lower equilibrium populations associated with lower temperatures. (The *V* and *I* concentrations could be reduced by recombination of the two defects with each other, or diffusion of the individual defects to the crystal surface.) Small dislocation loops can form during the cooling process as a result of the agglomeration of excess *V* or *I*. If, for example, the extra half-plane of atoms on the right of Figure 3-4 extended only over a small number of lattice sites, the dislocation bounding the extra plane would form a closed loop, often a circle, since a circle has the shortest dislocation length and lowest energy. Such a defect could be formed by a group of *I* forming the extra half-plane on the right of Figure 3-4, or by a group of *V* forming the missing half-plane on the left.

Dislocations are active defects in crystals, that is they can move when subjected to stresses or when excess point defects are present. The process of “climb” occurs when excess point defects are absorbed by the dislocation. It is easy to visualize in Figure 3-4, for example, the dislocation moving one lattice plane to the left by absorbing an *I*, or

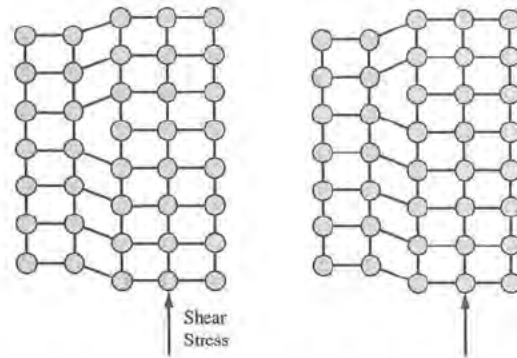


Figure 3-5 Movement of a dislocation by glide in response to shear stress.

moving one lattice plane to the right by absorbing a V . In fact these kinds of processes are often useful in determining whether a particular processing step increases or decreases V or I concentrations, by observing whether dislocations move during that step. We will discuss this idea more carefully in connection with stacking faults below. It is also possible for dislocations to move by “gliding” in response to shear stresses. Figure 3-5 illustrates this process. In silicon, the glide planes are $\{111\}$ planes, the most densely packed planes. The fact that dislocations prefer to propagate along these planes also makes it possible to “grow out” such defects during the initial stages of crystal growth.

The most common kind of 2D or area defect found in silicon is the stacking fault. Such a defect is schematically shown in Figure 3-4. Stacking faults always form along $\{111\}$ planes and are simply the insertion or removal of an extra $\{111\}$ plane. In the silicon crystal structure shown in Figure 3-3, there are actually 3 parallel $\{111\}$ planes in each unit cell commonly referred to as A, B, and C. In a perfect crystal, the stacking order is ABCABC, and so on. When a stacking fault is present, either an extra plane is inserted (ABCACBC, etc.) or a plane is missing (ABCABABC, etc.). Such faults are referred to as “extrinsic” if there is an extra plane of atoms, or “intrinsic” if a plane is missing. Stacking faults are bounded by dislocations and, when they intersect the wafer surface, are usually referred to as surface stacking faults. It is relatively easy to measure the density and the size of stacking faults in silicon using chemical etchants, which preferentially attack the highly stressed regions associated with the dislocations that bound the faults. This is discussed in more detail in Section 3.4.

Most of the stacking faults observed in silicon have been determined to be of the extrinsic type, implying that they formed in an environment of excess interstitials. Since the size of the faults can be determined experimentally, this measurement can be used as a means of determining whether a particular process increases or decreases point defect concentrations. If the faults are extrinsic, then they will grow by absorbing more I , or shrink by absorbing V . Experiments like this have been used to determine that oxidation injects excess interstitials into the underlying substrate, since stacking faults are observed to grow during oxidation. Stacking faults which are used as markers in such experiments are sometimes called Oxidation Induced Stacking Faults or OISF. We will discuss these experiments in more detail in Chapters 6 and 7 when we examine point defect based models for oxidation and diffusion.

Volume defects in crystals can include agglomerations of point defects such as a void caused by a collection of vacancies, precipitates of dopants or impurities, which may be crystalline or amorphous, or other 3D defects. One example is shown in Figure 3-4 where a cluster of doping atoms is shown, in this case as a crystalline precipitate (presumed in this example to extend into the third dimension). “Crystalline precipitate” is used in this case because each of the precipitated atoms occupies a lattice site. Such defects may nucleate either heterogeneously because of an existing crystal defect such as a dislocation, or homogeneously when through random diffusion, several of the species involved come close enough together to form a critical size nucleus for the precipitate. There are several such situations of importance in silicon technology. When doping concentrations are pushed too high (beyond the solubility of the dopant in the crystal), precipitation of dopant atoms in amorphous or crystalline clusters is often observed. The doping atoms in these clusters are usually not electrically active. Another example occurs in CZ silicon because of the relatively high oxygen impurity level. The oxygen is usually incorporated at a level corresponding to its solubility at the crystal growth temperature ($\approx 10^{18} \text{ cm}^{-3}$). This is far higher than oxygen’s solubility at normal device fabrication temperatures, with the result that oxygen precipitation and clustering can occur during normal device fabrication. In some situations, this can actually be a useful event, since if the resulting dislocations and damage are kept away from active device regions, these heavily faulted regions can actually preferentially getter or collect unwanted impurities like heavy metals, keeping them away from device active regions. This is usually accomplished by precipitating the oxygen in the bulk of the wafer, away from the surface where the active devices are located. We will discuss this process in more detail later in this chapter in Section 3.5.5.

We will find many examples throughout this book in which crystalline defects play important roles in process steps. In some cases the role is very beneficial; in others the defects are minimized as much as possible to prevent unwanted effects. In both cases, it has become very important in recent years to both understand and control these defects. We will deal with specific cases of how this is done as the need arises for specific technology steps throughout the book.

3.2.3 Raw Materials and Purification

Fortunately, silicon is a very abundant material, representing about 25.7% of the earth’s crust. Naturally occurring minerals containing silicon are very impure, however, which means that the silicon must be refined as well as be converted into the crystalline form. This is usually a multistage process, beginning with quartzite, a type of sand. Chemically quartzite is SiO_2 . The first step in the refining process is to convert the quartzite to Metallurgical Grade Silicon or MGS. This process usually takes place in a furnace in which a mixture of the quartzite and a carbon source (usually coal or coke) is heated to temperatures approaching 2000°C . The power source is electric, provided by submerged electrodes, which arc and heat the charge. A number of reactions take place in the furnace, the overall result of which is that liquid silicon is drawn off and CO is given off as a gas, as shown below.



9.1 Introduction

The CMOS process flow in Chapter 2 contained many layers above the silicon substrate. These layers include dielectrics, semiconductors, and metals. In most cases these layers must be deposited which implies that there are typically many deposition steps in an IC process. In this chapter we describe the different techniques used to deposit these thin films. In Chapter 10 the techniques used to etch thin films are discussed.

We begin with several issues related to thin films and their deposition. These issues are important when evaluating deposited films and choosing deposition methods. The first issue is the “quality” of the deposited film. By quality we mean composition, contamination levels, defect density, and mechanical and electrical properties. The composition of films may vary depending on the deposition conditions and it is usually important to achieve a specific composition. This is especially important when a range of compositions is possible, such as with TiN films which can be Ti or N rich, or with alloys of metals like Al/Cu. Generally there needs to be a low level of contaminants, such as unwanted metals, water, oxygen, or halogens. There should also be a minimum of pin holes or other structural defects. These often result from particle contamination on the surface and are a common problem in thin film deposition (recall Figure 4–1). There should also be a minimum of stress in the films, and the films should be mechanically stable during subsequent processing. Finally, good adhesion to underlying films is important.

A second issue is that of uniform thickness across a wafer, from wafer to wafer, and as the film crosses nonplanar topography. Figure 9–1 shows a metal layer deposited over nonplanar topography in which there is a step in the underlying oxide. In Figure 9–1(b) the film thickness decreases as it crosses a step, which can lead to high electrical resistance in metal lines and a greater chance of mechanical cracking and failure. Coverage on the side of a step in topography is called step coverage. It is often defined quantitatively as the minimum thickness deposited on the side of a step divided by the thickness deposited on the top horizontal surface. Conformal step coverage, or conformal coverage, refers to uniform film thickness on both horizontal and vertical surfaces, or a step coverage of one.

A related issue is that of filling spaces between or within topographical structures. This includes filling a via or contact hole with metal, as illustrated in Figure 9–2(a), and filling spaces or gaps in shallow trench isolation structures or between metal lines with an oxide (often called gap filling), as illustrated in Figure 9–2(b). In this figure incomplete filling is shown, leading to a void in the dielectric between the lines. A void in a metal layer can lead to high-contact or -sheet resistance, and in a dielectric layer can

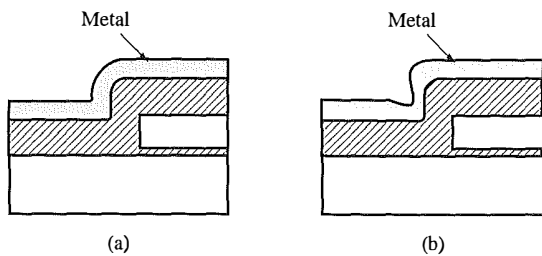


Figure 9-1 Step coverage of metal over nonplanar topography. (a) shows conformal step coverage, with constant thickness on horizontal and vertical surfaces. (b) shows poor step coverage.

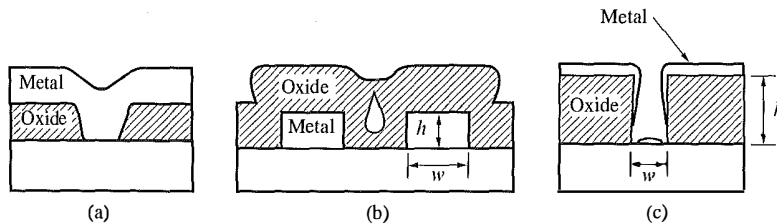


Figure 9-2 Thin film filling issues. (a) shows good metal filling of a via or contact hole in a dielectric layer. (b) shows silicon dioxide dielectric filling the space between metal lines, with poor filling leading to void formation. (c) shows poor filling of the bottom of a via hole with a barrier or contact metal.

result in cracking problems. In addition, a void may trap processing chemicals or moisture, leading to reliability problems. Even if complete filling of a space is not needed, good coverage or filling at the bottom of a space is often required. An example of this is the case of depositing contact and barrier layers, in which good filling of the bottom of contact or via holes is important. Figure 9-2(c) shows poor bottom coverage of a deep contact hole. The sides may also need adequate coverage of an adhesion layer.

An important parameter that can affect filling and bottom coverage is the Aspect Ratio (AR) of a feature, defined as the ratio of the height of a feature to its width:

$$AR = \frac{\text{height of feature}}{\text{width of feature}} = \frac{h}{w} \quad (9.1)$$

The feature could be a metal line, for example, or a space, such as a gap between two metal lines or a contact hole, as illustrated in Figure 9-2(b) and (c). A deep, narrow contact hole would have a large aspect ratio and would be harder to fill.

Two examples of poor filling and coverage are shown in Figure 9-3. Figure 9-3(a) shows poor step coverage of a TiW/Al/TiW metal stack layer over an oxide step. Figure 9-3(b) shows voids in an oxide layer for narrow spaces between metal lines. We will see that in filling or depositing in very deep contacts or vias, conditions that would lead to good step coverage may actually hinder adequate filling of the bottom of the contacts or vias.

We saw in the CMOS process flow in Chapter 2 that as more interconnect levels are used, another property that is desired for thin films is a smooth and flat top surface, as op-

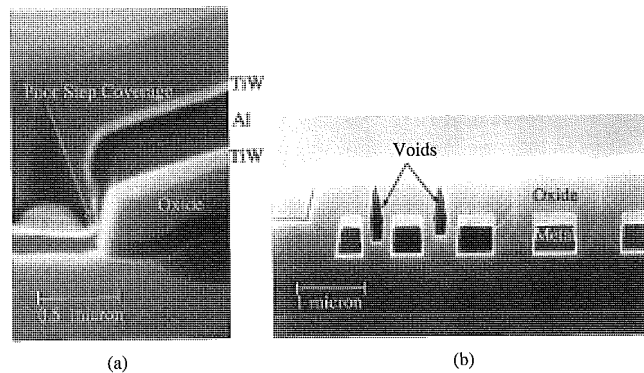


Figure 9-3 SEM images of coverage and filling problems. (a) shows poor step coverage of a TiW/Al/TiW metal stack layer, deposited by sputter deposition, over an oxide step. (b) shows voids in a Chemical Vapor Deposition (CVD) oxide layer for narrow spaces between metal lines. Photos courtesy of VLSI Technology, Inc.

posed to one that just follows the underlying topography. Figure 9-3(b) shows a relatively flat, planarized top surface of the oxide over the metal lines. While planarized topography can be achieved by postdeposition processing using etching or polishing techniques, we will also see how certain deposition techniques can also help attain this feature.

9.2 Historical Development and Basic Concepts

Methods of thin film deposition are usually separated into two main categories: Chemical Vapor Deposition (CVD) and Physical Vapor Deposition (PVD). In each case the silicon wafer is placed in a deposition chamber, and the constituents of the film are delivered through the gas phase to the surface of the substrate where they form the film. In the case of CVD, reactant gases are introduced into the deposition chamber, and chemical reactions between the reactant gases on the substrate surface are used to produce the film. In the case of PVD, physical methods are used to produce the constituent atoms which pass through a low-pressure gas phase and then condense on the substrate. The methods used to produce the atom flux in PVD include (1) heating a solid or molten source until it vaporizes, called “evaporation,” and (2) bombarding a solid source with energetic ions formed in a plasma, called “sputter deposition.” PVD is sometimes called “vacuum deposition” since very low-pressure environments are required for the transport of the gaseous species from the source to the film surface. (As we will see, low-pressure CVD is commonly used, but not at nearly as low pressures as in PVD.) CVD has historically been used in the integrated circuit industry mainly for silicon and dielectric deposition, primarily due to its good quality films and good step coverage. PVD has been used historically for metal deposition, mostly due to its ability to deposit a variety of metals and alloys not easily deposited by CVD.