

MAY
SZE

GARY S. MAY
SIMON M. SZE

Alexander Usenko
Silicon Wafer Technologies
240 King Blvd.
Newark, NJ 07102

This concise introduction to semiconductor fabrication technology covers everything professionals need to know, from crystal growth to integrated devices and circuits. Throughout, the authors address both theory and the practical aspects of each major fabrication step, including crystal growth, silicon oxidation, photolithography, etching, diffusion, ion implantation, and thin film deposition.

- The book integrates Computer Modeling & Simulation tools throughout.
- Process simulation is used as a tool for what-if analysis and discussion.
- Comprehensive coverage of process sequences helps readers connect individual steps into a cohesive whole.

About the Authors

Gary S. May, Ph.D. is Executive Assistant to the President and Motorola Foundation Professor of Microelectronics in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. Dr. May was a National Science Foundation National Young Investigator, Georgia Tech's Outstanding Young Alumnus, received Georgia Tech's Outstanding Service Award, and was named a Giant of Science by the Quality Education for Minorities network in 2001. He was a member of the NSF Engineering Advisory Committee, served on and chaired the NSF Committee for Equal Opportunity in Science and Engineering, and was Editor-In-Chief for *IEEE Transactions of Semiconductor Manufacturing* from 1997 to 2001. Dr. May currently serves as chair of the National Advisory Board for the National Society of Black Engineers.

Simon M. Sze, Ph.D. is UMC Chair Professor of National Chiao Tung University, and President of the National Nano Device Laboratories. He has received the IEEE Ebers Award, the Sun Yet-sen Award, the National Science and Technology Award, and the National Chair Professor Award. He is a Life Fellow of IEEE, a member of the Academia Sinica, the Chinese Academy of Engineering, and the US National Academy of Engineering. He has authored or coauthored over 150 technical papers, and has written, edited, and contributed to 24 books. His book, *Physics of Semiconductor Devices* (Wiley, 1969, 2nd Ed, 1981) is the *most* cited work in contemporary engineering and applied science publications (over 12,000 citations from ISI Press).

Find out more at: <http://www.wiley.com/college/may>

WILEY

www.wiley.com/college/may

ISBN 0-471-23279-3



9 780471 232797

Fundamentals of SEMICONDUCTOR FABRICATION

WILEY

Fundamentals of SEMICONDUCTOR FABRICATION



TSMC-1023
U.S. Patent 9,337,193

Fundamentals of Semiconductor Fabrication

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

SEMI 300

Fundamentals of Semiconductor Fabrication

GARY S. MAY

*Motorola Foundation Professor
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia*

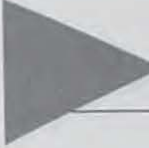
SIMON M. SZE

*UMC Chair Professor
National Chiao Tung University
National Nano Device Laboratories
Hsinchu, Taiwan*



WILEY

JOHN WILEY & SONS, INC.



Preface

This book provides an introduction to semiconductor fabrication technology, from crystal growth to integrated devices and circuits. It covers theoretical and practical aspects of all major steps in the fabrication sequence. It is intended as a textbook for senior undergraduates or first-year graduate students in physics, chemistry, electrical engineering, chemical engineering, and materials science. The book can be used conveniently in a semester-length course on integrated circuit fabrication. Such a course may or may not be accompanied by a corequisite laboratory. The text can also serve as a reference for practicing engineers and scientists in the semiconductor industry.

Chapter 1 gives a brief historical overview of major semiconductor devices and key technology developments, as well as an introduction to basic fabrication steps. Chapter 2 deals with crystal growth techniques. The next several chapters are organized according to a typical fabrication sequence. Chapter 3 presents silicon oxidation. Photolithography and etching are discussed in Chapters 4 and 5, respectively. Chapters 6 and 7 present the primary techniques for the introduction of dopants: diffusion and ion implantation. The final chapter on individual process steps, Chapter 8, covers various methods of thin film deposition. The final three chapters focus on broad, summative topics. Chapter 9 ties the individual process steps together by presenting the process flows for critical process technologies, integrated devices, and microelectrical mechanical systems (MEMS). Chapter 10 introduces high-level integrated circuit manufacturing issues, including electrical testing, packaging, process control, and yield. Finally, Chapter 11 discusses the future outlook and challenges for the semiconductor industry.

Each chapter begins with an introduction and a list of learning goals and concludes with a summary of important concepts. Solved example problems are provided throughout, and suggested homework problems appear at the end of the chapter. The concept of process simulation is presented in several chapters, using the popular SUPREM and PROLITH software packages as application vehicles. Mastery of this software is intended to supplement, but not replace, learning the fundamental concepts associated with microelectronics processing.

A complete set of detailed solutions to all end-of-chapter problems has been prepared. This instructor's manual is available to all adopting faculty. The figures in the text are also available, in electronic format, from the publisher at the following website: <http://www.wiley.com/college/may>.



Acknowledgments

The authors wish to thank Drs. T. C. Chang, T. S. Chao, M. C. Chiang, F. H. Ko, M. C. Liaw, and S. C. Wu of the National Nano Device Laboratories, and Prof. T. L. Li of the National Chia-Yi University for their helpful suggestions and discussions. We are further indebted to Mr. N. Erdos for technical editing of the manuscript, Ms. Iris Lin for typing the many revisions of the draft, and Ms. Y. G. Yang of the Semiconductor Laboratory, National Chiao Tung University who furnished the hundreds of technical illustrations used in the book.

At John Wiley and Sons, we wish to thank Mr. W. Zobrist, who encouraged us to undertake the project. S. M. Sze wishes also to acknowledge the Spring Foundation of the National Chiao Tung University for its financial support, and the United Microelectronic Corporation for the UMC Chair Professorship grant that provided the environment to work on this book. G. S. May would like to similarly acknowledge the Motorola Foundation Professorship.

Contents

CHAPTER 1

Introduction 1

- 1.1 Semiconductor Materials 2
- 1.2 Semiconductor Devices 2
- 1.3 Semiconductor Process Technology 5
 - 1.3.1 Key Semiconductor Technologies 5
 - 1.3.2 Technology Trends 8
- 1.4 Basic Fabrication Steps 11
 - 1.4.1 Oxidation 11
 - 1.4.2 Photolithography and Etching 13
 - 1.4.3 Diffusion and Ion Implantation 14
 - 1.4.4 Metallization 14
- 1.5 Summary 14
- References 15

CHAPTER 2

Crystal Growth 17

- 2.1 Silicon Crystal Growth from the Melt 18
 - 2.1.1 Starting Material 18
 - 2.1.2 The Czochralski Technique 18
 - 2.1.3 Distribution of Dopant 19
 - 2.1.4 Effective Segregation Coefficient 22
- 2.2 Silicon Float-Zone Process 24
- 2.3 GaAs Crystal Growth Techniques 26
 - 2.3.1 Starting Materials 26
 - 2.3.2 Crystal Growth Techniques 30
- 2.4 Material Characterization 31
 - 2.4.1 Wafer Shaping 31
 - 2.4.2 Crystal Characterization 33
- 2.5 Summary 38
- References 39
- Problems 39

CHAPTER 3

Silicon Oxidation 41

- 3.1 Thermal Oxidation Process 42
 - 3.1.1 Kinetics of Growth 42
 - 3.1.2 Thin Oxide Growth 49

- 3.2 Impurity Redistribution During Oxidation 50
- 3.3 Masking Properties of Silicon Dioxide 51
- 3.4 Oxide Quality 53
- 3.5 Oxide Thickness Characterization 54
- 3.6 Oxidation Simulation 54
- 3.7 Summary 57
- References 58
- Problems 59

CHAPTER 4

Photolithography 60

- 4.1 Optical Lithography 60
 - 4.1.1 The Clean Room 60
 - 4.1.2 Exposure Tools 62
 - 4.1.3 Masks 65
 - 4.1.4 Photoresist 67
 - 4.1.5 Pattern Transfer 70
 - 4.1.6 Resolution Enhancement Techniques 72
- 4.2 Next-Generation Lithographic Methods 73
 - 4.2.1 Electron Beam Lithography 73
 - 4.2.2 Extreme Ultraviolet Lithography 76
 - 4.2.3 X-Ray Lithography 78
 - 4.2.4 Ion Beam Lithography 79
 - 4.2.5 Comparison of Various Lithographic Methods 80
- 4.3 Photolithography Simulation 81
- 4.4 Summary 83
- References 83
- Problems 84

CHAPTER 5

Etching 85

- 5.1 Wet Chemical Etching 85
 - 5.1.1 Silicon Etching 86
 - 5.1.2 Silicon Dioxide Etching 87
 - 5.1.3 Silicon Nitride and Polysilicon Etching 88
 - 5.1.4 Aluminum Etching 88
 - 5.1.5 Gallium Arsenide Etching 88

- 5.2 Dry Etching 89
 - 5.2.1 Plasma Fundamentals 90
 - 5.2.2 Etch Mechanism, Plasma Diagnostics, and End-Point Control 91
 - 5.2.3 Reactive Plasma Etching Techniques and Equipment 93
 - 5.2.4 Reactive Plasma Etching Applications 97
- 5.3 Etch Simulation 101
- 5.4 Summary 102
 - References 103
 - Problems 103

CHAPTER 6

Diffusion 105

- 6.1 Basic Diffusion Process 106
 - 6.1.1 Diffusion Equation 107
 - 6.1.2 Diffusion Profiles 109
 - 6.1.3 Evaluation of Diffused Layers 113
- 6.2 Extrinsic Diffusion 114
 - 6.2.1 Concentration-Dependent Diffusivity 115
 - 6.2.2 Diffusion Profiles 117
- 6.3 Lateral Diffusion 118
- 6.4 Diffusion Simulation 120
- 6.5 Summary 121
 - References 122
 - Problems 122

CHAPTER 7

Ion Implantation 124

- 7.1 Range of Implanted Ions 125
 - 7.1.1 Ion Distribution 125
 - 7.1.2 Ion Stopping 127
 - 7.1.3 Ion Channeling 130
- 7.2 Implant Damage and Annealing 131
 - 7.2.1 Implant Damage 131
 - 7.2.2 Annealing 134
- 7.3 Implantation-Related Processes 136
 - 7.3.1 Multiple Implantation and Masking 136
 - 7.3.2 Tilt-Angle Ion Implantation 138
 - 7.3.3 High-Energy and High-Current Implantation 139
- 7.4 Ion Implantation Simulation 140
- 7.5 Summary 141
 - References 142
 - Problems 142

CHAPTER 8

Film Deposition 144

- 8.1 Epitaxial Growth Techniques 144
 - 8.1.1 Chemical Vapor Deposition 145
 - 8.1.2 Molecular Beam Epitaxy 148
- 8.2 Structures and Defects in Epitaxial Layers 152
 - 8.2.1 Lattice-Matched and Strained-Layer Epitaxy 152
 - 8.2.2 Defects in Epitaxial Layers 153
- 8.3 Dielectric Deposition 155
 - 8.3.1 Silicon Dioxide 156
 - 8.3.2 Silicon Nitride 160
 - 8.3.3 Low-Dielectric-Constant Materials 162
 - 8.3.4 High-Dielectric-Constant Materials 164
- 8.4 Polysilicon Deposition 165
- 8.5 Metallization 167
 - 8.5.1 Physical Vapor Deposition 167
 - 8.5.2 Chemical Vapor Deposition 168
 - 8.5.3 Aluminum Metallization 169
 - 8.5.4 Copper Metallization 173
 - 8.5.5 Silicide 175
- 8.6 Deposition Simulation 177
- 8.7 Summary 177
 - References 179
 - Problems 180

CHAPTER 9

Process Integration 182

- 9.1 Passive Components 184
 - 9.1.1 The Integrated Circuit Resistor 184
 - 9.1.2 The Integrated Circuit Capacitor 185
 - 9.1.3 The Integrated Circuit Inductor 187
- 9.2 Bipolar Technology 188
 - 9.2.1 The Basic Fabrication Process 189
 - 9.2.2 Dielectric Isolation 192
 - 9.2.3 Self-Aligned Double-Polysilicon Bipolar Structures 193
- 9.3 MOSFET Technology 196
 - 9.3.1 The Basic Fabrication Process 196
 - 9.3.2 Memory Devices 199
 - 9.3.3 CMOS Technology 203
 - 9.3.4 BiCMOS Technology 210
- 9.4 MESFET Technology 212

- 9.5 MEMS Technology 212
 - 9.5.1 Bulk Micromachining 215
 - 9.5.2 Surface Micromachining 215
 - 9.5.3 LIGA Process 215
- 9.6 Process Simulation 218
- 9.7 Summary 223
 - References 223
 - Problems 224

CHAPTER 10

IC Manufacturing 226

- 10.1 Electrical Testing 227
 - 10.1.1 Test Structures 227
 - 10.1.2 Final Test 228
- 10.2 Packaging 228
 - 10.2.1 Die Separation 230
 - 10.2.2 Package Types 230
 - 10.2.3 Attachment Methodologies 232
- 10.3 Statistical Process Control 237
 - 10.3.1 Control Charts for Attributes 237
 - 10.3.2 Control Charts for Variables 239
- 10.4 Statistical Experimental Design 242
 - 10.4.1 Comparing Distributions 242
 - 10.4.2 Analysis of Variance 243
 - 10.4.3 Factorial Designs 246
- 10.5 Yield 250
 - 10.5.1 Functional Yield 250
 - 10.5.2 Parametric Yield 254
- 10.6 Computer-Integrated Manufacturing 256
- 10.7 Summary 257
 - References 257
 - Problems 258

CHAPTER 11

Future Trends and Challenges 259

- 11.1 Challenges for Integration 259
 - 11.1.1 Ultrashallow Junction Formation 261
 - 11.1.2 Ultrathin Oxide 261
 - 11.1.3 Silicide Formation 261
 - 11.1.4 New Materials for Interconnection 261
 - 11.1.5 Power Limitations 261
 - 11.1.6 SOI Integration 262
- 11.2 System-on-a-Chip 262
- 11.3 Summary 264
 - References 264
 - Problems 264

► APPENDIX A

List of Symbols 265

► APPENDIX B

International System of Units (SI Units) 267

► APPENDIX C

Unit Prefixes 269

► APPENDIX D

Greek Alphabet 271

► APPENDIX E

Physical Constants 273

► APPENDIX F

Properties of Si and GaAs at 300 K 275

► APPENDIX G

Some Properties of the Error Function 277

► APPENDIX H

Basic Kinetic Theory of Gases 281

► APPENDIX I

SUPREM Commands 283

► APPENDIX J

Running PROLITH 287

► APPENDIX K

Percentage Points of the t Distribution 289

► APPENDIX L

Percentage Points of the F Distribution 291

Index 297

The most important device for advanced integrated circuits is the MOSFET (metal-oxide-semiconductor field-effect transistor), which was reported by Kahng and Atalla¹³ in 1960. Figure 1.3 shows the first device using a thermally oxidized silicon substrate. The device has a gate length of 20 μm and a gate oxide thickness of 100 nm ($1\text{ nm} = 10^{-7}\text{ cm}$). The two keyholes are the source and drain contacts, and the top elongated area is the aluminum gate evaporated through a metal mask. Although present-day MOSFETs have been scaled down to the deep-submicron regime, the choice of silicon and thermally grown silicon dioxide used in the first MOSFET remains the most important combination of materials. The MOSFET and related integrated circuits now constitute about 90% of the semiconductor device market. An ultrasmall MOSFET with a channel length of 15 nm has been demonstrated recently.¹⁴ This device can serve as the basis for the most advanced integrated circuit chips containing over one trillion ($>10^{12}$) devices.

In 1962, Hall et al.¹⁵ first achieved lasing in semiconductors. In 1963, Kroemer¹⁶ and Alferov and Kazarinov¹⁷ proposed the heterostructure laser. These proposals laid the foundation for modern laser diodes, which can be operated continuously at room temperature. Laser diodes are the key components for a wide range of applications, including digital video disks, optical-fiber communication, laser printing, and atmospheric pollution monitoring.



Figure 1.3 The first metal-oxide-semiconductor field-effect transistor.¹³ (Photograph courtesy of Bell Laboratories.)

Three important microwave devices were invented or realized in the next three years. The first device was the transferred-electron diode (TED; also called Gunn diode), invented by Gunn¹⁸ in 1963. The TED is used extensively in such millimeter-wave applications as detection systems, remote controls, and microwave test instruments. The second device is the IMPATT diode; its operation was first observed by Johnston et al.¹⁹ in 1965. IMPATT diodes can generate the highest continuous-wave (CW) power at millimeter-wave frequencies of all semiconductor devices. They are used in radar systems and alarm systems. The third device is the MESFET (metal-semiconductor field-effect transistor), invented by Mead²⁰ in 1966. It is a key device for monolithic microwave integrated circuits (MMIC).

An important semiconductor memory device was invented by Kahng and Sze²¹ in 1967. This is the nonvolatile semiconductor memory (NVSM), which can retain its stored information when the power supply is switched off. Although it is similar to a conventional MOSFET, the major difference is the addition of a "floating gate" in which semipermanent charge storage is possible. Because of its attributes of nonvolatility, high device density, low power consumption, and electrical rewritability (i.e., the stored charge can be removed by applying voltage to the control gate), the NVSM has become the dominant memory for portable electronic systems such as the cellular phone, notebook computer, digital camera, and smart card.

A limiting case of the floating-gate nonvolatile memory is the single-electron memory cell (SEMC), which is obtained by reducing the length of the floating gate to ultra-small dimensions (e.g., 10 nm). At this dimension, when an electron moves into the floating gate, the potential of the gate will be altered so that it will prevent the entrance of another electron. The SEMC is the ultimate floating-gate memory cell, since we need only one electron for information storage. The operation of a SEMC at room temperature was first demonstrated by Yano et al.²² in 1994. The SEMC can serve as the basis for the most advanced semiconductor memories, which can contain over one trillion bits.

The charge-coupled device (CCD) was invented by Boyle and Smith²³ in 1970. CCD is used extensively in video cameras and in optical sensing applications. The resonant tunneling diode (RTD) was first studied by Chang et al.²⁴ in 1974. RTD is the basis for most quantum-effect devices, which offer extremely high density, ultrahigh speed, and enhanced functionality, because it permits a greatly reduced number of devices to perform a given circuit function. In 1980, Mimura et al.²⁵ developed the MODFET (modulation-doped field-effect transistor). With the proper selection of heterojunction materials, the MODFET is the fastest field-effect transistor.

Since the invention of the bipolar transistor in 1947, the number and variety of semiconductor devices have increased tremendously as advanced technology, new materials, and broadened comprehension have been applied to the creation of new devices. However, one compelling question remains: What processes are required to construct these wondrous devices from basic semiconductor materials?

► 1.3 SEMICONDUCTOR PROCESS TECHNOLOGY

1.3.1 Key Semiconductor Technologies

Many important semiconductor technologies have been derived from processes invented centuries ago. For example, the growth of metallic crystals in a furnace was pioneered by Africans living on the western shores of Lake Victoria more than 2000 years ago.²⁶ This process was used to produce carbon steel in preheated forced-draft furnaces. Another example is the lithography process, which was invented in 1798. In this original process, the pattern, or image, was transferred from a stone plate (*litho*).²⁷ This section considers

the milestones of technologies that were applied for the first time to semiconductor processing or developed specifically for semiconductor device fabrication.

Some key semiconductor technologies are listed in Table 1.2 in chronological order. In 1918, Czochralski²⁸ developed a liquid-solid monocomponent growth technique. Czochralski growth is the process used to grow most of the crystals from which silicon wafers are produced. Another growth technique was developed by Bridgman²⁹ in 1925. The Bridgman technique has been used extensively for the growth of gallium arsenide and related compound semiconductor crystals. Although the semiconductor properties of silicon have been widely studied since early 1940, the study of semiconductor compounds was neglected for a long time. In 1952, Welker³⁰ noted that gallium arsenide and related III-V compounds were semiconductors. He was able to predict their characteristics and to prove them experimentally. Technology and devices using these compounds have since been actively studied.

The diffusion of impurity atoms in semiconductors is important for device processing. Basic diffusion theory was considered by Fick³¹ in 1855. The idea of using diffusion techniques to alter the type of conductivity in silicon was disclosed in a patent in 1952 by Pfann.³² In 1957, the ancient lithography process was applied to semiconductor device fabrication by Andrus.³³ He used photosensitive, etch-resistant polymers (photoresist) for pattern transfer. Lithography is a key technology for the semiconductor industry. The continued growth of the industry has been the direct result of improved lithographic technology. Lithography is also a significant economic factor, currently representing over 35% of IC manufacturing costs.

TABLE 1.2 Key Semiconductor Technologies

Year	Technology	Author(s)/Inventor(s)	Ref.
1918	Czochralski crystal growth	Czochralski	28
1925	Bridgman crystal growth	Bridgman	29
1952	III-V compounds	Welker	30
1952	Diffusion	Pfann	32
1957	Lithographic photoresist	Andrus	33
1957	Oxide masking	Frosch and Derrick	34
1957	Epitaxial CVD growth	Sheftal, Kokorish, and Krasilov	35
1958	Ion implantation	Shockley	36
1959	Hybrid integrated circuit	Kilby	37
1959	Monolithic integrated circuit	Noyce	38
1960	Planar process	Hoerni	39
1963	CMOS	Wanlass and Sah	40
1967	DRAM	Dennard	41
1969	Polysilicon self-aligned gate	Kerwin, Klein, and Sarace	42
1969	MOCVD	Manasevit and Simpson	43
1971	Dry etching	Irving, Lemons, and Bobos	44
1971	Molecular beam epitaxy	Cho	45
1971	Microprocessor (4004)	Hoff et al.	46
1982	Trench isolation	Rung, Momose, and Nagakubo	47
1989	Chemical mechanical polishing	Davari et al.	48
1993	Copper interconnect	Paraszcak et al.	49

CVD, chemical vapor deposition; CMOS, complementary metal-oxide-semiconductor field-effect transistor; DRAM, dynamic random access memory; MOCVD, metalorganic CVD.

The oxide masking method was developed by Frosch and Derrick³⁴ in 1957. They found that an oxide layer can prevent most impurity atoms from diffusing through it. In the same year, the epitaxial growth process based on the chemical vapor deposition technique was developed by Sheftal et al.³⁵ Epitaxy, derived from the Greek words *epi*, meaning "on," and *taxis*, meaning "arrangement," describes a technique of crystal growth to form a thin layer of semiconductor materials on the surface of a crystal that has a lattice structure identical to that of the crystal. This method is important for the improvement of device performance and the creation of novel device structures. In 1959, a rudimentary integrated circuit was made by Kilby.³⁷ It contained one bipolar transistor, three resistors, and one capacitor, all made in germanium and connected by wire bonding—a hybrid circuit. Also in 1959, Noyce³⁸ proposed the monolithic IC by fabricating all devices in a single semiconductor substrate (*monolith* means "single stone") and connecting the devices by aluminum metallization. Figure 1.4 shows the first monolithic IC of a flip-flop circuit containing six devices. The aluminum interconnection lines were obtained by etching evaporated aluminum layer over the entire oxide surface using the lithographic technique. These inventions laid the foundation for the rapid growth of the microelectronics industry.

The planar process was developed by Hoerni³⁹ in 1960. In this process, an oxide layer is formed on a semiconductor surface. With the help of a lithography process, portions of the oxide can be removed and windows cut in the oxide. Impurity atoms will diffuse only through the exposed semiconductor surface, and *p-n* junctions will form in the oxide window areas.

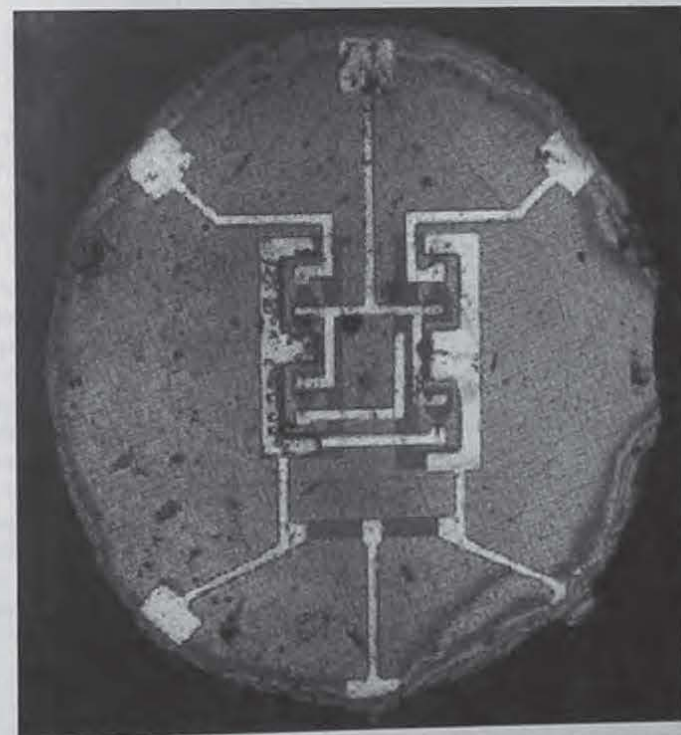


Figure 1.4 The first monolithic integrated circuit.³⁷ (Photograph courtesy of Dr. G. Moore.)

As the complexity of the IC increased, technology has moved from NMOS (*n*-channel MOSFET) to CMOS (complementary MOSFET), which employs both NMOS and PMOS (*p*-channel MOSFET) to form the logic elements. The CMOS concept was proposed by Wanless and Sah⁴⁰ in 1963. The advantage of CMOS technology is that logic elements draw significant current only during the transition from one state to another (e.g., from 0 to 1) and draw very little current between transitions, allowing power consumption to be minimized. CMOS technology is the dominant technology for advanced ICs.

In 1967, an important two-element circuit, the dynamic random access memory (DRAM), was invented by Dennard.⁴¹ The memory cell contains one MOSFET and one charge-storage capacitor. The MOSFET serves as a switch to charge or discharge the capacitor. Although a DRAM is volatile and consumes relatively high power, we expect that DRAMs will continue to be the first choice among various semiconductor memories for nonportable electronic systems in the foreseeable future.

To improve device performance, the polysilicon self-aligned gate process was proposed by Kerwin et al.⁴² in 1969. This process not only improved device reliability but also reduced parasitic capacitances. Also in 1969, the metalorganic chemical vapor deposition (MOCVD) method was developed by Manasevit and Simpson.⁴³ This is a very important epitaxial growth technique for compound semiconductors such as GaAs.

As the device dimensions were reduced, the dry etching technique was developed to replace wet chemical etching for high-fidelity pattern transfer. This technique was initiated by Irving et al.⁴⁴ in 1971 using a CF_4/O_2 gas mixture to etch silicon wafers. Another important technique developed in the same year by Cho⁴⁵ was molecular beam epitaxy. This technique has the advantage of near-perfect vertical control of composition and doping down to atomic dimensions. It is responsible for the creation of numerous photonic devices and quantum-effect devices.

In 1971, the first microprocessor was made by Hoff et al.,⁴⁶ who put the entire central processing unit (CPU) of a simple computer on one chip. It was a four-bit microprocessor (Intel 4004), shown in Figure 1.5, with a chip size of 3 mm by 4 mm, and it contained 2300 MOSFETs. It was fabricated by a *p*-channel polysilicon gate process using an 8- μ m design rule. This microprocessor performed as well as those in \$300,000 IBM computers of the early 1960s—each of which needed a CPU the size of a large desk. This was a major breakthrough for the semiconductor industry. Currently, microprocessors constitute the largest segment of the industry.

Since the early 1980s, many new technologies have been developed to meet the requirements of ever-shrinking minimum feature lengths. Three key technologies are trench isolation, chemical mechanical polishing, and copper interconnect. Trench isolation technology was introduced by Rung et al.⁴⁷ in 1982 to isolate CMOS devices. This approach eventually replaced all other isolation methods. In 1989, the chemical mechanical polishing method was developed by Davari et al.⁴⁸ for global planarization of the interlayer dielectrics. This is a key process for multilevel metallization. At submicron dimensions, a widely known failure mechanism is electromigration, which is the transport of metal ions through a conductor due to the passage of an electrical current. Although aluminum has been used since the early 1960s as interconnect material, it suffers from electromigration at high electrical current. Copper interconnect was introduced in 1993 by Paraszczak et al.⁴⁹ to replace aluminum for minimum feature lengths approaching 100 nm. This book considers all the technologies listed in Table 1.2.

1.3.2 Technology Trends

Since the beginning of the microelectronics era, the smallest linewidth (or the minimum feature length) of an integrated circuit has been reduced at a rate of about 13% per year.⁵⁰

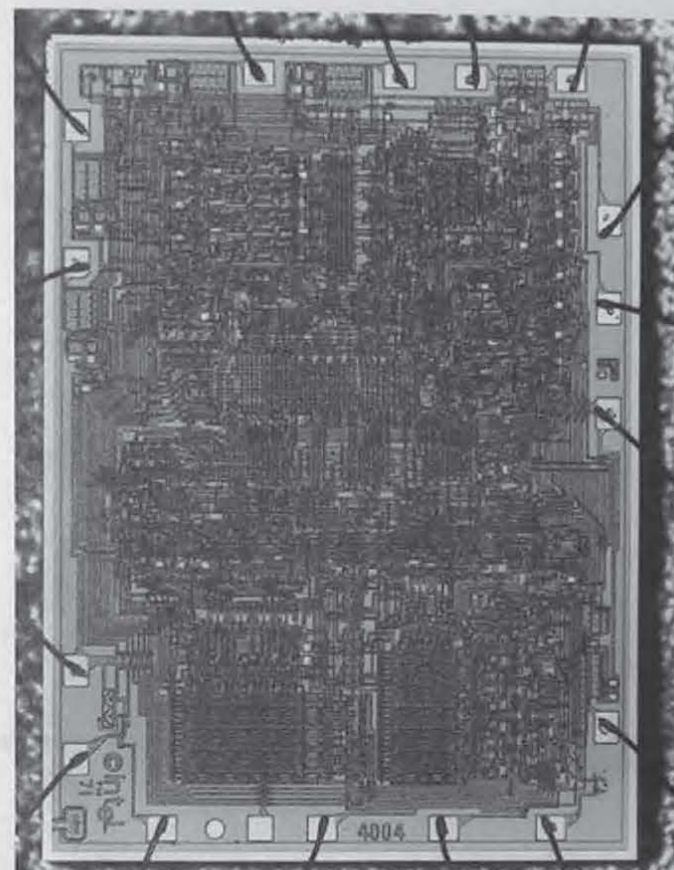


Figure 1.5 The first microprocessor.⁴⁶ (Photograph courtesy of Intel Corp.)

At that rate, the minimum feature length will shrink to about 50 nm in the year 2010. Device miniaturization results in reduced unit cost per circuit function. For example, the cost per bit of memory chips has halved every 2 years for successive generations of DRAM circuits. As device dimensions decrease, the intrinsic switching time decreases. Device speed has improved by four orders of magnitude since 1959. Higher speeds lead to expanded IC functional throughput rates. In the future, digital ICs will be able to perform data processing and numerical computation at terabit-per-second rates. As devices become smaller, they consume less power. Therefore, device miniaturization also reduces the energy used for each switching operation. The energy dissipated per logic gate has decreased by over one million times since 1959.

Figure 1.6 shows the exponential increase of the actual DRAM density versus the year of first production from 1978 to 2000. The density increases by a factor of 2 every 18 months. If trends continue, DRAM density will increase to 8 Gb in the year 2005 and to 64 Gb around the year 2012. Figure 1.7 shows the exponential increase of microprocessor computational power. Computational power also increases by a factor of 2 every 18 months. Currently, a Pentium-based personal computer has the same computational power as that of a CRAY 1 supercomputer of the late 1960s, yet it is three orders of magnitude smaller. If these trends continue, we will reach 100 GIP (billion instructions per second) in the year 2010.

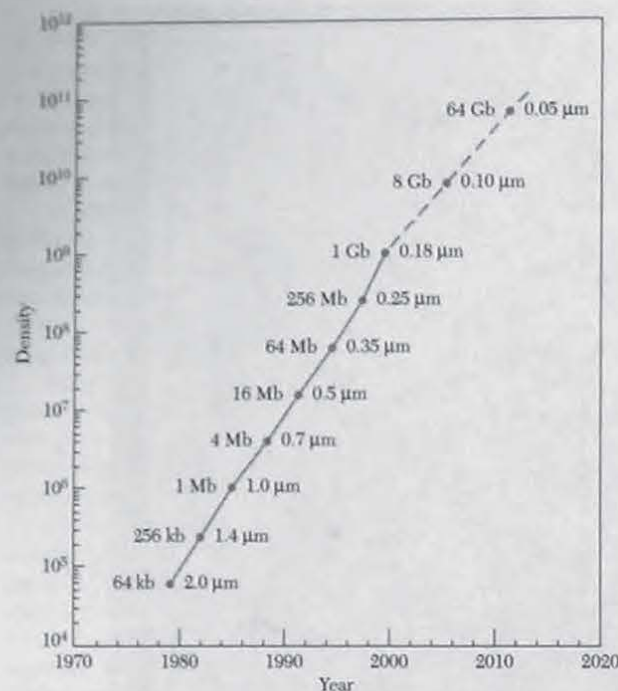


Figure 1.6 Exponential increase of dynamic random access memory density versus year based on the Semiconductor Industry Association roadmap.⁵⁰

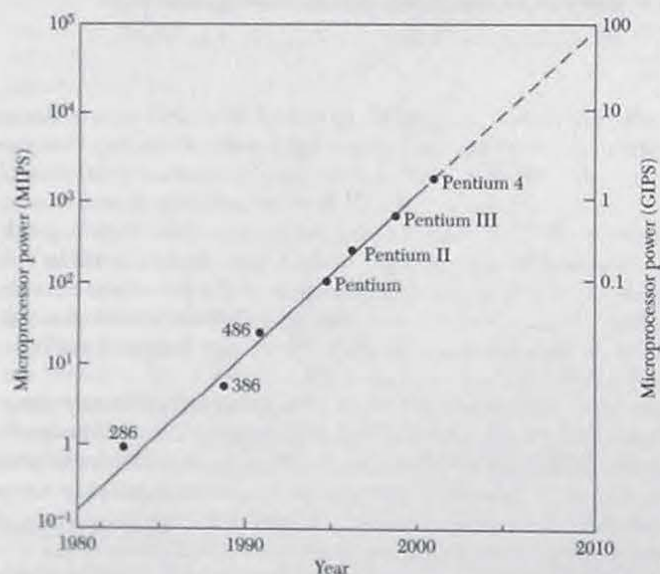


Figure 1.7 Exponential increase of microprocessor computational power versus year.

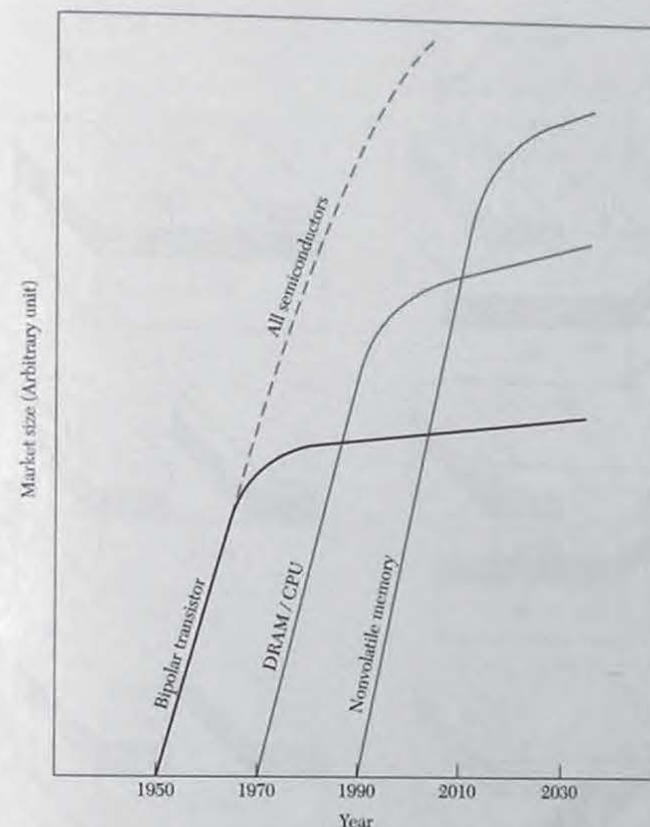


Figure 1.8 Growth curves for different technology drivers.⁵¹

Figure 1.8 illustrates the growth curves for different technology drivers.⁵¹ At the beginning of the modern electronic era (1950–1970), the bipolar transistor was the technology driver. From 1970 to 1990, the DRAM and the microprocessor based on MOS devices were the technology drivers because of the rapid growth of personal computers and advanced electronic systems. Since 1990, nonvolatile semiconductor memory has been the technology driver, mainly because of the rapid growth of portable electronic systems.

► 1.4 BASIC FABRICATION STEPS

Today, planar technology is used extensively for IC fabrication. Figures 1.9 and 1.10 show the major steps of a planar process. These steps include oxidation, photolithography, etching, ion implantation, and metallization. This section describes these steps briefly. More detailed discussions can be found in Chapters 3 to 8; Chapter 9 describes the integration of these unit process steps to form semiconductor devices.

1.4.1 Oxidation

The development of a high-quality silicon dioxide (SiO_2) has helped to establish the dominance of Si in the production of commercial ICs. Generally, SiO_2 functions as an insulator

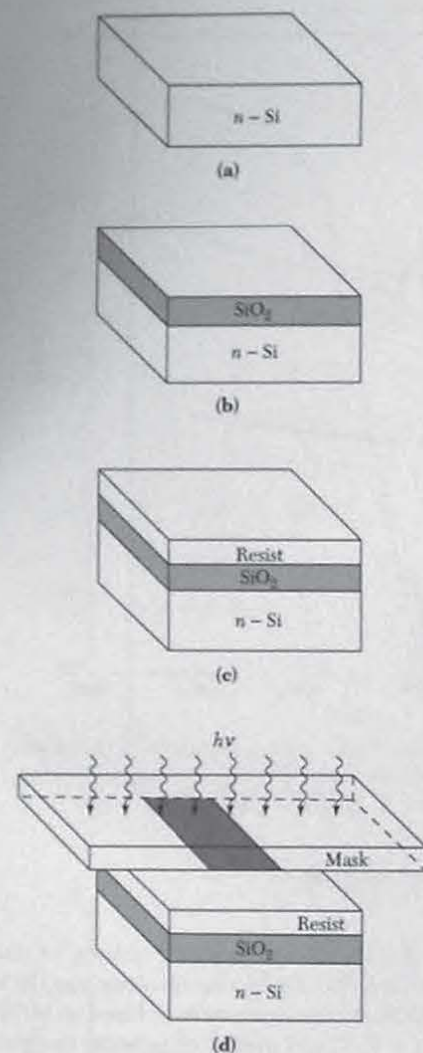


Figure 1.9 (a) A bare *n*-type Si wafer. (b) An oxidized Si wafer by dry or wet oxidation. (c) Application of resist. (d) Resist exposure through the mask.

in a number of device structures or as a barrier to diffusion or implantation during device fabrication. In the fabrication of a *p-n* junction (Fig. 1.9), the SiO₂ film is used to define the junction area.

There are two SiO₂ growth methods, dry and wet oxidation, depending on whether dry oxygen or water vapor is used. Dry oxidation is usually used to form thin oxides in a device structure because of its good Si-SiO₂ interface characteristics, whereas wet oxidation is used for thicker layers because of its higher growth rate. Figure 1.9a shows a section of a bare Si wafer ready for oxidation. After the oxidation process, a SiO₂ layer is formed all over the wafer surface. For simplicity, Figure 1.9b shows only the upper surface of an oxidized wafer. More details on oxidation may be found in Chapter 3.

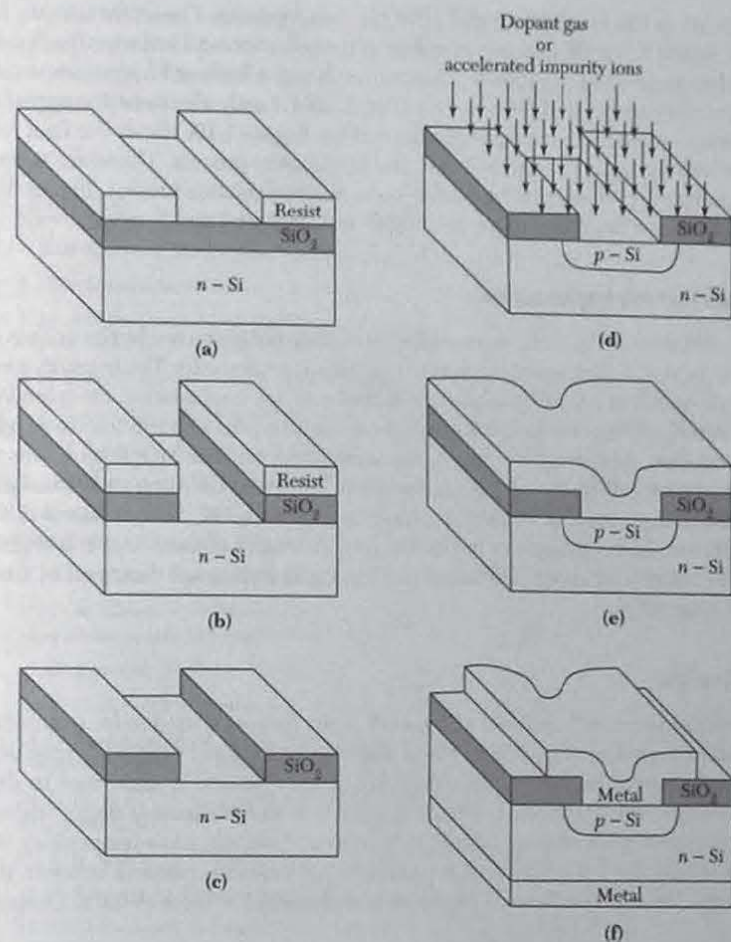


Figure 1.10 (a) The wafer after the development. (b) The wafer after SiO₂ removal. (c) The final result after a complete lithographic process. (d) A *p-n* junction is formed in the diffusion or implantation process. (e) The wafer after metallization. (f) A *p-n* junction after the complete processes.

1.4.2 Photolithography and Etching

Another technology, called *photolithography*, is used to define the geometry of the *p-n* junction. After the formation of SiO₂, the wafer is coated with an ultraviolet (UV) light-sensitive material called a *photoresist*, which is spun on the wafer surface by a high-speed spinner. Afterward (Fig. 1.9c), the wafer is baked at about 80°C to 100°C to drive the solvent out of the resist and to harden the resist for improved adhesion.

Figure 1.9d shows the next step, which is to expose the wafer through a patterned mask using a UV light source. The exposed region of the photoresist-coated wafer undergoes a chemical reaction depending on the type of resist. The area exposed to light becomes polymerized and difficult to remove in an etchant. The polymerized region remains when the wafer is placed in a developer, whereas the unexposed region (under the opaque area) dissolves and washes away.

Figure 1.10a shows the wafer after the development. The wafer is again baked to 120°C to 180°C for 20 minutes to enhance the adhesion and improve the resistance to the subsequent etching process. Then, an etch using buffered hydrofluoric acid (HF) removes the unprotected SiO₂ surface (Fig. 1.10b). Lastly, the resist is stripped away by a chemical solution or an oxygen plasma system. Figure 1.10c shows the final result of a region without oxide (a window) after the lithography process. The wafer is now ready for forming the *p-n* junction by a diffusion or ion implantation process. Photolithography and etching are described more thoroughly in Chapters 4 and 5, respectively.

1.4.3 Diffusion and Ion Implantation

In the diffusion method, the semiconductor surface not protected by the oxide is exposed to a source with a high concentration of opposite-type impurity. The impurity moves into the semiconductor crystal by solid-state diffusion. In ion implantation, the intended impurity is introduced into the semiconductor by accelerating the impurity ions to a high energy level and then implanting the ions in the semiconductor. The SiO₂ layer serves as a barrier to impurity diffusion or ion implantation. After the diffusion or implantation process, the *p-n* junction is formed, as shown in Figure 1.10d. Due to lateral diffusion of impurities or lateral straggle of implanted ions, the width of the *p*-region is slightly wider than the window opening. Diffusion and ion implantation are discussed in Chapters 6 and 7, respectively.

1.4.4 Metallization

After diffusion or ion implantation, a metallization process is used to form ohmic contacts and interconnections (Fig. 1.10e). Metal films can be formed by physical vapor deposition or chemical vapor deposition. The photolithography process is again used to define the front contact, which is shown in Figure 1.10f. A similar metallization step is performed on the back contact without using a lithography process. Normally, a low-temperature ($\leq 500^\circ\text{C}$) anneal would also be performed to promote low-resistance contacts between the metal layers and the semiconductor. Metallization is discussed in more detail in Chapter 8.

1.5 SUMMARY

Semiconductor devices have an enormous impact on our society and the global economy because they serve as the foundation of the largest industry in the world—the electronics industry.

This introductory chapter has presented a historical review of major semiconductor devices, from the first study of the metal-semiconductor contact in 1874 to the fabrication of an ultrasmall 15-nm MOSFET in 2001. Of particular importance are the invention of the bipolar transistor in 1947, which ushered in the modern electronics era; the development of the MOSFET in 1960, which is the most important device for integrated circuits; and the invention of the nonvolatile semiconductor memory in 1967, which has been the technology driver of the electronics industry since 1990.

This chapter also described key semiconductor technologies. The origins of these technologies can be traced back as far as two millennia. Of particular importance are the development of the lithographic photoresist in 1957, which established the basic pattern transfer process for semiconductor devices; the invention of the integrated circuit in 1959, which was seminal to the rapid growth of the microelectronics industry; and the developments of the DRAM in 1967 and the microprocessor in 1971, which constitute the two largest segments of the semiconductor industry.

In this book, each chapter deals with a key IC fabrication process step or sequence of steps. Each chapter is presented in a clear and coherent fashion without heavy reliance on the original literature. However, a few important papers are listed at the end of each chapter for reference and for further reading.

REFERENCES

1. 2000 *Electronic Market Data Book*, Electron. Ind. Assoc., Washington, DC, 2000.
2. 2000 *Semiconductor Industry Report*, Ind. Technol. Res. Inst., Hsinchu, Taiwan, 2000.
3. Most of the classic device papers are collected in S. M. Sze, Ed., *Semiconductor Devices: Pioneering Papers*, World Sci., Singapore, 1991.
4. K. K. Ng, *Complete Guide to Semiconductor Devices*, McGraw-Hill, New York, 1995.
5. F. Braun, "Über die Stromleitung durch Schwefelmetalle," *Ann. Phys. Chem.*, **153**, 556 (1874).
6. H. J. Round, "A Note on Carborundum," *Electron. World*, **19**, 309 (1907).
7. J. Bardeen and W. H. Brattain, "The Transistor, a Semiconductor Triode," *Phys. Rev.*, **71**, 230 (1948).
8. W. Shockley, "The Theory of *p-n* Junction in Semiconductors and *p-n* Junction Transistors," *Bell Syst. Tech. J.*, **28**, 435 (1949).
9. J. Ebers, "Four Terminal *p-n-p-n* Transistors," *Proc. IRE*, **40**, 1361 (1952).
10. D. M. Chapin, C. S. Fuller, and G. L. Pearson, "A New Silicon *p-n* Junction Photocell for Converting Solar Radiation into Electrical Power," *J. Appl. Phys.*, **25**, 676 (1954).
11. H. Kroemer, "Theory of a Wide-Gap Emitter for Transistors," *Proc. IRE*, **45**, 1535 (1957).
12. L. Esaki, "New Phenomenon in Narrow Germanium *p-n* Junctions," *Phys. Rev.*, **109**, 603 (1958).
13. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Surface Device," in *IRE Device Research Conference*, Pittsburgh, 1960. (The paper can be found in Ref. 3.)
14. B. Yu, et al., "15 nm Gate Length Planar CMOS Transistor," *IEEE IEDM Technical Digest*, Washington, DC, p. 937 (2001).
15. R. N. Hall, et al., "Coherent Light Emission from GaAs Junctions," *Phys. Rev. Lett.*, **9**, 366 (1962).
16. H. Kroemer, "A Proposed Class of Heterojunction Injection Lasers," *Proc. IEEE*, **51**, 1782 (1963).
17. I. Alferov and R. F. Kazarinov, "Semiconductor Laser with Electrical Pumping," U.S.S.R. Patent 181, 737 (1963).
18. J. B. Gunn, "Microwave Oscillations of Current in III-V Semiconductors," *Solid State Commun.*, **1**, 88 (1963).
19. R. L. Johnston, B. C. DeLoach, Jr., and B. G. Cohen, "A Silicon Diode Microwave Oscillator," *Bell Syst. Tech. J.*, **44**, 369 (1965).
20. C. A. Mead, "Schottky Barrier Gate Field Effect Transistor," *Proc. IEEE*, **54**, 307 (1966).
21. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.*, **46**, 1283 (1967).
22. K. Yano, et al., "Room Temperature Single-Electron Memory," *IEEE Trans. Electron Devices*, **41**, 1628 (1994).
23. W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," *Bell Syst. Tech. J.*, **49**, 587 (1970).
24. L. L. Chang, L. Esaki, and R. Tsu, "Resonant Tunneling in Semiconductor Double Barriers," *Appl. Phys. Lett.*, **24**, 593 (1974).
25. T. Mimura, et al., "A New Field-Effect Transistor with Selectively Doped GaAs/*n*-AlGa_{1-x} as Heterojunction," *Jpn. J. Appl. Phys.*, **19**, L225 (1980).
26. D. Shore, "Steel-Making in Ancient Africa," in I. Van Sertima, Ed., *Blacks in Science: Ancient and Modern*, New Brunswick, NJ: Transaction Books, 157 (1986).
27. M. Hefner, "The Photoresist Story," *J. Photo. Sci.*, **12**, 181 (1964).

28. J. Czechralski, "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle," *Z. Phys. Chem.*, **92**, 219 (1918).
29. F. W. Bridgman, "Certain Physical Properties of Single Crystals of Tungsten, Antimony, Bismuth, Tellurium, Cadmium, Zinc, and Tin," *Proc. Am. Acad. Arts Sci.*, **60**, 303 (1925).
30. H. Welker, "Über Neue Halbleitende Verbindungen," *Z. Naturforsch.*, **7a**, 744 (1952).
31. A. Fick, "Ueber Diffusion," *Ann. Phys. Lpz.*, **170**, 59 (1855).
32. W. G. Pfann, "Semiconductor Signal Translating Device," U.S. Patent 2,597,028 (1952).
33. J. Andrus, "Fabrication of Semiconductor Devices," U.S. Patent 3,122,817 (filed 1957; granted 1964).
34. C. J. Froesch and L. Derrick, "Surface Protection and Selective Masking during Diffusion in Silicon," *J. Electrochem. Soc.*, **104**, 547 (1957).
35. N. N. Sheftal, N. P. Kokorish, and A. V. Kraslov, "Growth of Single-Crystal Layers of Silicon and Germanium from the Vapor Phase," *Bull. Acad. Sci. U.S.S.R., Phys. Ser.*, **21**, 140 (1957).
36. W. Shockley, "Forming Semiconductor Device by Ionic Bombardment," U.S. Patent 2,787,564 (1958).
37. J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. Electron Devices*, **ED-23**, 648 (1976); U.S. Patent 3,138,743 (filed 1959; granted 1964).
38. R. N. Noyce, "Semiconductor Device-and-Lead Structure," U.S. Patent 2,981,877 (filed 1959; granted 1961).
39. J. A. Hoerni, "Planar Silicon Transistors and Diodes," *IRE Int. Electron Devices Meet.*, Washington, DC (1960).
40. F. M. Wanlass and C. T. Sah, "Nanowatt Logics Using Field-Effect Metal-Oxide Semiconductor Triodes," *Tech. Dig. IEEE Int. Solid-State Circuit Conf.*, p. 32 (1963).
41. R. M. Dennard, "Field Effect Transistor Memory," U.S. Patent 3,387,286 (filed 1967; granted 1968).
42. R. E. Kerwin, D. L. Klein, and J. C. Sarace, "Method for Making MIS Structure," U.S. Patent 3,475,234 (1969).
43. H. M. Manasevit and W. I. Simpson, "The Use of Metal-Organic in the Preparation of Semiconductor Materials. I. Epitaxial Gallium-V Compounds," *J. Electrochem. Soc.*, **116**, 1725 (1969).
44. S. M. Irving, K. E. Lemons, and G. E. Bobos, "Gas Plasma Vapor Etching Process," U.S. Patent 3,615,956 (1971).
45. A. Y. Cho, "Film Deposition by Molecular Beam Technique," *J. Vac. Sci. Technol.*, **8**, S31 (1971).
46. The inventors of the microprocessor are M. E. Hoff, F. Faggin, S. Mazor, and M. Shima. For a profile of M. E. Hoff, see *Portraits in Silicon* by R. Slater, p. 175, MIT Press, Cambridge, 1987.
47. R. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 237 (1982).
48. B. Davari, et al., "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polish (CMP)," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 61 (1989).
49. J. Paraszczak, et al., "High Performance Dielectrics and Processes for ULSI Interconnection Technologies," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 261 (1993).
50. *The International Technology Roadmap for Semiconductor*, Semiconductor Ind. Assoc., San Jose, 1999.
51. F. Masuoka, "Flash Memory Technology," *Proc. Int. Electron Devices Mater. Symp.*, 83, Hsinchu, Taiwan (1996).

Crystal Growth

The two most important semiconductors for discrete devices and integrated circuits are silicon and gallium arsenide. This chapter describes the common techniques for growing single crystals of these two semiconductors. The basic process flow from starting materials to polished wafers is shown in Figure 2.1. The starting materials—silicon dioxide for a silicon wafer, and gallium and arsenic for a gallium arsenide wafer—are chemically processed to form a high-purity polycrystalline semiconductor from which single crystals are grown. The single-crystal ingots are shaped to define the diameter of the material and are sawed into wafers. These wafers are etched and polished to provide smooth, specular surfaces upon which devices will be made. This chapter covers the following topics:

- Basic techniques to grow silicon and GaAs single-crystal ingots
- Wafer-shaping steps from ingots to polished wafers
- Wafer characterization in terms of its electrical and mechanical properties

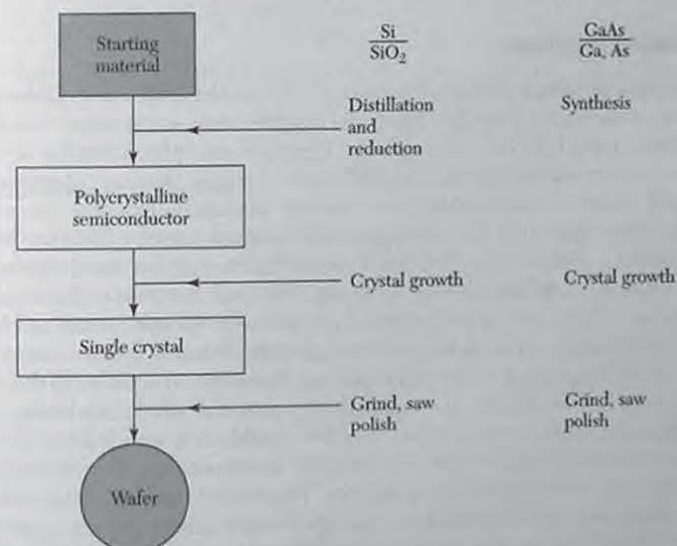


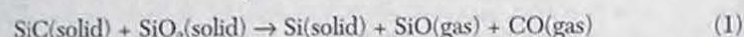
Figure 2.1 Process flow from starting material to polished wafer.

2.1 SILICON CRYSTAL GROWTH FROM THE MELT

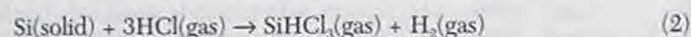
The basic technique for silicon crystal growth from the melt, which is material in liquid form, is the *Czochralski technique*. A substantial percentage (>90%) of silicon crystals for the semiconductor industry are prepared by the Czochralski technique, and so is virtually all the silicon used for fabricating integrated circuits.

2.1.1 Starting Material

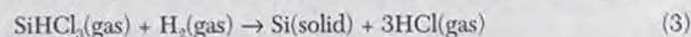
The starting material for silicon is a relatively pure form of sand (SiO_2) called *quartzite*. This is placed in a furnace with various forms of carbon (coal, coke, and wood chips). Although a number of reactions take place in the furnace, the overall reaction is



This process produces metallurgical-grade silicon with a purity of about 98%. Next, the silicon is pulverized and treated with hydrogen chloride (HCl) at 300°C to form trichlorosilane (SiHCl_3):



The trichlorosilane is a liquid at room temperature (it has a boiling point of 32°C). Fractional distillation of the liquid removes the unwanted impurities. The purified SiHCl_3 is then used in a hydrogen reduction reaction to prepare the electronic-grade silicon (EGS):



This reaction takes place in a reactor containing a resistance-heated silicon rod, which serves as the nucleation point for the deposition of silicon. EGS, a polycrystalline material of high purity, is the raw material used to prepare device-quality, single-crystal silicon. Pure EGS generally has impurity concentrations in the parts-per-billion range.¹

2.1.2 The Czochralski Technique

The Czochralski technique uses an apparatus called a *crystal puller*. A simplified version of this device is shown in Figure 2.2. The puller has three main components: (a) a furnace, which includes a fused-silica (SiO_2) crucible, a graphite susceptor, a rotation mechanism (clockwise as shown), a heating element, and a power supply; (b) a crystal-pulling mechanism, which includes a seed holder and a rotation mechanism (counterclockwise); and (c) an ambient control, which includes a gas source (such as argon), a flow control, and an exhaust system. In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps. Various sensors and feedback loops allow the control system to respond automatically, reducing operator intervention.

In the crystal-growing process, polycrystalline silicon (EGS) is placed in the crucible, and the furnace is heated above the melting temperature of silicon. A suitably oriented seed crystal (e.g., <111>) is suspended over the crucible in a seed holder. The seed is inserted into the melt. Part of it melts, but the tip of the remaining seed crystal still touches the liquid surface. It is then slowly withdrawn. Progressive freezing at the solid-liquid interface yields a large, single crystal. A typical pull rate is a few millimeters per minute. For large-diameter silicon ingots, an external magnetic field is applied to the basic Czochralski puller. The purpose of the external magnetic field is to control the concentration of defects, impurities, and oxygen.² Figure 2.3 shows a 300-mm (12 in.) and a 400-mm (16 in.) Czochralski-grown silicon ingot.

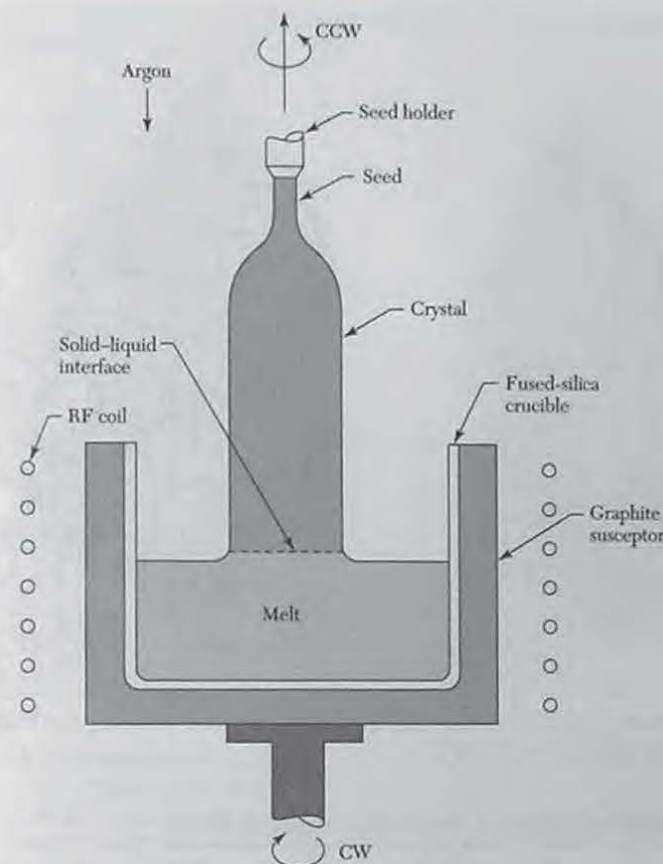


Figure 2.2 Czochralski crystal puller. CW, clockwise; CCW, counterclockwise.

2.1.3 Distribution of Dopant

In crystal growth, a known amount of dopant is added to the melt to obtain the desired doping concentration in the grown crystal. For silicon, boron and phosphorus are the most common dopants for *p*- and *n*-type materials, respectively.

As a crystal is pulled from the melt, the doping concentration incorporated into the crystal (solid) is usually different from the doping concentration of the melt (liquid) at the interface. The ratio of these two concentrations is defined as the *equilibrium segregation coefficient*, k_0 :

$$k_0 \equiv \frac{C_s}{C_l} \quad (4)$$

where C_s and C_l are, respectively, the equilibrium concentrations of the dopant in the solid and liquid near the interface. Table 2.1 lists values of k_0 for the commonly used dopants for silicon. Note that most values are below 1, which means that during growth the dopants are rejected into the melt. Consequently, the melt becomes progressively enriched with the dopant as the crystal grows.

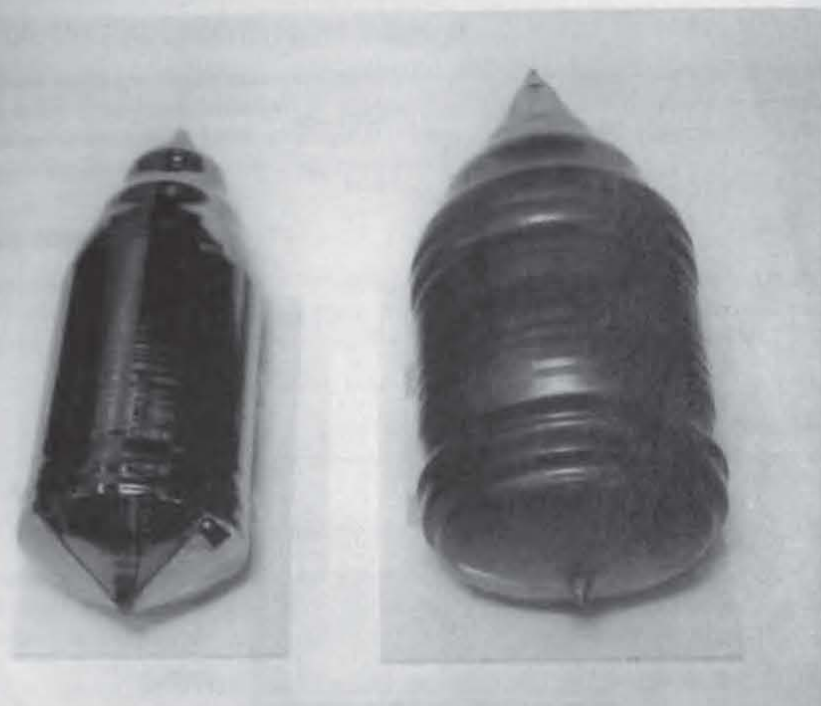


Figure 2.3 300-mm (12 in.) and 400-mm (16 in.) Czocharalski-grown silicon ingots. (Photo courtesy of Shin-Etsu Handotai Co., Tokyo.)

TABLE 2.1 Equilibrium Segregation Coefficients for Dopants in Silicon

Dopant	k_0	Type	Dopant	k_0	Type
B	8×10^{-1}	<i>p</i>	As	3.0×10^{-1}	<i>n</i>
Al	2×10^{-3}	<i>p</i>	Sb	2.3×10^{-2}	<i>n</i>
Ga	8×10^{-3}	<i>p</i>	Te	2.0×10^{-4}	<i>n</i>
In	4×10^{-4}	<i>p</i>	Li	1.0×10^{-2}	<i>n</i>
O	1.25	<i>n</i>	Cu	4.0×10^{-4}	— ^a
C	7×10^{-2}	<i>n</i>	Au	2.5×10^{-5}	— ^a
P	0.35	<i>n</i>			

^aDeep-lying impurity level.

Consider a crystal being grown from a melt having an initial weight M_0 with an initial doping concentration C_0 in the melt (i.e., the weight of the dopant per 1 g of melt). At a given point of growth when a crystal of weight M has been grown, the amount of dopant remaining in the melt (by weight) is S . For an incremental amount of the crystal with weight dM , the corresponding reduction of the dopant ($-dS$) from the melt is $C_s dM$, where C_s is the doping concentration in the crystal (by weight):

$$-dS = C_s dM \quad (5)$$

Now, the remaining weight of the melt is $M_0 - M$, and the doping concentration in the liquid (by weight), C_l , is given by

$$C_l = \frac{S}{M_0 - M} \quad (6)$$

Combining Eqs. 5 and 6 and substituting $C_s/C_l = k_0$ yields

$$\frac{dS}{S} = -k_0 \left(\frac{dM}{M_0 - M} \right) \quad (7)$$

Given the initial weight of the dopant, $C_0 M_0$, we can integrate Eq. 7:

$$\int_{C_0 M_0}^S \frac{dS}{S} = k_0 \int_0^M \frac{-dM}{M_0 - M} \quad (8)$$

Solving Eq. 8 and combining with Eq. 6 gives

$$C_s = k_0 C_0 \left(1 - \frac{M}{M_0} \right)^{k_0 - 1} \quad (9)$$

Figure 2.4 illustrates the doping distribution as a function of the fraction solidified (M/M_0) for several segregation coefficients.^{3,4} As crystal growth progresses, the composition initially at $k_0 C_0$ will increase continually for $k_0 < 1$ and decrease continually for $k_0 > 1$. When $k_0 \equiv 1$, a uniform impurity distribution can be obtained.

EXAMPLE 1

A silicon ingot, which should contain 10^{16} boron atoms/cm³, is to be grown by the Czochralski technique. What concentration of boron atoms should be in the melt to give the required concentration in the ingot? If the initial load of silicon in the crucible is 60 kg, how many grams of boron (atomic weight 10.8) should be added? The density of molten silicon is 2.53 g/cm³.

SOLUTION Table 2.1 shows that the segregation coefficient k_0 for boron is 0.8. We assume that $C_s = k_0 C_l$ throughout the growth. Thus, the initial concentration of boron in the melt should be

$$\frac{10^{16}}{0.8} = 1.25 \times 10^{16} \text{ boron atoms/cm}^3$$

Since the amount of boron concentration is so small, the volume of melt can be calculated from the weight of silicon. Therefore, the volume of 60 kg of silicon is

$$\frac{60 \times 10^3}{2.53} = 2.37 \times 10^4 \text{ cm}^3$$

The total number of boron atoms in the melt is

$$1.25 \times 10^{16} \text{ atoms/cm}^3 \times 2.37 \times 10^4 \text{ cm}^3 = 2.96 \times 10^{20} \text{ boron atoms}$$

so that

$$\frac{2.96 \times 10^{20} \text{ atoms} \times 10.8 \text{ g/mol}}{6.02 \times 10^{23} \text{ atoms/mol}} = 5.31 \times 10^{-3} \text{ g of boron} = 5.31 \text{ mg of boron}$$

Note the small amount of boron needed to dope such a large load of silicon.

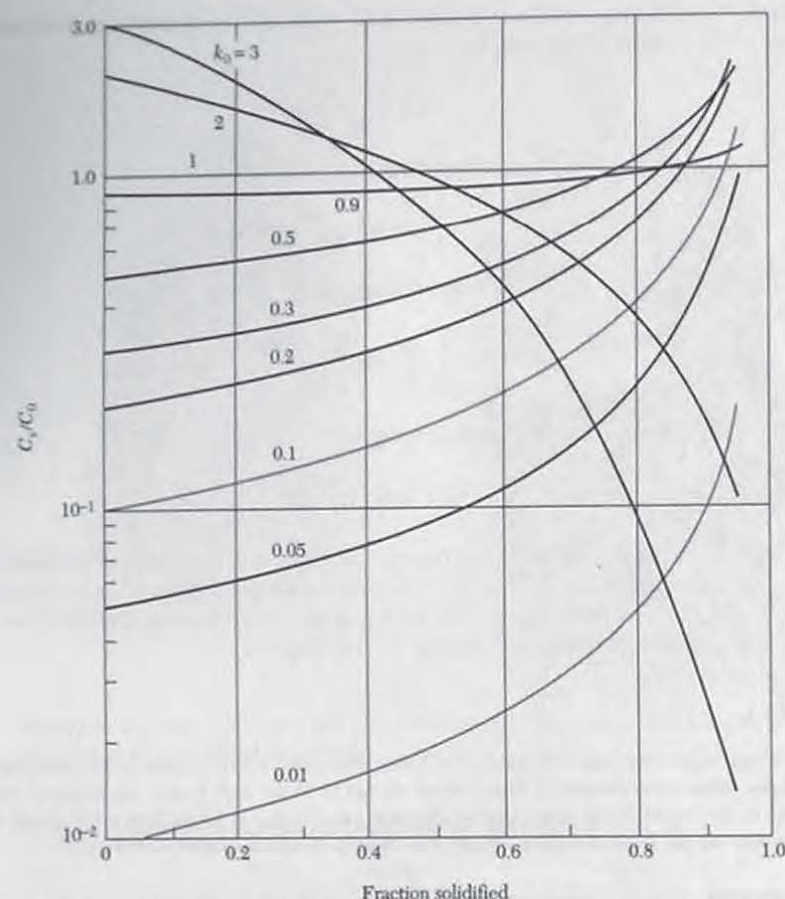


Figure 2.4 Curves for growth from the melt showing the doping concentration in a solid as a function of the fraction solidified.⁴

2.1.4 Effective Segregation Coefficient

While the crystal is growing, dopants are constantly being rejected into the melt (for $k_0 > 1$). If the rejection rate is higher than the rate at which the dopant can be transported away by diffusion or stirring, then a concentration gradient will develop at the interface, as illustrated in Figure 2.5. The segregation coefficient (given in Section 2.1.3) is $k_0 = C_s/C_l(0)$. We can define an effective segregation coefficient k_e , which is the ratio of C_s and the impurity concentration far away from the interface:

$$k_e \equiv \frac{C_s}{C_l} \quad (10)$$

Consider a small, virtually stagnant layer of melt with width δ in which the only flow is that required to replace the crystal being withdrawn from the melt. Outside this stagnant layer, the doping concentration has a constant value C_l . Inside the layer, the doping concentration can be described by the steady-state continuity equation:

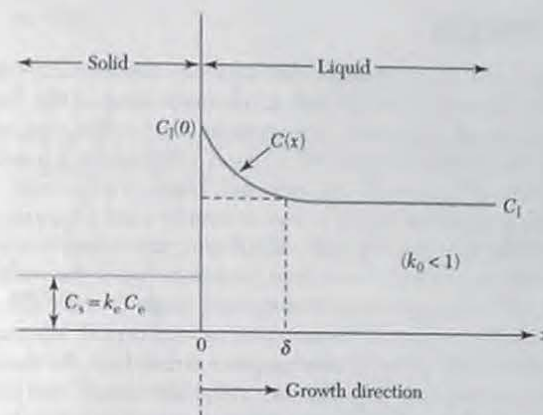


Figure 2.5 Doping distribution near the solid-melt interface.

$$0 = v \frac{dC}{dx} + D \frac{d^2C}{dx^2} \quad (11)$$

where D is the dopant diffusion coefficient in the melt, v is the crystal growth velocity, and C is the doping concentration in the melt.

The solution of Eq. 11 is

$$C = A_1 e^{-vx/D} + A_2 \quad (12)$$

where A_1 and A_2 are constants to be determined by the boundary conditions. The first boundary condition is that $C = C_l(0)$ at $x = 0$. The second boundary condition is the conservation of the total number of dopants; that is, the sum of the dopant fluxes at the interface must be zero. By considering the diffusion of dopant atoms in the melt (neglecting diffusion in the solid), we have

$$D \left(\frac{dC}{dx} \right)_{x=0} + [C_l(0) - C_l]v = 0 \quad (13)$$

Substituting these boundary conditions into Eq. 12 and noting that $C = C_l$ at $x = \delta$ gives

$$e^{-v\delta/D} = \frac{C_l - C_s}{C_l(0) - C_s} \quad (14)$$

Therefore,

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1 - k_0)e^{-v\delta/D}} \quad (15)$$

The doping distribution in the crystal is given by the same expression as in Eq. 9, except that k_0 is replaced by k_e . Values of k_e are larger than those of k_0 and can approach 1 for large values of the growth parameter $v\delta/D$. Uniform doping distribution ($k_e \rightarrow 1$) in the crystal can be obtained by employing a high pull rate and a low rotation speed (since δ is inversely proportional to the rotation speed). Another approach to achieve uniform doping is to add ultrapure polycrystalline silicon continuously to the melt so that the initial doping concentration is maintained.

2.2 SILICON FLOAT-ZONE PROCESS

The *float-zone process* can be used to grow silicon that has lower contamination than that normally obtained from the Czochralski technique. A schematic setup of the float-zone process is shown in Figure 2.6a. A high-purity polycrystalline rod with a seed crystal at the bottom is held in a vertical position and rotated. The rod is enclosed in a quartz envelope within which an inert atmosphere (argon) is maintained. During the operation, a small zone (a few centimeters in length) of the crystal is kept molten by a radio-frequency (RF) heater, which is moved from the seed upward so that this *floating zone* traverses the length of the rod. The molten silicon is retained by surface tension between the melting and growing solid-silicon faces. As the floating zone moves upward, single-crystal silicon freezes at the zone's retreating end and grows as an extension of the seed crystal. Materials with higher resistivities can be obtained from the float-zone process than from the Czochralski process because the former can be used to purify the crystal more easily. Furthermore, since no crucible is used in the float-zone process, there is no contamination from the crucible (as with Czochralski growth). At the present time, float-zone crystals are used mainly for high-power, high-voltage devices, where high-resistivity materials are required.

To evaluate the doping distribution of a float-zone process, consider a simplified model, as shown in Figure 2.6b. The initial uniform doping concentration in the rod is C_0 (by weight). L is the length of the molten zone at a distance x along the rod, A is the cross-sectional area of the rod, ρ_d is the specific density of silicon, and S is the amount of dopant present in the molten zone. As the zone traverses a distance dx , the amount of dopant added to it at its advancing end is $C_0 \rho_d A dx$, whereas the amount of dopant removed from it at the retreating end is $k_e S dx/L$, where k_e is the effective segregation coefficient. Thus,

$$dS = C_0 \rho_d A dx - \frac{k_e S}{L} dx = \left(C_0 \rho_d A - \frac{k_e S}{L} \right) dx \quad (16)$$

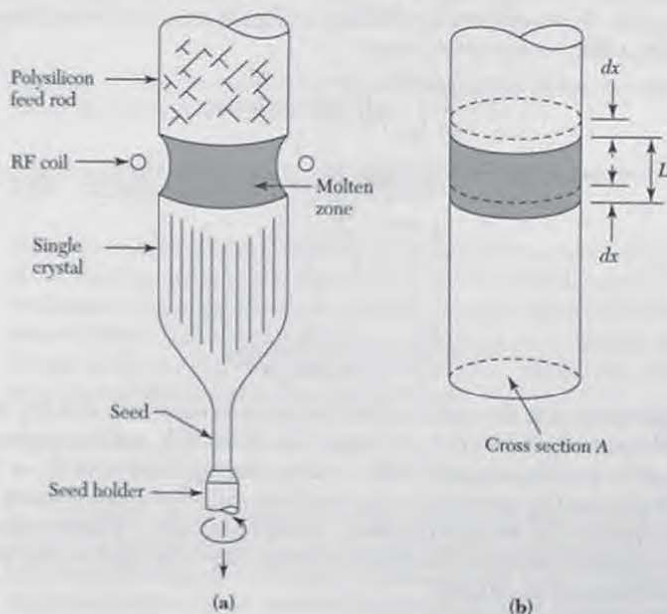


Figure 2.6 Float-zone process. (a) Schematic setup. (b) Simple model for doping evaluation.

so that

$$\int_0^x dx = \int_{S_0}^S \frac{dS}{C_0 \rho_d A - (k_e S/L)} \quad (16a)$$

where $S_0 = C_0 \rho_d AL$ is the amount of dopant in the zone when it was first formed at the front end of the rod. From Eq. 16a, we obtain

$$\exp\left(\frac{k_e x}{L}\right) = \frac{C_0 \rho_d A - (k_e S_0/L)}{C_0 \rho_d A - (k_e S/L)} \quad (17)$$

or

$$S = \frac{C_0 \rho_d AL}{k_e} \left[1 - (1 - k_e)^{-k_e x/L} \right] \quad (17a)$$

Since C_s (the doping concentration in the crystal at the retreating end) is given by $C_s = k_e (S/A\rho_d L)$, then

$$C_s = C_0 \left[1 - (1 - k_e)^{-k_e x/L} \right] \quad (18)$$

Figure 2.7 shows the doping concentration versus the solidified zone length for various values of k_e .

These two crystal growth techniques can also be used to remove impurities. A comparison of Figure 2.7 with Figure 2.4 shows that a single pass in the float-zone process does not produce as much purification as a single Czochralski growth. For example, for $k_0 = k_e = 0.1$, C_s/C_0 is smaller over most of the solidified ingot made by the Czochralski growth. However, multiple float-zone passes can be performed on a rod much more easily than a crystal can be grown, the end region cropped off, and regrown from the melt. Figure 2.8 shows the impurity distribution for an element with $k_e = 0.1$ after a number of successive passes of the zone along the length of the rod.⁴ Note that there is a substantial reduction of impurity concentration in the rod after each pass. Therefore, the float-zone process is ideally suited for crystal purification. This process is also called the *zone-refining technique*, which can provide a very high purity level of the raw material.

If it is desirable to dope the rod rather than purify it, consider the case in which all the dopants are introduced in the first zone ($S_0 = C_1 A \rho_d L$) and the initial concentration C_0 is negligibly small. Equation 17 gives

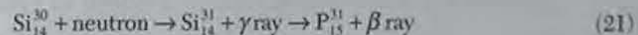
$$S_0 = S \exp\left(\frac{k_e x}{L}\right) \quad (19)$$

Since $C_s = k_e (S/A\rho_d L)$, we obtain the following from Eq. 19:

$$C_s = k_e C_1 e^{-k_e x/L} \quad (20)$$

Therefore, if $k_e x/L$ is small, C_s will remain nearly constant with distance except at the end that is last to solidify.

For certain switching devices, such as high-voltage thyristors, large chip areas are used—frequently an entire wafer for a single device. This size imposes stringent requirements on the uniformity of the starting material. To obtain homogeneous distribution of dopants, a float-zone silicon slice that has an average doping concentration well below the required amount is used. The slice is then irradiated with thermal neutrons. This process, called *neutron irradiation*, gives rise to fractional transmutation of silicon into phosphorus and dopes the silicon n-type:



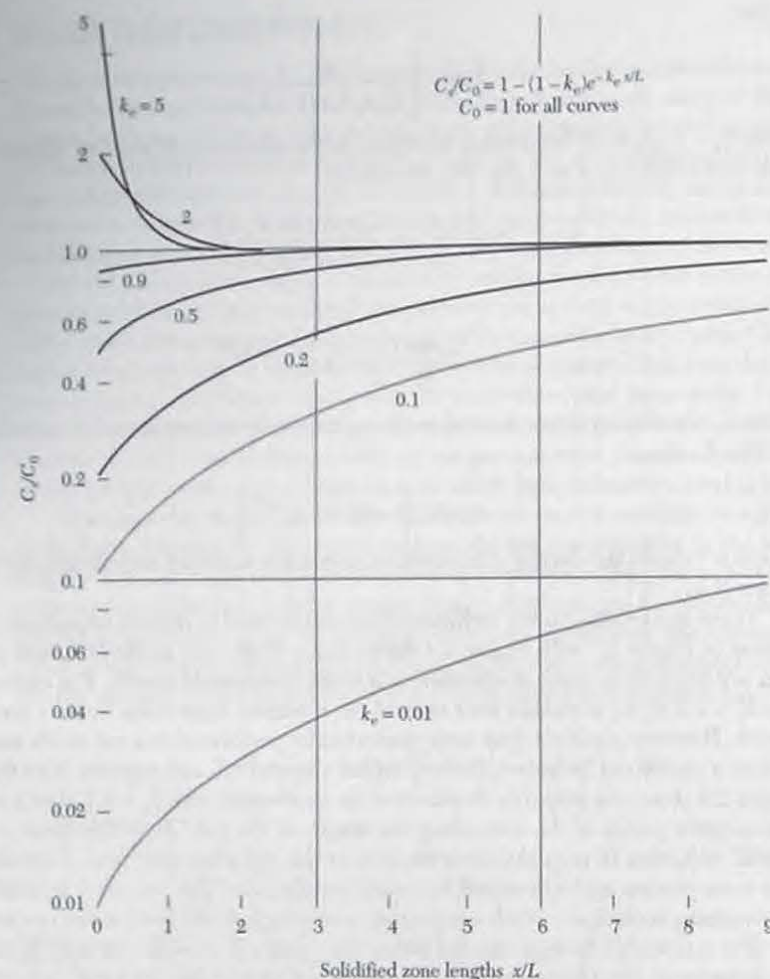


Figure 2.7 Curves for the float-zone process showing doping concentration in the solid as a function of solidified zone lengths.⁴

The half-life of the intermediate element Si_{14}^{31} is 2.62 hours. Because the penetration depth of neutrons in silicon is about 100 cm, doping is very uniform throughout the slice. Figure 2.9 compares the lateral resistivity distributions in conventionally doped silicon and in silicon doped by neutron irradiation.⁵ Note that the resistivity variations for the neutron-irradiated silicon are much smaller than those for the conventionally doped silicon.

2.3 GaAs CRYSTAL GROWTH TECHNIQUES

2.3.1 Starting Materials

The starting materials for the synthesis of polycrystalline gallium arsenide are the elemental, chemically pure gallium and arsenic. Because gallium arsenide is a combination of two materials, its behavior is different from that of a single material such as silicon. The behavior of a combination can be described by a phase diagram. A *phase* is a state

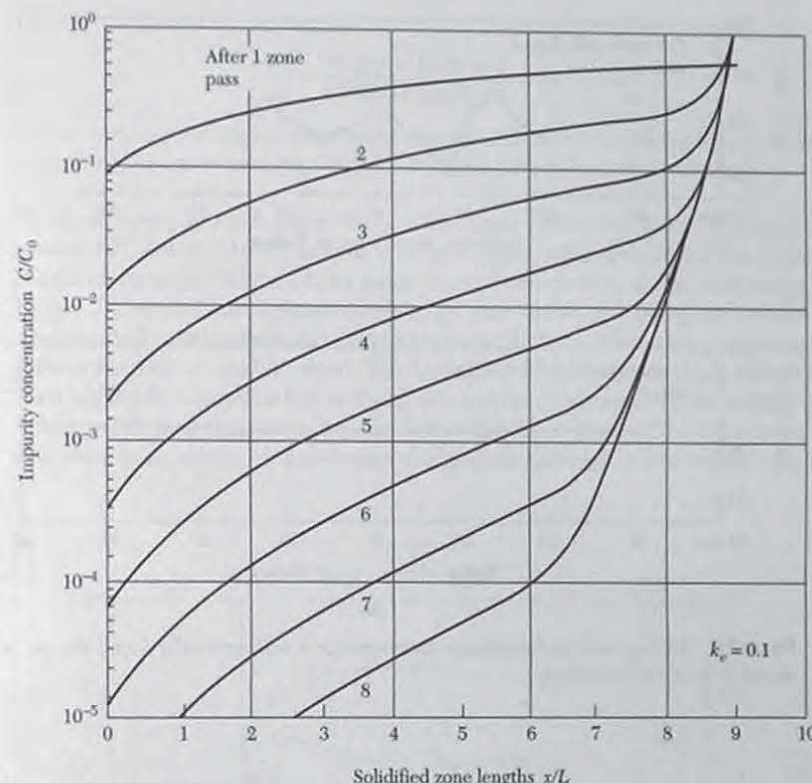


Figure 2.8 Relative impurity concentration versus zone length for a number of passes. L denotes the zone length.⁴

(e.g., solid, liquid, or gaseous) in which a material may exist. A *phase diagram* shows the relationship between the two components, gallium and arsenic, as a function of temperature.

Figure 2.10 shows the phase diagram of the gallium–arsenic system. The abscissa represents various compositions of the two components in terms of atomic percent (lower scale) or weight percent (upper scale).^{6,7} Consider a melt that is initially of composition x (e.g., 85 atomic percent arsenic, as shown in Fig. 2.10). When the temperature is lowered, its composition will remain fixed until the liquidus line is reached. At the point (T_1, x) , material of 50 atomic percent arsenic (i.e., gallium arsenide) will begin to solidify.

EXAMPLE 2

In Figure 2.10, consider a melt of initial composition C_m (weight percent scale) that is cooled from T_a (on the liquidus line) to T_b . Find the fraction of the melt that will be solidified.

SOLUTION At T_b , M_l is the weight of the liquid, M_s is the weight of the solid (i.e., GaAs), and C_l and C_s are the concentrations of dopant in the liquid and the solid, respectively. Therefore, the weights of arsenic in the liquid and solid are $M_l C_l$ and $M_s C_s$, respectively. Because the total arsenic weight is $(M_l + M_s)C_m$, we have

$$M_l C_l + M_s C_s = (M_l + M_s) C_m$$

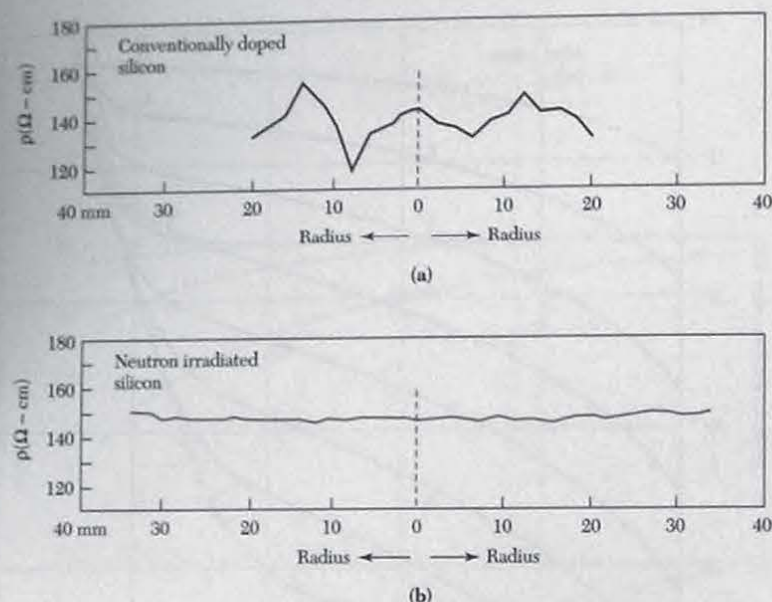


Figure 2.9 (a) Typical lateral resistivity distribution in a conventionally doped silicon. (b) Silicon doped by neutron irradiation.⁵

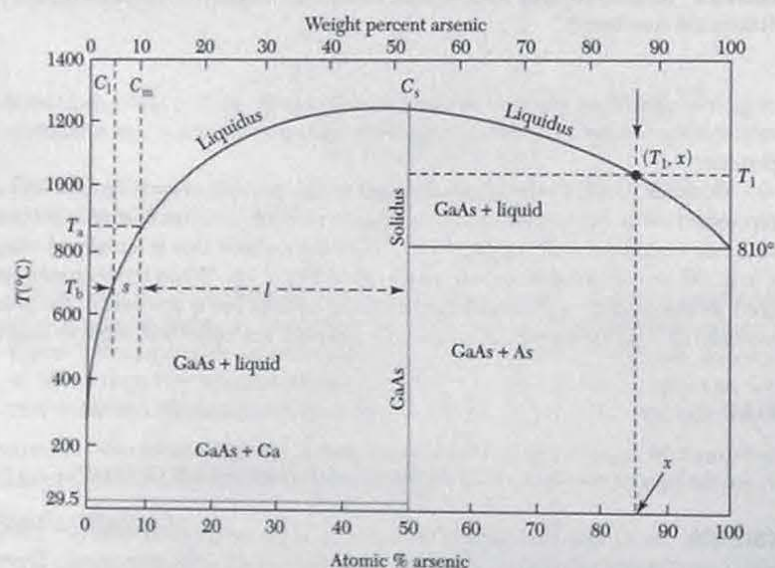


Figure 2.10 Phase diagram for the gallium-arsenic system.⁶

or

$$\frac{M_s}{M_l} = \frac{\text{Weight of GaAs at } T_b}{\text{Weight of liquid at } T_b} = \frac{C_m - C_l}{C_s - C_m} = \frac{s}{l}$$

where s and l are the lengths of the two lines measured from C_m to the liquidus and solidus line, respectively. As can be seen from Figure 2.10, about 10% of the melt is solidified.

Unlike silicon, which has a relatively low vapor pressure at its melting point (approximately 10^{-6} atm at 1412°C), arsenic has much higher vapor pressures at the melting point of gallium arsenide (1240°C). In its vapor phase, arsenic has As_2 and As_4 as its major species. Figure 2.11 shows the vapor pressures of gallium and arsenic along the liquidus curve.⁸ Also shown for comparison is the vapor pressure of silicon. The vapor pressure curves for gallium arsenide are double valued. The dashed curves are for arsenic-rich gallium arsenide melt (right side of liquidus line in Fig. 2.10), and the solid curves are for gallium-rich gallium arsenide melt (left side of liquidus line in Fig. 2.10). Because there is a larger amount of arsenic in an arsenic-rich melt than in a gallium-rich melt, more arsenic (As_2 and As_4)

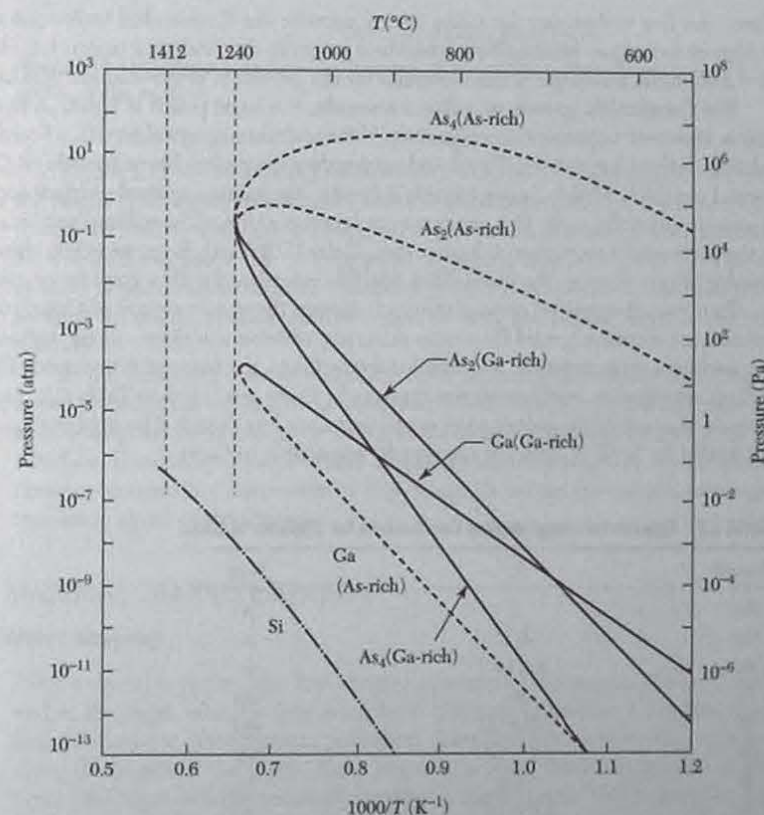


Figure 2.11 Partial pressure of gallium and arsenic over gallium arsenide as a function of temperature.⁸ Also shown is the partial pressure of silicon.

will be vaporized from the arsenic-rich melt, thus resulting in a higher vapor pressure. A similar argument can explain the higher vapor pressure of gallium in a gallium-rich melt. Note that long before the melting point is reached, the surface layers of liquid gallium arsenide may decompose into gallium and arsenic. Because the vapor pressures of gallium and arsenic are different, there is a preferential loss of the more volatile arsenic species, and the liquid becomes gallium rich.

To synthesize gallium arsenide, an evacuated, sealed quartz tube system with a two-temperature furnace is commonly used. The high-purity arsenic is placed in a graphite boat and heated to 610°C to 620°C, whereas the high-purity gallium is placed in another graphite boat and heated to slightly above the gallium arsenide melting temperature (1240–1260°C). Under these conditions, an overpressure of arsenic is established (a) to cause the transport of arsenic vapor to the gallium melt, converting it into gallium arsenide; and (b) to prevent decomposition of the gallium arsenide while it is being formed in the furnace. When the melt cools, a high-purity polycrystalline gallium arsenide results. This serves as the raw material to grow single-crystal gallium arsenide.⁷

2.3.2 Crystal Growth Techniques

There are two techniques for GaAs crystal growth: the Czochralski technique and the Bridgman technique. Most gallium arsenide is grown by the Bridgman technique. However, the Czochralski technique is more popular for the growth of larger-diameter GaAs ingots.

For Czochralski growth of gallium arsenide, the basic puller is identical to that for silicon. However, to prevent decomposition of the melt during crystal growth, a liquid encapsulation method is employed. The liquid encapsulant is a molten boron trioxide (B_2O_3) layer about 1 cm thick. Molten boron trioxide is inert to the gallium arsenide surface and serves as a cap to cover the melt. This cap prevents decomposition of the gallium arsenide as long as the pressure on its surface is higher than 1 atm (760 Torr). Because boron trioxide can dissolve silicon dioxide, the fused-silica crucible is replaced with a graphite crucible.

To obtain the desired doping concentration in the grown crystal of GaAs, cadmium and zinc are commonly used for *p*-type materials, whereas selenium, silicon, and tellurium are used for *n*-type materials. For semiinsulating GaAs, the material is undoped. The equilibrium segregation coefficients for dopants in GaAs are listed in Table 2.2. Similar to those in Si, most of the segregation coefficients are less than 1. The expressions derived previously for Si (Eqs. 4 to 15) are equally applicable to GaAs.

TABLE 2.2 Equilibrium Segregation Coefficients for Dopants in GaAs

Dopant	k_0	Type
Be	3	<i>p</i>
Mg	0.1	<i>p</i>
Zn	4×10^{-1}	<i>p</i>
C	0.8	<i>n/p</i>
Si	1.85×10^{-1}	<i>n/p</i>
Ge	2.8×10^{-2}	<i>n/p</i>
S	0.5	<i>n</i>
Se	5.0×10^{-1}	<i>n</i>
Sn	5.2×10^{-2}	<i>n</i>
Te	6.8×10^{-2}	<i>n</i>
Cr	1.03×10^{-4}	Semiinsulating
Fe	1.0×10^{-3}	Semiinsulating

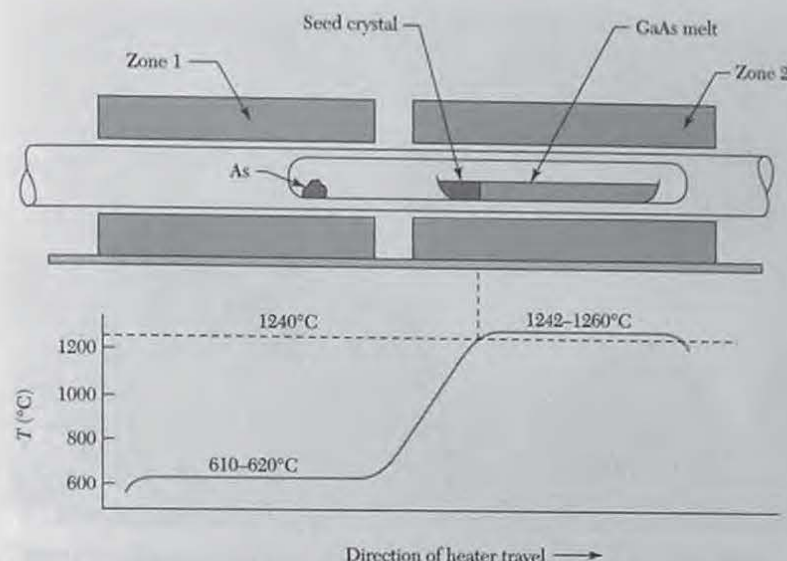


Figure 2.12 Bridgman technique for growing single-crystal gallium arsenide, and a temperature profile of the furnace.

Figure 2.12 shows a Bridgman system in which a two-zone furnace is used for growing single-crystal gallium arsenide. The left-hand zone is held at a temperature (approximately 610°C) to maintain the required overpressure of arsenic, whereas the right-hand zone is held just above the melting point of gallium arsenide (1240°C). The sealed tube is made of quartz, and the boat is made of graphite. In operation, the boat is loaded with a charge of polycrystalline gallium arsenide, with the arsenic kept at the other end of the tube.

As the furnace is moved toward the right, the melt cools at one end. Usually, there is a seed placed at the left end of the boat to establish a specific crystal orientation. The gradual freezing (solidification) of the melt allows a single crystal to propagate at the liquid–solid interface. Eventually, a single crystal of gallium arsenide is grown. The impurity distribution can be described essentially by Eqs. 9 and 15, where the growth rate is given by the traversing speed of the furnace.

► 2.4 MATERIAL CHARACTERIZATION

2.4.1 Wafer Shaping

After a crystal is grown, the first shaping operation is to remove the seed and the other end of the ingot, which is last to solidify.¹ The next operation is to grind the surface so that the diameter of the material is defined. After that, one or more flat regions are ground along the length of the ingot. These regions, or *flats*, mark the specific crystal orientation of the ingot and the conductivity type of the material. The largest flat, the *primary flat*, allows a mechanical locator in automatic processing equipment to position the wafer and to orient the devices relative to the crystal. Other smaller flats, called *secondary flats*, are ground to identify the orientation and conductivity type of the crystal, as shown in Figure 2.13. For crystals with diameters equal to or larger than 200 mm, no flats are ground. Instead, a small groove is ground along the length of the ingot.

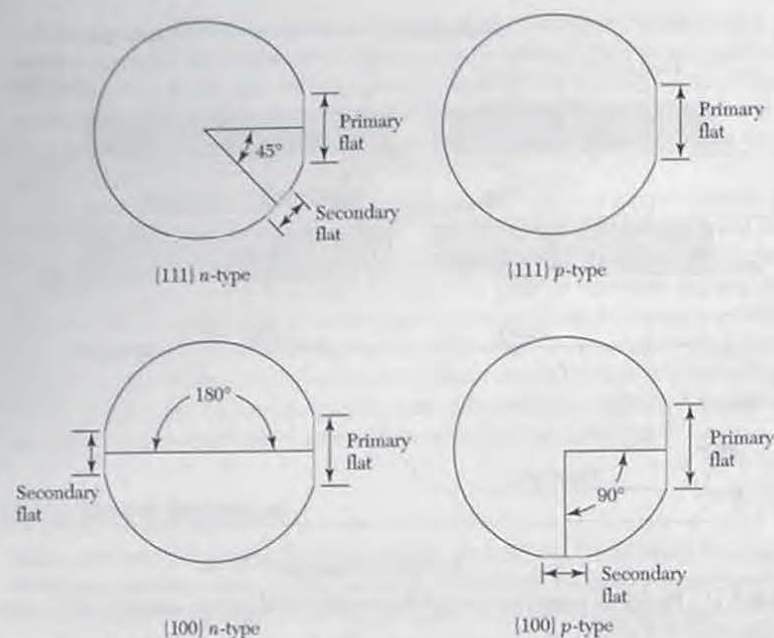


Figure 2.13 Identifying flats on a semiconductor wafer.

The ingot is then ready to be sliced by diamond saw into wafers. Slicing determines four wafer parameters: *surface orientation* (e.g., $\langle 111 \rangle$ or $\langle 100 \rangle$); *thickness* (e.g., 0.5–0.7 mm, depending on wafer diameter); *taper*, which is the wafer thickness variations from one end to another; and *bow*, which is the surface curvature of the wafer, measured from the center of the wafer to its edge.

After slicing, both sides of the wafer are lapped using a mixture of Al_2O_3 and glycerine to produce a typical flatness uniformity within 2 μm . The lapping operation usually leaves the surface and edges of the wafer damaged and contaminated. The damaged and contaminated regions can be removed by chemical etching (see Chapter 5). The final step of wafer shaping is polishing. Its purpose is to provide a smooth, specular surface where device features can be defined by photolithographic processes (see Chapter 4). Figure 2.14 shows 200-mm (8 in.) and 400-mm (16 in.) polished silicon wafers in cassettes. Table 2.3 shows the specifications for 125-, 150-, 200-, and 300-mm diameter polished silicon wafers from the Semiconductor Equipment and Materials Institute (SEMI). As mentioned previously, for large crystals (≥ 200 mm diameter), no flats are ground; instead, a groove is made on the edge of the wafer for positioning and orientation purposes.

Gallium arsenide is a more fragile material than silicon. Although the basic shaping operation of gallium arsenide is essentially the same as that for silicon, greater care must be exercised in gallium arsenide wafer preparation. The state of gallium arsenide technology is relatively primitive compared with that of silicon. However, the technology of group III-V compounds has advanced partly because of the advances in silicon technology.

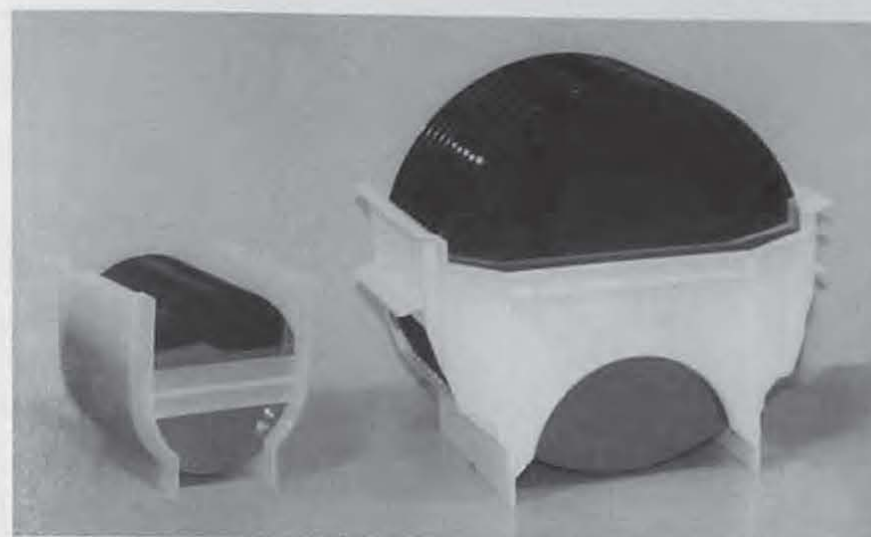


Figure 2.14 200-mm (8 in.) and 400-mm (16 in.) polished silicon wafers in cassettes. (Photo courtesy of Shin-Etsu Handotai Co., Tokyo.)

TABLE 2.3 Specifications for Polished Monocrystalline Silicon Wafers

Parameter	125 mm	150 mm	200 mm	300 mm
Diameter (mm)	125 $\pm$ 1	150 $\pm$ 1	200 $\pm$ 1	300 $\pm$ 1
Thickness (mm)	0.6–0.65	0.65–0.7	0.715–0.735	0.755–0.775
Primary flat length (mm)	40–45	55–60	NA	NA
Secondary flat length (mm)	25–30	35–40	NA	NA
Bow (μm)	70	60	30	< 30
Total thickness variation (μm)	65	50	10	< 10
Surface orientation	(100) $\pm$ 1°	Same	Same	Same
	(111) $\pm$ 1°	Same	Same	Same

NA, not available.

2.4.2 Crystal Characterization

Crystal Defects

A real crystal (such as a silicon wafer) differs from the ideal crystal in important ways. It is finite; thus, surface atoms are incompletely bonded. Furthermore, it has defects, which strongly influence the electrical, mechanical, and optical properties of the semiconductor. There are four categories of defects: point defects, line defects, area defects, and volume defects.

Figure 2.15 shows several forms of *point defects*.¹⁹ Any foreign atom incorporated into the lattice at either a substitutional site [i.e., at a regular lattice site (Fig. 2.15a)] or an interstitial site [i.e., between regular lattice sites (Fig. 2.15b)] is a point defect. A missing atom in the lattice creates a vacancy, also considered a point defect (Fig. 2.15c).

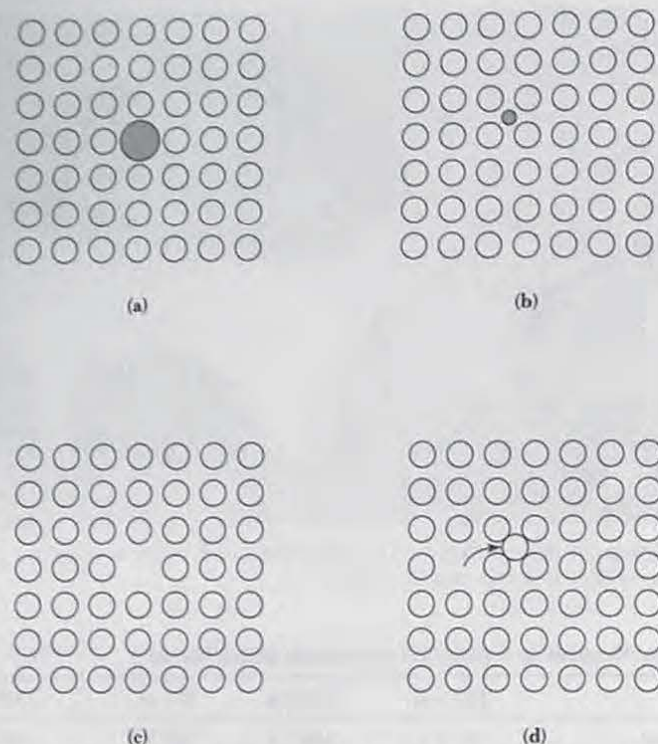


Figure 2.15 Point defects. (a) Substitutional impurity. (b) Interstitial impurity. (c) Lattice vacancy. (d) Frenkel-type defect.⁹

A host atom that is situated between regular lattice sites and adjacent to a vacancy is called a *Frenkel defect* (Fig. 2.15d). Point defects are particularly important subjects in the kinetics of oxidation and diffusion processes. These topics are considered in Chapters 3 and 6, respectively.

The next class of defects is the *line defect*, also called a *dislocation*.¹⁰ There are two types of dislocations: the edge and screw types. Figure 2.16a is a schematic representation of an edge dislocation in a cubic lattice. An extra plane of atoms AB is inserted into the lattice. The line of the dislocation would be perpendicular to the plane of the page. The screw dislocation may be considered as being produced by cutting the crystal part-way through and pushing the upper part one lattice spacing over, as shown in Figure 2.16b. Line defects in devices are undesirable because they act as precipitation sites for metallic impurities, which may degrade device performance.

Area defects represent a large-area discontinuity in the lattice. Typical defects are twins and grain boundaries. *Twinning* represents a change in the crystal orientation across a plane. A grain boundary is a transition between crystals having no particular orientational relationship to one another. Such defects appear during crystal growth. Another area defect is the *stacking fault*.⁹ In this defect, the stacking sequence of an atomic layer is interrupted. In Figure 2.17 the sequence of atoms in the stack is ABCABC.... If a part of layer C is missing, the defect is called an intrinsic stacking fault (Fig. 2.17a). If an extra plane A is inserted between layers B and C, it is an extrinsic stacking fault (Fig. 2.17b). Such defects may appear during crystal growth. Crystals having these area defects are not usable for integrated circuit manufacture and are discarded.

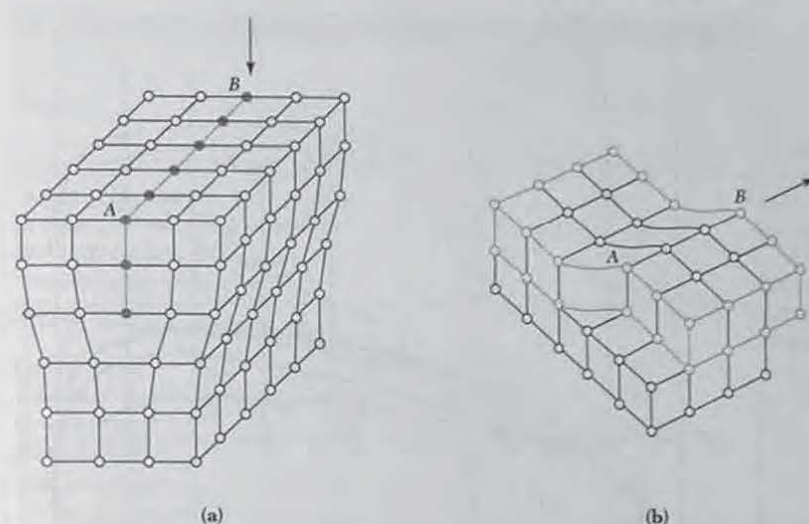


Figure 2.16 (a) Edge and (b) screw dislocation formation in cubic crystals.¹⁰

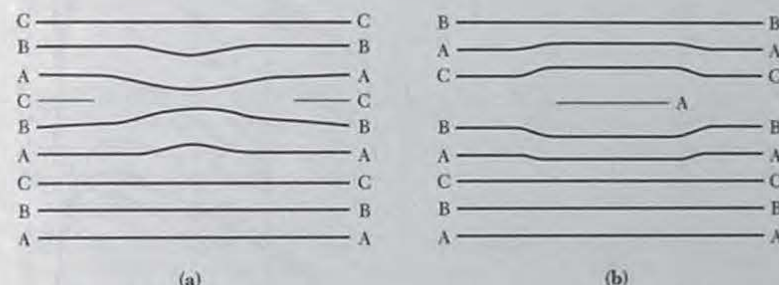


Figure 2.17 Stacking faults in semiconductor. (a) Intrinsic stacking fault. (b) Extrinsic stacking fault.⁹

Precipitates of impurities or dopant atoms make up the fourth class of defects: *volume defects*. These defects arise because of the inherent solubility of the impurity in the host lattice. There is a specific concentration of impurity that the host lattice can accept in a solid solution of itself and the impurity. Figure 2.18 shows solubility versus temperature for a variety of elements in silicon.¹¹ The solubility of most impurities decreases with decreasing temperature. Thus, if an impurity is introduced to the maximum concentration allowed by its solubility at a given temperature, and the crystal is then cooled to a lower temperature, the crystal can only achieve an equilibrium state by precipitating the impurity atoms in excess of the solubility level. However, the volume mismatch between the host lattice and the precipitates results in dislocations.

Material Properties

Table 2.4 compares silicon characteristics and the requirements for integrated circuit technology having more than 10^7 components, which is referred to as *ultralarge-scale integration* (ULSI).^{12,13} The semiconductor material properties listed in Table 2.4 can be measured by various methods. The resistivity is measured by the *four-point probe* method,¹⁴

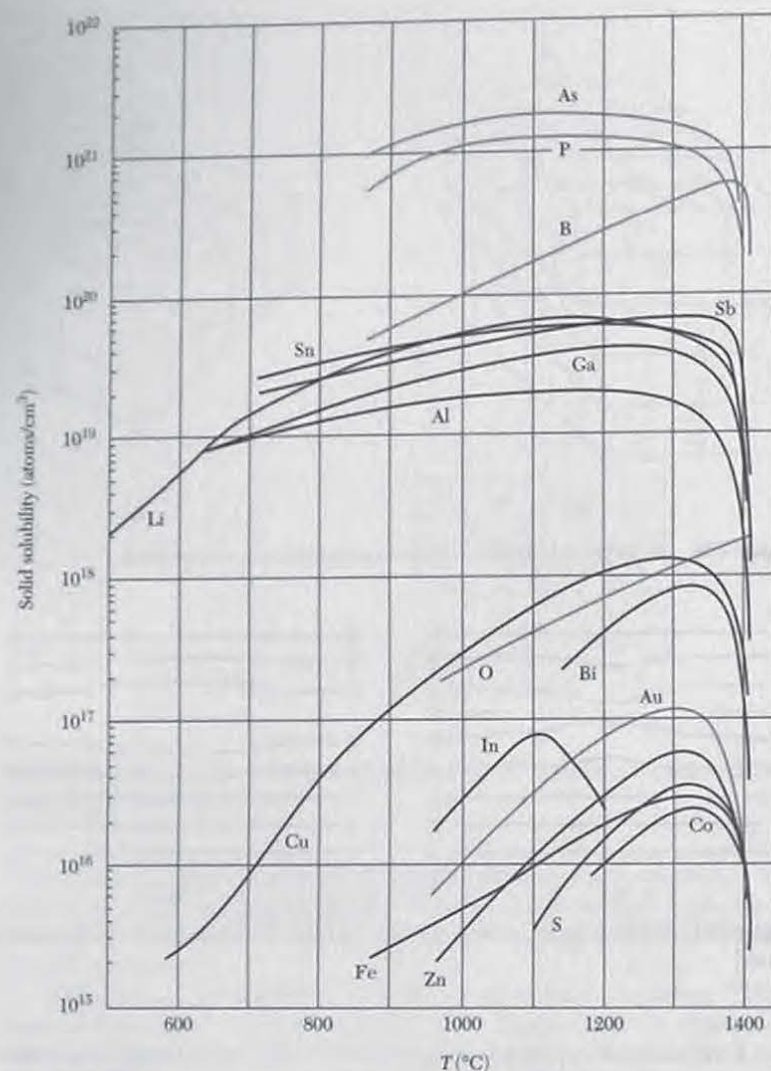


Figure 2.18 Solid solubilities of impurity elements in silicon.¹¹

and trace impurities such as oxygen and carbon in silicon can be analyzed by the secondary ion mass spectroscopic (SIMS) technique described in Chapter 6. Note that although current capabilities can meet most of the wafer specifications listed in Table 2.3, many improvements are needed to satisfy the stringent requirements for ULSI technology.¹³

The oxygen and carbon concentrations are substantially higher in Czochralski crystals than in float-zone crystals because of the dissolution of oxygen from the silica crucible and transport of carbon to the melt from the graphite susceptor during crystal growth. Typical carbon concentrations range from 10^{16} to about 10^{17} atoms/cm³, and carbon atoms in silicon occupy substitutional lattice sites. The presence of carbon is undesirable because

TABLE 2.4 Comparison of Silicon Material Characteristics and Requirements for ULSI

Property	Characteristics		Requirements for ULSI
	Czochralski	Float Zone	
Resistivity (phosphorus) <i>n</i> -type (ohm-cm)	1–50	1–300 and up	5–50 and up
Resistivity (antimony) <i>n</i> -type (ohm-cm)	0.005–10	—	0.001–0.02
Resistivity (boron) <i>p</i> -type (ohm-cm)	0.005–50	1–300	5–50 and up
Resistivity gradient (four-point probe) (%)	5–10	20	< 1
Minority carrier lifetime (μs)	30–300	50–500	300–1000
Oxygen (ppma)	5–25	Not detected	Uniform and controlled
Carbon (ppma)	1–5	0.1–1	< 0.1
Dislocation (before processing) (per cm ²)	≤ 500	≤ 500	≤ 1
Diameter (mm)	Up to 200	Up to 100	Up to 300
Slice bow (μm)	≤ 25	≤ 25	< 5
Slice taper (μm)	≤ 15	≤ 15	< 5
Surface flatness (μm)	≤ 5	≤ 5	< 1
Heavy-metal impurities (ppba)	≤ 1	≤ 0.01	< 0.001

ppma, parts per million atoms; ppba, parts per billion atoms.

it aids the formation of defects. Typical oxygen concentrations range from 10^{17} to 10^{18} atoms/cm³. Oxygen, however, has both deleterious and beneficial effects. It can act as a donor, distorting the resistivity of the crystal caused by intentional doping. On the other hand, oxygen in an interstitial lattice site can increase the yield strength of silicon.

In addition, the precipitates of oxygen due to the solubility effect can be used for gettering. *Gettering* is a general term meaning a process that removes harmful impurities or defects from the region in a wafer where devices are fabricated. When the wafer is subjected to high-temperature treatment (e.g., 1050°C in N₂), oxygen evaporates from the surface. This lowers the oxygen content near the surface. The treatment creates a defect-free (or *denuded*) zone for device fabrication, as shown in the inset of Figure 2.19.¹ Additional thermal cycles can be used to promote the formation of oxygen precipitates in the interior of the wafer for gettering of impurities. The depth of the defect-free zone depends on the time and temperature of the thermal cycle and on the diffusivity of oxygen in silicon. Measured results for the denuded zone are shown in Figure 2.19.¹ It is possible to obtain Czochralski crystals of silicon that are virtually free of dislocations.

Commercial melt-grown materials of gallium arsenide are heavily contaminated by the crucible. However, for photonic applications, most requirements call for heavily doped materials (between 10^{17} and 10^{18} cm⁻³). For integrated circuits or for discrete MESFET devices, undoped gallium arsenide can be used as the starting material with a resistivity of 10^9 Ω-cm. Oxygen is an undesirable impurity in GaAs because it can form a deep donor level, which contributes to a trapping charge in the bulk of the substrate and increases its resistivity. Oxygen contamination can be minimized by using graphite crucibles for melt growth. The dislocation content for Czochralski-grown gallium arsenide crystals is about two orders of magnitude higher than that for silicon. For Bridgman GaAs crystals, the dislocation density is about an order of magnitude lower than that for Czochralski-grown GaAs crystals.

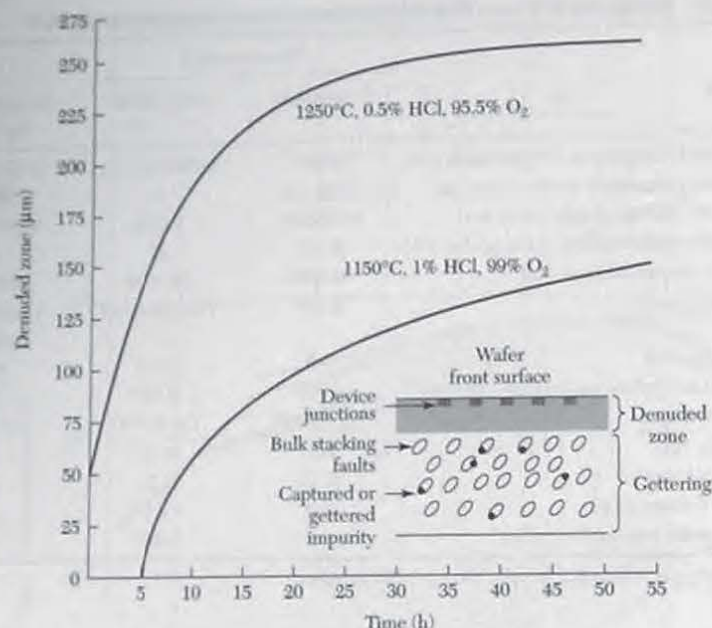


Figure 2.19 Denuded zone width for two sets of processing conditions. Inset shows a schematic of the denuded zone and gettering sites in a wafer cross section.¹

► 2.5 SUMMARY

Several techniques are available to grow single crystals of silicon and gallium arsenide. For silicon crystals, sand (SiO_2) is used to produce polycrystalline silicon, which then serves as the raw material in a Czochralski puller. A seed crystal with the desired orientation is used to grow a large ingot from the melt. Over 90% of silicon crystals are prepared by this technique. During crystal growth, the dopant in the crystal will redistribute. A key parameter is the segregation coefficient, that is, the ratio of the dopant concentration in the solid to that in the melt. Since most of the coefficients are less than 1, the melt becomes progressively enriched with the dopant as the crystal grows.

Another growth technique for silicon is the float-zone process. It offers lower contamination than that normally obtained from the Czochralski technique. Float-zone crystals are used mainly for high-power, high-voltage devices where high-resistivity materials are required.

To make GaAs, chemically pure gallium and arsenic are used as the starting materials that are synthesized to form polycrystalline GaAs. Single crystals of GaAs can be grown by the Czochralski technique. However, a liquid encapsulant (e.g., B_2O_3) is required to prevent decomposition of GaAs at the growth temperature. Another technique is the Bridgman process, which uses a two-zone furnace for gradual solidification of the melt.

After a crystal is grown, it usually goes through wafer-shaping operations to give an end product of highly polished wafers with a specified diameter, thickness, and surface orientation. For example, 200-mm silicon wafers for a MOSFET fabrication line should have a diameter of 200 ± 1 mm, a thickness of 0.725 ± 0.01 mm, and a surface orientation of $(100) \pm 1^\circ$. Wafers with diameters larger than 200 mm are being manufactured for future integrated circuits. Their specifications are listed in Table 2.3.

A real crystal has defects that influence the electrical, mechanical, and optical properties of the semiconductor. These defects are point defects, line defects, area defects, and volume defects. This chapter also discussed means to minimize such defects. For the more demanding ULSI applications, the dislocation density must be less than 1 per square centimeter. Other important requirements are listed in Table 2.4.

► REFERENCES

1. C. W. Pearce, "Crystal Growth and Wafer Preparation" and "Epitaxy," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
2. T. Abe, "Silicon Crystals for Giga-Bit Scale Integration," in T. S. Moss, Ed., *Handbook on Semiconductors*, Vol. 3, Elsevier Science B. V., Amsterdam/New York, 1994.
3. W. R. Runyan, *Silicon Semiconductor Technology*, McGraw-Hill, New York, 1965.
4. W. G. Pfann, *Zone Melting*, 2nd Ed., Wiley, New York, 1966.
5. E. W. Hass and M. S. Schnoller, "Phosphorus Doping of Silicon by Means of Neutron Irradiation," *IEEE Trans. Electron Devices*, **ED-23**, 803 (1976).
6. M. Hansen, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.
7. S. K. Ghandhi, *VLSI Fabrication Principles*, Wiley, New York, 1983.
8. J. R. Arthur, "Vapor Pressures and Phase Equilibria in the GaAs System," *J. Phys. Chem. Solids*, **28**, 2257 (1967).
9. B. El-Kareh, *Fundamentals of Semiconductor Processing Technology*, Kluwer Academic, Boston, 1995.
10. C. A. Wert and R. M. Thomson, *Physics of Solids*, McGraw-Hill, New York, 1964.
11. F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell Syst. Tech. J.*, **39**, 205 (1960); R. Hull, *Properties of Crystalline Silicon*, INSPEC, London, 1999.
12. Y. Matsushita, "Trend of Silicon Substrate Technologies for 0.25 μm Devices," *Proc. VLSI Technol. Workshop*, Honolulu (1996).
13. *The International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, San Jose, CA, 2001.
14. W. F. Beadle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Engineers*, Wiley, New York, 1985.

► PROBLEMS

SECTION 2.1: SILICON CRYSTAL GROWTH FROM THE MELT

1. Plot the doping distribution of arsenic at distances of 10, 20, 30, 40, and 45 cm from the seed in a silicon ingot 50 cm long that has been pulled from a melt with an initial doping concentration of 10^{17} cm^{-3} .
2. In silicon, the lattice constant is 5.43 Å. Assume a hard-sphere model. (a) Calculate the radius of a silicon atom. (b) Determine the density of silicon atoms in atoms/cm³. (c) Use the Avogadro constant to find the density of silicon.
3. Assuming that a 10-kg pure silicon charge is used, what is the amount of boron that must be added to get the boron-doped silicon to have a resistivity of 0.01 $\Omega\text{-cm}$ when one-half of the ingot is grown?
4. A silicon wafer 1 mm thick having a diameter of 200 mm contains 5.41 mg of boron uniformly distributed in substitutional sites. Find (a) the boron concentration in atoms/cm³ and (b) the average distance between boron atoms.
5. The seed crystal used in the Czochralski process is usually necked down to a small diameter (5.5 mm) as a means to initiate dislocation-free growth. If the critical-yield strength of silicon is $2 \times 10^6 \text{ g/cm}^2$, calculate the maximum length of a silicon ingot 200 mm in diameter that can be supported by such a seed.

- Plot the curve of the C/C_0 value for $k_0 = 0.05$ in the Czochralski technique.
- A Czochralski-grown crystal is doped with boron. Why is the boron concentration larger at the tail end of the crystal than at the seed end?
- Why is the impurity concentration larger in the center of the wafer than at its perimeter?

SECTION 2.2: SILICON FLOAT-ZONE PROCESS

- We use the float-zone process to purify a silicon ingot that contains a uniform gallium concentration of $5 \times 10^{16} \text{ cm}^{-3}$. One pass is made with a molten zone 2 cm long. Over what distance is the resulting gallium concentration below $5 \times 10^{15} \text{ cm}^{-3}$?
- From Eq. 18 find the C/C_0 value at $x/L = 1$ and 2 with $k_0 = 0.3$.
- If $p+n$ abrupt-junction diodes are fabricated using the silicon materials shown in Figure 2.9, find the percentage change of breakdown voltages for the conventionally doped silicon and the neutron-irradiated silicon.

SECTION 2.3: GAAS CRYSTAL GROWTH TECHNIQUES

- From Figure 2.10, if C_m is 20%, what fraction of the liquid will remain at T_b ?
- From Figure 2.11, explain why the GaAs liquid always becomes gallium rich.

SECTION 2.4: MATERIAL CHARACTERIZATION

- The equilibrium density of vacancy n_v is given by $N \exp(-E_v/kT)$, where N is the density of semiconductor atoms and E_v is the energy of formation. Calculate n_v in silicon at 27°C, 900°C, and 1200°C. Assume $E_v = 2.3 \text{ eV}$.
- Assume the energy of formation (E_f) of a Frenkel-type defect to be 1.1 eV and estimate the defect density at 27°C and 900°C. The equilibrium density of Frenkel-type defects is given by

$$n_f = \sqrt{NN'} e^{-E_f/2kT}$$

where N is the atomic density of silicon (cm^{-3}), and N' is the density of available interstitial sites (cm^{-3}) and is represented by $N' = 1 \times 10^{27} e^{-3.8 \text{ eV}/kT} \text{ cm}^{-3}$.

- How many chips of area 400 mm^2 can be placed on a wafer 300 mm in diameter? Explain your assumptions regarding the chip shape and unused wafer perimeter.

CHAPTER 3

Silicon Oxidation

Many different kinds of thin films are used to fabricate discrete devices and integrated circuits, including thermal oxides, dielectric layers, polycrystalline silicon, and metal films. Figure 3.1 shows a schematic view of a conventional silicon n -channel MOSFET that uses all four groups of films. The first important thin film from the thermal oxide group is the gate oxide layer, under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

This chapter covers the following topics:

- The thermal oxidation process used to form silicon dioxide (SiO_2)
- Impurity redistribution during oxidation
- Material properties and thickness measurement techniques for SiO_2 films

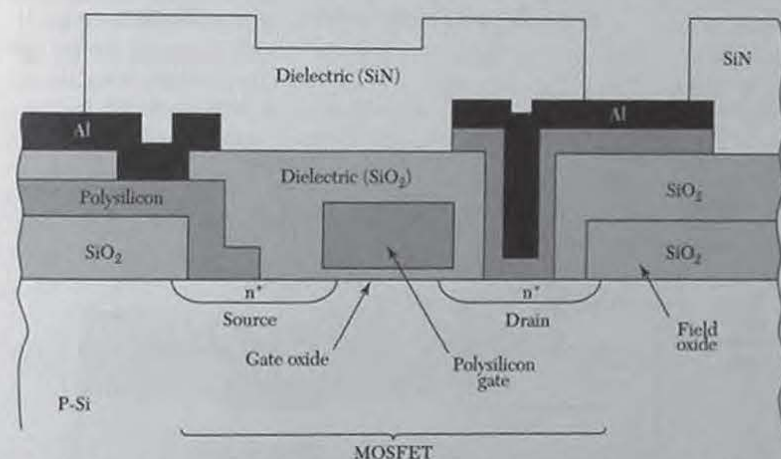


Figure 3.1 Schematic cross section of a metal-oxide-semiconductor field-effect transistor (MOSFET).

3.1 THERMAL OXIDATION PROCESS

Semiconductors can be oxidized by various methods. These include thermal oxidation, electrochemical anodization, and plasma-enhanced chemical vapor deposition (PECVD, see Chapter 8). Among these methods, thermal oxidation is by far the most important for silicon devices. It is a key process in modern silicon integrated circuit technology. For gallium arsenide, however, thermal oxidation results in generally nonstoichiometric films. The oxides provide poor electrical insulation and semiconductor surface protection; hence, these oxides are rarely used in gallium arsenide technology. Consequently, this chapter concentrates on the thermal oxidation of silicon.

The basic thermal oxidation apparatus is shown in Figure 3.2.¹ The reactor consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. The loading end of the furnace tube protrudes into a vertical flow hood where a filtered flow of air is maintained. Flow is directed as shown by the arrow in Figure 3.2. The hood reduces dust and particulate matter in the air surrounding the wafers and minimizes contamination during wafer loading. The oxidation temperature is generally in the range of 900°C to 1200°C, and the typical gas flow rate is about 1 L/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature so that the wafers will not warp due to sudden temperature change, to maintain the oxidation temperature to within $\pm 1^\circ\text{C}$, and to ramp the temperature down when oxidation is completed.

3.1.1 Kinetics of Growth

The following chemical reactions describe the thermal oxidation of silicon in oxygen (dry oxidation) or water vapor (wet oxidation):

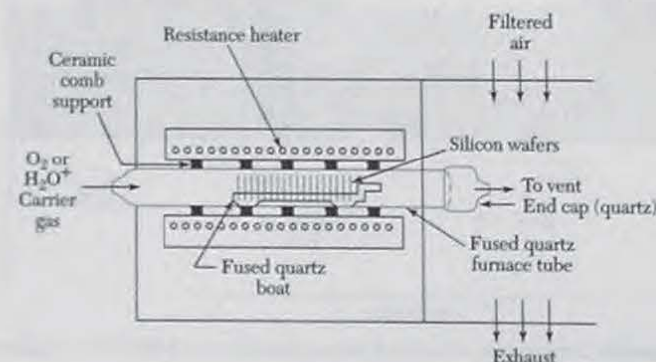
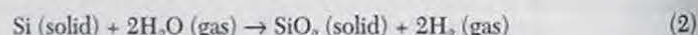
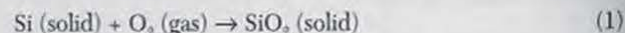


Figure 3.2 Schematic cross section of a resistance-heated oxidation furnace.

The silicon-silicon dioxide interface moves into the silicon during the oxidation process. This creates a fresh interface region, with surface contamination on the original silicon ending up on the oxide surface. The densities and molecular weights of silicon and silicon dioxide are used in the following example to show that growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Fig. 3.3).

EXAMPLE 1

If a silicon oxide layer of thickness x is grown by thermal oxidation, what is the thickness of silicon being consumed? The molecular weight of Si is 28.9 g/mol, and the density of Si is 2.33 g/cm³. The corresponding values for SiO₂ are 60.08 g/mol and 2.21 g/cm³.

SOLUTION The volume of 1 mol of silicon is

$$\frac{\text{Molecular weight of Si}}{\text{Density of Si}} = \frac{28.9 \text{ g/mol}}{2.33 \text{ g/cm}^3} = 12.06 \text{ cm}^3/\text{mol}$$

The volume of 1 mol of silicon dioxide is

$$\frac{\text{Molecular weight of SiO}_2}{\text{Density of SiO}_2} = \frac{60.08 \text{ g/mol}}{2.21 \text{ g/cm}^3} = 27.18 \text{ cm}^3/\text{mol}$$

Since 1 mol of silicon is converted to 1 mol of silicon dioxide,

$$\frac{\text{Thickness of Si} \times \text{area}}{\text{Thickness of SiO}_2 \times \text{area}} = \frac{\text{Volume of 1 mol of Si}}{\text{Volume of 1 mol of SiO}_2}$$

$$\frac{\text{Thickness of Si}}{\text{Thickness of SiO}_2} = \frac{12.06}{27.18} = 0.44$$

Thickness of silicon = 0.44 (thickness of SiO₂). For example, to grow a silicon dioxide layer of 100 nm, a layer of 44 nm of silicon is consumed.

The basic structural unit of thermally grown silicon dioxide is a silicon atom surrounded tetrahedrally by four oxygen atoms, as illustrated in Figure 3.4a.¹ The silicon-to-oxygen internuclear distance is 1.6 Å, and the oxygen-to-oxygen internuclear distance is 2.27 Å. These tetrahedra are joined at their corners by oxygen bridges in a variety of ways to form the various phases or structures of silicon dioxide (also called *silica*). Silica has several crystalline structures (e.g., quartz) and an amorphous structure. When silicon is thermally oxidized, the silicon dioxide structure is amorphous. Typically, amorphous silica has a density of 2.21 g/cm³, compared with 2.65 g/cm³ for quartz.

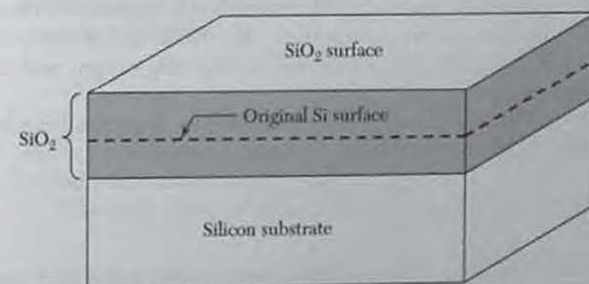


Figure 3.3 Growth of silicon dioxide by thermal oxidation.

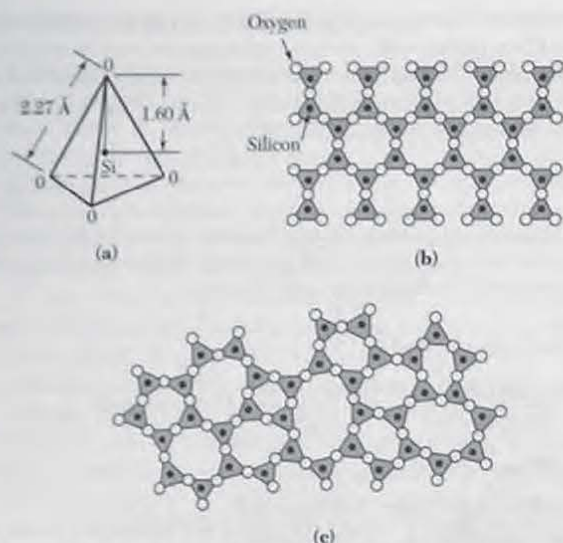


Figure 3.4 (a) Basic structural unit of silicon dioxide. (b) Two-dimensional representation of a quartz crystal lattice. (c) Two-dimensional representation of the amorphous structure of silicon dioxide.¹

The basic difference between the crystalline and amorphous structures is that the former is a periodic structure, extending over many molecules, whereas the latter has no periodic structure at all. Figure 3.4b is a two-dimensional schematic diagram of a quartz crystalline structure made up of rings with six silicon atoms. Figure 3.4c is a two-dimensional schematic diagram of an amorphous structure shown for comparison. In the amorphous structure, there is still a tendency to form characteristic rings with six silicon atoms. Note that the amorphous structure in Figure 3.4c is quite open because only 43% of the space is occupied by silicon dioxide molecules. The relatively open structure accounts for the lower density and allows a variety of impurities (such as sodium) to enter and diffuse readily through the silicon dioxide layer.

The kinetics of thermal oxidation of silicon can be studied based on a simple model illustrated in Figure 3.5.² A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm³ for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface. At 1000°C and at a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} molecules/cm³ for dry oxygen and 3×10^{19} molecules/cm³ for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D \frac{dC}{dx} \equiv \frac{D(C_0 - C_s)}{x} \quad (3)$$

where D is the diffusion coefficient of the oxidizing species, and x is the thickness of the oxide layer already present.

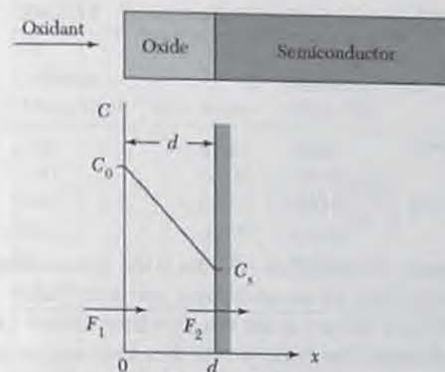


Figure 3.5 Basic model for the thermal oxidation of silicon.²

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction is proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = \kappa C_s \quad (4)$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. 3 and 4 gives

$$F = \frac{DC_0}{x + (D/\kappa)} \quad (5)$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm³ in the oxide, and we add one oxygen molecule (O_2) to each silicon dioxide molecule, whereas we add two water molecules (H_2O) to each SiO_2 molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm⁻³, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm⁻³). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)} \quad (6)$$

We can solve this differential equation subject to the initial condition, $x(0) = d_0$, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. 6 yields the general relationship for the oxidation of silicon,

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t + \tau) \quad (7)$$

where $\tau \equiv (d_0^2 + 2Dd_0/\kappa)C_1/2DC_0$, which represents a time coordinate shift to account for the initial oxide layer d_0 .

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{\kappa} \left[\sqrt{1 + \frac{2C_0\kappa^2(t + \tau)}{DC_1}} - 1 \right] \quad (8)$$

For small values of t , Eq. 8 reduces to

$$x \equiv \frac{C_0 \kappa}{C_1} (t + \tau) \quad (9)$$

and for larger values of t , it reduces to

$$x \equiv \sqrt{\frac{2DC_0}{C_1}} (t + \tau) \quad (10)$$

During the early stages of oxide growth, when surface reaction is the rate-limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon-silicon dioxide interface, and the reaction becomes diffusion limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation 7 is often written in a more compact form:

$$x^2 = Ax = B(t + \tau) \quad (11)$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$, and $B/A = \kappa C_0/C_1$. Using this form, Eqs. 9 and 10 can be written as

$$x = \frac{B}{A} (t + \tau) \quad (12)$$

for the linear region and as

$$x^2 = B(t + \tau) \quad (13)$$

for the parabolic region. For this reason, the term B/A is referred to as the *linear rate constant*, and B is the *parabolic rate constant*. Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \approx 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 25 nm. Thus, the use of Eq. 11 for dry oxidation on bare silicon requires a value for τ that can be generated using this initial thickness. Table 3.1 lists the values of the rate constants for wet oxidation of silicon, and Table 3.2 lists the values for dry oxidation.

The temperature dependence of the linear rate constant B/A is shown in Figure 3.6 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers.² The linear rate constant varies as $\exp(-E_a/kT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This closely agrees with the energy required to break silicon-silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related

TABLE 3.1 Rate Constants for Wet Oxidation of Silicon

Oxidation Temperature (°C)	A (μm)	Parabolic Rate Constant B (μm ² /h)	Linear Rate Constant B/A (μm/h)	τ (h)
1200	0.05	0.720	14.40	0
1100	0.11	0.510	4.64	0
1000	0.226	0.287	1.27	0
920	0.50	0.203	0.406	0

TABLE 3.2 Rate Constants for Dry Oxidation of Silicon

Oxidation Temperature (°C)	A (μm)	Parabolic Rate Constant B (μm ² /h)	Linear Rate Constant B/A (μm/h)	τ (h)
1200	0.040	0.045	1.12	0.027
1100	0.090	0.027	0.30	0.076
1000	0.165	0.0117	0.071	0.37
920	0.235	0.0049	0.0208	1.40
800	0.370	0.0011	0.0030	9.0
700	...	...	0.00026	81.0

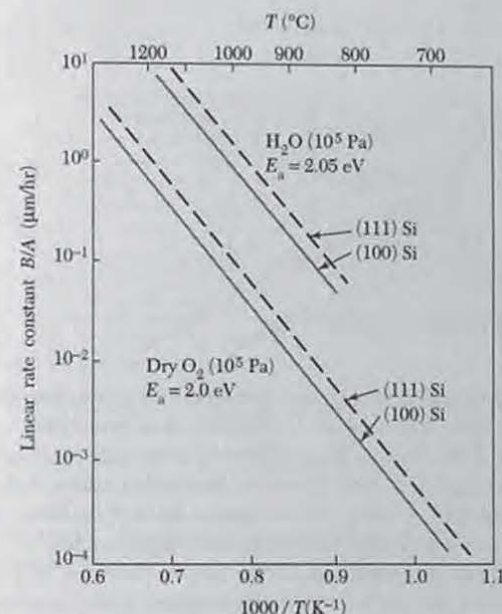


Figure 3.6 Linear rate constant versus temperature.²

to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation dependent. Because the density of available bonds on the (111) plane is higher than that on the (100) plane, the linear rate constant for (111) silicon is larger.

Figure 3.7 shows the temperature dependence of the parabolic rate constant B , which can also be described by $\exp(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because it is a measure of the diffusion process of the oxidizing species through a random network layer of amorphous silica.

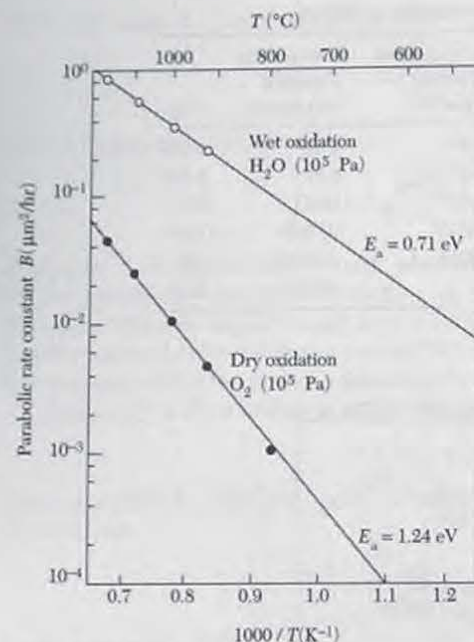


Figure 3.7 Parabolic rate constant versus temperature.²

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically ≤ 20 nm), dry oxidation is used. However, for thicker oxides such as field oxides (≥ 20 nm) in MOS integrated circuits, and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

Figure 3.8 shows the experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations.³ Under a given oxidation condition, the oxide thickness grown on a (111) substrate is larger than that grown on a (100) substrate because of the larger linear rate constant of the (111) orientation. Note that for a given temperature and time, the oxide film obtained using wet oxidation is about 5 to 10 times thicker than that using dry oxidation.

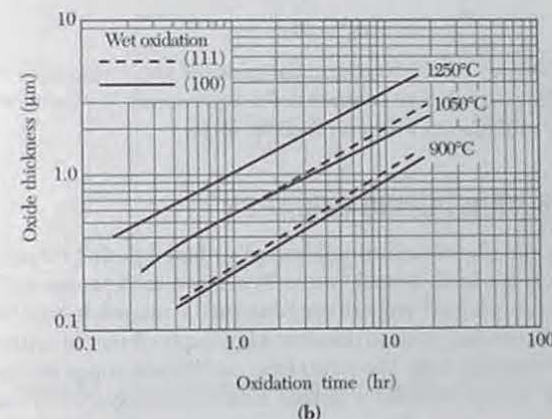
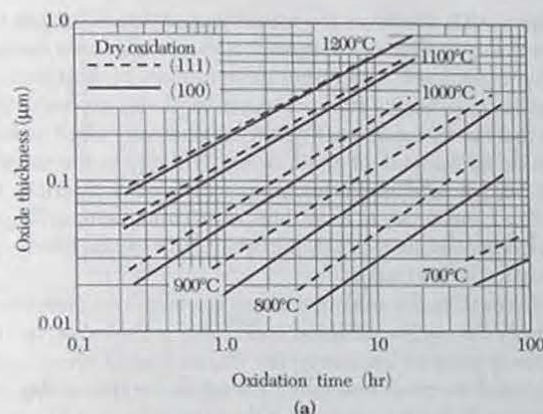


Figure 3.8 Experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (a) Growth in dry oxygen. (b) Growth in steam.³

Since $d_0 = 0.196$ μm from the first step, we have

$$\tau = (d_0^2 + 2Dd_0/k)C_1/2DC_0 = \frac{d_0^2 + Ad_0}{B} = 0.067 \text{ h}$$

The final desired thickness is $x = d_0 + 0.1$ $\mu\text{m} = 0.296$ μm . Using these parameters in Eq. 11, we obtain an additional oxidation time of

$$t = 0.76 \quad h = 4.53 \text{ min}$$

3.1.2 Thin Oxide Growth

Relatively slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. Various approaches to achieve such slow growth rates have been reported, including growth in dry O_2 at atmospheric pressure and lower temperatures (800°C to 900°C); growth at pressures lower than atmospheric pressure; growth in reduced partial pressures of O_2 by using a diluent inert gas, such as N_2 , Ar, or He, together with the gas containing the oxidizing species; and the use of composite oxide films, with the gate oxide films consisting of a layer of thermally grown SiO_2 and an overlayer of SiO_2 grown by

EXAMPLE 2

A silicon sample is oxidized in dry O_2 at 1200°C for 1 hour. (a) What is the thickness of the oxide grown? (b) How much additional time is required to grow 0.1 μm more oxide in wet O_2 at 1200°C ?

SOLUTION (a) From Table 3.2, the values of the rate constants for dry O_2 at 1200°C are

$$A = 0.04 \text{ } \mu\text{m} \quad B = 0.045 \text{ } \mu\text{m}^2/\text{h}$$

and $\tau = 0.027$ h. Using these parameters in Eq. 11, we obtain an oxide thickness of

$$x = 0.196 \text{ } \mu\text{m}$$

(b) From Table 3.1, the values of the rate constants for wet O_2 at 1200°C are

$$A = 0.05 \text{ } \mu\text{m} \quad B = 0.72 \text{ } \mu\text{m}^2/\text{h}$$

chemical vapor deposition (CVD). However, the mainstream approach for gate oxides 10 to 15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800°C to 900°C). With this approach, processing using modern vertical oxidation furnaces can grow reproducible, high-quality 10-nm oxides to within 0.1 nm across the wafer.

It was noted earlier that for dry oxidation, there is an apparently rapid oxidation that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model presented in Section 3.1.1 is not valid for dry oxidation with an oxide thickness less than or equal to 20 nm. For ULSI, the ability to grow thin (~5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important. This section briefly considers the growth mechanisms of such thin oxides.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress is reduced because of the viscous flow of silica, and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/k may be sufficiently small that we can neglect the term Ax in Eq. 11 and obtain

$$x^2 - d_0^2 = Bt \quad (14)$$

where d_0 is equal to $\sqrt{2DC_0\tau/C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously. We therefore expect the initial growth in dry oxidation to follow a parabolic form.

3.2 IMPURITY REDISTRIBUTION DURING OXIDATION

Dopant impurities near the silicon surface will be redistributed during thermal oxidation. The redistribution depends on several factors. When two solid phases are brought together, an impurity in one solid will redistribute between the two solids until it reaches equilibrium. This is similar to the previous discussion in Chapter 2 on impurity redistribution in crystal growth from the melt. The ratio of the equilibrium concentration of the impurity in the silicon to that in the silicon dioxide is called the *segregation coefficient* and is defined as

$$k = \frac{\text{Equilibrium concentration of impurity in silicon}}{\text{Equilibrium concentration of impurity in SiO}_2} \quad (15)$$

A second factor that influences impurity distribution is that the impurity may diffuse rapidly through the silicon dioxide and escape to the gaseous ambient. If the diffusivity of the impurity in silicon dioxide is large, this factor will be important. A third factor in the redistribution process is that the oxide is growing, and thus the boundary between the silicon and the oxide is advancing into the silicon as a function of time. The relative rate of this advance compared with the diffusion rate of the impurity through the oxide is important in determining the extent of the redistribution. Note that even if the segregation coefficient of an impurity equals unity, some redistribution of the impurity in the silicon will still take place. As indicated in Figure 3.3, the oxide layer will be about twice as thick as the silicon layer it replaced. Therefore, the same amount of impurity will now be distributed in a larger volume, resulting in depletion of the impurity from the silicon.

Four possible redistribution processes are illustrated in Figure 3.9.⁴ These processes can be classified into two groups. In one group, the oxide takes up the impurity (Figs. 3.9a and b for $k < 1$), and in the other the oxide rejects the impurity (Fig. 3.9c and d for $k > 1$). In each case, what happens depends on how rapidly the impurity can diffuse through the oxide. In group 1, the silicon surface is depleted of impurities; an example is boron, with k approximately equal to 0.3. Rapid diffusion of the impurity through the silicon dioxide

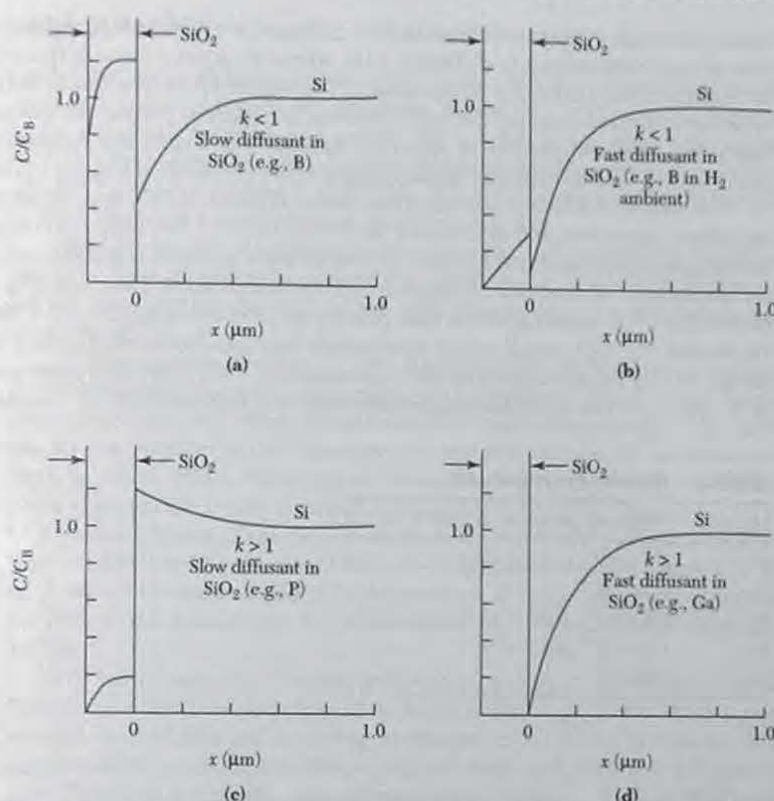


Figure 3.9 Four different cases of impurity redistribution in silicon due to thermal oxidation.⁴

increases the amount of depletion; an example is boron-doped silicon heated in a hydrogen ambient, because hydrogen in silicon dioxide enhances the diffusivity of boron. In group 2, k is greater than unity, so the oxide rejects the impurity. If diffusion of the impurity through the silicon dioxide is relatively slow, the impurity piles up near the silicon surface; an example is phosphorus, with k approximately equal to 10. When diffusion through the silicon dioxide is rapid, so much impurity may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity; an example is gallium, with k approximately equal to 20.

The redistributed dopant impurities in silicon dioxide are seldom electrically active. However, redistribution in silicon has an important effect on processing and device performance. For example, nonuniform dopant distribution will modify the interpretation of the measurements of interface trap properties, and the change of the surface concentration will modify the threshold voltage and device contact resistance.

3.3 MASKING PROPERTIES OF SILICON DIOXIDE

A silicon dioxide layer can also provide a selective mask against the diffusion of dopants at elevated temperatures, a very useful property in IC fabrication. Predeposition of dopants (see Chapter 6), whether it be by ion implantation, chemical diffusion, or spin-on techniques, typically results in a dopant source at or near the surface of the oxide. During a

subsequent high-temperature drive-in step, diffusion in oxide-masked regions must be slow enough with respect to diffusion in the silicon to prevent dopants from diffusing through the oxide mask to the silicon surface. The required thickness may be determined experimentally by measuring the oxide thickness necessary to prevent the inversion of a lightly doped silicon substrate of opposite conductivity at a particular temperature and time. Typically, oxides used for masking common impurities are 0.5 to 1.0 μm thick.

The values of diffusion constants for various dopants in SiO_2 depend on the concentration, properties, and structure of the oxide. Table 3.3 lists diffusion constants for various common dopants, and Figure 3.10 gives the oxide thickness required to mask boron and phosphorus as a function of diffusion time and temperature. Note that SiO_2 is much more effective for masking boron than phosphorus. Nevertheless, the diffusivities of P, Sb, As, and B in SiO_2 are all orders of magnitude less than their corresponding values in silicon, so they are all compatible with oxide masking. This is not true, however, for Ga or Al. Silicon nitride is used as an alternative masking material for these elements.

TABLE 3.3 Diffusion Constants in SiO_2

Dopants	Diffusion Constants at 1100°C (cm^2/s)
B	3.4×10^{-17} to 2.0×10^{-14}
Ga	5.3×10^{-11}
P	2.9×10^{-16} to 2.0×10^{-13}
As	1.2×10^{-16} to 3.5×10^{-15}
Sb	9.9×10^{-17}

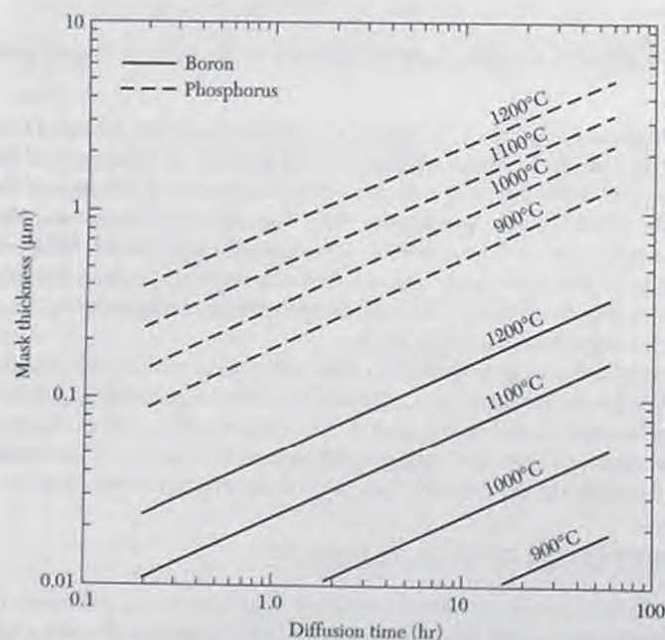


Figure 3.10 Thickness of silicon dioxide needed to mask boron and phosphorus diffusions as a function of diffusion time and temperature.

3.4 OXIDE QUALITY

Oxides used for masking are usually grown by wet oxidation. A typical growth cycle consists of a sequence of dry-wet-dry oxidations. Most of the growth in such a sequence occurs in the wet phase, since the SiO_2 growth rate is much higher when water is used as the oxidant. Dry oxidation, however, results in a higher-quality oxide that is denser and has a higher breakdown voltage (5–10 MV/cm). It is for these reasons that the thin gate oxides in MOS devices are usually formed using dry oxidation.

MOS devices are also affected by charges in the oxide and traps at the SiO_2 -Si interface. The basic classification of these traps and charges is shown in Figure 3.11. They are the interface trapped charge, fixed oxide charge, oxide trapped charge, and mobile ionic charge.³

Interface trapped charges (Q_{it}) are due to the SiO_2 -Si interface properties and are dependent on the chemical composition of this interface. The traps are located at the SiO_2 -Si interface, with energy states in the silicon forbidden bandgap. The interface trap density (i.e., number of interface traps per unit area and per eV) is orientation dependent. In silicon with a $\langle 100 \rangle$ crystal orientation, the interface trap density is about an order of magnitude smaller than that in the $\langle 111 \rangle$ orientation. Present-day MOS devices with thermally grown silicon dioxide on silicon have most of the interface trapped charges passivated by low-temperature (450°C) hydrogen annealing (see Chapter 7). The value of Q_{it} for $\langle 100 \rangle$ -oriented silicon can be as low as 10^{10} cm^{-2} , which amounts to about one interface trapped charge per 10^3 surface atoms. For $\langle 111 \rangle$ -oriented silicon, Q_{it} is about 10^{11} cm^{-2} .

The fixed charge (Q_f) is located within approximately 3 nm of the SiO_2 -Si interface. This charge is fixed and very difficult to charge or discharge. Generally, Q_f is positive and depends on oxidation and annealing conditions and on silicon orientation. It has been suggested that when the oxidation is stopped, some ionic silicon is left near the interface. These ions, along with uncompleted silicon bonds (e.g., Si-Si or Si-O bonds) at the surface, may result in the positive interface charge. Q_f can be regarded as a charge sheet

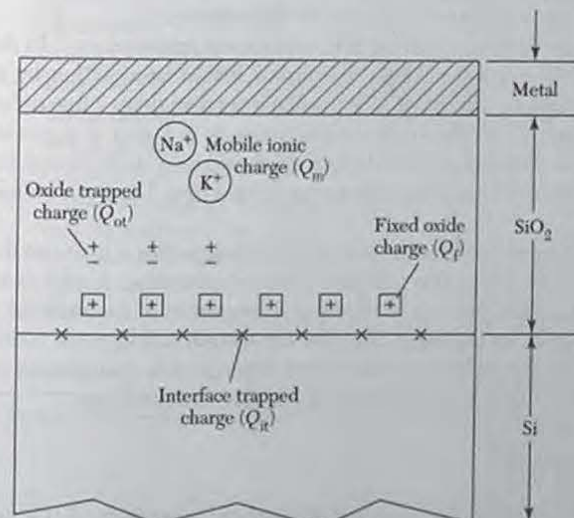


Figure 3.11 Terminology for the charges associated with thermally oxidized silicon.³

located at the SiO_2 -Si interface. Typical fixed oxide charge densities for a carefully treated SiO_2 -Si interface system are about 10^{10} cm^{-2} for a $\langle 100 \rangle$ surface and about $5 \times 10^{10} \text{ cm}^{-2}$ for a $\langle 111 \rangle$ surface. Because of the lower values of Q_n and Q_p , the $\langle 100 \rangle$ orientation is preferred for silicon MOSFETs.

Oxide trapped charges (Q_{ot}) are associated with defects in the silicon dioxide. These charges can be created, for example, by x-ray radiation or high-energy electron bombardment. The traps are distributed inside the oxide layer. Most process-related Q_{ot} can be removed by low-temperature annealing.

Mobile ionic charges (Q_m), which result from contamination from sodium or other alkali ions, are mobile within the oxide under raised temperatures (e.g., $>100^\circ\text{C}$) and high electric field operations. Trace contamination by alkali metal ions may cause stability problems in semiconductor devices operated under high-bias and high-temperature conditions. Under these conditions mobile ionic charges can move back and forth through the oxide layer and cause threshold voltage shifts. Therefore, special attention must be paid to the elimination of mobile ions in device fabrication. For example, the effects of sodium contamination can be reduced by adding chlorine during oxidation. Chlorine immobilizes the sodium ions. A small amount (6% or less) of anhydrous HCl in the oxidizing gas can accomplish this, but the presence of chlorine during dry oxidation increases both the linear and parabolic rate constants, leading to a higher growth rate.

► 3.5 OXIDE THICKNESS CHARACTERIZATION

Perhaps the simplest method for determining the thickness of an oxide is to compare the color of the wafer with a reference color chart, such as the one in Table 3.4.⁶ When an oxide-coated wafer is illuminated with white light perpendicular to the surface, the light penetrates the oxide and is reflected by the underlying silicon wafer. Constructive interference leads to enhancement of a certain wavelength of reflected light, and the color of the wafer corresponds to that wavelength. For example, a wafer with a 500-nm silicon dioxide layer will appear blue green.

Clearly, color chart comparisons are subjective and are therefore not the most accurate mechanism for determining oxide thickness. A more accurate measurement can be obtained using techniques such as profilometry or ellipsometry.

Profilometry is a very common method of film thickness measurement. In this technique, a step feature in the grown or deposited film is first created, either by masking during deposition or by etching afterward. The profilometer then drags a fine stylus across the film surface (see Fig. 3.12)⁷. When the stylus encounters a step, a signal variation indicates the step height. This information is then displayed on a chart recorder or CRT screen. Films of thicknesses of less than 100 nm to greater than 5 μm can be measured with this instrument.

Ellipsometry is another widely used measurement technique that is based on the polarization changes that occur when light is reflected from or transmitted through a medium. Changes in polarization are a function of the optical properties of the material (i.e., its complex refractive indices), its thickness, and the wavelength and angle of incidence of the light beam relative to the surface normal. These differences in polarization are measured by an ellipsometer, and the oxide thickness can then be calculated.

► 3.6 OXIDATION SIMULATION

As trends toward miniaturization continue and the scale of integration of ICs increases, accurate knowledge of one-, two-, and three-dimensional structural information regarding

TABLE 3.4 Color Chart for Thermally Grown SiO_2 Films Observed Perpendicularly under Daylight Fluorescent Lighting

Film Thickness (μm)	Color and Comments	Film Thickness (μm)	Color and Comments
0.05	Tan	0.68	"Bluish" (not blue but borderline between violet and blue green; appears more like a mixture between violet red and blue green and looks grayish)
0.07	Brown	0.72	Blue green to green (quite broad)
0.10	Dark violet to red violet	0.77	"Yellowish"
0.12	Royal blue	0.80	Orange (rather broad for orange)
0.15	Light blue to metallic blue	0.82	Salmon
0.17	Metallic to very light yellow green	0.85	Dull, light red violet
0.20	Light gold or yellow; slightly metallic	0.86	Violet
0.22	Gold with slight yellow orange	0.87	Blue violet
0.25	Orange to melon	0.89	Blue
0.27	Red violet	0.92	Blue green
0.30	Blue to violet blue	0.95	Dull yellow green
0.31	Blue	0.97	Yellow to "yellowish"
0.32	Blue to blue green	0.99	Orange
0.34	Light green	1.00	Carnation pink
0.35	Green to yellow green	1.02	Violet red
0.36	Yellow green	1.05	Red violet
0.37	Green yellow	1.06	Violet
0.39	Yellow	1.07	Blue violet
0.41	Light orange	1.10	Green
0.42	Carnation pink	1.11	Yellow green
0.44	Violet red	1.12	Green
0.46	Red violet	1.18	Violet
0.47	Violet	1.19	Red violet
0.48	Blue violet	1.21	Violet red
0.49	Blue	1.24	Carnation pink to salmon
0.50	Blue green	1.25	Orange
0.52	Green (broad)	1.28	"Yellowish"
0.54	Yellow green	1.32	Sky blue to green blue
0.56	Green yellow	1.40	Orange
0.57	Yellow to "yellowish" (not yellow but is in the position where yellow is to be expected; at times appears to be light creamy gray or metallic)	1.45	Violet
0.58	Light orange or yellow to pink	1.46	Blue violet
0.60	Carnation pink	1.50	Blue
0.63	Violet red	1.54	Dull yellow green

Source: Copyright 1964 by International Business Machines Corporation; reprinted with permission from Ref. 6.

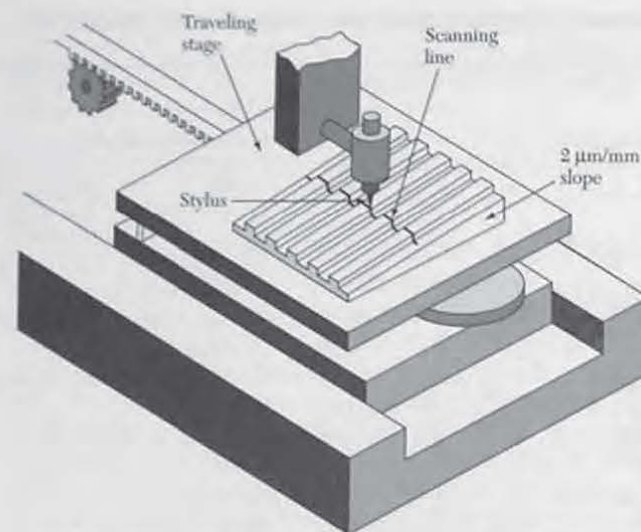


Figure 3.12 Schematic drawing of a surface profilometer.⁷

the features being fabricated becomes more and more critical. However, actual experimental examination and characterization of such structures is both time-consuming and expensive. Therefore, computer simulation is an important tool for investigating ULSI fabrication processes. Sophisticated simulation programs are needed to solve the generalized differential equations that model various processes.

Perhaps the most widely used example of such process simulation software is the Stanford University Process Engineering Modeling (SUPREM) program. SUPREM can accurately simulate multidimensional oxide growth, accounting for the moving Si-SiO₂ boundary, impurity segregation during growth, and other physical phenomena. In addition, SUPREM can predict the results of the various deposition, diffusion, epitaxial growth, and ion implantation processes discussed in subsequent chapters.

SUPREM performs oxidation simulations based on the kinetic growth model described in Section 3.1.1. The package incorporates Arrhenius functions to describe the linear and parabolic rate coefficients for wet and dry oxidation, as well as a rudimentary model for chlorinated oxidation. Oxidation is simulated using the command DIFFUSION, with either WET02 or DRY02 included as parameters indicating wet or dry oxidation, respectively.⁸ SUPREM also requires the specification of process conditions such as times, temperature profiles, and so forth. In the thin oxide regime, SUPREM uses an empirical model of the form⁹

$$\frac{dx}{dt} = \frac{B}{2x + A} + Ce^{-x/L} \quad (16)$$

where B and A are the oxidation rate coefficients, and C and L are empirical constants.

To run SUPREM, an input deck must be provided. This file contains a series of statements and comments. A description of a few commonly used statements is given in Appendix I. The deck begins with a TITLE statement, which is merely a comment repeated on each page of the program output. The next command, INITIALIZE, is a control statement that sets the substrate type, orientation, and doping. This command can also be

used to specify the thickness of the region to be simulated and establish a grid. After the substrate and materials are established, a series of statements is used to specify the sequence of process steps as they occur. Finally, the output of the simulation can be printed or plotted using PRINT or PLOT statements, respectively. Simulation ends with a STOP statement. Several COMMENT statements will typically appear throughout the deck. The user is encouraged to use these statements to better facilitate documentation of the process flow. These ideas are illustrated in Example 3.

EXAMPLE 3

Suppose we want to perform a dry-wet-dry oxidation sequence on a <100> silicon wafer at 1100°C for 5 minutes in dry O₂, 2 hours in wet O₂, and finally, for 5 more minutes in dry O₂. If the silicon substrate is doped with phosphorus at a level of 10¹⁶ cm⁻³, use SUPREM to determine the final oxidation thickness and the phosphorus doping profile in the oxide and silicon layers.

SOLUTION The SUPREM input listing is as follows:

TITLE	Oxidation Example
COMMENT	Initialize silicon substrate
INITIALIZE	<100> Silicon Phosphor Concentration=1e16
COMMENT	Ramp furnace up to 1100 C over 10 minutes in N2
DIFFUSION	Time=10 Temperature=900 Nitrogen T.rate=20
COMMENT	Oxidize the wafers for 5 minutes at 1100 C in dry O2
DIFFUSION	Time=5 Temperature=1100 DryO2
COMMENT	Oxidize the wafers for 120 minutes at 1100 C in wet O2
DIFFUSION	Time=120 Temperature=1100 WetO2
COMMENT	Oxidize the wafers for 5 minutes at 1000 C in dry O2
DIFFUSION	Time=5 Temperature=1100 DryO2
COMMENT	Ramp furnace down to 900 C over 10 minutes in N2
DIFFUSION	Time=10 Temperature=1100 Nitrogen T.rate=20
PRINT	Layers Chemical Concentration Phosphor
PLOT	Active Net Cmin=1e14
STOP	End oxidation example

We assume the furnace has an idle temperature of 900°C when the wafer is inserted, so we use a 10-minute ramp up at 20°C/minute to 1100°C at the beginning of processing and a 10-minute ramp down at -20°C/minute back to 900°C at the end. The ramp up and ramp down are performed in a nitrogen ambient.

After oxidation is complete, we print and plot the phosphorus concentration as a function of depth into the silicon substrate. The results are shown in Figure 3.13, which indicates a final oxide thickness of 0.909 μm and depicts the phosphorus incorporation in the oxide layer.

3.7 SUMMARY

Silicon dioxide is a high-quality insulator that can be thermally grown on silicon wafers. It can also serve as a barrier layer during impurity diffusion or implantation, and it is a key component of MOS devices and circuits. These factors have contributed significantly to silicon's current status as the dominant semiconductor material in use today.

This chapter described the mechanism of thermal oxidation of silicon and presented a kinetic model of oxide growth. This model accurately predicts oxide growth rate for a wide range of process conditions. The chapter also discussed dopant redistribution and the masking properties of oxides. Oxide characterization methods and oxide quality were

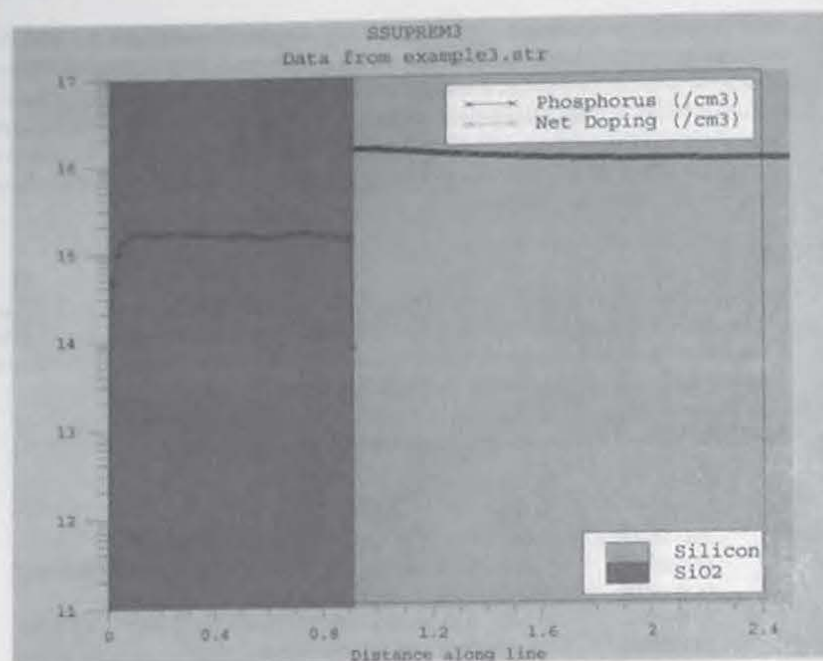


Figure 3.13 Plot of phosphorus concentration as a function of depth into the silicon substrate, using SUPREM software.

described as well. Finally, the process simulation software package SUPREM was introduced. The use of SUPREM, however, is not limited to oxidation, and it will be revisited in subsequent chapters.

► REFERENCES

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
2. B. E. Deal and A. S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, **36**, 3770 (1965).
3. J. D. Meindl, et al., "Silicon Epitaxy and Oxidation," in F. Van de Wiele, W. L. Engl, and P. O. Jespers, Eds., *Process and Device Modeling for Integrated Circuit Design*, Noordhoff, Leyden, 1977.
4. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
5. B. E. Deal, "Standardized Terminology for Oxide Charge Associated with Thermally Oxidized Silicon," *IEEE Trans. Electron Devices*, **ED-27**, 606 (1980).
6. W. Pliskin and E. Conrad, "Nondestructive Determination of Thickness and Refractive Index of Transparent Films," *IBM J. Res. Develop.*, **8**, 43-51 (1964).
7. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 2000.
8. *SSUPREM3 User's Manual*, Silvaco International, Santa Clara, CA, 1995.
9. H. Massoud, C. Ho, and J. Plummer, in J. Plummer, Ed., *Computer Aided Design of Integrated Circuit Fabrication Processes for VLSI Devices*, Stanford University Technical Report, Stanford, CA, 1982.

► PROBLEMS

Asterisks denote difficult problems.

1. A p-type <100>-oriented silicon wafer with a resistivity of $10 \Omega\text{-cm}$ is placed in a wet oxidation system to grow a field oxide of $0.45 \mu\text{m}$ at 1050°C . Determine the time required to grow the oxide.
- *2. After the first oxidation as given in Problem 1, a window is opened in the oxide to grow a gate oxide at 1000°C for 20 minutes in dry oxidation. Find the thicknesses of the gate oxide and the total field oxide.
3. Show that Eq. 11 reduces to $x^2 = Bt$ for long times and to $x = B/A(t + \tau)$ for short times.
4. Determine the diffusion coefficient D for dry oxidation of <100>-oriented silicon samples at 980°C and 1 atm.
5. Define segregation coefficient.
6. Assume that the Cu concentration in the SiO_2 layer is $5 \times 10^{13} \text{ atoms/cm}^3$ after vapor phase deposition and is measured with atomic absorption spectrometry. The Cu concentration in the Si layer is $3 \times 10^{11} \text{ atoms/cm}^3$ after $\text{HF}/\text{H}_2\text{O}_2$ dissolution. Calculate the segregation coefficient of Cu in SiO_2/Si layers.
- *7. A bare and undoped <100> silicon sample is oxidized for 1 hour at 1100°C in dry O_2 . It is then covered and has the oxide removed over half the wafer. Next, it is re-oxidized in wet O_2 at 1000°C for 30 minutes. Use SUPREM to determine the thickness in the two regions. How high are the step on the surface and the step in the substrate?

Photolithography

Photolithography is the process of transferring patterns of geometric shapes on a mask to a thin layer of photosensitive material (called *photoresist*) covering the surface of a semiconductor wafer.¹ These patterns define the various regions in an integrated circuit, such as the implantation regions, the contact windows, and the bonding-pad areas. The resist patterns defined by the lithographic process are not permanent elements of the final device, but only replicas of circuit features. To produce circuit features, these resist patterns must be transferred once more into the underlying layers comprising the device. Pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer (see Chapter 5).² A brief description of pattern transfer was given in Section 1.4.2. The present chapter covers the following topics:

- The importance of a clean room for lithography
- The most widely used lithographic method—optical lithography—and its resolution enhancement techniques
- Advantages and limitations of other lithographic methods

▶ 4.1 OPTICAL LITHOGRAPHY

The vast majority of lithographic equipment for IC fabrication is optical equipment using ultraviolet light (wavelength or $\lambda \approx 0.2\text{--}0.4\ \mu\text{m}$). This section considers the exposure tools, masks, resists, and resolution enhancement techniques used for optical lithography. It also considers the pattern transfer process, which serves as a basis for other lithographic systems. The section first briefly discusses the clean room, because all lithographic processes must be performed in an ultraclean environment.

4.1.1 The Clean Room

An IC fabrication facility requires a clean processing room, especially in the area used for photolithography. The need for such a clean room arises because dust particles in the air can settle on semiconductor wafers and lithographic masks and can cause defects in the devices, which result in circuit failure. For example, a dust particle on a semiconductor surface can disrupt the single-crystal growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into the gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in the lithographic area. When dust particles adhere to the surface of a photomask, they behave as opaque patterns on the mask, and these patterns will be transferred to the underlying layer along with the circuit patterns on the mask. Figure 4.1 shows three dust particles on a photomask.³ Particle 1 may result in the formation of

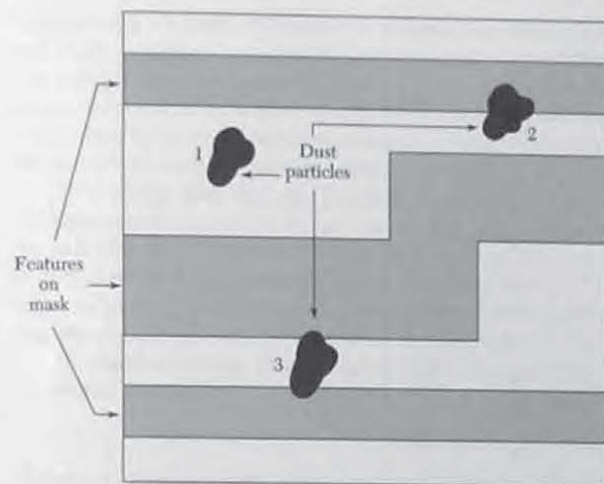


Figure 4.1 Various ways in which dust particles can interfere with photomask patterns.³

a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause a constriction of current flow in a metal runner. Particle 3 can lead to a short circuit between the two conducting regions and render the circuit useless.

In a clean room, the total number of dust particles per unit volume must be tightly controlled, along with the temperature and humidity. Figure 4.2 shows the particle-size distribution curves for various classes of clean rooms. Two systems are used to define the classes of clean room.⁴ In the English system, the numerical designation of the class is taken from the maximum allowable number of particles that are $0.5\ \mu\text{m}$ and larger per cubic foot of air. In the metric system, the class is taken from the logarithm (base 10) of the maximum allowable number of particles that are $0.5\ \mu\text{m}$ and larger per cubic meter. For example, a class 100 clean room (English system) has a dust count of 100 particles/ ft^3 with particle diameters of $0.5\ \mu\text{m}$ and larger, whereas a class M 3.5 clean room (metric system) has a dust count of $10^{3.5}$, or about 3500 particles/ m^3 with particle diameters of $0.5\ \mu\text{m}$ or larger. Since $100\ \text{particles}/\text{ft}^3 = 3500\ \text{particles}/\text{m}^3$, a class 100 in the English system corresponds to a class M 3.5 in the metric system.

Because the number of dust particles increases as particle size decreases, a more stringent control of the clean room environment is required when the minimum feature lengths of ICs are reduced to the deep submicron range. For most IC fabrication areas, a class 100 clean room is required; that is, the dust count must be about four orders of magnitude lower than that of ordinary room air. However, for the lithography area, a class 10 clean room or one with a lower dust count is required.

EXAMPLE 1

If we expose a 200-mm wafer for 1 minute to an air stream under a laminar-flow condition at 30 m/min, how many dust particles will land on the wafer in a class 10 clean room?

SOLUTION For a class 10 clean room, there are 350 particles ($0.5\ \mu\text{m}$ and larger) per cubic meter. The air volume that goes over the wafer in 1 minute is

$$(30\ \text{m/min}) \times \pi \left(\frac{0.2\ \text{m}}{2} \right)^2 \times 1\ \text{minute} = 0.942\ \text{m}^3$$

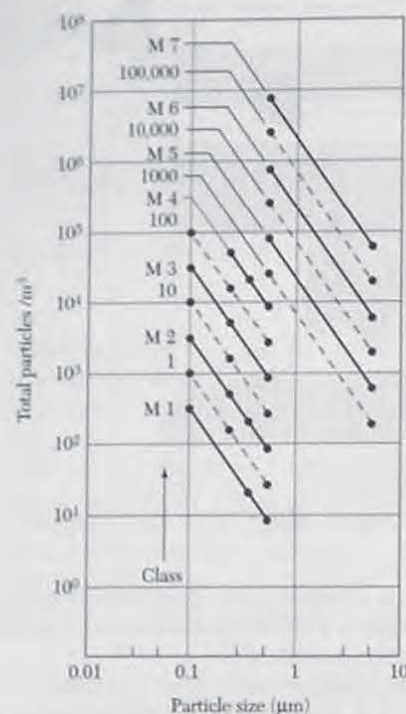


Figure 4.2 Particle-size distribution curve for English (---) and metric (—) classes of clean rooms.⁴

The number of dust particles (0.5 μm and larger) contained in the air volume is $350 \times 0.942 = 330$ particles.

Therefore, if there are 400 IC chips on the wafer, the particle count amounts to one particle on each of 82% of the chips. Fortunately, only a fraction of the particles that land adhere to the wafer surface, and of those only a fraction are at a circuit location critical enough to cause a failure. However, the calculation indicates the importance of the clean room.

4.1.2 Exposure Tools

The pattern transfer process is accomplished by using a lithographic exposure tool. The performance of an exposure tool is determined by three parameters: resolution, registration, and throughput. *Resolution* is the minimum feature dimension that can be transferred with high fidelity to a resist film on a semiconductor wafer. *Registration* is a measure of how accurately patterns on successive masks can be aligned (or overlaid) with respect to previously defined patterns on the wafer. *Throughput* is the number of wafers that can be exposed per hour for a given mask level.

There are basically two optical exposure methods: shadow printing and projection printing.^{5,6} Shadow printing may have the mask and wafer in direct contact with one another (as in *contact printing*) or in close proximity (as in *proximity printing*). Figure 4.3a shows a basic setup for contact printing, in which a resist-coated wafer is brought into physical contact with a mask, and the resist is exposed by a nearly collimated beam of ultraviolet

light through the back of the mask for a fixed time. The intimate contact between resists and mask provides a resolution of approximately 1 μm . However, contact printing suffers a major drawback caused by dust particles. A dust particle or a speck of silicon dust on the wafer can be imbedded into the mask when the mask makes contact with the wafer. The imbedded particle causes permanent damage to the mask and results in defects in the wafer with each succeeding exposure.

To minimize mask damage, the proximity exposure method is used. Figure 4.3b shows the basic setup. It is similar to the contact printing method, except that there is a small gap (10–50 μm) between the wafer and the mask during exposure. The small gap results in optical diffraction at feature edges on the photomask; that is, when light passes by the edges of an opaque mask feature, fringes are formed and some light penetrates into the shadow region. As a result, the resolution is degraded to the 2 to 5- μm range.

In shadow printing, the minimum linewidth [or critical dimension (CD)] that can be printed is roughly

$$\text{CD} \cong \sqrt{\lambda g} \quad (1)$$

where λ is the wavelength of the exposure radiation and g is the gap between the mask and the wafer and includes the thickness of the resist. For $\lambda = 0.4 \mu\text{m}$ and $g = 50 \mu\text{m}$, the CD is 4.5 μm . If we reduce λ to 0.25 μm (a wavelength range of 0.2 to 0.3 μm is in the deep UV spectral region) and g to 15 μm , the CD becomes 2 μm . Thus, there is an advantage in reducing both λ and g . However, for a given distance g , any dust particle with a diameter larger than g potentially can cause mask damage.

To avoid the mask damage problem associated with shadow printing, projection-printing exposure tools have been developed to project an image of the mask patterns onto a resist-coated wafer many centimeters away from the mask. To increase resolution, only a small portion of the mask is exposed at a time. The small image area is scanned or stepped over the wafer to cover the entire wafer surface. Figure 4.4a shows a 1:1 wafer scan projection system.^{6,7} A narrow, arc-shaped image field approximately 1 mm in width serially transfers the slit image of the mask onto the wafer. The image size on the wafer is the same as that on the mask.

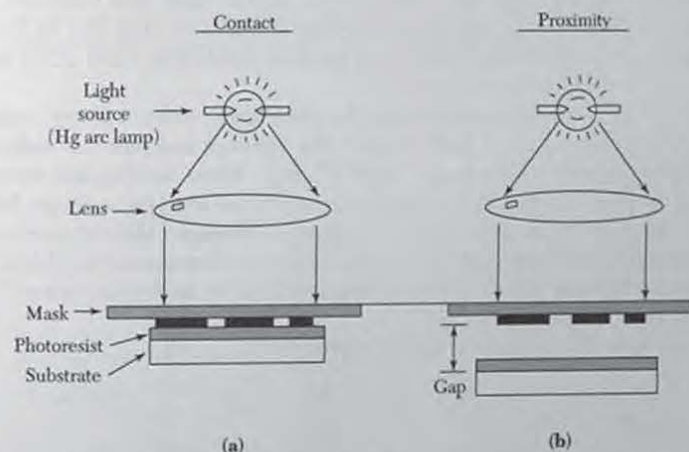


Figure 4.3 Schematic of optical shadow printing techniques.¹ (a) Contact printing. (b) Proximity printing.

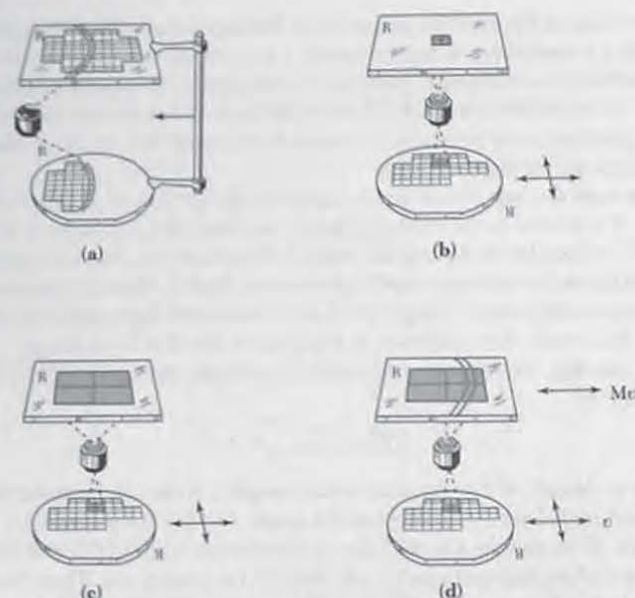


Figure 4.4 Image partitioning techniques for projection printing. (a) Annual-field wafer scan. (b) 1:1 step-and-repeat. (c) M:1 reduction step-and-repeat. (d) M:1 reduction step-and-scan.^{6,7}

The small image field can also be stepped over the surface of the wafer by two-dimensional translations of the wafer only, while the mask remains stationary. After the exposure of one chip site, the wafer is moved to the next chip site and the process is repeated. Figures 4.4b and 4.4c show the partitioning of the wafer image by *step-and-repeat projection* with a ratio of 1:1 or at a demagnification ratio of M:1 (e.g., 10:1 for a 10 times reduction on the wafer), respectively. The demagnification ratio is an important factor in our ability to produce both the lens and the mask from which we wish to print. The 1:1 optical systems are easier to design and fabricate than 10:1 or 5:1 reduction systems, but it is much more difficult to produce defect-free masks at 1:1 than it is at a 10:1 or a 5:1 demagnification ratio.

Reduction projection lithography can also print larger wafers without redesigning the stepper lens, as long as the field size (i.e., the exposure area onto the wafer) of the lens is large enough to contain one or more IC chips. When the chip size exceeds the field size of the lens, further partitioning of the image on the reticle is necessary. In Figure 4.4d, the image field on the reticle can be a narrow arc shape for M:1 step-and-scan projection lithography. The step-and-scan system yields two-dimensional translations of the wafer with speed v , and one-dimensional translation of the mask with a speed M times that of the wafer speed.

The resolution of a projection system is given by

$$l_m = k_1 \frac{\lambda}{NA} \quad (2)$$

where λ is again the exposure wavelength, k_1 is a process-dependent factor, and NA is the numerical aperture, which is given by

$$NA = \bar{n} \sin \theta \quad (3)$$

where $\bar{n}$ is the index of refraction in the image medium (usually air, where $\bar{n} = 1$), and θ is the half-angle of the cone of light converging to a point image at the wafer, as shown in Figure 4.5.⁵ Also shown in the figure is the depth of focus (DOF), which can be expressed as

$$DOF = \frac{\pm l_m/2}{\tan \theta} \approx \frac{\pm l_m/2}{\sin \theta} = k_2 \frac{\lambda}{(NA)^2} \quad (4)$$

where k_2 is another process-dependent factor.

Equation 2 indicates that resolution can be improved (i.e., smaller l_m) by either reducing the wavelength or increasing NA or both. However, Eq. 4 indicates that the DOF degrades much more rapidly by increasing NA than by decreasing λ . This explains the trend toward shorter-wavelength sources in optical lithography.

The high-pressure mercury-arc lamp is widely used in exposure tools because of its high intensity and reliability. The mercury-arc spectrum is composed of several peaks, as shown in Figure 4.6. The terms *G-line*, *H-line*, and *I-line* refer to the peaks at 436 nm, 405 nm, and 365 nm, respectively. I-line lithography with 5:1 step-and-repeat projection can offer a resolution of 0.3 μm with resolution enhancement techniques (see Section 4.1.6). Advanced exposure tools such as the 248-nm lithographic system using a KrF excimer laser, the 193-nm lithographic system using an ArF excimer laser, and the 157-nm lithographic system using a F₂ excimer laser have been developed for mass production with a resolution of 0.18 μm (180 nm), 0.10 μm (100 nm), and 0.07 μm (70 nm), respectively.

4.1.3 Masks

Masks used for IC manufacturing are usually reduction reticles. The first step in mask making is to use a computer-aided design (CAD) system in which designers can completely describe the circuit patterns electrically. The digital data produced by the CAD system then drives a pattern generator, which is an electron beam lithographic system (see Section 4.2.1) that transfers the patterns directly to electron-sensitized mask. The mask consists of a fused-silica substrate covered with a chromium layer. The circuit pattern is first transferred to the electron-sensitized layer (electron resist), which is transferred once more

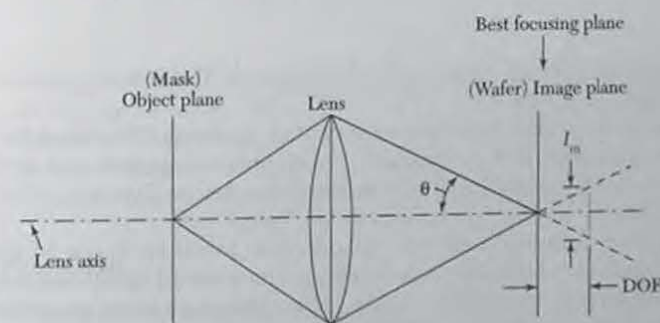


Figure 4.5 Simple image system.⁵

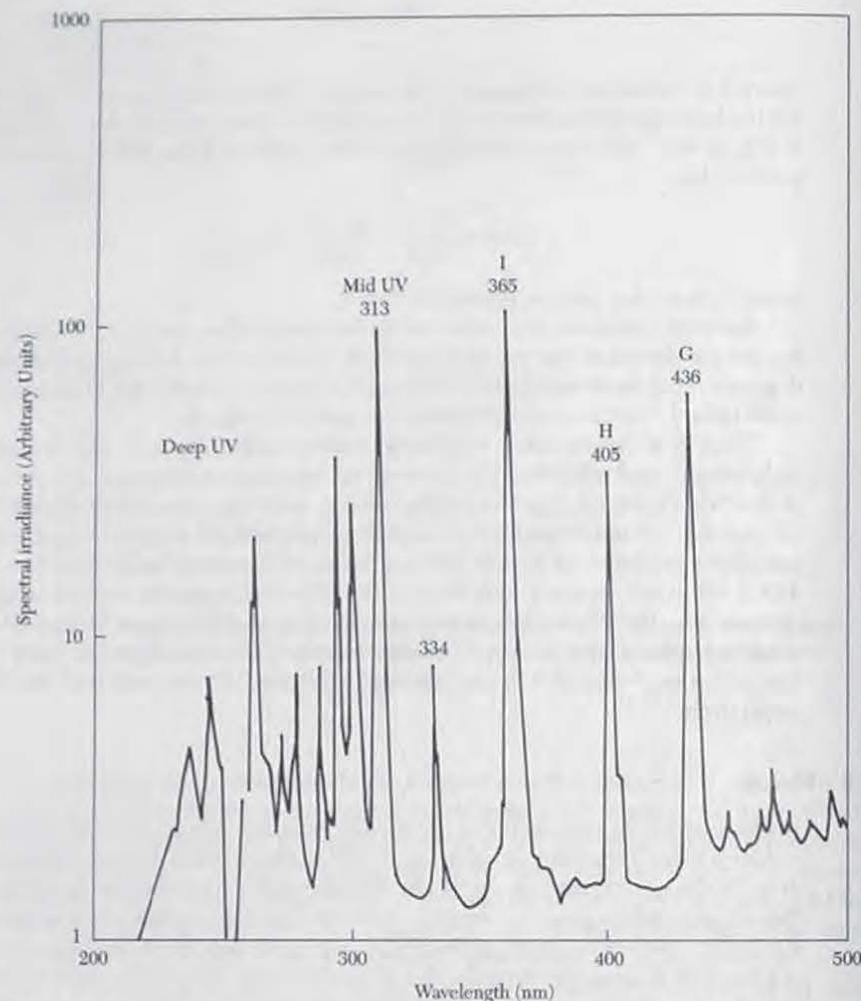


Figure 4.6 Typical high-pressure mercury-arc lamp spectrum.

into the underlying chromium layer for the finished mask. The details of pattern transfer are considered in Section 4.1.5.

The patterns on a mask represent one level of an IC design. The composite layout is broken into mask levels that correspond to the IC process sequence, such as the isolation region on one level, the gate region on another, and so on. Typically, 15 to 20 different mask levels are required for a complete IC process cycle.

The standard-size mask substrate is a fused-silica plate 15×15 cm square, 0.6 cm thick. The size is needed to accommodate the lens field sizes for 4:1 or 5:1 optical exposure tools, whereas the thickness is required to minimize pattern placement errors due to substrate distortion. The fused-silica plate is needed for its low coefficient of thermal expansion, its high transmission at shorter wavelengths, and its mechanical strength. Figure 4.7 shows a

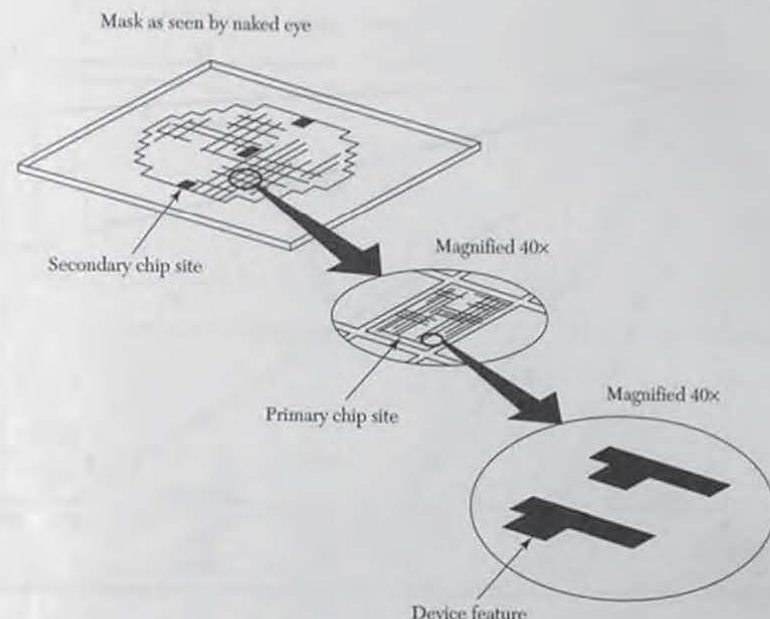


Figure 4.7 An integrated circuit photomask.¹

mask on which patterns of geometric shapes have been formed. A few secondary chip sites, used for process evaluation, are also included in the mask.

One of the major concerns about masks is defect density. Mask defects can be introduced during the manufacture of the mask or during subsequent lithographic processes. Even a small mask-defect density has a profound effect on the final IC yield. Yield is defined as the ratio of good chips per wafer to the total number of chips per wafer (see Chapter 10). As a first-order approximation, the yield Y for a given masking level can be expressed as

$$Y \cong e^{-D_0 A_c} \quad (5)$$

where D_0 is the average number of "fatal" defects per unit area, and A_c is the defect-sensitive area (or "critical area") of the IC chip. If D_0 remains the same for all mask levels (e.g., $N = 10$ levels), then the final yield becomes

$$Y \cong e^{-N D_0 A_c} \quad (6)$$

Figure 4.8 shows the mask limit yield for a 10-level lithographic process as a function of chip size for various values of defect densities. For example, for $D_0 = 0.25$ defect/cm², the yield is 10% for a chip size of 90 mm², and it drops to about 1% for a chip size of 180 mm². Therefore, inspection and cleaning of masks are important to achieve high yields on large chips. Of course, an ultraclean processing area is mandatory for lithographic processing.

4.1.4 Photoresist

Photoresist is a radiation-sensitive compound that can be classified as positive or negative, depending on how it responds to radiation. For *positive resists*, the exposed regions become more soluble and are thus more easily removed in the development process. The

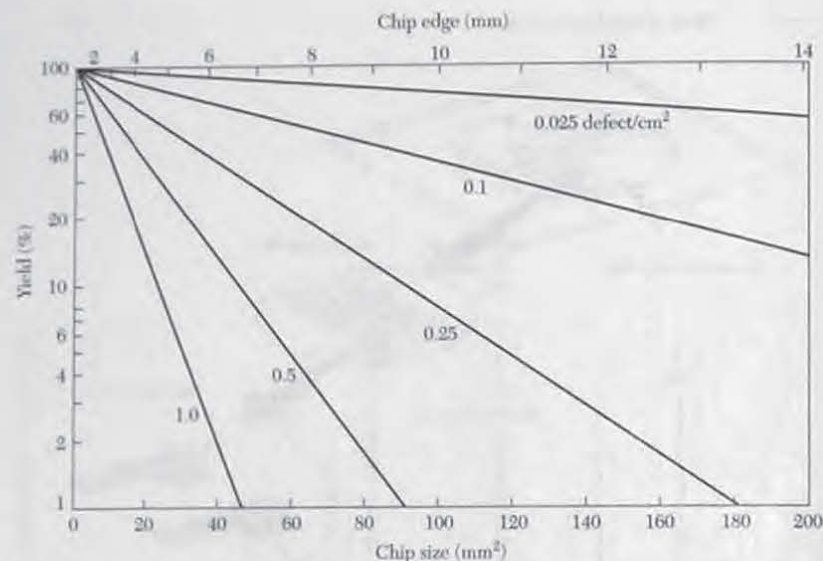


Figure 4.8 Yield for a 10-mask lithographic process with various defect densities per level.

net result is that the patterns formed (also called images) in the positive resist are the same as those on the mask. For *negative resists*, the exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns.

Positive photoresists consist of three components: a photosensitive compound, a base resin, and an organic solvent. Prior to exposure, the photosensitive compound is insoluble in the developer solution. After exposure, the photosensitive compound absorbs radiation in the exposed pattern areas, changes its chemical structure, and becomes soluble in the developer solution. After development, the exposed areas are removed.

Negative photoresists are polymers combined with a photosensitive compound. After exposure, the photosensitive compound absorbs the optical energy and converts it into chemical energy to initiate a polymer cross-linking reaction. This reaction causes cross linking of the polymer molecules. The cross-linked polymer has a higher molecular weight and becomes insoluble in the developer solution. After development, the unexposed areas are removed. One major drawback of a negative photoresist is that in the development process, the whole resist mass swells by absorbing developer solvent. This swelling action limits the resolution of negative photoresists.

Figure 4.9a shows a typical exposure response curve and image cross section for a positive resist.¹ The response curve describes the percentage of resist remaining after exposure and development versus the exposure energy. Note that the resist has a finite solubility in its developer, even without exposure to radiation. As the exposure energy increases, the solubility gradually increases until at a threshold energy E_T , the resist becomes completely soluble. The sensitivity of a positive resist is defined as the energy required to produce complete solubility in the exposed region. Thus, E_T corresponds to the sensitivity. In addition to E_T , a parameter γ , the contrast ratio, is defined to characterize the resist:

$$\gamma = \left[\ln \left(\frac{E_T}{E_i} \right) \right]^{-1} \quad (7)$$

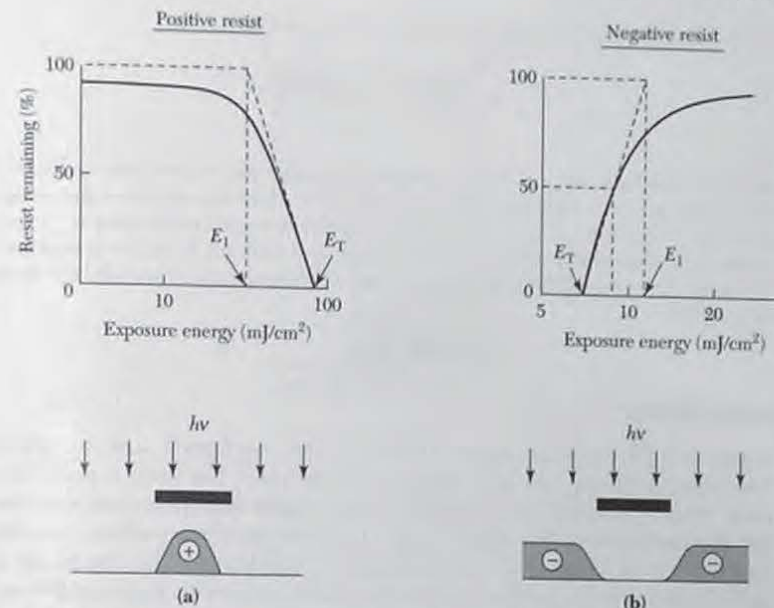


Figure 4.9 Exposure response curve and cross section of the resist image after development.¹ (a) Positive photoresist. (b) Negative photoresist.

where E_i is the energy obtained by drawing the tangent at E_T to reach 100% resist thickness, as shown in Figure 4.9a. A larger γ implies a higher solubility of the resist with an incremental increase of exposure energy and results in sharper images.

The image cross section in Figure 4.9a illustrates the relationship between the edges of a photomask image and the corresponding edges of the resist images after development. The edges of the resist image are generally not at the vertically projected positions of the mask edges because of *diffraction*. The edge of the resist image corresponds to the position where the total absorbed optical energy equals the threshold energy E_T .

Figure 4.9b shows the exposure response curve and image cross section for a negative resist. The negative resist remains completely soluble in the developer solution for exposure energies lower than E_T . Above E_T , more of the resist film remains after development. At exposure energies twice the threshold energy, the resist film becomes essentially insoluble in the developer. The sensitivity of a negative resist is defined as the energy required to retain 50% of the original resist film thickness in the exposed region. The parameter γ is defined similarly to γ in Eq. 7, except that E_i and E_T are interchanged. The image cross section for the negative resist (Fig. 4.9b) is also influenced by the diffraction effect.

EXAMPLE 2

Find the parameter γ for the photoresists shown in Figure 4.9.

SOLUTION For the positive resist, $E_T = 90 \text{ mJ/cm}^2$ and $E_i = 45 \text{ mJ/cm}^2$, so

$$\gamma = \left[\ln \left(\frac{E_T}{E_i} \right) \right]^{-1} = \left[\ln \left(\frac{90}{45} \right) \right]^{-1} = 1.4$$

For the negative resist, $E_T = 7 \text{ mJ/cm}^2$ and $E_i = 12 \text{ mJ/cm}^2$, so

$$\gamma = \left[\ln \left(\frac{E_i}{E_T} \right) \right]^{-1} = \left[\ln \left(\frac{12}{7} \right) \right]^{-1} = 1.9$$

For deep UV lithography (e.g., 248 and 193 nm), we cannot use conventional photoresists because these resists require a high-dose exposure in deep UV, which will cause lens damage and lower throughput. *Chemical-amplified resist* (CAR) has been developed for the deep UV process. CAR consists of a photo-acid generator, a resin polymer, and a solvent. CAR is very sensitive to deep UV radiation, and the exposed and unexposed regions differ greatly in their solubility in the developer solution.

4.1.5 Pattern Transfer

Figure 4.10 illustrates the steps to transfer IC patterns from a mask to a silicon wafer that has an insulating SiO_2 layer formed on its surface.⁵ The wafer is placed in a clean room, which typically is illuminated with yellow light, since photoresists are not sensitive to wavelengths greater than $0.5 \mu\text{m}$. To ensure satisfactory adhesion of the resist, the surface must be changed from hydrophilic to hydrophobic. This change can be made by the application of an adhesion promoter, which can provide a chemically compatible surface for the resist. The most common adhesion promoter for silicon ICs is hexamethyl-disilazane (HMDS). After the application of this adhesion layer, the wafer is held on a vacuum spindle, and 2 to 3 cm^3 of liquidous resist is applied to the center of the wafer. The wafer is then rapidly accelerated to a constant rotational speed, which is maintained for about 30 seconds. Spin speed is generally in the range of 1000 to 10,000 rpm to coat a uniform film about 0.5 to $1 \mu\text{m}$ thick, as shown in Figure 4.10a. The thickness of photoresist is correlated with its viscosity.

After the spinning step, the wafer is "soft baked" (typically at 90 – 120°C for 60–120 seconds) to remove the solvent from the photoresist film and to increase resist adhesion to the wafer. The wafer is aligned with respect to the mask in an optical lithographic system, and the resist is exposed to UV light, as shown in Figure 4.10b. If a positive photoresist is used, the exposed resist is dissolved in the developer, as shown on the left side of Figure 4.10c. Photoresist development is usually done by flooding the wafer with the developer solution. The wafer is then rinsed and dried. After development, "post baking" at approximately 100°C to 180°C may be required to increase the adhesion of the resist to the substrate. The wafer is then put in an ambient that etches the exposed insulation layer but does not attack the resist, as shown in Figure 4.10d. Finally, the resist is stripped (e.g., using solvents or plasma oxidation), leaving behind an insulator image (or pattern) that is the same as the opaque image on the mask (left side of Fig. 4.10e).

For negative photoresist, the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Figure 4.10e) is the reverse of the opaque image on the mask.

The insulator image can be used as a mask for subsequent processing. For example, *ion implantation* (see Chapter 7) can be done to dope the exposed semiconductor region, but not the area covered by the insulator. The dopant pattern is a duplicate of the design pattern on the photomask for a negative photoresist or is its complementary pattern for a positive photoresist. The complete circuit is fabricated by aligning the next mask in the sequence to the previous pattern and repeating the lithographic transfer process.

A related pattern transfer process is the *liftoff* technique, shown in Figure 4.11. A positive resist is used to form the resist pattern on the substrate (Figs. 4.11a and 4.11b).

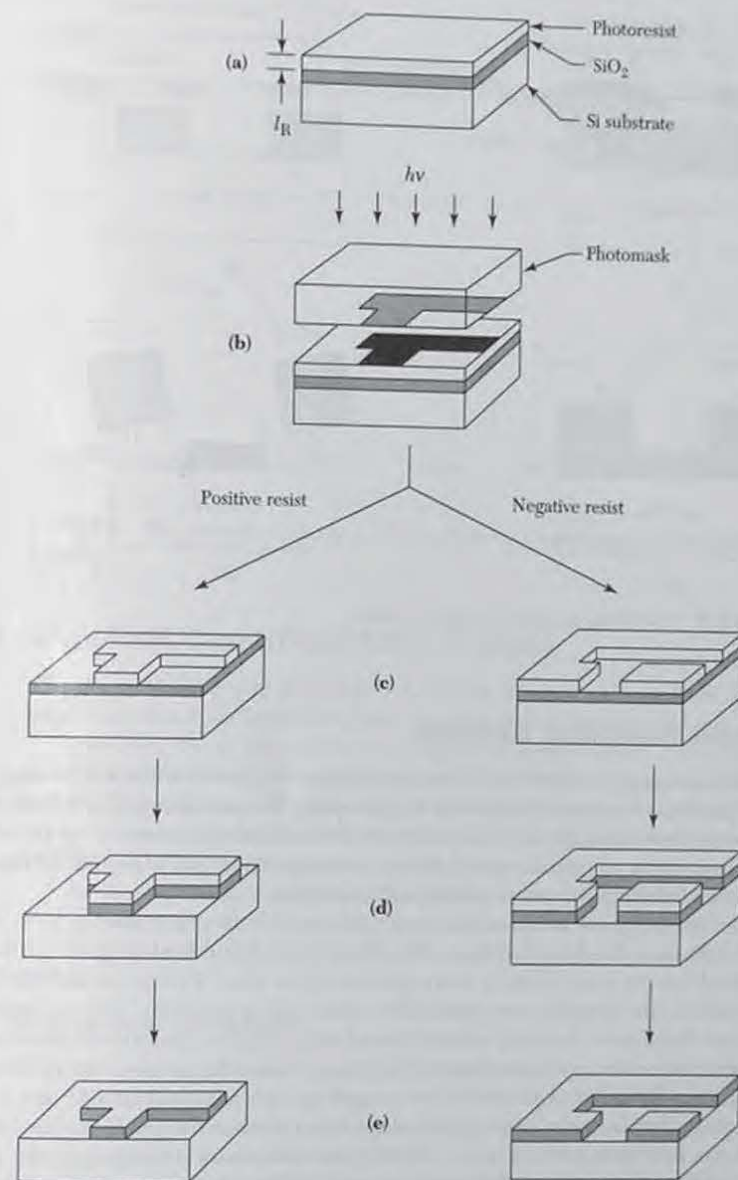


Figure 4.10 Details of the optical lithographic pattern transfer process.⁵

The film (e.g., aluminum) is deposited over the resist and the substrate (Fig. 4.11c). The film thickness must be smaller than that of the resist. Those portions of the film on the resist are removed by selectively dissolving the resist layer in an appropriate liquid etchant so that the overlying film is lifted off and removed (Fig. 4.11d). The liftoff technique is capable of high resolution and is used extensively for discrete devices such as high-power MESFETs. However, it is not as widely applicable for ultralarge-scale integration, in which dry etching is the preferred technique.

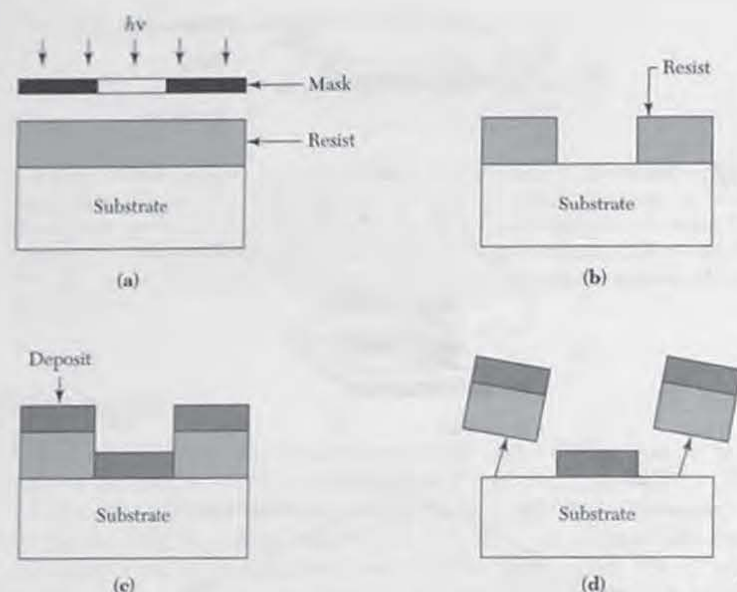


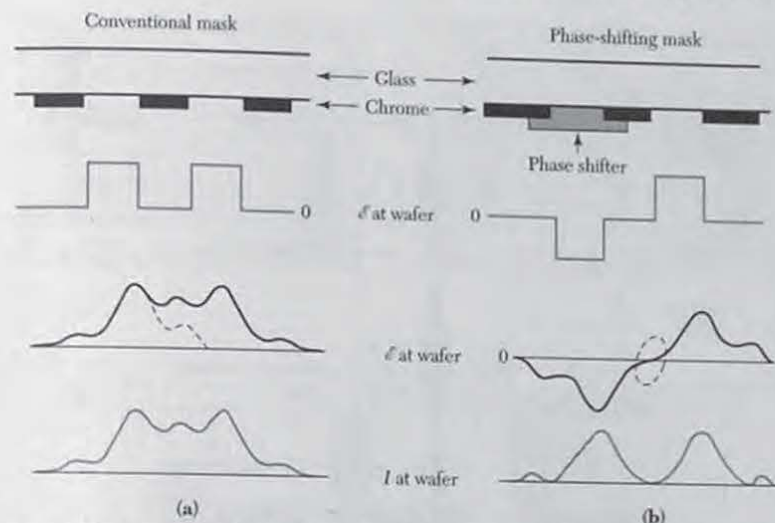
Figure 4.11 The lift-off process for pattern transfer.

4.1.6 Resolution Enhancement Techniques

Optical lithography has been continuously challenged to provide better resolution, greater DOF, and wider exposure latitude in IC processing. These challenges have been met by reducing the wavelength of the exposure tools and developing new resists. In addition, many resolution enhancement techniques have been developed to extend the capability of optical lithography to even smaller feature lengths.

An important resolution enhancement technique is the *phase-shifting mask* (PSM). The basic concept is shown in Figure 4.12.⁹ For a conventional mask (Fig. 4.12a), the electric field has the same phase at every aperture (clear area). Diffraction and the limited resolution of the optical system spread the electric field at the wafer, as shown by the dotted lines. Interference between waves diffracted by the adjacent apertures enhances the field between them. Because the intensity (I) is proportional to the square of the electric field, it becomes difficult to separate the two images that are projected close to one another. The phase-shift layer that covers adjacent apertures reverses the sign of the electric field, as shown in Figure 4.12b. Because the intensity at the mask is unchanged, the electric field of the images at the wafer can be cancelled. Therefore, images that are projected close to one another can be separated. A 180° phase change can be obtained by using a transparent layer of thickness $d = \lambda/2(\bar{n} - 1)$, where $\bar{n}$ is the refractive index and λ is the wavelength that covers one aperture, as shown in Figure 4.12b.

Another resolution enhancement technique is *optical proximity correction* (OPC), which uses modified shapes of adjacent subresolution geometry to improve imaging capability. For example, a square contact hole with dimensions near the resolution limit will print nearly as a circle. Modifying the contact-hole pattern with additional geometry at the corners will help to print a more accurate square hole.

Figure 4.12 The principle of phase-shift technology. (a) Conventional technology. (b) Phase-shift technology.⁹

4.2 NEXT-GENERATION LITHOGRAPHIC METHODS

Why is optical lithography so widely used and what makes it such a promising method? The reasons are that it has high throughput, good resolution, low cost, and ease in operation.

However, due to deep-submicron IC process requirements, optical lithography has some limitations that have not yet been solved. Although we can use PSM or OPC to extend its useful span, the complexity of mask production and mask inspection cannot be easily resolved. In addition, the cost of masks is very high. Therefore, we need to find alternatives to optical lithography to process deep-submicron or nanometer ICs.

Various types of next-generation lithographic methods for IC fabrication are discussed in this section. Electron beam lithography, extreme UV lithography, x-ray lithography, and ion beam lithography are considered, as are the differences among these methods.

4.2.1 Electron Beam Lithography

Electron beam (or *e-beam*) lithography is primarily used to produce photomasks. Relatively few tools are dedicated to direct exposure of the resist by a focused electron beam without a mask. Figure 4.13 shows a schematic of an electron beam lithography system.¹⁰ The electron gun is a device that can generate a beam of electrons with a suitable current density. A tungsten thermionic emission cathode or single-crystal lanthanum hexaboride (LaB_6) is used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size 10 to 25 nm in diameter. Beam-blanking plates, which turn the electron beam on and off, and beam deflection coils are computer controlled and operated at MHz or higher rates to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1 cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron beam lithography include the generation of submicron resist geometries, highly automated and precisely controlled operation, greater depth

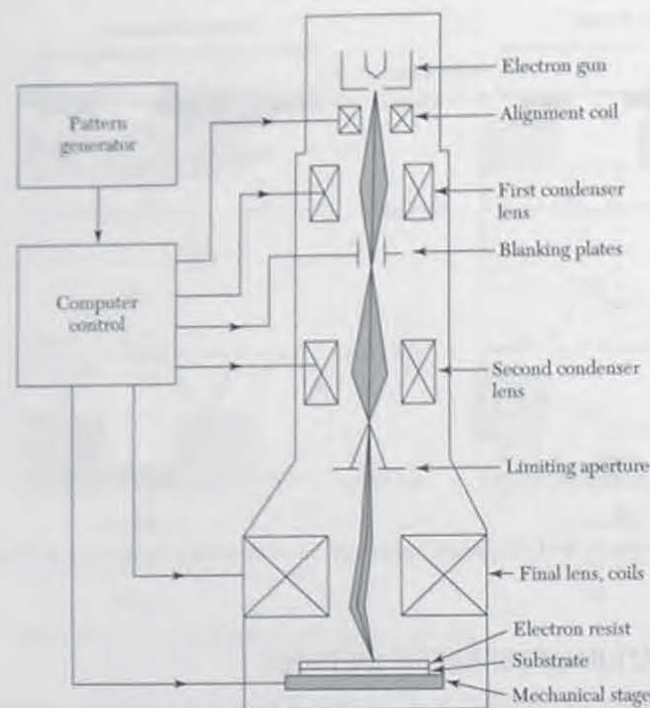


Figure 4.13 Schematic of an electron beam lithography machine.¹⁰

of focus than that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron beam lithography machines have low throughput—approximately 10 wafers per hour at less than $0.25\text{ }\mu\text{m}$ resolution. This throughput is adequate for the production of photomasks, for situations that require small numbers of custom circuits, and for design verification. However, for maskless direct writing, the machine must have the highest possible throughput, and therefore the largest beam diameter possible consistent with the minimum device dimensions.

There are basically two ways to scan the focused electron beam: raster scan and vector scan.¹¹ In a *raster scan* system, resist patterns are written by a vertically oriented beam that moves through a regular mode, as shown in Figure 4.14a. The beam scans sequentially over every possible location on the mask and is blanked (turned off) where no exposure is required. All patterns on the area to be written must be subdivided into individual addresses, and a given pattern must have a minimum incremental interval that is evenly divisible by the beam address size.

In a *vector scan* system, shown in Figure 4.14b, the beam is directed only to the requested pattern features and jumps from feature to feature, rather than scanning the whole chip, as in raster scan. For many chips, the average exposed region is only 20% of the chip area, so time is saved using a vector scan system.

Figure 4.14c shows several types of electron beams employed for e-beam lithography: the Gaussian spot beam (round beam), the variable-shaped beam, and cell projection. In the variable-shaped beam system, the patterning beam has a rectangular cross section of variable size and aspect ratio. It offers the advantage of exposing several address units simultaneously. Therefore, the vector scan method using a variable-shaped beam

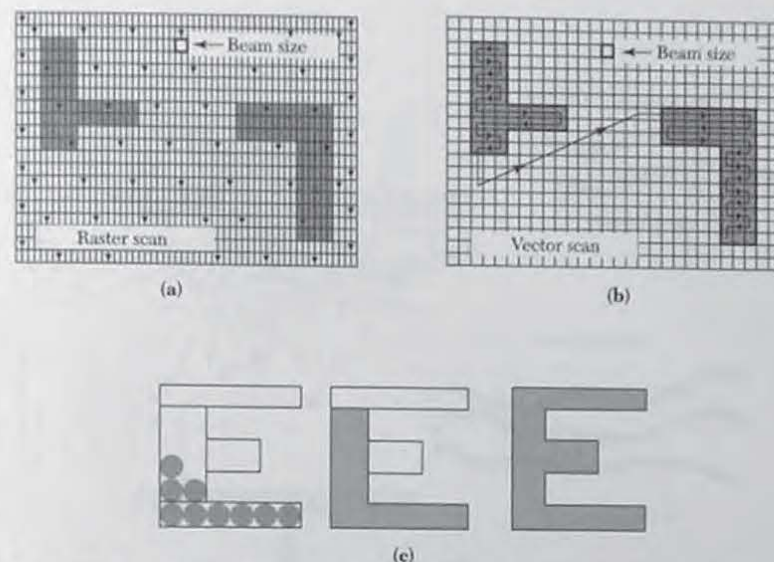


Figure 4.14 (a) Raster scan writing scheme. (b) Vector scan writing schemes. (c) Shapes of electron beam: round, variable, and cell projection.¹²

has higher throughput than the conventional Gaussian spot beam. It is also possible to pattern a complex geometric shape in one exposure with an electron beam system; this is called *cell projection*, as shown in the far right of Figure 4.14c. The cell projection technique¹³ is particularly suitable for highly repetitive designs, as in MOS memory cells, since several memory cell patterns can be exposed at once. Cell projection has not yet achieved the throughput of optical exposure tools.

Electron Resist

Electron resists are polymers. The behavior of an electron beam resist is similar to that of a photoresist; that is, a chemical or physical change is induced in the resist by irradiation. This change allows the resist to be patterned. For a positive electron resist, the polymer-electron interaction causes chemical bonds to be broken (chain scission) to form shorter molecular fragments, as shown in Figure 4.15a.¹⁴ As a result, the molecular weight is reduced in the irradiated area, which can be dissolved subsequently in a developer solution that attacks the low-molecular-weight material. Common positive electron resists include poly-methyl methacrylate (PMMA) and poly-butene-1 sulfone (PBS). Positive electron resists can achieve resolution of $0.1\text{ }\mu\text{m}$ or better.

For a negative electron resist, the irradiation causes radiation-induced polymer linking, as shown in Figure 4.15b. The cross linking creates a complex three-dimensional structure with a molecular weight higher than that of the nonirradiated polymer. The nonirradiated resist can be dissolved in a developer solution that does not attack the high-molecular-weight material. Poly-glycidyl methacrylate-co-ethyl-acrylate (COP) is a common negative electron resist. COP, like most negative photoresists, also swells during development, so resolution is limited to about $1\text{ }\mu\text{m}$.

The Proximity Effect

In optical lithography, the resolution is limited by diffraction of light. In electron beam lithography, the resolution is not limited by diffraction (because the wavelengths associated

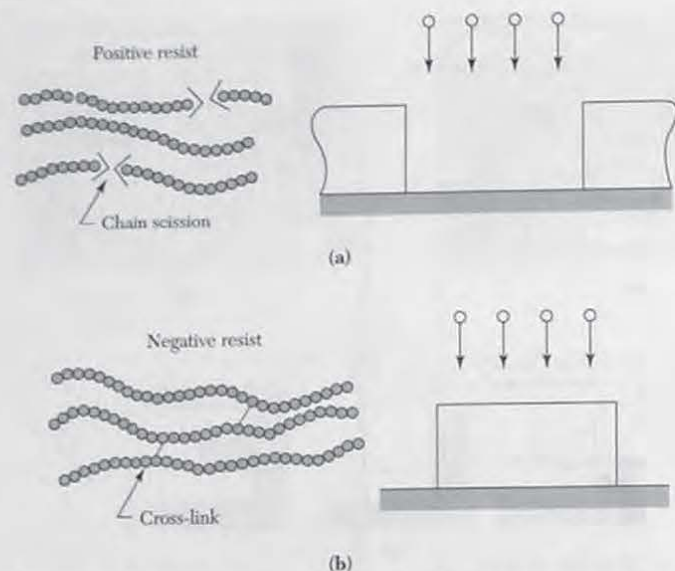


Figure 4.15 Schematic of positive and negative resists used in electron beam lithography.¹³

with electrons of a few keV and higher energies are less than 0.1 nm) but by electron scattering. When electrons penetrate the resist film and underlying substrate, they undergo collisions. These collisions lead to energy losses and path changes. Thus, the incident electrons spread out as they travel through the material until either all of their energy is lost or they leave the material because of backscattering.

Figure 4.16a shows computed electron trajectories of 100 electrons with initial energy of 20 keV incident at the origin of a 0.4- μm PMMA film on a thick silicon substrate.¹⁴ The electron beam is incident along the z -axis, and all trajectories have been projected onto the xz plane. This figure shows qualitatively that the electrons are distributed in an oblong pear-shaped volume with a diameter on the same order of magnitude as the electron penetration depth ($\sim 3.5 \mu\text{m}$). Also, many electrons undergo backscattering collisions and travel backward from the silicon substrate into the PMMA resist film and leave the material.

Figure 4.16b shows the normalized distributions of the forward-scattering and backscattering electrons at the resist-substrate interface. Because of backscattering, electrons effectively can irradiate several micrometers away from the center of the exposure beam. Since the dose of a resist is given by the sum of the irradiations from all surrounding areas, electron beam irradiation at one location will affect the irradiation in neighboring locations. This phenomenon is called the *proximity effect*. The proximity effect places a limit on the minimum spacings between pattern features. To correct for the proximity effect, patterns are divided into smaller segments. The incident electron dose in each segment is adjusted so that the integrated dose from all its neighboring segments is the correct exposure dose. This approach further decreases the throughput of the electron beam system because of the additional computer time required to expose the subdivided resist patterns.

4.2.2 Extreme Ultraviolet Lithography

Extreme ultraviolet (EUV) lithography is a promising next-generation lithographic technology to extend the minimum linewidths to 30 nm without throughput losses.¹⁵ Figure 4.17 shows

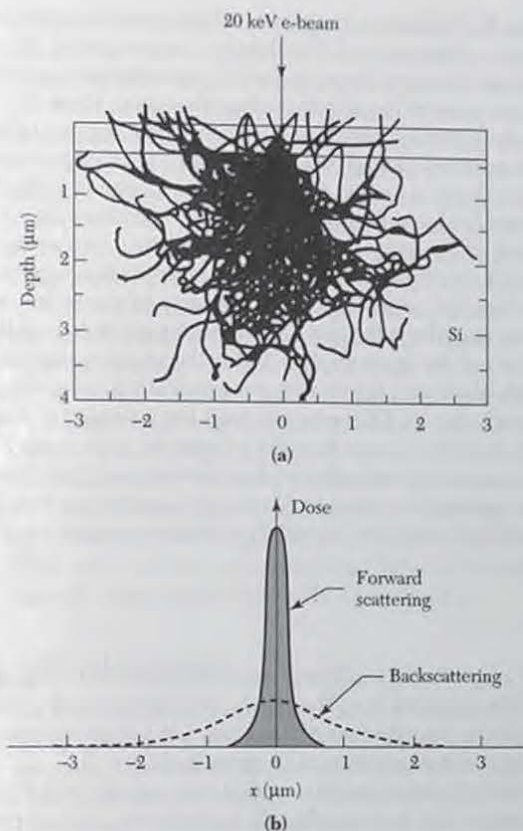


Figure 4.16 (a) Simulated trajectories of 100 electrons in PMMA for a 20-keV electron beam.¹⁴ (b) Dose distribution for forward scattering and backscattering at the resist-substrate interface.

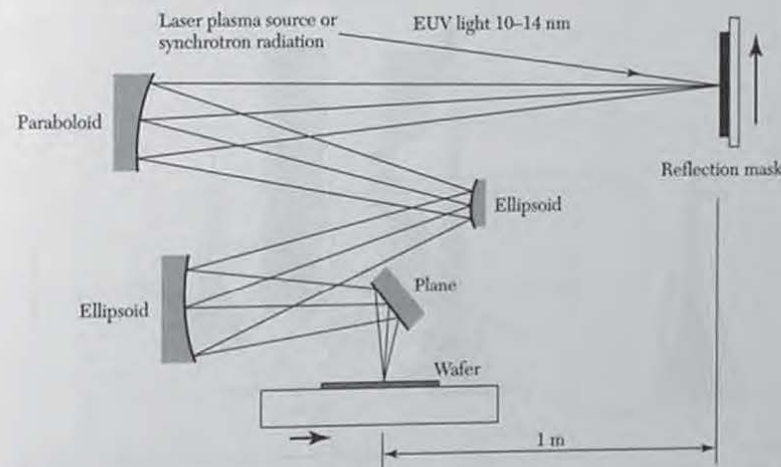


Figure 4.17 Schematic representation of an extreme ultraviolet (EUV) lithography system.¹⁵

a schematic diagram of an EUV lithography system. A laser-produced plasma or synchrotron radiation can serve as the source of EUV having a wavelength of 10 to 14 nm. The EUV radiation is reflected by a mask that is produced by patterning an absorber material deposited on a multilayer coated flat silicon or glass-plate mask blank. EUV radiation is reflected from the nonpatterned regions (i.e., nonabsorbing regions) of the mask through a 4 $\times$ reduction camera and imaged into a thin layer of resist on the wafer.

Since the EUV radiation beam is narrow, the mask must be scanned by the beam to illuminate the entire pattern field that describes the circuit mask layer. Also, for a 4 $\times$ 4 mirror (i.e., the one-paraboloid, two-ellipsoid, and one-plane mirror) reduction camera, the wafer must be scanned at one-fourth the mask speed in a direction opposite to the mask movement to reproduce the image field on all chip sites on the wafer surface. A precision system is required to perform the chip-site alignment and to control the wafer and mask stage movements and the exposure dose during the scanning process.

EUV lithography is capable of printing 50-nm features with PMMA resist using 13-nm radiation. However, the production of EUV exposure tools has a number of challenges. Since EUV is strongly absorbed in all materials, the lithography process must be performed in a vacuum. The camera must use reflective lens elements, and the mirrors must be coated with multilayer coatings that produce distributed quarter-wave Bragg reflectors. In addition, the mask blank must also be multilayer coated to maximize its reflectivity at λ of 10 to 14 nm.

4.2.3 X-Ray Lithography

X-ray lithography¹⁶ (XRL) is a potential candidate to succeed optical lithography for the fabrication of integrated circuits at 100 nm. The synchrotron storage ring is the choice of x-ray source for high-volume manufacturing. It can provide a large amount of collimated flux and can easily accommodate 10 to 20 exposure tools.

XRL uses a shadow printing method similar to optical proximity printing. Figure 4.18 shows a schematic XRL system. The x-ray wavelength is about 1 nm, and the printing is through a 1 $\times$ mask in close proximity (10–40 μ m) to the wafer. Since x-ray absorption

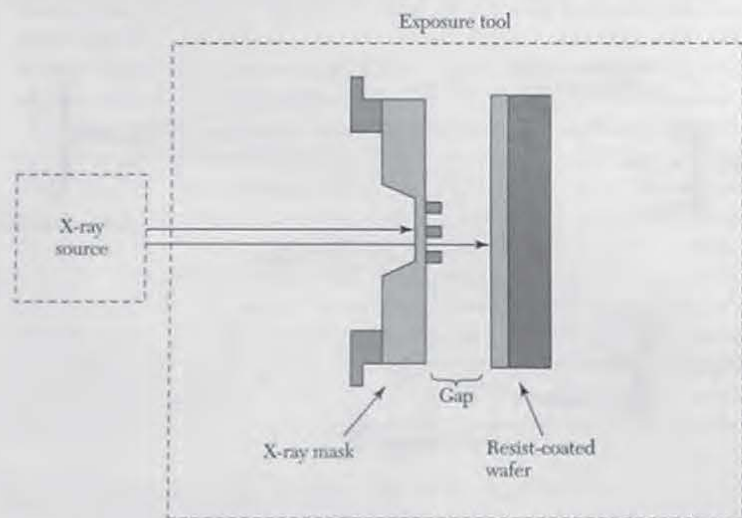


Figure 4.18 Schematic representation of a proximity x-ray lithography system.¹⁷

depends on the atomic number of the material and most materials have low transparency at $\lambda \approx 1$ nm, the mask substrate must be a thin membrane (1–2 μ m thick) made of low-atomic-number material, such as silicon carbide or silicon. The pattern itself is defined in a thin (~ 0.5 μ m), relatively high-atomic-number material, such as tantalum, tungsten, gold, or one of their alloys, which is supported by the thin membrane.

Masks are the most difficult and critical element of an XRL system, and the construction of an x-ray mask is much more complicated than that of a photomask. To avoid absorption of the x-rays between the source and mask, the exposure generally takes place in a helium environment. The x-rays are produced in vacuum, which is separated from the helium by a thin vacuum window (usually of beryllium). The mask substrate will absorb 25% to 35% of the incident flux and must therefore be cooled. An x-ray resist 1 μ m thick will absorb about 10% of the incident flux. There are no reflections from the substrate to create standing waves, so antireflection coatings are unnecessary.

We can use electron beam resists as x-ray resists because when an x-ray is absorbed by an atom, the atom goes to an excited state with the emission of an electron. The excited atom returns to its ground state by emitting an x-ray having a different wavelength than the incident x-ray. This x-ray is absorbed by another atom, and the process repeats. Since all the processes result in the emission of electrons, a resist film under x-ray irradiation is equivalent to one being irradiated by a large number of secondary electrons from any of the other processes. Once the resist film is irradiated, chain cross linking or chain scission will occur, depending on the type of resist.

4.2.4 Ion Beam Lithography

Ion beam lithography can achieve higher resolution than optical, x-ray, or electron beam lithographic techniques because ions have a higher mass and therefore scatter less than electrons. The most important application is the repair of masks for optical lithography, a task for which commercial systems are available.

Figure 4.19 shows the computer-simulated trajectories of 50 H⁺ ions implanted at 60 keV into PMMA and various substrates.¹⁷ Note that the spread of the ion beam at a depth of 0.4 μ m is only 0.1 μ m in all cases (compare with Fig. 4.16a for electrons). Backscattering is completely absent for the silicon substrate, and there is only a small

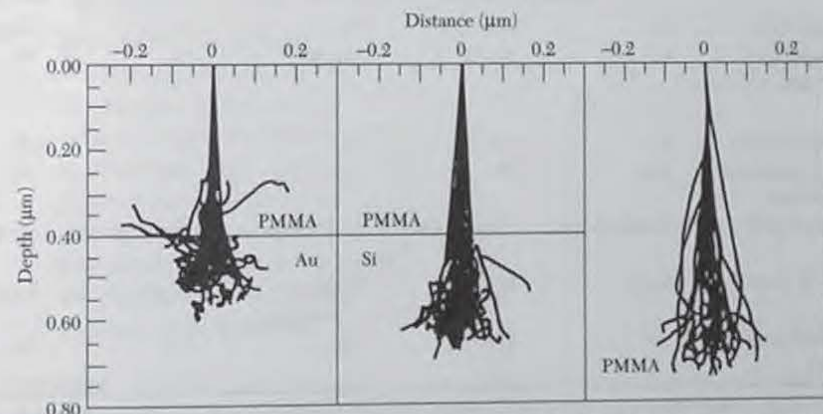


Figure 4.19 Trajectories of 60-keV H⁺ ions traveling through PMMA into Au, Si, and PMMA.¹⁷

amount of backscattering for the gold substrate. However, ion beam lithography may suffer from random (or stochastic) space-charge effects, causing broadening of the ion beam.

There are two types of ion beam lithography systems: a scanning focused-beam system and a mask-beam system. The former system is similar to the electron beam machine (Fig. 4.13), in which the ion source can be Ga^+ or H^+ . The latter system is similar to an optical 5 $\times$ reduction projection step-and-repeat system, which projects 100-keV light ions such as H_2^+ through a stencil mask.

4.2.5 Comparison of Various Lithographic Methods

The lithographic methods discussed earlier all have 100 nm or better resolution. A comparison of various lithographic technologies is shown in Table 4.1. However, each method has its own limitations: the diffraction effect in optical lithography, the proximity effect in electron beam lithography, mask fabrication complexities in x-ray lithography, difficulty in mask blank production for EUV lithography, and stochastic space charge in ion beam lithography.

For IC fabrication, many mask levels are involved. However, it is not necessary to use the same lithographic method for all levels. A mix-and-match approach can take advantage of the unique features of each lithographic process to improve resolution and to maximize throughput. For example, a 4:1 EUV method can be used for the most critical mask levels, whereas a 4:1 or 5:1 optical system can be used for the rest.

According to the Semiconductor Industry Association's *International Technology Roadmap for Semiconductors*, IC manufacturing technology will reach the 50 nm generation around 2010.¹⁸ With each new technology generation, lithography has become

TABLE 4.1 Comparison of Various Lithographic Technologies

	Optical 248/193 nm	SCALPEL	EUV	X-ray	Ion Beam
Exposure Tool					
Source	Laser	Filament	Laser plasma	Synchrotron	Multicusp
Diffraction limited	Yes	No	Yes	Yes	No
Optics	Refractive	Refractive	Refractive	No optics	Full-field refractive
Step and scan	Yes	Yes	Yes	Yes	Stepper
Throughput of 200-mm wafers/hr	40	30–35	20–30	30	30
Mask					
Demagnification	4 $\times$	4 $\times$	4 $\times$	1 $\times$	4 $\times$
Optical proximity correction	Yes	No	Yes	Yes	No
Radiation path	Transmission	Transmission	Reflection	Transmission	Stencil
Resist					
Single or multilayer	Single	Single	Surface imaging	Single	Single
Chemical-amplified resist	Yes	Yes	No	Yes	No

SCALPEL, Scattering with angular limitation projection electron beam lithography; EUV, extreme ultraviolet.

an even more important key driver for the semiconductor industry because of the requirements of smaller feature size and tighter overlay tolerance. In addition, lithography tool costs have become higher relative to the total equipment costs for an IC manufacturing facility. Currently, the technology development of next-generation lithography is conducted by multinational research projects or by industrial partners.

4.3 PHOTOLITHOGRAPHY SIMULATION

As is the case with oxidation (see Chapter 3), computer simulation is also an important tool for studying the photolithography process. The SUPREM package, unfortunately, is not capable of photolithographic simulation. However, another popular tool, PROLITH, does provide this capability.

PROLITH is a Windows-based program that uses a positive/negative photoresist optical lithography model originally developed by Chris Mack.¹⁹ PROLITH simulates the complete one- and two-dimensional optical lithography process from aerial image formation through resist exposure and development. The output of the program is an accurate prediction of the final resist profile, which is presented in a wide variety of images, plots, graphs, and calculations. In particular, PROLITH is able to simulate the following:

- Formation of an image of a mask feature by an optical projection system
- Exposure of photoresist by this image
- Diffusion of the image
- Development of the exposed photoresist

PROLITH accepts lithography information in the form of data files and input parameters and uses this information to simulate standard and advanced lithography processes. To run PROLITH, the user simply clicks on the PROLITH icon from the Windows Start menu. After a successful license search, the Imaging Tool parameters window appears (see Fig. 4.20). As the user makes choices from the View menu, PROLITH displays windows in which parameters may be entered in order to view simulation results. These can be observed from the Graphs menu. These ideas are illustrated in Example 3.

EXAMPLE 3

Use PROLITH to view the resist profile for the cylindrical mask feature in Figure 4.20 after exposure and development. Assume the following process conditions:

Photoresist type = SPR 500
 Pre-bake temperature = 95°C
 Pre-bake time = 60 seconds
 Numerical aperture of the lens = 0.5
 Exposure wavelength = 365 nm
 Exposure energy = 150 mJ/cm²
 Post-exposure bake temperature = 110°C
 Post-exposure bake time = 60 seconds
 Development time = 60 seconds
 Developer = MFT 245/501

SOLUTION All of the given values can be entered from the Parameters menu or by clicking on the appropriate icons on the toolbar. The resulting resist profile is shown in Figure 4.21.

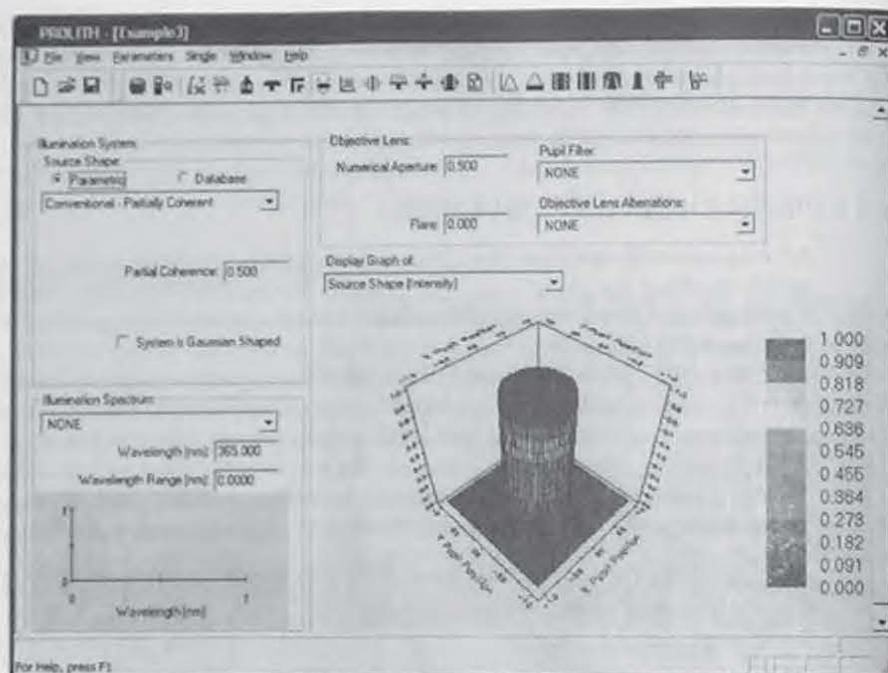


Figure 4.20 The Imaging Tools windows in PROLITH.

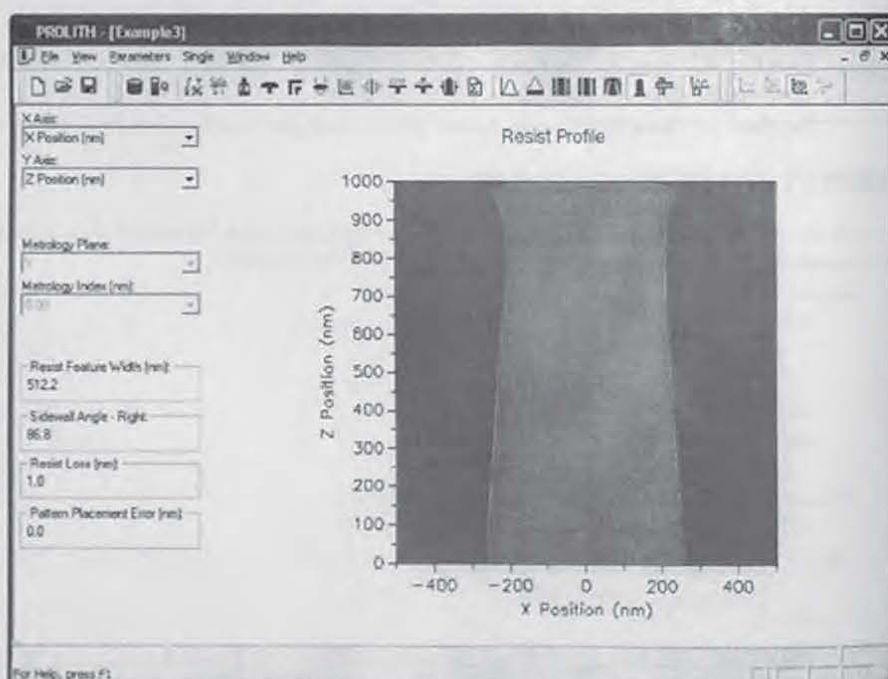


Figure 4.21 The resist profile for the mask feature specified in Fig. 4.20 and Example 3.

► 4.4 SUMMARY

The continued growth of the semiconductor industry is a direct result of the capability to transfer smaller and smaller circuit patterns onto semiconductor wafers. Currently, the vast majority of lithographic equipment is optical systems. This chapter considered various exposure tools, masks, photoresists, and the clean room for optical lithography. The primary factor limiting resolution in optical lithography is diffraction. However, because of advancements in excimer lasers, photoresist chemistry, and resolution enhancement techniques such as PSM and OPC, optical lithography will remain the mainstream technology, at least to the 100 nm generation.

Electron beam lithography is the technology of choice for mask making and nanofabrication, in which new device concepts are explored. Other lithographic processing technologies are EUV, x-ray, and ion beam lithography. Although all these have 100 nm or better resolution, each process has its own limitations: proximity effect in electron beam lithography, mask blank production difficulties in EUV lithography, mask fabrication complexity in x-ray lithography, and stochastic space charge in ion beam lithography.

At the present time, no obvious successor to optical lithography can be identified unambiguously. However, a mix-and-match approach can take advantage of the unique features of each lithography process to improve resolution and to maximize throughput.

► REFERENCES

1. For a more detailed discussion on lithography, see (a) K. Nakamura, "Lithography," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996. (b) P. Rai-Choudhury, *Handbook of Microlithography, Micromachining, and Microfabrication*, Vol. 1, SPIE, Washington, DC, 1997. (c) D. A. McGillis, "Lithography," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
2. For a more detailed discussion on etching, see Y. J. T. Liu, "Etching," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
3. J. M. Duffalo and J. R. Monkowski, "Particulate Contamination and Device Performance," *Solid State Technol.*, **27**, 3, 109 (1984).
4. H. P. Tseng and R. Jansen, "Cleanroom Technology," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
5. M. C. King, "Principles of Optical Lithography," in N. G. Einspruch, Ed., *VLSI Electronics*, Vol. 1, Academic, New York, 1981.
6. J. H. Bruning, "A Tutorial on Optical Lithography," in D. A. Doane, et al., Eds., *Semiconductor Technology*, Electrochemical Soc., Pennington, 1982.
7. R. K. Watts and J. H. Bruning, "A Review of Fine-Line Lithographic Techniques: Present and Future," *Solid State Technol.*, **24**, 5, 99 (1981).
8. W. C. Till and J. T. Luxon, *Integrated Circuits, Materials, Devices, and Fabrication*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
9. M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, "Improving Resolution in Photolithography with a Phase-Shift Mask," *IEEE Trans. Electron Devices*, **ED-29**, 18-28 (1982).
10. D. P. Kern, et al., "Practical Aspects of Microfabrication in the 100-nm Region," *Solid State Technol.*, **27**, 2, 127 (1984).
11. J. A. Reynolds, "An Overview of e-Beam Mask-Making," *Solid State Technol.*, **22**, 8, 87 (1979).
12. Y. Sameda, et al., "Electron-Beam Cell Projection Lithography: Its Accuracy and Its Throughput," *J. Vac. Sci. Technol.*, **B12** (6), 3399 (1994).
13. W. L. Brown, T. Venkatesan, and A. Wagner, "Ion Beam Lithography," *Solid State Technol.*, **24**, 8, 60 (1981).
14. D. S. Kyser and N. W. Viswanathan, "Monte Carlo Simulation of Spatially Distributed Beams in Electron-Beam Lithography," *J. Vac. Sci. Technol.*, **12**, 1305 (1975).

15. Charles Gwyn, et al., "Extreme Ultraviolet Lithography," White Paper, Sematech, Next Generation Lithography Workshop, Colorado Springs, Dec. 7–10, 1998.
16. J. P. Silverman, "Proximity X-Ray Lithography," White Paper, Sematech, Next Generation Lithography Workshop, Colorado Springs, Dec. 7–10, 1998.
17. L. Karapiperis, et al., "Ion Beam Exposure Profiles in PMMA-Computer Simulation," *J. Vac. Sci. Technol.*, **19**, 1259 (1981).
18. *The International Technology Roadmap for Semiconductors*, Semiconductor Ind. Assoc., San Jose, CA, 2001.
19. *PROLITH2 User's Manual*, FINLE Technologies, Austin, TX, 1998.

► PROBLEMS

SECTION 4.1: OPTICAL LITHOGRAPHY

1. For a class 100 clean room, find the number of dust particles per cubic meter with particle sizes (a) between 0.5 and 1 μm , (b) between 1 and 2 μm , and (c) above 2 μm .
2. Find the final yield for a nine-mask-level process in which the average fatal defect density per cm^2 is 0.1 for four levels, 0.25 for four levels, and 1.0 for one level. The chip area is 50 mm^2 .
3. An optical lithographic system has an exposure power of 0.3 mW/cm^2 . The required exposure energy for a positive photoresist is 140 mJ/cm^2 , and for a negative photoresist is 9 mJ/cm^2 . Assuming negligible times for loading and unloading wafers, compare the wafer throughput for positive photoresist and negative photoresist.
4. (a) For an ArF excimer laser 193-nm optical lithographic system with $\text{NA} = 0.65$, $k_1 = 0.60$, and $k_2 = 0.50$, what are the theoretical resolution and depth of focus for this tool? (b) What can we do in practice to adjust NA , k_1 , and k_2 parameters to improve resolution? (c) What parameter does the phase-shift mask (PSM) technique change to improve resolution?
5. The plots in Figure 4.9 are called *response curves* in microlithography. (a) What are the advantages and disadvantages of using resists with high γ values? (b) Conventional resists cannot be used for 248-nm or 193-nm lithography. Why not?

SECTION 4.2: NEXT-GENERATION LITHOGRAPHIC METHODS

6. (a) Explain why a shaped beam promises higher throughput than a Gaussian beam in e-beam lithography. (b) How can alignment be performed for e-beam lithography? Why is alignment in x-ray lithography so difficult? (c) What are the potential advantages of x-ray lithography over e-beam lithography?
7. Why has the operating mode of optical lithographic systems evolved from proximity printing to 1:1 projection printing and finally to 5:1 projection step-and-repeat? (b) Is it possible to build a step-and-scan x-ray lithographic system? Why or why not?

SECTION 4.3: PHOTOLITHOGRAPHY SIMULATION

8. Repeat Example 3 with the following revised process conditions:

Pre-bake temperature = 100°C
 Pre-bake time = 5 minutes
 Exposure energy = 50 mJ/cm^2
 Post-exposure bake temperature = 120°C
 Post-exposure bake time = 15 minutes
 Development time = 60 seconds
 Developer = MF 319

Explain any differences in the resist profile.

Etching

As discussed in the previous chapter, lithography is the process of transferring patterns to photoresist covering the surface of a semiconductor wafer. To produce circuit features, these resist patterns must be transferred into the underlying layers comprising the device. The pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer.¹ A brief description of etching was given in Section 1.4.2. The present chapter covers the following topics:

- Mechanisms for wet chemical etching of semiconductors, insulators, and metal films
- Plasma-assisted etching (also called dry etching) for high-fidelity pattern transfer

► 5.1 WET CHEMICAL ETCHING

Wet chemical etching is used extensively in semiconductor processing. Starting from semiconductor wafers sliced from an ingot (Chapter 2), chemical etchants are used for lapping and polishing to give an optically flat, damage-free surface. Prior to thermal oxidation (Chapter 3) or epitaxial growth (Chapter 8), semiconductor wafers are chemically cleaned to remove contamination that results from handling and storing. Wet chemical etching is especially suitable for blanket etches (i.e., over the whole wafer surface) of polysilicon, oxide, nitride, metals, and III-V compounds.

The mechanisms for wet chemical etching involve three essential steps, as illustrated in Figure 5.1: The reactants are transported by diffusion to the reacting surface, chemical reactions occur at the surface, and the products from the surface are removed by diffusion. Both agitation and the temperature of the etchant solution influence the etch rate, which is the amount of film removed by etching per unit time. In IC processing, most wet chemical etches proceed by immersing the wafers in a chemical solution or by spraying the wafers with the etchant solution. For immersion etching, the wafer is immersed in the etch solution, and mechanical agitation is usually required to ensure etch uniformity and a consistent etch rate. Spray etching has gradually replaced immersion etching because it greatly increases the etch rate and uniformity by constantly supplying fresh etchant to the wafer surface.

For semiconductor production lines, highly uniform etch rates are important. Etch rates must be uniform across a wafer, from wafer to wafer, from run to run, and for any variations in feature sizes and pattern densities. Etch rate uniformity is given by the following equation:

$$\text{Etch rate uniformity (\%)} = \frac{(\text{Maximum etch rate} - \text{Minimum etch rate})}{(\text{Maximum etch rate} + \text{Minimum etch rate})} \times 100\% \quad (1)$$

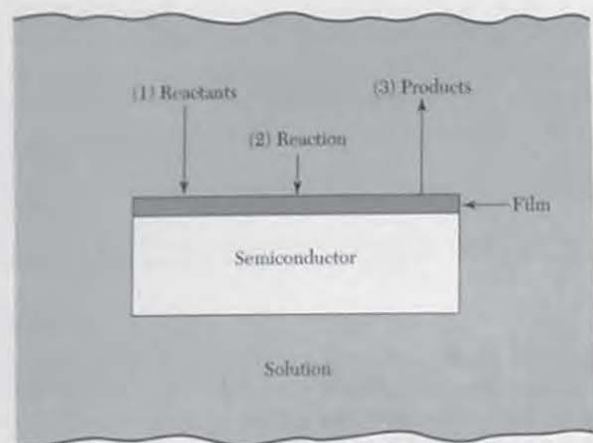


Figure 5.1 Basic mechanisms in wet chemical etching.

EXAMPLE 1

Calculate the Al average etch rate and etch rate uniformity on a 200-mm diameter silicon wafer, assuming the etch rates at the center, left, right, top, and bottom of the wafer are 750, 812, 765, 743, and 798 nm/min, respectively.

SOLUTION

$$\text{Al average etch rate} = (750 + 812 + 765 + 743 + 798) / 5 = 773.6 \text{ nm/min}$$

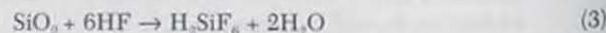
$$\text{Etch rate uniformity} = (812 - 743) / (812 + 743) \times 100\% = 4.4\%$$

5.1.1 Silicon Etching

For semiconductor materials, wet chemical etching usually proceeds by oxidation, followed by the dissolution of the oxide by a chemical reaction. For silicon, the most commonly used etchants are mixtures of nitric acid (HNO_3) and hydrofluoric acid (HF) in water or acetic acid (CH_3COOH). Nitric acid oxidizes silicon to form a SiO_2 layer.² The oxidation reaction is



Hydrofluoric acid is used to dissolve the SiO_2 layer. The reaction is



Water can be used as a diluent for this etchant. However, acetic acid is preferred because it reduces the dissolution of the nitric acid.

Some etchants dissolve a given crystal plane of single-crystal silicon much faster than another plane; this results in orientation-dependent etching.³ For a silicon lattice, the (111) plane has more available bonds per unit area than the (110) and (100) planes; therefore, the etch rate is expected to be slower for the (111) plane. A commonly used orientation-dependent etch for silicon consists of a mixture of KOH in water and isopropyl alcohol.

For example, a solution with 19 wt % KOH in deionized (DI) water at about 80°C removes the (100) plane at a much higher rate than the (110) and (111) planes. The ratio of the etch rates for the (100), (110), and (111) planes is 100:16:1.

Orientation-dependent etching of $\langle 100 \rangle$ -oriented silicon through a patterned silicon dioxide mask creates precise V-shaped grooves,⁴ the edges being the (111) planes at an angle of 54.7° from the (111) surface, as shown at the left of Figure 5.2a. If the window in the mask is sufficiently large or if the etching time is short, a U-shaped groove will be formed, as shown at the right of Figure 5.2a. The width of the bottom surface is given by

$$W_b = W_0 - 2l \cot 54.7^\circ$$

or

$$W_b = W_0 - \sqrt{2}l \quad (4)$$

where W_0 is the width of the window on the wafer surface and l is the etched depth. If $\langle 110 \rangle$ -oriented silicon is used, essentially straight-walled grooves with sides of (111) planes can be formed, as shown in Figure 5.2b. We can use the large orientation dependence in the etch rates to fabricate device structures with submicron feature lengths.

5.1.2 Silicon Dioxide Etching

The wet etching of silicon dioxide is commonly achieved in a dilute solution of HF with or without the addition of ammonium fluoride (NH_4F). Adding NH_4F creates what is

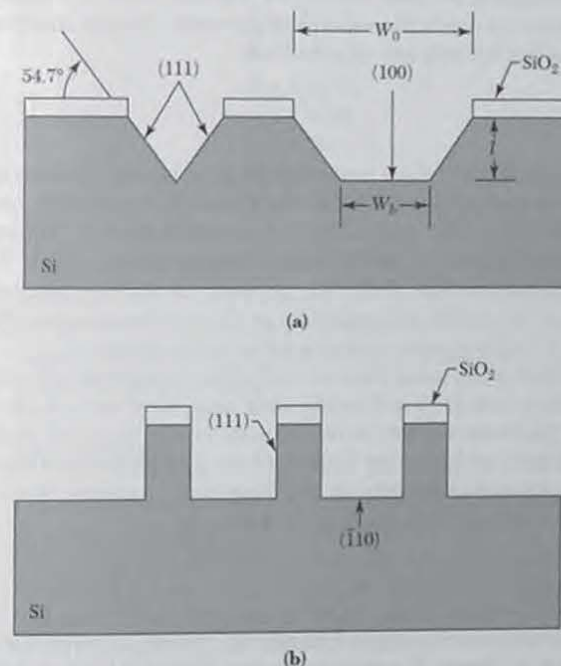


Figure 5.2 Orientation-dependent etching.⁴ (a) Through window patterns on $\langle 100 \rangle$ -oriented silicon. (b) Through window patterns on $\langle 110 \rangle$ -oriented silicon.

referred to as a *buffered HF solution* (BHF), also called *buffered oxide etch* (BOE). The addition of NH_4F to HF controls the pH value and replenishes the depletion of the fluoride ions, thus maintaining stable etching performance. The overall reaction for SiO_2 etching is the same as that in Eq. 3. The etch rate of SiO_2 etching depends on etchant solution, etchant concentration, agitation, and temperature. In addition, density, porosity, microstructure, and the presence of impurities in the oxide influence the etch rate. For example, a high concentration of phosphorus in the oxide results in a rapid increase in the etch rate, and a loosely structured oxide formed by chemical vapor deposition (CVD) or sputtering exhibits a faster etch rate than thermally grown oxide.

Silicon dioxide can also be etched in vapor-phase HF. Vapor-phase HF oxide etch technology has a potential for submicron feature etching because the process can be well controlled.

5.1.3 Silicon Nitride and Polysilicon Etching

Silicon nitride films are etchable at room temperature in concentrated HF or buffered HF and in a boiling H_3PO_4 solution. Selective etching of nitride to oxide is done with 85% H_3PO_4 at 180°C because this solution attacks silicon dioxide very slowly. The etch rate is typically 10 nm/min for silicon nitride, but less than 1 nm/min for silicon dioxide. However, photoresist adhesion problems are encountered when etching nitride with boiling H_3PO_4 solution. Better patterning can be achieved by depositing a thin oxide layer on top of the nitride film before resist coating. The resist pattern is transferred to the oxide layer, which then acts as a mask for subsequent nitride etching.

Etching polysilicon is similar to etching single-crystal silicon. However, the etch rate is considerably faster because of the grain boundaries. The etch solution is usually modified to ensure that it does not attack the underlying gate oxide. Dopant concentrations and temperature may affect the etch rate of polysilicon.

5.1.4 Aluminum Etching

Aluminum and aluminum alloy films are generally etched in heated solutions of phosphoric acid, nitric acid, acetic acid, and DI water. The typical etchant is a solution of 73% H_3PO_4 , 4% HNO_3 , 3.5% CH_3COOH , and 19.5% DI water at 30°C to 80°C . The wet etching of aluminum proceeds as follows: HNO_3 oxidizes aluminum, and H_3PO_4 then dissolves the oxidized aluminum. The etch rate depends on etchant concentration, temperature, agitation of the wafers, and impurities or alloys in the aluminum film. For example, the etch rate is reduced when copper is added to the aluminum.

Wet etching of insulating and metal films is usually done with similar chemicals that dissolve these materials in bulk form and involve their conversion into soluble salts or complexes. Generally, film materials will be etched more rapidly than their bulk counterparts. Also, the etch rates are higher for films that have a poor microstructure, built-in stress, or departure from stoichiometry, or that have been irradiated. Some useful etchants for insulating and metal films are listed in Table 5.1.

5.1.5 Gallium Arsenide Etching

A wide variety of etches have been investigated for gallium arsenide; however, few of them are truly isotropic.⁵ This is because the surface activities of the (111) Ga and (111) As faces are very different. Most etches give a polished surface on the arsenic face, but the gallium face tends to show crystallographic defects and etches more slowly. The most commonly used etchants are the $\text{H}_2\text{SO}_4\text{-H}_2\text{O}_2\text{-H}_2\text{O}$ and $\text{H}_3\text{PO}_4\text{-H}_2\text{O}_2\text{-H}_2\text{O}$ systems. For

TABLE 5.1 Etchants for Insulators and Conductors

Material	Etchant Composition	Etch Rate (nm/min)
SiO_2	28 ml HF	100
	170 ml HF	
	113 g NH_4F	
	15 ml HF	12
	10 ml HNO_3	
	300 ml H_2O	
Si_3N_4	Buffered HF	0.5
Al	H_3PO_4	10
	4 ml HNO_3	30
	3.5 ml CH_3COOH	
	73 ml H_3PO_4	
	19.5 ml H_2O	1000
	4 g KI	
Au	1 g I_2	500
	40 ml H_2O	
	5 ml H_3PO_4	
Mo	2 ml HNO_3	50
	4 ml CH_3COOH	
	150 ml H_2O	
Pt	1 ml HNO_3	50
	7 ml HCl	
	8 ml H_2O	
W	34 g KH_2PO_4	160
	13.4 g KOH	
	33 g $\text{K}_3\text{Fe}(\text{CN})_6$	
	H_2O to make 1 liter	

an etchant with an 8:1:1 volume ratio of $\text{H}_2\text{SO}_4\text{:H}_2\text{O}_2\text{:H}_2\text{O}$, the etch rate is $0.8\text{ }\mu\text{m/min}$ for the <111> Ga face and $1.5\text{ }\mu\text{m/min}$ for all other faces. For an etchant with a 3:1:50 volume ratio of $\text{H}_3\text{PO}_4\text{:H}_2\text{O}_2\text{:H}_2\text{O}$, the etch rate is $0.4\text{ }\mu\text{m/min}$ for the <111> Ga face and $0.8\text{ }\mu\text{m}$ for all other faces.

5.2 DRY ETCHING

In pattern transfer operations, a resist pattern is defined by a lithographic process to serve as a mask for etching of its underlying layer (Fig. 5.3a).⁶ Most of the layer materials (e.g., SiO_2 , Si_3N_4 , and deposited metals) are amorphous or polycrystalline thin films. If they are etched in a wet chemical etchant, the etch rate is generally isotropic (i.e., the lateral and vertical etch rates are the same), as illustrated in Figure 5.3b. If h_t is the thickness of the layer material and l the lateral distance etched underneath the resist mask, we can define the degree of anisotropy (A_t) by

$$A_t \equiv 1 - \frac{l}{h_t} = 1 - \frac{R_l t}{R_v t} = 1 - \frac{R_l}{R_v} \quad (5)$$

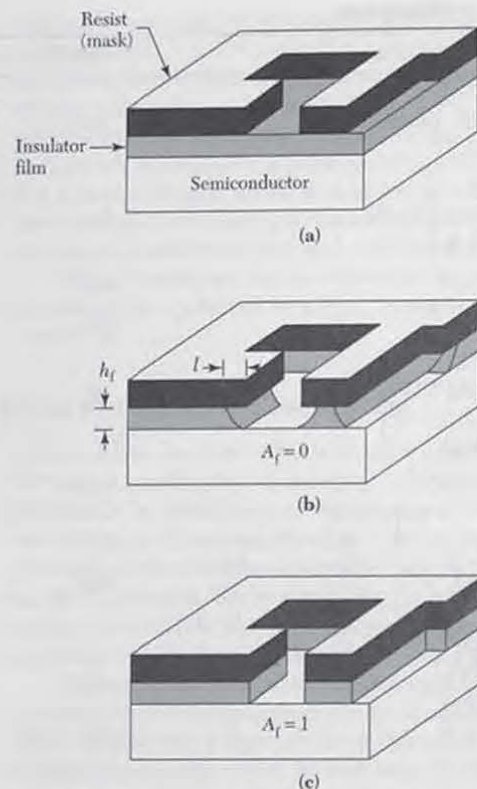


Figure 5.3 Comparison of wet chemical etching and dry etching for pattern transfer.⁹

where t is time and R_l and R_v are the lateral and vertical etch rates, respectively. For isotropic etching, $R_l = R_v$ and $A_t = 0$.

The major disadvantage of wet chemical etching in pattern transfer is the undercutting of the layer underneath the mask, resulting in a loss of resolution in the etched pattern. In practice, for isotropic etching, the film thickness should be about one-third or less of the resolution required. If patterns are required with resolutions much smaller than the film thickness, anisotropic etching (i.e., $1 \geq A_t > 0$) must be used. In practice, the value of A_t is chosen to be close to unity. Figure 5.3c shows the limiting case where $A_t = 1$, corresponding to $l = 0$ (or $R_l = 0$).

To achieve a high-fidelity transfer of the resist patterns required for ultralarge-scale integration processing, dry etching methods have been developed. Dry etching is synonymous with *plasma-assisted etching*, which denotes several techniques that use plasma in the form of low-pressure discharges. Dry etch methods include plasma etching, reactive ion etching (RIE), sputter etching, magnetically enhanced RIE (MERIE), reactive ion beam etching, and high-density plasma (HDP) etching.

5.2.1 Plasma Fundamentals

A plasma is a fully or partially ionized gas composed of equal numbers of positive and negative charges and a different number of unionized molecules. A plasma is produced

when an electric field of sufficient magnitude is applied to a gas, causing the gas to break down and become ionized. The plasma is initiated by free electrons that are released by some means, such as field emission from a negatively biased electrode. The free electrons gain kinetic energy from the electric field. In the course of their travel through the gas, the electrons collide with gas molecules and lose their energy. The energy transferred in the collision causes the gas molecules to be ionized (i.e., to free electrons). The free electrons gain kinetic energy from the field, and the process continues. Therefore, when the applied voltage is larger than the breakdown potential, a sustained plasma is formed throughout the reaction chamber.

The electron concentrations in the plasma for dry etchings are relatively low, typically on the order of 10^9 to 10^{12} cm⁻³. At a pressure of 1 Torr, the concentrations of gas molecules are 10^4 to 10^7 times higher than the electron concentrations. This results in an average gas temperature in the range of 50°C to 100°C. Therefore, plasma-assisted dry etching is a low-temperature process.

EXAMPLE 2

The electron densities in RIE and HDP systems range from 10^9 to 10^{10} cm⁻³ and 10^{11} to 10^{12} cm⁻³, respectively. Assuming the RIE chamber pressure is 200 mTorr and HDP chamber pressure is 5 mTorr, calculate the ionization efficiency in RIE reactors and HDP reactors at room temperature. The ionization efficiency is the ratio of the electron density to the density of molecules.

SOLUTION

$$PV = nRT$$

where P is the pressure in atm (1 atm = 760,000 mTorr), V is the volume in liters, n is the number of moles, R is the gas constant (0.082 liter-atm/mol-K), and T is the absolute temperature in K, respectively.

For the RIE system,

$$\begin{aligned} n/V &= P/RT = (200/760,000)/(0.082 \times 300) = 1.06 \times 10^{-3} \text{ (mol/liter)} \\ &= 1.06 \times 10^{-3} \times 6.02 \times 10^{23} + 1000 \\ &= 6.38 \times 10^{13} \text{ (cm}^{-3}\text{)} \\ \text{Ionization efficiency} &= (10^9 - 10^{10})/(6.38 \times 10^{13}) \\ &= 1.56 \times 10^{-7} - 1.56 \times 10^{-6} \end{aligned}$$

For the HDP system,

$$\begin{aligned} n/V &= P/RT = (5/760,000)/(0.082 \times 300) = 2.66 \times 10^{-7} \text{ (mol/liter)} \\ &= 2.66 \times 10^{-7} \times 6.02 \times 10^{23} + 1000 \\ &= 1.6 \times 10^{14} \text{ (cm}^{-3}\text{)} \\ \text{Ionization efficiency} &= (10^{11} - 10^{12})/(1.6 \times 10^{14}) \\ &= 6.25 \times 10^{-4} - 6.25 \times 10^{-3} \end{aligned}$$

Therefore, HDP has much higher ionization efficiency than RIE.

5.2.2 Etch Mechanism, Plasma Diagnostics, and End-Point Control

Plasma etching is a process in which a solid film is removed by a chemical reaction with ground-state or excited-state neutral species. Plasma etching is often enhanced or induced

by energetic ions generated in a gaseous discharge. The basic etch mechanism, plasma diagnostics, and end-point control are introduced briefly in this section.

Etch Mechanism

Plasma etching proceeds in five steps, as illustrated in Figure 5.4. First, the etchant species is generated in the plasma. The reactant is then transported by diffusion through a stagnant gas layer to the surface. Next, the reactant is adsorbed on the surface. A chemical reaction (along with physical effects such as ion bombardment) follows to form volatile compounds. Finally, the compounds are desorbed from the surface, diffused into the bulk gas, and pumped out by the vacuum system.⁷

Plasma etching is based on the generation of plasma in a gas at low pressure. Two basic types of methods are used: physical methods and chemical methods. The former includes sputter etching, and the latter includes pure chemical etching. In physical etching, positive ions bombard the surface at high speed; small amounts of negative ions formed in the plasma cannot reach the wafer surface and therefore play no direct role in plasma etching. In chemical etching, neutral reactive species generated by the plasma interact with the material surface to form volatile products. Chemical and physical etch mechanisms have different characteristics. Chemical etching exhibits a high etch rate and good selectivity (i.e., the ratio of etch rates for different materials) and produces low ion bombardment-induced damage, but yields isotropic profiles. Physical etching can yield anisotropic profiles, but is associated with low etch selectivity and high bombardment-induced damage. Combinations of chemical and physical etching give anisotropic etch profiles, reasonably good selectivity, and moderate bombardment-induced damage. An example is the RIE process, which uses a physical method to assist chemical etching or creates reactive ions to participate in chemical etching.

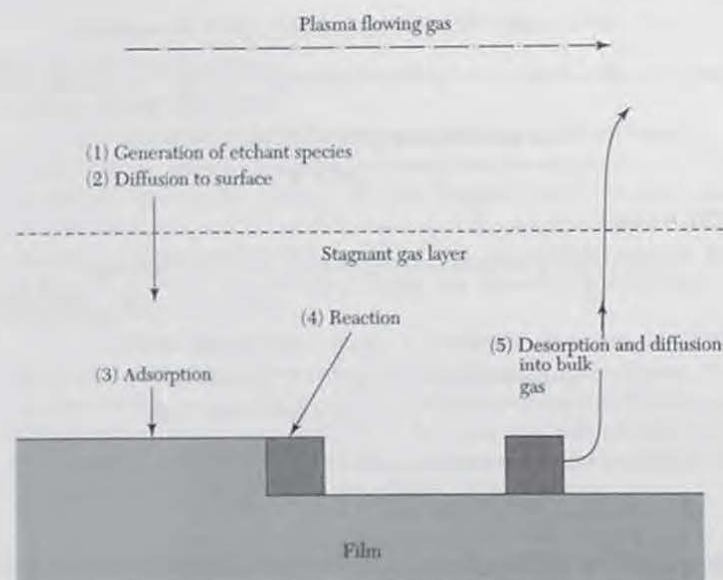


Figure 5.4 Basic steps in a dry etching process.⁷

Plasma Diagnostics

Most processing plasmas emit radiation in the range from infrared to ultraviolet. A simple analytical technique is to measure the intensity of these emissions versus wavelength with the aid of optical emission spectroscopy (OES). Using observed spectral peaks, it is usually possible to determine the presence of neutral and ionic species by correlating these emissions with previously determined spectral series. Relative concentrations of the species can be obtained by correlating changes in intensity with the plasma parameters. The emission signal derived from the primary etchant or by-product begins to rise or fall at the end of the etch cycle.

End-Point Control

Dry etching differs from wet chemical etching in that dry etching has less etch selectivity to the underlying layer. Therefore, the plasma reactor must be equipped with a monitor that indicates when the etching process is to be terminated (i.e., an *end-point detection system*). Laser interferometry of the wafer surface is used to continuously monitor etch rates and to determine the end point. During etching, the intensity of laser light reflected off a thin film surface oscillates. This oscillation occurs because of the phase interference between the light reflected from the outer and inner interfaces of the etching layer. This layer must therefore be optically transparent or semitransparent to observe the oscillation. Figure 5.5 shows a typical signal from a silicide/polycrystalline Si gate etch. The period of the oscillation is related to the change in film thickness by

$$\Delta d = \lambda / 2\bar{n} \quad (6)$$

where Δd is the change in film thickness for one period of reflected light, λ is the wavelength of the laser light, and $\bar{n}$ is the refractive index of the layer being etched. For example, Δd for polysilicon is 80 nm, measured by using a helium-neon laser source for which $\lambda = 632.8$ nm.

5.2.3 Reactive Plasma Etching Techniques and Equipment

Plasma reactor technology in the IC industry has changed dramatically since the first application of plasma processing to photoresist stripping. A reactor for plasma etching contains a vacuum chamber, pump system, power supply generators, pressure sensors, gas flow control units, and end-point detector. Table 5.2 shows the similarities and differences

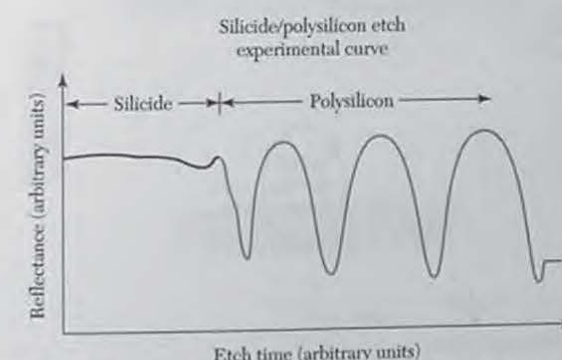


Figure 5.5 The relative reflectance of the etching surface of a composite silicide/poly-Si layer. The end point of the etch is indicated by the cessation of the reflectance oscillation.

TABLE 5.2 Etch Mechanisms and Pressure Ranges of Plasma Reactors

Etch Tool Configuration	Etch Mechanism	Pressure Range (Torr)
Barrel etching	Chemical	0.1–10
Downstream plasma etching	Chemical	0.1–10
Reactive ion etching (RIE)	Chemical and physical	0.01–1
Magnetic enhanced RIE	Chemical and physical	0.01–1
Magnetic confinement triode RIE	Chemical and physical	0.001–0.1
Electron cyclotron resonance plasma etch	Chemical and physical	0.001–0.1
Inductively coupled plasma or transformer-coupled plasma	Chemical and physical	0.001–0.1
Surface wave coupled plasma or helicon plasma etching	Chemical and physical	0.001–0.1

in the types of etching equipment that are commercially available. A comparison of pressure operating ranges and ion energies for different types of reactors is shown in Figure 5.6. Each etch tool is designed empirically and uses a particular combination of pressure, electrode configuration and type, and source frequency to control the two primary etch mechanisms—chemical and physical. Higher etch rates and tool automation are required for most etchers used in manufacturing.

Reactive Ion Etching

RIE has been extensively used in the microelectronics industry. In a parallel-plate diode system, a radio frequency capacitively coupled bottom electrode holds the wafer. This allows the grounded electrode to have a significantly larger area because it is, in fact, the chamber itself. The larger grounded area combined with the lower operating pressure (<500 mTorr) causes the wafers to be subjected to a heavy bombardment of energetic ions from the plasma as a result of the large negative self-bias at the wafer surface.

The etch selectivity of this system is relatively low compared with traditional barrel-etch systems because of strong physical sputtering. However, selectivity can be improved

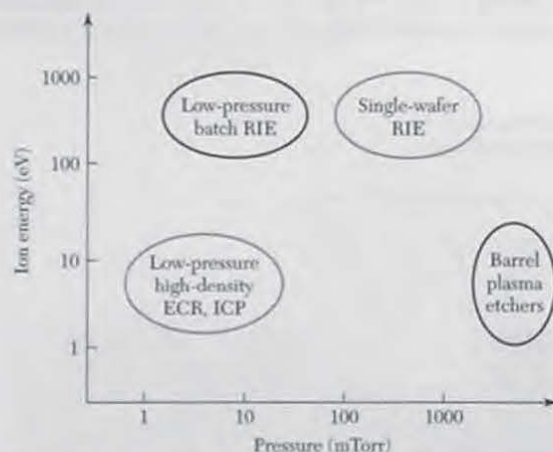


Figure 5.6 Comparison of ion energy and operating pressure ranges for different types of plasma reactors.

by choosing the proper etch chemistry; for example, by polymerizing the silicon surface with fluorocarbon polymers to obtain selectivity of SiO_2 over silicon. Alternatively, a triode-configuration RIE etch, as shown in Figure 5.7, can separate plasma generation from ion transport. Ion energy is controlled through a separate bias on the wafer electrode, thereby minimizing the loss of selectivity and the ion bombardment-induced damage observed in most traditional RIE systems.

Electron Cyclotron Resonance Plasma Etching

Most parallel-plate plasma etchers, except triode RIE, do not provide the ability to control plasma parameters such as electron energy, plasma density, and reactant density independently. As a result, ion bombardment-induced damage becomes a serious problem. The *electron cyclotron resonance* (ECR) reactor combines microwave power with a static magnetic field to force electrons to circulate around the magnetic field lines at an angular frequency. When this frequency equals the applied microwave frequency, a resonance coupling occurs between the electron energy and the applied electric field that results in a high degree of dissociation and ionization (10^{-2} for ECR compared with 10^{-6} for RIE). Figure 5.8 shows an ECR reaction chamber configuration. Microwave power is coupled through a microwave window into the ECR source region. The magnetic field is supplied from the magnetic coils. ECR plasma systems can also be used in thin film deposition. High efficiency in exciting the reactants in ECR plasmas allows the deposition of films at room temperature without the need for thermal activation.

Other High-Density Plasma Etchers

As feature sizes for ULSI continue to decrease, the limits of the conventional RIE system are being approached. In addition to the ECR system, other types of high-density plasma sources, such as the inductively coupled plasma (ICP) source, the transformer-coupled plasma (TCP) source, and surface wave-coupled plasma (SWP) sources, have been developed. These etchers have high plasma density (10^{11} – 10^{12} cm^{-3}) and low processing pressure (<20 mTorr). In addition, they allow the wafer platen to be powered independently of the source, providing significant decoupling between the ion energy (wafer bias) and the ion flux (plasma density, primarily driven by source power). The primary

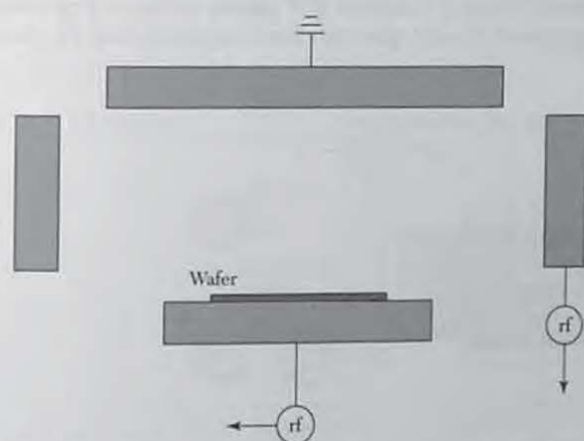


Figure 5.7 Schematic of a triode reactive ion etch reactor. The ion energy is separately controlled by a bias voltage on the bottom electrode. rf, radio frequency.

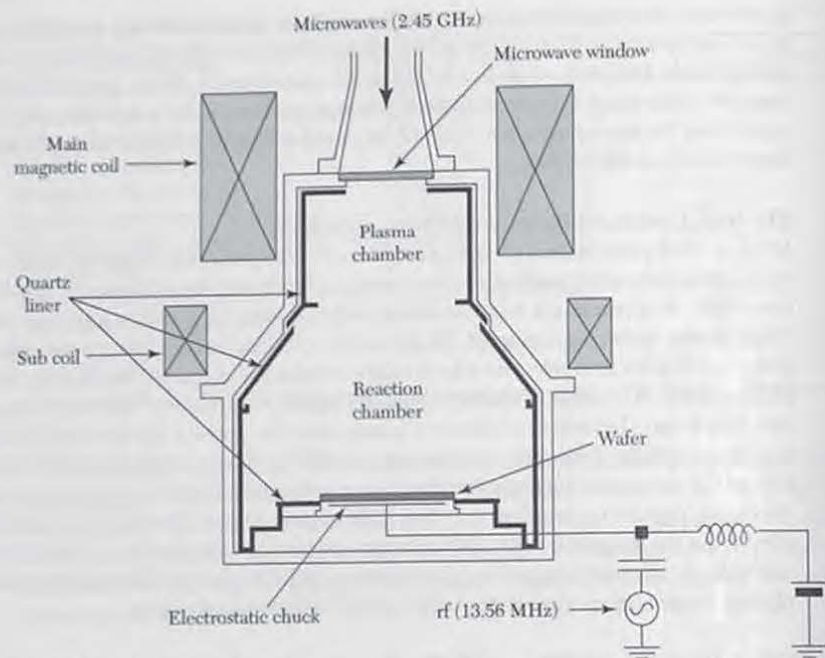


Figure 5.8 Schematic of an electron cyclotron resonance reactor.

processing advantages of HDP sources are better critical dimension (CD) control, higher etching rates, and better selectivity.

In addition, HDP sources provide low substrate damage (because of independent biasing of the substrate and the side electrode potentials) and high anisotropy (because of low pressure, yet high active species density). However, because of their complexity and higher cost, these systems may not be used for less critical applications, such as spacer etching or planarization.⁸ Figure 5.9 shows a TCP plasma reactor. A high-density, low pressure plasma is generated by a flat spiral coil that is separated from the plasma by a

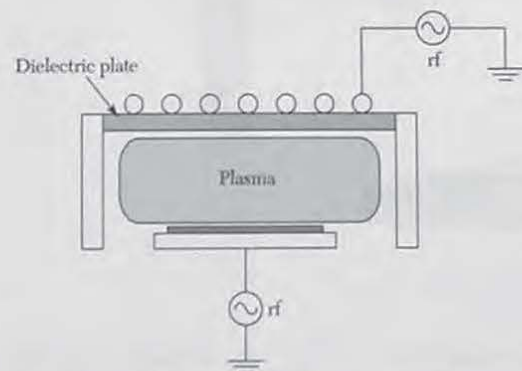


Figure 5.9 Schematic of a transformer-coupled plasma reactor.

dielectric plate on the top of the reactor. The wafer is located away from the coil, so it is not affected by the electromagnetic field generated by the coil. There is little plasma density loss because plasma is generated only a few mean free paths away from the wafer surface. Therefore, a high-density plasma and high etch rates are achieved.

Clustered Plasma Processing

Semiconductor wafers are processed in clean rooms to minimize exposure to ambient particulate contamination. As device dimensions shrink, particulate contamination becomes a more serious problem. To minimize particulate contamination, clustered plasma tools use a wafer handler to pass wafers from one process chamber to another in a vacuum environment. Clustered plasma processing tools can also increase throughput. Figure 5.10 shows the multilayer metal (TiW/AlCu/TiW) etching process with clustered tools in an AlCu etch chamber, a TiW etch chamber, and a strip passivation chamber. Clustered tools provide an economic advantage by having high chip yield because the wafer is exposed to less ambient contamination and is handled less.

5.2.4 Reactive Plasma Etching Applications

Plasma etching has rapidly evolved from simple batch resist stripping to large and single-wafer processing. Etching systems continue to be improved, from the conventional RIE tool to the high-density plasma tool for pattern transfer of deep-submicron devices. Aside from the etching tool, etch chemistry also plays a critical role in the performance of the etch process. Table 5.3 lists some etch chemistries for different etch processes. Developing an etch process usually means optimizing etch rate, selectivity, profile control, critical dimension, damage, and so forth by adjusting a large number of process parameters.

Silicon Trench Etching

As device feature size decreases, a corresponding decrease is needed in the wafer surface area occupied by the isolation between circuit elements and the storage capacitor

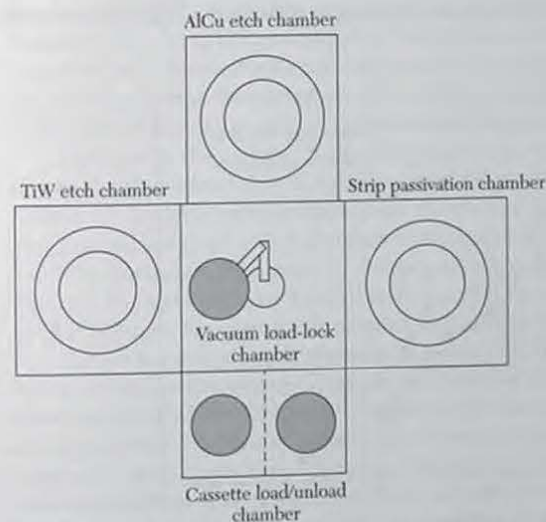


Figure 5.10 Cluster reactive ion etch tool for multilayer metal (TiW/AlCu/TiW) interconnect etching.²

TABLE 5.3 Etch Chemistries of Different Etch Processes

Material Being Etched	Etching Chemistry
Deep Si trench	HBr/NF ₃ /O ₂ /SF ₆
Shallow Si trench	HBr/Cl ₂ /O ₂
Poly Si	HBr/Cl ₂ /O ₂ , HBr/O ₂ , BCl ₃ /Cl ₂ , SF ₆
Al	BCl ₃ /Cl ₂ , SiCl ₄ /Cl ₂ , HBr/Cl ₂
AlSiCu	BCl ₃ /Cl ₂ /N ₂
W	SF ₆ only, NF ₃ /Cl ₂
TiW	SF ₆ only
WSi ₂ , TiSi ₂ , CoSi ₂	CCl ₄ /F ₂ /NF ₃ , CF ₄ /Cl ₂ , Cl ₂ /N ₂ /C ₂ F ₆
SiO ₂	CF ₄ /CHF ₃ /Ar, C ₂ F ₆ , C ₃ F ₈ , C ₄ F ₈ /CO, C ₅ F ₈ , CH ₂ F ₂
Si ₃ N ₄	CHF ₃ /O ₂ , CH ₂ F ₂ , CH ₃ CHF ₂

of a DRAM cell. This surface area can be reduced by etching trenches into the silicon substrate and filling them with suitable dielectric or conductive materials. Deep trenches, usually with a depth greater than 5 μm , are used mainly for forming storage capacitors. Shallow trenches, usually with a depth less than 1 μm , are often used for isolation.

Chlorine-based and bromine-based chemistries have a high silicon etch rate and high etch selectivity to the silicon dioxide mask. The combination of HBr + NF₃ + SF₆ + O₂ gas mixtures is used to form a trench capacitor with a depth of approximately 7 μm . This combination is also used for shallow trench isolation etching. *Aspect ratio-dependent etching* (i.e., variation in etch rate with aspect ratio) is often observed in deep silicon trench etching, caused by limited ion and neutral transport within the trench. Figure 5.11 shows the dependence of average silicon trench etch rate on aspect ratio. Trenches with large aspect ratios are etched more slowly than trenches with small aspect ratios.

Polysilicon and Polycide Gate Etching

Polysilicon or polycide (i.e., low-resistance metal silicides over polysilicon) is usually used as a gate material for MOS devices. Anisotropic etching and high etch selectivity to the

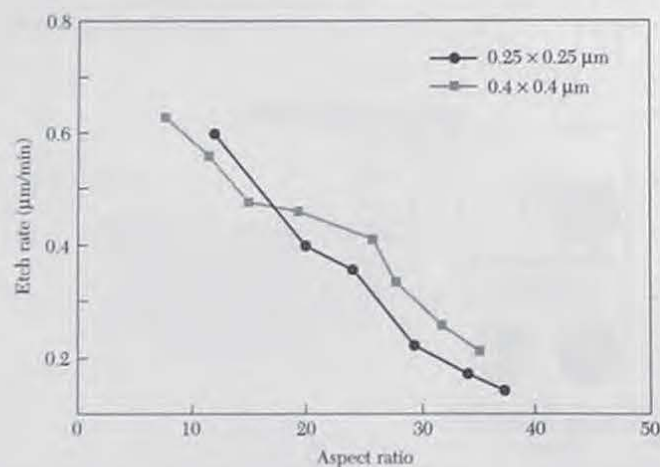


Figure 5.11 Dependence of average silicon trench etch rate on aspect ratio.²

gate oxide are the most important requirements for gate etching. For example, the selectivity required in 1G DRAM is more than 150 (i.e., the ratio of etch rates for polycide and gate oxide is 150:1). Achieving high selectivity and etch anisotropy at the same time is difficult for most ion-enhanced etching processes. Therefore, multistep processing is used, in which different etch steps in the process are optimized for etch anisotropy and selectivity. On the other hand, the trend in plasma technology for anisotropic etching and high selectivity is to utilize a low-pressure, high-density plasma using a relatively low power. Most chlorine-based and bromine-based chemistries can be used for gate etching to achieve the required etch anisotropy and selectivity.

Dielectric Etching

The patterning of dielectrics, especially silicon dioxide and silicon nitride, is a key process in the manufacture of modern semiconductor devices. Because of their higher bonding energies, dielectric etching requires aggressive ion-enhanced, fluorine-based plasma chemistry. Vertical profiles are achieved by sidewall passivation, typically by introducing a carbon-containing fluorine species to the plasma (e.g., CF₄, CHF₃, C₂F₆). High ion-bombardment energies are required to remove this polymer layer from the oxide, as well as to mix the reactive species into the oxide surface to form SiF_x products.

A low-pressure, high-density plasma is advantageous for aspect ratio-dependent etching. However, the HDP generates high-temperature electrons and subsequently generates a high degree of dissociation of ions and radicals. It generates far more active radicals and ions than RIE or MERIE plasmas. In particular, a high F concentration worsens the selectivity to silicon. Various methods have been tried to enhance the selectivities in the high-density plasma. First, a parent gas with a high C/F ratio, such as C₂F₆, C₃F₈, or C₄F₈, was attempted. Other methods to scavenge F radicals have also been developed.⁹

Interconnect Metal Etching

Etching of a metallization layer is a very important step in IC fabrication. Aluminum, copper, and tungsten are the most popular materials used for interconnection. These materials usually require anisotropic etching. The reaction of aluminum with fluorine results in nonvolatile AlF₃, which has a vapor pressure of only 1 Torr at 1240°C. Chlorine-based chemistry (e.g., a Cl₂/BCl₃ mixture) has been widely used for aluminum etching. Chlorine has a very high chemical etch rate with aluminum and tends to produce an undercut during etching. Carbon-containing gas (e.g., CHF₃) or N₂ is added to form sidewall passivation during aluminum etching to obtain anisotropic etching.

Exposure to the ambient is another problem in aluminum etching. Residual chlorine on the aluminum sidewall and the photoresist tends to react with atmospheric water to form HCl, which corrodes aluminum. An *in situ* exposure of the wafer to a CF₄ discharge to exchange Cl with F and then to an oxygen discharge to remove the resist, followed by immediate immersion in deionized water, can eliminate aluminum corrosion. Figure 5.12 shows 0.35- μm TiN/Al/Ti lines and spaces on a wafer maintained at ambient for 72 hours. No corrosion is present even after prolonged exposure to the ambient.

Copper has drawn much attention as a new metallization material in ULSI circuits because of its low resistivity ($\sim 1.7 \Omega\text{-cm}$) and superior resistance to electromigration compared with Al or Al alloys. However, because of the low volatility of copper halides, plasma etching at room temperature is difficult. Process temperatures higher than 200°C are required to etch copper films. Therefore, the *damascene* process is used to form Cu interconnection without dry etching. Damascene processing involves the creation of interconnect lines by first etching a trench or canal in a planar dielectric layer and then filling that trench with metal, such as aluminum or copper. In dual damascene processing (Fig. 5.13), a second level is involved where a series of holes (i.e., contacts or vias) are etched and filled

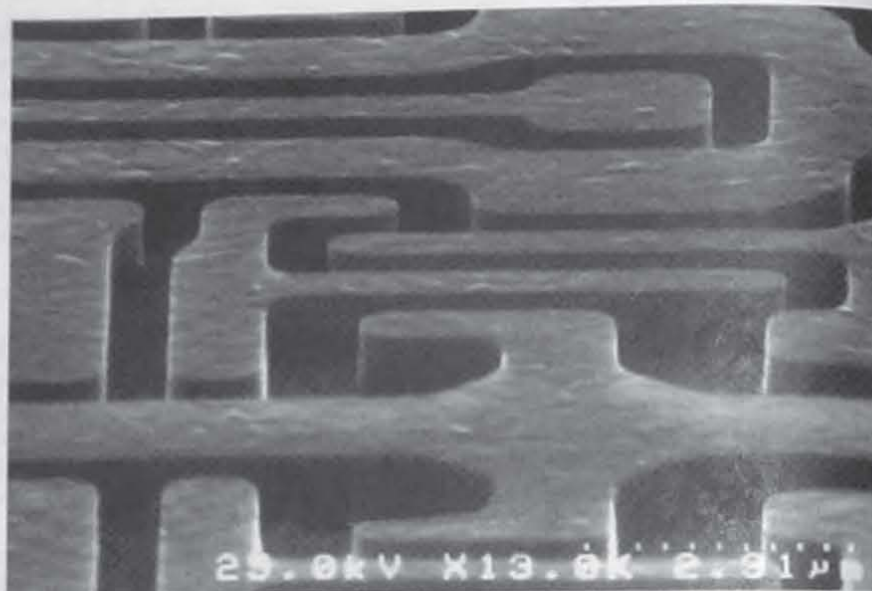


Figure 5.12 0.35- μm TiN/Al/TiN lines and spaces on a wafer maintained at ambient for 72 hours after a microwave strip are not corroded.

in addition to the trench. After filling, the metal and dielectric are planarized by *chemical mechanical polishing* (CMP, see Chapter 8). The advantage of damascene processing is that it eliminates the need for metal etch. This is an important concern as the industry moves from aluminum to copper interconnections.

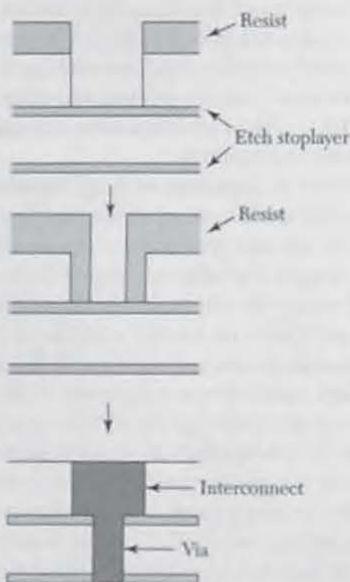


Figure 5.13 Process sequence for the dual damascene process.

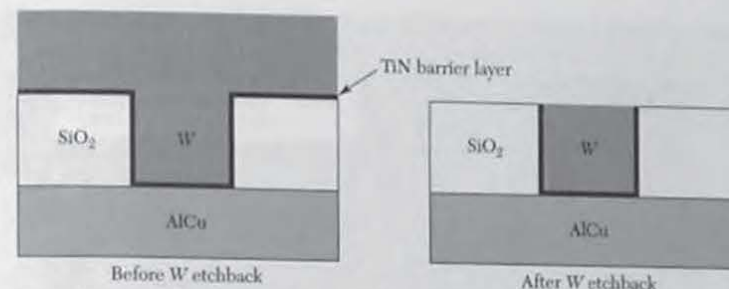


Figure 5.14 Formation of tungsten plug in a contact hole by depositing blanket low-pressure chemical vapor deposition W and then using reaction ion etching etchback.

Low-pressure CVD (LPCVD) tungsten (W) has been widely used for filling contact holes and first-level metallization because of its excellent deposition conformability. Both fluorine- and chlorine-based chemistries etch W and form volatile etch products. An important tungsten etch process is the blanket W etchback to form a W plug. The blanket LPCVD W is deposited on top of a TiN barrier layer, as shown in Figure 5.14. A two-step process is usually used. First, 90% of the W is etched at a high etch rate, and then the etch rate is reduced to remove the remaining W with an etchant that has a high W-to-TiN selectivity.

► 5.3 ETCH SIMULATION

SUPREM may be used to simulate the etching process. However, etch simulation using SUPREM is rudimentary at best. Simulation results may be achieved using the ETCH command, which allows the user to etch all or part of any given layer at the top of the current structure. If the material at the top of the structure is not the material specified, then no etching takes place. If the amount to be etched is not specified, then the entire layer is removed.

EXAMPLE 3

Suppose we want to simulate the etching of 0.3 μm of the oxide grown after the dry-wet-dry sequence in Example 3 of Chapter 3.

SOLUTION The SUPREM input listing is as follows:

TITLE	Etching Example
COMMENT	Initialize silicon substrate
INITIALIZE	<100> Silicon Phosphor Concentration=1e16
COMMENT	Ramp furnace up to 1100 C over 10 minutes in N2
DIFFUSION	Time=10 Temperature=900 Nitrogen T.rate=20
COMMENT	Oxidize the wafers for 5 minutes at 1100 C in dry O2
DIFFUSION	Time=5 Temperature=1100 DryO2
COMMENT	Oxidize the wafers for 120 minutes at 1100 C in wet O2
DIFFUSION	Time=120 Temperature=1100 WetO2
COMMENT	Oxidize the wafers for 5 minutes at 1000 C in dry O2
DIFFUSION	Time=5 Temperature=1100 DryO2
COMMENT	Ramp furnace down to 900 C over 10 minutes in N2
DIFFUSION	Time=10 Temperature=1100 Nitrogen T.rate=20

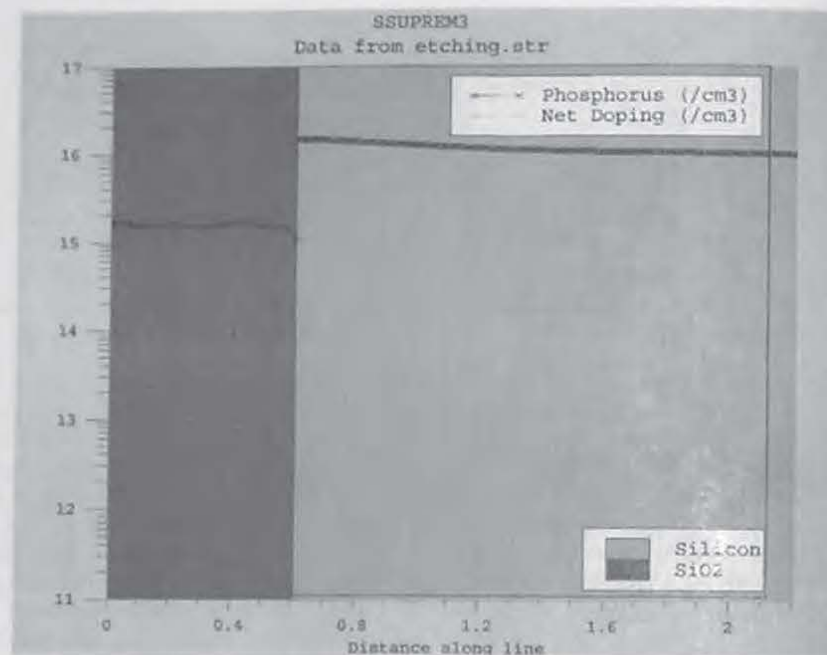


Figure 5.15 Plot of phosphorus concentration as a function of depth into the silicon substrate, using SUPREM.

```

ETCH      Oxide Thickness = 0.3
PRINT     Layers Chemical Concentration Phosphor
PLOT      Active Net Cmin=1e14
STOP      End etching example

```

After oxidation is complete, we print and plot the phosphorus concentration as a function of depth into the silicon substrate. The results are shown in Figure 5.15, which indicates a final oxide thickness of 0.609 μm and depicts the phosphorus incorporation in the oxide layer.

5.4 SUMMARY

The two major processes to transfer patterns in IC fabrication are photolithography and etching. Wet chemical etching is used extensively in semiconductor processing. It is particularly suitable for blanket etching. This chapter discussed wet chemical etching processes for silicon and gallium arsenide, insulators, and metal interconnections. Wet chemical etching was used for pattern transfer. However, undercutting of the layer underneath the mask resulted in loss of resolution in the etched pattern.

Dry etching methods are used to achieve high-fidelity pattern transfer. Dry etching is synonymous with plasma-assisted etching. This chapter considered plasma fundamentals and various dry etching systems, which have grown from relatively simple, parallel-plate configurations to complex chambers with multiple frequency generators and a variety of process-control sensors.

The challenges for future etching technology are high etch selectivity, better dimensioned control, low aspect ratio-dependent etching, and low plasma-induced damage. Low-pressure, high-density plasma reactors are necessary to meet these requirements.

As processing evolves from 200-mm to 300-mm and even larger diameter wafers, continued improvements are required for etch uniformity across the wafer. New gas chemistries must be developed to provide the improved selectivity necessary for advanced integration schemes.

REFERENCES

1. For a more detailed discussion on etching, see Y. J. T. Liu, "Etching," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
2. H. Robbins and B. Schwartz, "Chemical Etching of Silicon II, the System HF , HNO_3 , H_2O and $\text{HC}_2\text{H}_3\text{O}_2$," *J. Electrochem. Soc.*, **107**, 108 (1960).
3. K. E. Bean, "Anisotropic Etching in Silicon," *IEEE Trans. Electron Devices*, **ED-25**, 1185 (1978).
4. D. P. Kern, et al., "Practical Aspects of Microfabrication in the 100-nm Region," *Solid State Technol.*, **27**, 127 (1984).
5. S. Iida and K. Ito, "Selective Etching of Gallium Arsenide Crystal in H_2SO_4 - H_2O_2 - H_2O System," *J. Electrochem. Soc.*, **118**, 768 (1971).
6. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Technol.*, **24**, 5, 65 (1981).
7. J. A. Mucha and D. W. Hess, "Plasma Etching," in L. F. Thompson and C. G. Willson, Eds., *Microcircuit Processing: Lithography and Dry Etching*, American Chemical Society, Washington, DC, 1984.
8. M. Armacost, et al., "Plasma-Etching Processes for ULSI Semiconductor Circuits," *IBM J. Res. Dev.*, **43**, 39 (1999).
9. C. O. Jung, et al., "Advanced Plasma Technology in Microelectronics," *Thin Solid Films*, **341**, 112 (1999).

PROBLEMS

Asterisks denote difficult problems.

SECTION 5.1: WET CHEMICAL ETCHING

1. If the mask and the substrate cannot be etched by a particular etchant, sketch the edge profile of an isotropically etched feature in a film of thickness h_i for (a) etching just to completion, (b) 100% overetch, and (c) 200% overetch.
2. A $\langle 100 \rangle$ -oriented silicon crystal is etched in a KOH solution through a $1.5 \mu\text{m} \times 1.5 \mu\text{m}$ window defined in silicon dioxide. The etch rate normal to $\langle 100 \rangle$ planes is $0.6 \mu\text{m}/\text{min}$. The etch rate ratios are 100:16:1 for the $\langle 100 \rangle$: $\langle 110 \rangle$: $\langle 111 \rangle$ planes. Show the etched profile after 20 seconds, 40 seconds, and 60 seconds.
3. Repeat the previous problem with a $\langle 110 \rangle$ -oriented silicon etched with a thin SiO_2 mask in KOH solution. Show the etched pattern profiles on $\langle 110 \rangle$ Si.
4. A $\langle 100 \rangle$ -oriented silicon wafer 150 mm in diameter is 625 μm thick. The wafer has 1000 $\mu\text{m} \times 1000 \mu\text{m}$ ICs on it. The IC chips are to be separated by orientation-dependent etching. Describe two methods for doing this and calculate the fraction of the surface area that is lost in these processes.

SECTION 5.2: DRY ETCHING

- *5. The average distance traveled by particles between collisions is called the mean free path (λ), $\lambda \approx 5 \times 10^{-3}/P$ (cm), where P is pressure in Torr. In typical plasmas of interest, the chamber pressure ranges from 1 Pa to 150 Pa. What are the corresponding density of gas molecules (cm^{-3}) and the mean free path?
6. Fluorine (F) atoms etch Si at a rate given by

$$\text{Etch rate (nm/min)} = 2.86 \times 10^{-13} n_F \times T^{1/2} \exp(-E_a/RT)$$

where n_F is the concentration of F atoms (cm^{-3}), T the temperature (K), and E_a and R the activation energy (2.48 kcal/mol) and gas constant (1.987 cal-K), respectively. If n_F is 3×10^{15} , calculate the etch rate of Si at room temperature.

7. SiO_2 etched by F atoms could also be expressed by

$$\text{Etch rate (nm/min)} = 0.614 \times 10^{-13} n_F \times T^{1/2} \exp(-E_a/RT)$$

where n_F is 3×10^{15} (cm^{-3}) and E_a is 3.76 kcal/mol. Calculate the etch rate of SiO_2 and etch selectivity of SiO_2 over Si at room temperature.

8. A multiple-step etch process is required for etching a polysilicon gate with thin gate oxide. How do you design an etch process that has no micromasking, has an anisotropic etch profile, and is selective to thin gate oxide?
9. Find the etch selectivity required to etch a 400-nm polysilicon layer without removing more than 1 nm of its underlying gate oxide, assuming that the polysilicon is etched with a process having a 10% etch-rate uniformity.
10. A 1- μm Al film is deposited over a flat field oxide region and patterned with photoresist. The metal is then etched with a mixture of BCl_3/Cl_2 gases at a temperature of 70°C in a Helicon etcher. The selectivity of Al over photoresist is maintained at 3. Assuming a 30% overetch, what is the minimum photoresist thickness required to ensure that the top metal surface is not attacked?
11. In an ECR plasma, a static magnetic field B forces electrons to circulate around the magnetic field lines at an angular frequency, ω_e , that is given by

$$\omega_e = qB/m_e$$

where q is the electronic charge and m_e the electron mass. If the microwave frequency is 2.45 GHz, what is the required magnetic field?

12. What are the major distinctions between traditional reactive ion etching and high-density plasma etching (ECR, ICP, etc.)?
13. Describe how to eliminate the corrosion issues in Al lines after etching with chlorine-based plasma.

Diffusion

Impurity doping is the introduction of controlled amounts of impurity dopants into semiconductors. The practical use of impurity doping mainly has been to change the electrical properties of the semiconductors. Diffusion and ion implantation are the two key methods of impurity doping. Both diffusion and ion implantation are used for fabricating discrete devices and integrated circuits because these processes generally complement each other.^{1,2} For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS), whereas ion implantation, which is discussed in Chapter 7, is used to form a shallow junction (e.g., a source/drain junction of a MOSFET).

Until the early 1970s, impurity doping was done mainly by diffusion at elevated temperatures, as shown in Figure 6.1. In this method the dopant atoms are placed on or near the surface of the wafer by deposition from the gas phase of the dopant or by using doped oxide sources. The doping concentration decreases monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and diffusion time.

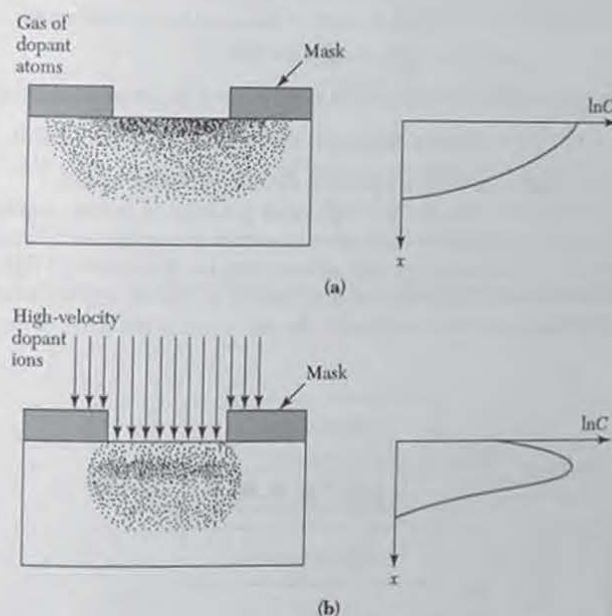


Figure 6.1 Comparison of (a) diffusion and (b) ion-implantation techniques for the selective introduction of dopants into the semiconductor substrate.

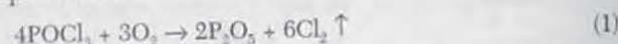
This chapter focuses on the diffusion process. Ion implantation is covered in Chapter 7. Specifically, this chapter covers the following topics:

- The movement of impurity atoms in the crystal lattice under high-temperature and high-concentration-gradient conditions
- Impurity profiles for constant diffusivity and concentration-dependent diffusivity
- The impact of lateral diffusion and impurity redistribution on device characteristics
- The simulation of diffusion using SUPREM

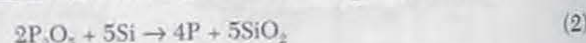
6.1 BASIC DIFFUSION PROCESS

Diffusion of impurities is typically done by placing semiconductor wafers in a carefully controlled, high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The temperature usually ranges between 800°C and 1200°C for silicon and 600°C and 1000°C for gallium arsenide. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture.

For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These three elements are highly soluble in silicon; they have solubilities above $5 \times 10^{20} \text{ cm}^{-3}$ in the diffusion temperature range. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3 , AsCl_3 , and POCl_3), and gaseous sources (B_2H_6 , AsH_3 , and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Figure 6.2. This arrangement is similar to that used for thermal oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is



The P_2O_5 forms a glass-on-silicon wafer and is then reduced to phosphorus by silicon,



The phosphorus is released and diffuses into the silicon, and Cl_2 is vented.

For diffusion in gallium arsenide, the high vapor pressure of arsenic requires special methods to prevent the loss of arsenic by decomposition or evaporation.² These methods include diffusion in sealed ampules with an overpressure of arsenic and diffusion in an open-tube furnace with a doped oxide capping layer (e.g., silicon nitride). Most of the studies on *p*-type diffusion have been confined to the use of zinc in the forms of Zn-Ga-As

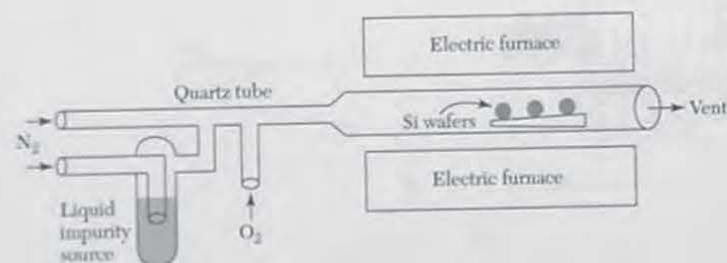


Figure 6.2 Schematic diagram of a typical open-tube diffusion system.

alloys and ZnAs_2 for the sealed-ampule approach or ZnO-SiO_2 for the open-tube approach. The *n*-type dopants in gallium arsenide include selenium and tellurium.

6.1.1 Diffusion Equation

Diffusion in a semiconductor can be visualized as the atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. Figure 6.3 shows the two basic atomic diffusion models in a solid.^{1,3} The open circles represent the host atoms occupying the equilibrium lattice positions. The solid dots represent impurity atoms. At elevated temperatures, the lattice atoms vibrate around the equilibrium lattice sites. There is a finite probability that a host atom will acquire sufficient energy to leave the lattice site and become an interstitial atom, thereby creating a vacancy. When a neighboring impurity atom migrates to the vacancy site, as illustrated in Figure 6.3a, the mechanism is called *vacancy diffusion*. If an interstitial atom moves from one place to another without occupying a lattice site (Fig. 6.3b), the mechanism is *interstitial diffusion*. An atom smaller than the host atom often moves interstitially.

The basic diffusion process of impurity atoms is similar to that of charge carriers (electrons and holes). We define a flux F as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. We then have

$$F = -D \frac{\partial C}{\partial x} \quad (3)$$

where the proportionality constant D is the *diffusion coefficient* or *diffusivity*. Note that the basic driving force of the diffusion process is the concentration gradient $\partial C / \partial x$. The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

If we substitute Eq. 3 into the one-dimensional continuity equation under the condition that no materials are formed or consumed in the host semiconductor, we obtain

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right) \quad (4)$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered to be independent of doping concentration, and Eq. 4 becomes

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} \quad (5)$$

Equation 5 is often referred to as *Fick's diffusion equation* or *Fick's law*.

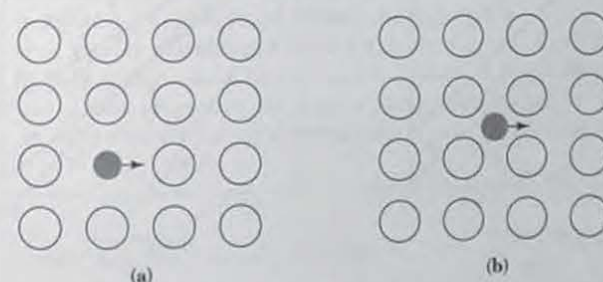


Figure 6.3 Atomic diffusion mechanisms for a two-dimensional lattice.^{1,3} (a) Vacancy mechanism. (b) Interstitial mechanism.

Figure 6.4 shows the measured diffusion coefficients for low concentrations of various dopant impurities in silicon and gallium arsenide.^{4,5} The logarithm of the diffusion coefficient plotted against the reciprocal of the absolute temperature gives a straight line in most of the cases. This implies that over the temperature range, the diffusion coefficients can be expressed as

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) \quad (6)$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature, and E_a is the activation energy in eV.

For the interstitial diffusion model, E_a is related to the energies required to move dopant atoms from one interstitial site to another. The values of E_a are found to be between 0.5 and 2 eV in both silicon and gallium arsenide. For the vacancy diffusion model, E_a is related to both the energies of motion and the energies of formation of vacancies. Thus, E_a for vacancy diffusion is larger than that for interstitial diffusion, usually between 3 and 5 eV.

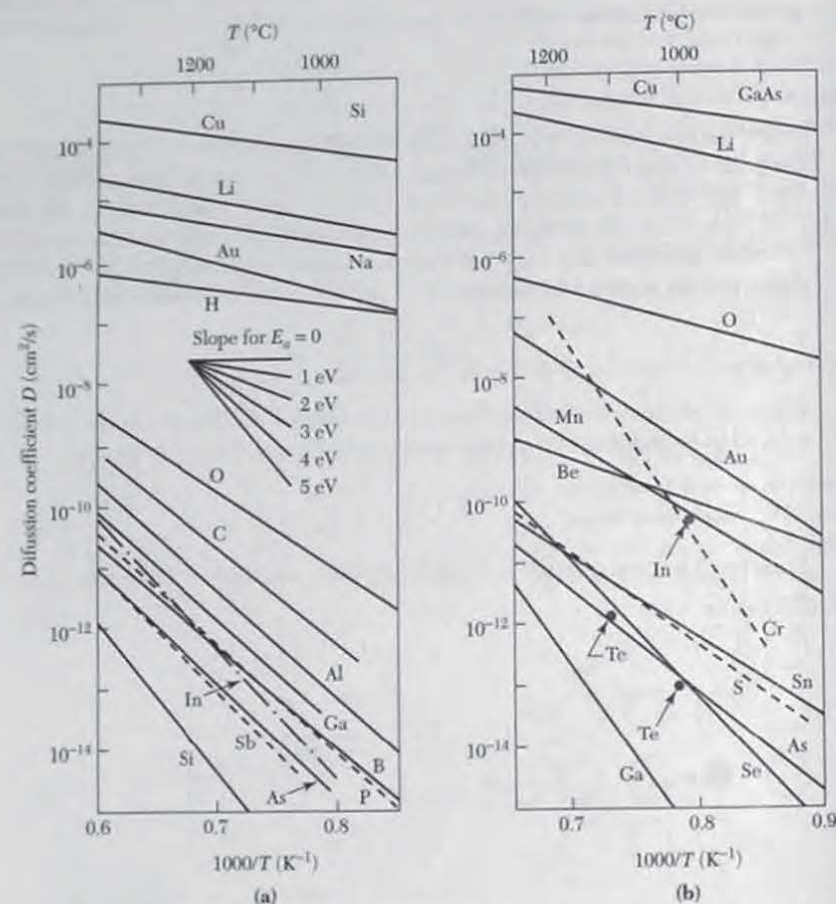


Figure 6.4 Diffusion coefficient (also called diffusivity) as a function of the reciprocal of temperature for (a) silicon and (b) gallium arsenide.^{4,5}

For fast diffusants, such as Cu in Si and GaAs, shown in the upper portion of Figures 6.4a and b, the measured activation energies are less than 2 eV, and interstitial atomic movement is the dominant diffusion mechanism. For slow diffusants, such as As in Si and GaAs, shown in the lower portion of Figures 6.4a and b, E_a is larger than 3 eV, and vacancy diffusion is the dominant mechanism.

6.1.2 Diffusion Profiles

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. This subsection considers two important cases, namely, constant surface concentration diffusion and constant total dopant diffusion. In the first case, impurity atoms are transported from a vapor source onto the semiconductor surface and diffused into the semiconductor wafers. The vapor source maintains a constant level of surface concentration during the entire diffusion period. In the second case, a fixed amount of dopant is deposited onto the semiconductor surface and is subsequently diffused into the wafers.

Constant Surface Concentration

The initial condition at $t = 0$ is

$$C(x, 0) = 0 \quad (7)$$

which states that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are

$$C(0, t) = C_s \quad (8a)$$

and

$$C(\infty, t) = 0 \quad (8b)$$

where C_s is the surface concentration (at $x = 0$), which is independent of time. The second boundary condition states that at large distances from the surface there are no impurity atoms.

The solution of Fick's diffusion equation that satisfies the initial and boundary conditions is given by⁶

$$C(x, t) = C_s \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \quad (9)$$

where erfc is the complementary error function and $\sqrt{Dt}$ is the diffusion length. The definition of erfc and some properties of the function are summarized in Table 6.1. The diffusion profile for the constant surface concentration condition is shown in Figure 6.5a, which plots, on both linear (upper) and logarithmic (lower) scales, the normalized concentration as a function of depth for three values of the diffusion length corresponding to three consecutive diffusion times and a fixed D for a given diffusion temperature. Note that as time progresses, the dopant penetrates deeper into the semiconductor.

The total number of dopant atoms per unit area of the semiconductor is given by

$$Q(t) = \int_0^\infty C(x, t) dx \quad (10)$$

Substituting Eq. 9 into Eq. 10 yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \approx 1.13 C_s \sqrt{Dt} \quad (11)$$

TABLE 6.1 Error Function Algebra

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$$

$$\operatorname{erf}(0) = 0$$

$$\operatorname{erf}(\infty) = 1$$

$$\operatorname{erf}(x) \approx \frac{2}{\sqrt{\pi}} x \text{ for } x \ll 1$$

$$\operatorname{erfc}(x) \approx \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{x} \text{ for } x \gg 1$$

$$\frac{d}{dx} \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2}$$

$$\frac{d^2}{dx^2} \operatorname{erf}(x) = -\frac{4}{\sqrt{\pi}} x e^{-x^2}$$

$$\int_0^x \operatorname{erfc}(y') dy' = x \operatorname{erfc}(x) + \frac{1}{\sqrt{\pi}} (1 - e^{-x^2})$$

$$\int_0^\infty \operatorname{erfc}(x) dx = \frac{1}{\sqrt{\pi}}$$

This expression can be interpreted as follows. The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Figure 6.5a. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \approx C_s \sqrt{Dt}$, which is close to the exact result obtained from Eq. 11.

A related quantity is the gradient of the diffusion profile $\partial C/\partial x$. The gradient can be obtained by differentiating Eq. 9:

$$\left. \frac{\partial C}{\partial x} \right|_{x_j} = \frac{C_s}{\sqrt{\pi Dt}} e^{-x_j^2/4Dt} \quad (12)$$

EXAMPLE 1

For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at 10^{19} cm^{-3} and the diffusion time is 1 hour. Find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches 10^{15} cm^{-3} .

SOLUTION The diffusion coefficient of boron at 1000°C, as obtained from Figure 6.4, is about $2 \times 10^{-14} \text{ cm}^2/\text{s}$, so the diffusion length is

$$\sqrt{Dt} = \sqrt{2 \times 10^{-14} \times 3600} = 8.48 \times 10^{-6} \text{ cm}$$

$$Q(t) = 1.13 C_s \sqrt{Dt} = 1.13 \times 10^{19} \times 8.48 \times 10^{-6} = 9.5 \times 10^{13} \text{ atoms/cm}^2$$

$$\left. \frac{dC}{dx} \right|_{x=0} = -\frac{C_s}{\sqrt{\pi Dt}} = \frac{-10^{19}}{\sqrt{\pi} \times 8.48 \times 10^{-6}} = -6.7 \times 10^{23} \text{ cm}^{-4}$$

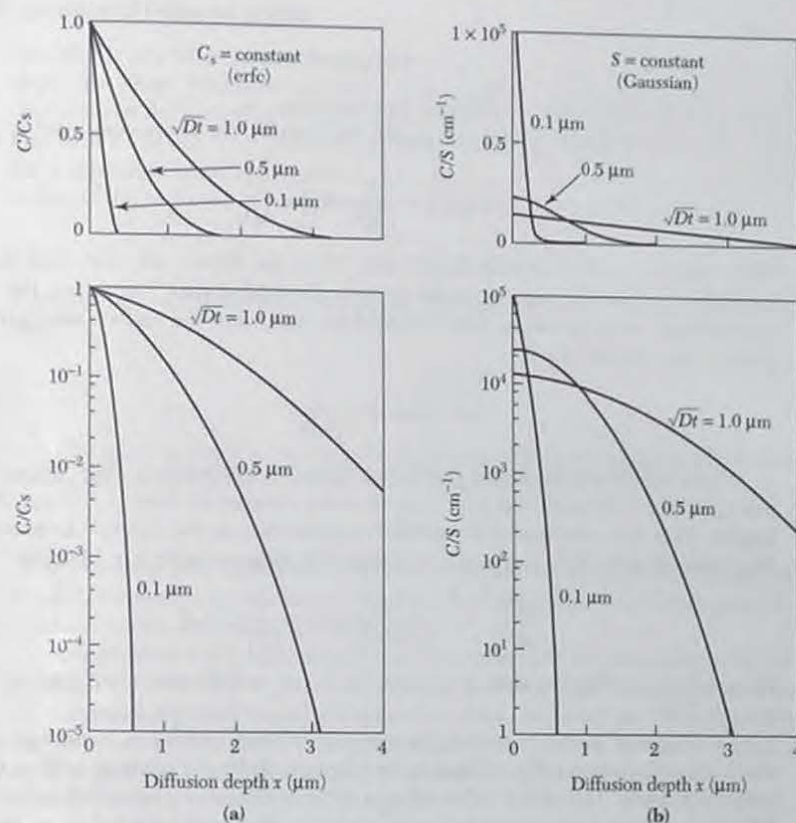


Figure 6.5 Diffusion profiles. (a) Normalized complementary error function versus distance for successive diffusion times. (b) Normalized Gaussian function versus distance.

When $C = 10^{15} \text{ cm}^{-3}$, the corresponding distance x_j is given by Eq. 9, or

$$x_j = 2\sqrt{Dt} \operatorname{erfc}^{-1}\left(\frac{10^{15}}{10^{19}}\right) = 2\sqrt{Dt}(2.75) = 4.66 \times 10^{-5} \text{ cm} = 0.466 \mu\text{m}$$

$$\left. \frac{dC}{dx} \right|_{x=0.466 \mu\text{m}} = -\frac{C_s}{\sqrt{\pi Dt}} e^{-x_j^2/4Dt} = -3.5 \times 10^{20} \text{ cm}^{-4}$$

Constant Total Dopant

For this case, a fixed (or constant) amount of dopant is deposited onto the semiconductor surface in a thin layer, and the dopant subsequently diffuses into the semiconductor. The initial condition is the same as in Eq. 7. The boundary conditions are

$$\int_0^\infty C(x, t) dx = S \quad (13a)$$

and

$$C(\infty, t) = 0 \quad (13b)$$

where S is the total amount of dopant per unit area.

The solution of the diffusion equation that satisfies the above conditions is

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \quad (14)$$

This expression is the Gaussian distribution. Since the dopant will move into the semiconductor as time increases, in order to keep the total dopant S constant, the surface concentration must decrease. This is indeed the case, since the surface concentration is given by Eq. 14 with $x = 0$:

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \quad (15)$$

Figure 6.5b shows the dopant profile for a Gaussian distribution in which the normalized concentration (C/S) is plotted as a function of the distance for three increasing diffusion lengths. Note the reduction of the surface concentration as the diffusion time increases. The gradient of the diffusion profile is obtained by differentiating Eq. 14 and is

$$\left. \frac{dC}{dx} \right|_{x=0} = -\frac{xS}{2\sqrt{\pi}(Dt)^{3/2}} = -\frac{x}{2Dt} C(x, t) \quad (16)$$

The gradient (or slope) is zero at $x = 0$ and at $x = \infty$, and the maximum gradient occurs at $x = \sqrt{2Dt}$.

In integrated circuit processing, a two-step diffusion process is commonly used, in which a *predeposition* diffused layer is first formed under the constant surface concentration condition. This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under a constant total dopant condition. For most practical cases, the diffusion length $\sqrt{Dt}$ for the predeposition diffusion is much smaller than the diffusion length for the drive-in diffusion. Therefore, the predeposition profile can be considered a delta function at the surface, and the extent of the penetration of the predeposition profile can be regarded to be negligibly small compared with that of the final profile that results from the drive-in step.

EXAMPLE 2

Arsenic was predeposited by arsine gas, and the resulting total amount of dopant per unit area was 1×10^{14} atoms/cm². How long would it take to drive the arsenic in to a junction depth of 1 μm ? Assume a background doping of $C_B = 1 \times 10^{15}$ atoms/cm³, and a drive-in temperature of 1200°C. For As diffusion, $D_0 = 24$ cm²/s, and $E_a = 4.08$ eV.

SOLUTION

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) = 24 \exp\left(\frac{-4.08}{8.614 \times 10^{-5} \times 1473}\right) = 2.602 \times 10^{-13} \text{ cm}^2/\text{s}$$

$$x_j^2 = 10^{-8} = 4Dt \ln\left(\frac{S}{C_B \sqrt{Dt}}\right) = 1.04 \times 10^{-12} t \ln\left(\frac{1.106 \times 10^5}{\sqrt{t}}\right)$$

$$t \cdot \log t - 10.09t + 8350 = 0$$

The solution to this equation can be determined by the cross point of equation $y = t \cdot \log t$ and $y = 10.09t - 8350$. Therefore, $t = 1190$ seconds, or approximately 20 minutes.

6.1.3 Evaluation of Diffused Layers

The results of a diffusion process can be evaluated by three measurements: the junction depth, the sheet resistance, and the dopant profile of the diffused layer. The junction depth can be delineated by cutting a groove into the semiconductor and etching the surface with a solution (e.g., 100 cm³ HF and a few drops of HNO₃ for silicon) that stains the p -type region darker than the n -type region, as illustrated in Figure 6.6a. If R_0 is the radius of the tool used to form the groove, then the junction depth x_j is given by

$$x_j = \sqrt{R_0^2 - b^2} = \sqrt{R_0^2 - a^2} \quad (17)$$

where a and b are indicated in the figure. In addition, if R_0 is much larger than a and b , then

$$x_j \approx \frac{a^2 - b^2}{2R_0} \quad (18)$$

The junction depth x_j , as illustrated in Figure 6.6b, is the position where the dopant concentration equals the substrate concentration C_B , or

$$C(x_j) = C_B \quad (19)$$

Thus, if the junction depth and C_B are known, the surface concentration C_s and the impurity distribution can be calculated, provided the diffusion profile follows one or the other simple equation derived in Section 6.1.2.

The resistance of a diffused layer can be measured by the *four-point probe* technique shown in Figure 6.7. The probes are equally spaced. A small current I from a constant-current source is passed through the outer two probes, and a voltage V is measured between the inner two probes. For a thin semiconductor sample with thickness W that is much smaller than the sample diameter d , the *resistivity* (ρ) is given by

$$\rho = \frac{V}{I} \cdot W \cdot CF \text{ } \Omega\text{-cm} \quad (20)$$

where CF is known as the correction factor. The correction factor depends on the ratio of d/s , where s is the probe spacing. When $d/s > 20$, the correction factor approaches 4.54.

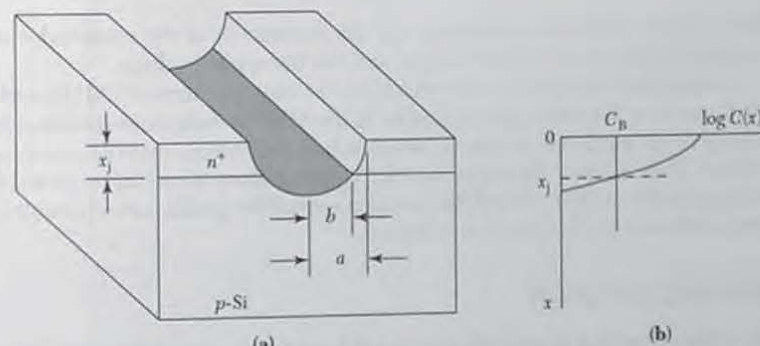


Figure 6.6 Junction-depth measurement. (a) Grooving and staining. (b) Position in which dopant and substrate concentrations are equal.

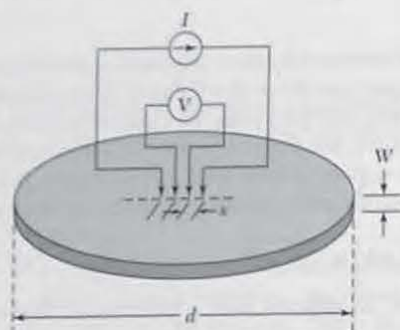


Figure 6.7 Measurement of resistivity using a four-point probe.³

The sheet resistance (R_s) is related to the junction depth (x_j), the carrier mobility (μ , which is a function of the total impurity concentration), and the impurity distribution $C(x)$ by the following expression:⁷

$$R_s = \frac{1}{q \int_0^j \mu C(x) dx} \quad (21)$$

For a given diffusion profile, the average resistivity ($\bar{\rho} = R_s x_j$) is uniquely related to the surface concentration (C_s) and the substrate doping concentration for an assumed diffusion profile. Design curves relating C_s and $\bar{\rho}$ have been calculated for simple diffusion profiles, such as the *erfc* or Gaussian distribution.⁸ To use these curves correctly, we must be sure that the diffusion profiles agree with the assumed profiles. For low concentration and deep diffusions, the diffusion profiles generally can be represented by the aforementioned simple functions. However, as discussed in the next section, for high concentration and shallow diffusions, the diffusion profiles cannot be represented by these simple functions.

The diffusion profile can be measured using a capacitance-voltage technique. The majority carrier profile (n), which is equal to the impurity profile if impurities are fully ionized, can be determined by measuring the reverse-bias capacitance of a *p-n* junction or a Schottky barrier diode as a function of the applied voltage. This is due to the relationship⁹

$$n = \frac{2}{q\epsilon_s} \left[\frac{-1}{d(1/C'^2)/dV} \right] \quad (22)$$

where q is the charge of an electron, ϵ_s is the permittivity of the semiconductor, C' is capacitance per unit area of the sample, and V is the applied voltage.

A more elaborate method is the secondary ion mass spectroscopic (SIMS) technique, which measures the total impurity profile. In the SIMS technique, an ion beam sputters material off the surface of a semiconductor, and the ion component is detected and mass analyzed. This technique has high sensitivity to many elements, such as boron and arsenic, and is an ideal tool for providing the precision needed for profile measurements in high-concentration or shallow-junction diffusions.¹⁰

6.2 EXTRINSIC DIFFUSION

The diffusion profiles described in Section 6.1 are for constant diffusivities. These profiles occur when the doping concentration is lower than the intrinsic carrier concentra-

tion (n_i) at the diffusion temperature. For example, at $T = 1000^\circ\text{C}$, n_i equals $5 \times 10^{18} \text{ cm}^{-3}$ for silicon and $5 \times 10^{17} \text{ cm}^{-3}$ for gallium arsenide. The diffusivity at low concentrations less than $n_i(T)$ are in the intrinsic diffusion region, as indicated in the left side of Figure 6.8. In this region, the resulting dopant profiles of sequential or simultaneous diffusions of *n*- and *p*-type impurities can be determined by superposition; that is, the diffusions can be treated independently. However, when the impurity concentration, including both the substrate and the dopant, is greater than $n_i(T)$, the semiconductor becomes extrinsic, and the diffusivity is considered to be extrinsic. In the extrinsic diffusion region, the diffusivity becomes concentration dependent.¹¹ In the extrinsic diffusion region, the diffusion profiles are more complicated, and there are interactions and cooperative effects among the sequential or simultaneous diffusions.

6.2.1 Concentration-Dependent Diffusivity

As mentioned previously, when a host atom acquires sufficient energy from the lattice vibration to leave its lattice site, a vacancy is created. Depending on the charges associated with a vacancy, we can have a neutral vacancy V^0 , an acceptor vacancy V^- , a double-charged acceptor vacancy V^{2-} , a donor vacancy V^+ , and so forth. We expect that the vacancy density of a given charge state (i.e., the number of vacancies per unit volume, C_v) has a temperature dependence similar to that of the carrier density, that is,

$$C_v = C_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad (23)$$

where C_i is the intrinsic vacancy density, E_F is the Fermi level, and E_i is the intrinsic Fermi level.

If the dopant diffusion is dominated by the vacancy mechanism, the diffusion coefficient is expected to be proportional to the vacancy density. At low doping concentrations

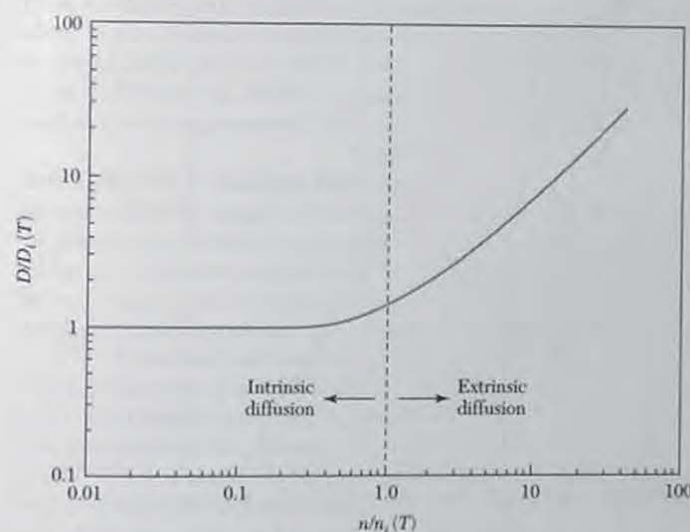


Figure 6.8 Donor impurity diffusivity versus electron concentration showing regions of intrinsic and extrinsic diffusion.¹¹

($n < n_i$), the Fermi level coincides with the intrinsic Fermi level ($E_F = E_i$). The vacancy density is equal to C_i and is independent of doping concentration. The diffusion coefficient, which is proportional to C_i , also is independent of doping concentration. At high concentrations ($n > n_i$), the Fermi level will move toward the conduction band edge (for donor-type vacancies), and the term $[\exp(E_F - E_i)/kT]$ becomes larger than unity. This causes C_v to increase, which in turn causes the diffusion coefficient to increase, as shown in the right side of Figure 6.8.

When the diffusion coefficient varies with dopant concentration, Eq. 4 should be used as the diffusion equation instead of Eq. 5, in which D is independent of C . We consider the case where the diffusion coefficient can be written as

$$D = D_s \left(\frac{C}{C_s} \right)^\gamma \quad (24)$$

where C_s is the surface concentration, D_s is the diffusion coefficient at the surface, and γ is a parameter to describe the concentration dependence. For such a case, we can write Eq. 4 as an ordinary differential equation and solve it numerically.

Figure 6.9 shows the solutions¹² for constant surface concentration diffusion with different values of γ . For $\gamma = 0$, we have the case of constant diffusivity, and the profile is the same as that shown in Figure 6.5a. For $\gamma > 0$, the diffusivity decreases as the concentration decreases, and increasingly steep and boxlike concentration profiles result for increasing γ . Therefore, highly abrupt junctions are formed when diffusions are made into a background of an opposite impurity type. The abruptness of the doping profile results in a junction depth virtually independent of the background concentration. Note that the junction depth (see Fig. 6.9) is given by

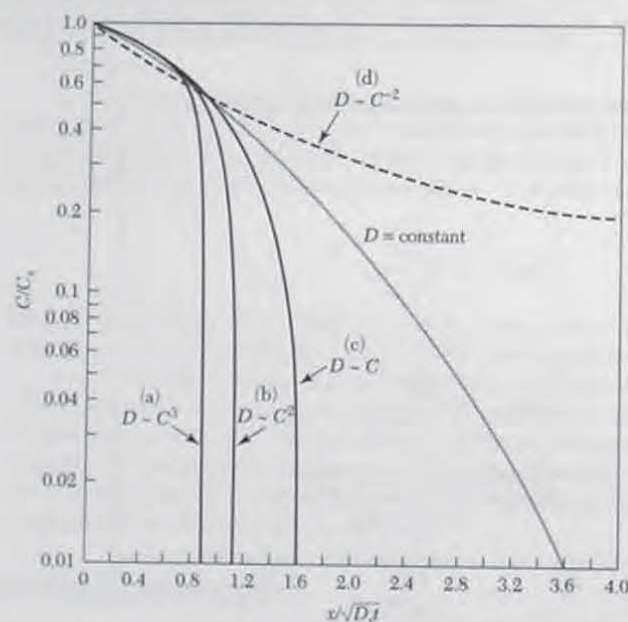


Figure 6.9 Normalized diffusion profiles for extrinsic diffusion where the diffusion coefficient becomes concentration dependent.¹²

$$\begin{aligned} x_j &= 1.6\sqrt{D_s t} \text{ for } D \sim C \quad (\gamma = 1) \\ x_j &= 1.1\sqrt{D_s t} \text{ for } D \sim C^2 \quad (\gamma = 2) \\ x_j &= 0.87\sqrt{D_s t} \text{ for } D \sim C^3 \quad (\gamma = 3) \end{aligned} \quad (25)$$

In the case of $\gamma = -2$, the diffusivity increases with decreasing concentration, which leads to a concave profile, as opposed to the convex profiles for other cases.

6.2.2 Diffusion Profiles

Diffusion in Silicon

The measured diffusion coefficients of boron and arsenic in silicon have a concentration dependence with $\gamma \approx 1$. Their concentration profiles are abrupt, as depicted in curve *c* of Figure 6.9. For gold and platinum diffusion in silicon, γ is close to -2 , and their concentration profiles have the concave shape shown in curve *d* of Figure 6.9.

The diffusion of phosphorus in silicon is associated with the doubly charged acceptor vacancy V^{2-} , and the diffusion coefficient at high concentration varies as C^2 . We would expect that the diffusion profile of phosphorus resembles that shown in curve *b* of Figure 6.9. However, because of a dissociation effect, the diffusion profile exhibits anomalous behavior.

Figure 6.10 shows phosphorus diffusion profiles for various surface concentrations after diffusion into silicon for 1 hour at 1000°C .¹³ When the surface concentration is low, corresponding to the intrinsic diffusion region, the diffusion profile is given by an erfc (curve *a*). As the concentration increases, the profile begins to deviate from the simple expression (curves *b* and *c*). At very high concentration (curve *d*), the profile near the surface is indeed similar to that shown in curve *b* of Figure 6.9. However, at concentration n_c , a kink occurs and is followed by a rapid diffusion in the tail region. The concentration n_c corresponds to a Fermi level 0.11 eV below the conduction band. At this energy level, the coupled impurity-vacancy pair (P^+V^{2-}) dissociates to P^+ , V^- , and an electron. Thus, the dissociation generates a large number of singly charged acceptor vacancies V^- , which in turn enhances the diffusion in the tail region of the profile. The diffusivity in the tail region is over $10^{-12} \text{ cm}^2/\text{s}$, which is about two orders of magnitude larger than the intrinsic diffusivity at 1000°C . Because of its high diffusivity, phosphorus is commonly used to form deep junctions, such as the *n*-tubs in CMOS.

Zinc Diffusion in Gallium Arsenide

We expect diffusion in gallium arsenide to be more complicated than that in silicon because the diffusion of impurities may involve atomic movements on both the gallium and arsenic sublattices. Vacancies play a dominant role in diffusion processes in gallium arsenide because both *p*- and *n*-type impurities must ultimately reside in lattice sites. However, the charge states of the vacancies have not been established.

Zinc is the most extensively studied diffusant in gallium arsenide. Its diffusion coefficient is found to vary as C^2 . Therefore, the diffusion profiles are steep, as shown in Figure 6.11,¹³ and resemble curve *b* of Figure 6.9. Note that even for the case of the lowest surface concentration, the diffusion is in the extrinsic diffusion region, because n_i for GaAs at 1000°C is less than 10^{18} cm^{-3} . As seen in Figure 6.11, the surface concentration has a profound effect on the junction depth. The diffusivity varies linearly with the partial pressure of the zinc vapor, and the surface concentration is proportional to the square root of the partial pressure. Therefore, from Eq. 25, the junction depth is linearly proportional to the surface concentration.

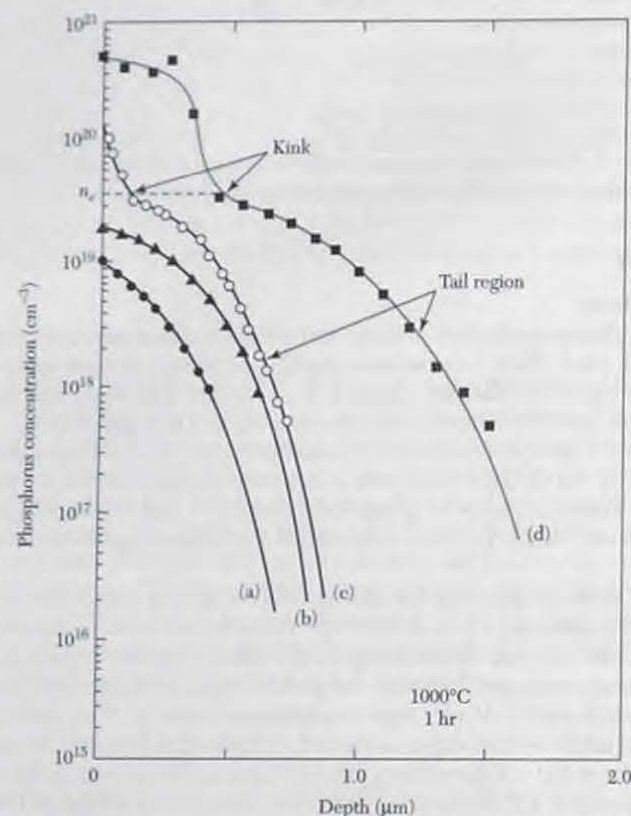


Figure 6.10 Phosphorus diffusion profiles¹³ for various surface concentrations after diffusion into silicon for 1 hour at 1000°C.

► 6.3 LATERAL DIFFUSION

The one-dimensional diffusion equation discussed previously can describe the diffusion process satisfactorily, except at the edge of the mask window. Here the impurities will diffuse downward and sideways (i.e., laterally). In this case, we must consider a two-dimensional diffusion equation and use a numerical technique to obtain the diffusion profiles under different initial and boundary conditions.

Figure 6.12 shows the contours of constant doping concentration for a constant surface concentration diffusion condition, assuming that the diffusivity is independent of concentration.¹⁴ At the far right of the figure, the variation of the dopant concentration from $0.5 C_s$ to $10^{-4} C_s$ (where C_s is the surface concentration) corresponds to the erfc distribution given by Eq. 9. The contours are in effect a map of the location of the junctions created by diffusing into various background concentrations. For example, at $C/C_s = 10^{-4}$ (i.e., the background doping is 10^4 times lower than the surface concentration), we see from this constant concentration curve that the vertical penetration is about $2.8 \mu\text{m}$, whereas the lateral penetration is about $2.3 \mu\text{m}$ (i.e., the penetration along the diffusion mask-semiconductor interface). Therefore, the lateral penetration is about 80% of the

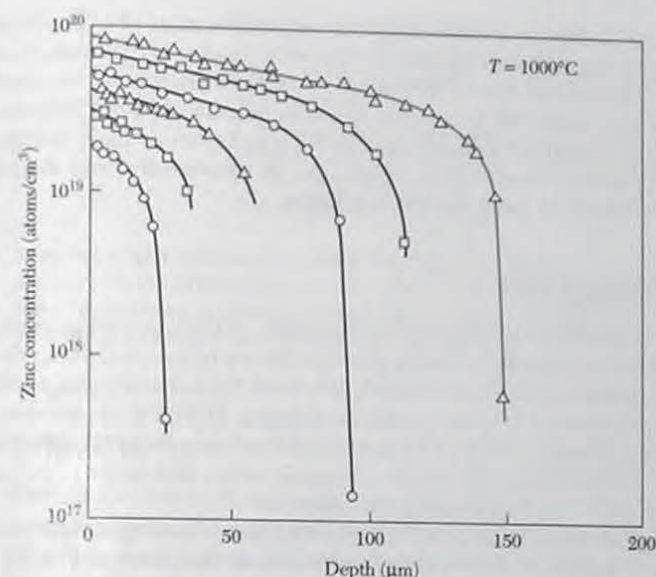


Figure 6.11 Diffusion profiles¹³ of zinc in GaAs after annealing at 1000°C for 2.7 hours. The different surface concentrations are obtained by maintaining the Zn source at temperatures in the range 600°C to 800°C.

penetration in the vertical direction for concentrations three or more orders of magnitude below the surface concentration. Similar results are obtained for a constant total dopant diffusion condition. The ratio of lateral to vertical penetration is about 75%. For concentration-dependent diffusivities, the ratio is found to be reduced slightly, to about 65% to 70%.

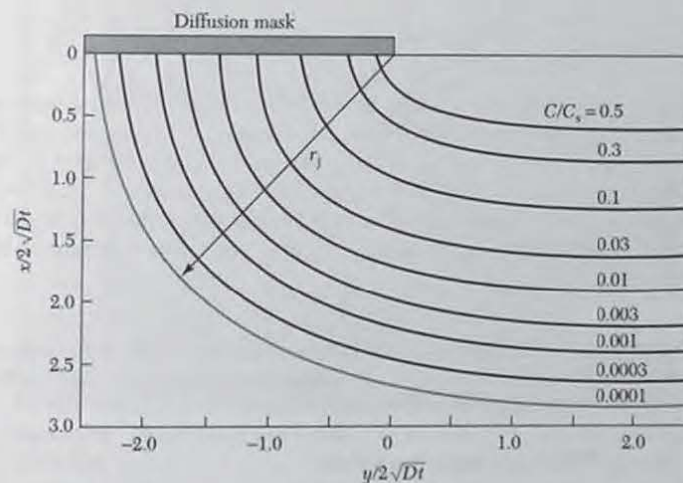


Figure 6.12 Diffusion contours at the edge of an oxide window, where r_j is the radius of curvature.¹⁴

Because of the lateral diffusion effect, the junction consists of a central plane (or flat) region with approximately cylindrical edges with a radius of curvature r_j , as shown in Figure 6.12. In addition, if the diffusion mask contains sharp corners, the shape of the junction near the corner will be roughly spherical because of lateral diffusion. Since the electric field intensities are higher for cylindrical and spherical junction regions, the avalanche breakdown voltages of such regions can be substantially lower than that of a plane junction having the same background doping.

► 6.4 DIFFUSION SIMULATION

The various complications that arise in the computation of diffusion profiles (such as concentration-dependent diffusivity) tend to preclude the use of analytical hand calculations for all but the simplest examples. Fortunately, the SUPREM software package introduced in Chapter 3 also includes complete models for diffusion. SUPREM can simulate one- or two-dimensional diffusion profiles. This is accomplished using the DIFFUSION command. The output of the program is typically the chemical, carrier, and vacancy concentrations as functions of depth into the semiconductor substrate.

All diffusion process simulators, including SUPREM, are based upon three basic equations.¹⁵ The first equation is for the flux (J), which in one dimension is given by

$$J_i = -D_i \frac{dC_i}{dx} + Z_i \mu_i C_i E \quad (26)$$

where Z_i is the charge state, μ_i is the mobility of the impurity, and E is the electric field. The subscript i indicates the SUPREM grid location. The second relationship is the continuity equation, which is given by

$$\frac{dC_i}{dt} + \frac{dJ_i}{dx} = G_i \quad (27)$$

where G_i is the generation/recombination rate of the impurity. The final key relationship is Poisson's equation, which in one dimension is given by

$$\frac{d}{dx} [\epsilon_s E] = q(p - n + N_D^+ - N_A^-) \quad (28)$$

where ϵ_s is the permittivity, n and p are the electron and hole concentrations, and N_D^+ and N_A^- are the concentrations of ionized donors and acceptors, respectively. SUPREM solves Eqs. 26 to 28 simultaneously over a one-dimensional grid specified by the user. The diffusivity values used by SUPREM are based on the vacancy model of Fair.¹¹ The values of E_s and D_0 for B, Sb, and As are included in a look-up table. Empirical models are used to account for field-aided, oxidation-enhanced, and oxidation-retarded diffusion.

EXAMPLE 3

Suppose we want to simulate the predeposition of boron into an n -type $\langle 100 \rangle$ silicon wafer at 850°C for 15 minutes. If the silicon substrate is doped with phosphorus at a level of 10^{16} cm^{-3} , use SUPREM to determine the boron doping profile and the junction depth.

SOLUTION The SUPREM input listing is as follows:

```
TITLE      Predeposition Example
COMMENT    Initialize silicon substrate
INITIALIZE <100> Silicon Phosphor Concentration=1e16
COMMENT    Diffuse boron
DIFFUSION  Time=15 Temperature=850 Boron Solidsol
PRINT      Layers Chemical Concentration Phosphorus Boron Net
PLOT       Active Net Cmin=1e15
STOP       End predeposition example
```

Note that surface concentration of the boron is set to the solid solubility limit by the Solidsol parameter in the DIFFUSION command. After predeposition is complete, we print and plot the boron concentration as a function of depth into the silicon substrate. The results are shown in Figure 6.13, which indicates a junction depth of $0.0555 \mu\text{m}$.

► 6.5 SUMMARY

Diffusion is a key method of impurity doping. This chapter first considered the basic diffusion equation for constant diffusivity. The complementary error function (erfc) and the Gaussian function were obtained for the constant surface concentration case and constant total dopant cases, respectively. The results of a diffusion process can be evaluated by measurements of the junction depth, the sheet resistance, and the dopant profile.



Figure 6.13 Plot of boron concentration as a function of depth into the silicon substrate, using SUPREM.

When the doping concentration is higher than the intrinsic carrier concentration n_i at the diffusion temperature, the diffusivity becomes concentration dependent. This dependence has a profound effect on the resulting doping profile. For example, arsenic and boron diffusivity in silicon vary linearly with the impurity concentration. Their doping profiles are much more abrupt than the erfc profile. Phosphorus diffusivity in silicon varies as the square of concentration. This dependence and a dissociation effect give rise to a phosphorus diffusivity that is 100 times larger than its intrinsic diffusivity.

Lateral diffusion at the edge of a mask and impurity redistribution during oxidation are two processes in which diffusion can have an important impact on device performance. The former can substantially reduce the breakdown voltage, and the latter will influence the threshold voltage as well as the contact resistance.

REFERENCES

1. S. M. Sze, Ed., *VLSI Technology*, 2nd Ed., McGraw-Hill, New York, 1988, Ch. 7, 8.
2. S. K. Ghandhi, *VLSI Fabrication Principles*, 2nd Ed., Wiley, New York, 1994, Ch. 4, 6.
3. W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, Boston, 1990, Ch. 8.
4. H. C. Casey and G. L. Pearson, "Diffusion in Semiconductors," in J. H. Crawford and L. M. Slifkin, Eds., *Point Defects in Solids*, Vol. 2, Plenum, New York, 1975.
5. J. P. Joly, "Metallic Contamination of Silicon Wafers," *Microelectron. Eng.*, **40**, 285 (1998).
6. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
7. ASTM Method F374-88, "Test Method for Sheet Resistance of Silicon Epitaxial, Diffused, and Ion-Implanted Layers Using a Collinear Four-Probe Array," V10, 249 (1993).
8. J. C. Irvin, "Evaluation of Diffused Layers in Silicon," *Bell Syst. Tech. J.*, **41**, 2 (1962).
9. S. M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd Ed., Wiley, New York, 2002, Ch. 7.
10. ASTM Method E1438-91, "Standard Guide for Measuring Width of Interfaces in Sputter Depth Profiling Using SIMS," V10, 578 (1993).
11. R. B. Fair, "Concentration Profiles of Diffused Dopants," in F. F. Y. Wang, Ed., *Impurity Doping Processes in Silicon*, North-Holland, Amsterdam, 1981.
12. L. R. Weisberg and J. Blanc, "Diffusion with Interstitial-Substitutional Equilibrium, Zinc in GaAs," *Phys. Rev.*, **131**, 1548 (1963).
13. F. A. Cunnell and C. H. Gooch, "Diffusion of Zinc in Gallium Arsenide," *J. Phys. Chem. Solid*, **15**, 127 (1960).
14. D. P. Kennedy and R. R. O'Brien, "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar p-n Junction," *IBM J. Res. Dev.*, **9**, 179 (1965).
15. S. A. Campbell, *The Science and Engineering of Microelectronic Fabrication*, 2nd Ed., Oxford University Press, New York, 2001, Ch. 3.

PROBLEMS

Asterisks denote difficult problems.

SECTION 6.1: BASIC DIFFUSION PROCESS

1. Calculate the junction depth and the total amount of dopant introduced after boron predeposition performed at 950°C for 30 minutes in a neutral ambient. Assume the substrate is n-type silicon with $N_D = 1.8 \times 10^{16} \text{ cm}^{-3}$ and the boron surface concentration is $C_s = 1.8 \times 10^{20} \text{ cm}^{-3}$.
2. If the sample in Problem 1 is subjected to a neutral drive-in at 1050°C for 60 minutes, calculate the diffusion profile and the junction depth.

3. Assume the measured phosphorus profile can be represented by a Gaussian function with a diffusivity $D = 2.3 \times 10^{-13} \text{ cm}^2/\text{s}$. The measured surface concentration is $1 \times 10^{18} \text{ atoms/cm}^3$, and the measured junction depth is $1 \mu\text{m}$ at a substrate concentration of 1×10^{15} . Calculate the diffusion time and the total dopant in the diffused layer.
- *4. To avoid wafer warp due to a sudden reduction in temperature, the temperature in a diffusion furnace is decreased linearly from 1000°C to 500°C in 20 minutes. What is the effective diffusion time at the initial diffusion temperature for a phosphorus diffusion in silicon?
- *5. For a low-concentration phosphorus drive-in diffusion in silicon at 1000°C, find the percentage change of surface concentration for 1% variation in diffusion time and temperature.
6. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 1100°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at $4 \times 10^{18} \text{ atoms/cm}^3$? What are the diffusion length and junction depth?

SECTION 6.2: EXTRINSIC DIFFUSION

7. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 900°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at $4 \times 10^{18} \text{ atoms/cm}^3$? What is the junction depth? Assume the following:

$$D_0 = 45.8 \text{ cm}^2/\text{s} \quad E_a = 4.05 \text{ eV} \quad x_j = 1.6\sqrt{Dt}$$

8. Explain the meaning of *intrinsic diffusion* and *extrinsic diffusion*.

SECTION 6.4: DIFFUSION SIMULATION

9. Use SUPREM to perform a drive-in step for 6 hours at 1175°C following the predeposition described in Example 3. Plot the boron profile and give the new junction depth.
- *10. After the boron drive-in step in Problem 9, suppose phosphorus is subsequently predeposited and driven in. The phosphorus predeposition occurs at 850°C for 30 minutes, and the drive-in occurs at 1000°C for 30 minutes. Use SUPREM to plot the phosphorus and boron impurity profiles, and determine the junction depth(s).

Ion Implantation

As discussed in Chapter 6, diffusion and ion implantation are the two key methods of impurity doping. Since the early 1970s, many doping operations have been performed by ion implantation, as shown in Figure 7.1. In this process the dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor, and the profile of the dopant distribution is determined mainly by the ion mass and the implanted-ion energy. This chapter discusses the following topics:

- The process and advantages of ion implantation
- Ion distributions in the crystal lattice and how to remove lattice damage caused by ion implantation
- Implantation-related processes, such as masking, high-energy implantation, and high-current implantation
- The simulation of ion implantation using SUPREM

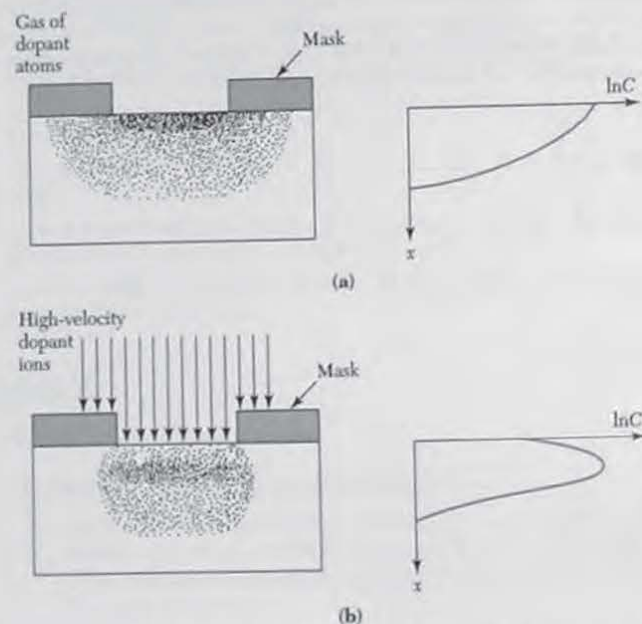


Figure 7.1 Comparison of (a) diffusion and (b) ion-implantation techniques for the selective introduction of dopants into the semiconductor substrate.

7.1 RANGE OF IMPLANTED IONS

Ion implantation is the introduction of energetic, charged particles into a substrate such as silicon. Implantation energies are between 1 keV and 1 MeV, resulting in ion distributions with average depths ranging from 10 nm to 10 μm . Ion doses vary from 10^{12} ions/ cm^2 for threshold voltage adjustment to 10^{18} ions/ cm^2 for the formation of buried insulating layers. Note that the dose is expressed as the number of ions implanted into 1 cm^2 of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature compared with those of the diffusion process.

Figure 7.2 schematically shows a medium-energy ion implanter.¹ The ion source has a heated filament to break up source gases such as BF_3 or AsH_3 into charged ions (B^+ or As^+). An extraction voltage, around 40 kV, causes the charged ions to move out of the ion-source chamber into a mass analyzer. The magnetic field of the analyzer is chosen such that only ions with the desired mass-to-charge ratio can travel through it without being filtered. The selected ions then enter the acceleration tube, where they are accelerated to the implantation energy as they move from high voltage to ground. Apertures ensure that the ion beam is well collimated. The pressure in the implanter is kept below 10^{-4} Pa to minimize ion scattering by gas molecules. The ion beam is then scanned over the wafer surface using electrostatic deflection plates and is implanted into the semiconductor substrate.

The energetic ions lose energy through collisions with electrons and nuclei in the substrate and finally come to rest at some depth within the lattice. The average depth can be controlled by adjusting the acceleration energy. The dopant dose can be controlled by monitoring the ion current during implantation. The principle side effect is disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove this damage.

7.1.1 Ion Distribution

The total distance that an ion travels in coming to rest is called its *range* (R) and is illustrated in Figure 7.3a.² The projection of this distance along the axis of incidence is called the *projected range* (R_p). Because the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having the same mass and the same initial energy. The statistical fluctuations in the projected range are called the *projected straggle* (σ_p). There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, which is called the *lateral straggle* (σ_l).

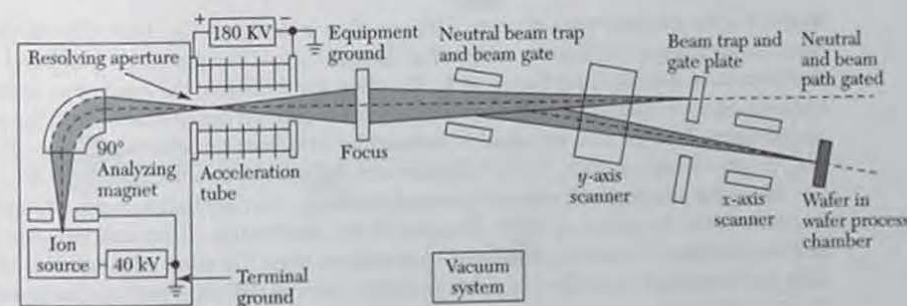


Figure 7.2 Schematic of a medium-current ion implanter.

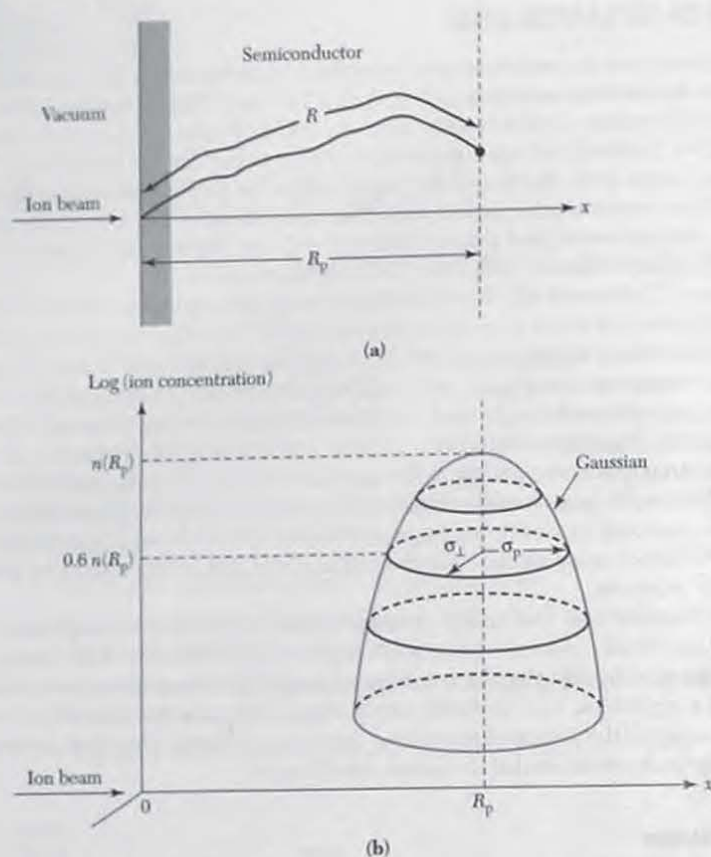


Figure 7.3 (a) Schematic of the ion range R and projected range R_p . (b) Two-dimensional distribution of the implanted ions.²

Figure 7.3b shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function:

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x - R_p)^2}{2\sigma_p^2}\right] \quad (1)$$

where S is the ion dose per unit area. This equation is similar to Eq. 14 in Chapter 6 for constant total dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$ and the distribution is shifted along the x -axis by R_p . Thus, for diffusion, the maximum concentration is at $x = 0$, whereas for ion implantation the maximum concentration is at the projected range. The ion concentration is reduced by 40% from its peak value at $(x - R_p) = \pm\sigma_p$, by one decade at $\pm 2\sigma_p$, by two decades at $\pm 3\sigma_p$, and by five decades at $\pm 4.8\sigma_p$.

Along the axis perpendicular to the axis of incidence, the distribution is also a Gaussian function of the form $\exp(-y^2/2\sigma_{\perp}^2)$. Because of this distribution, there will be some lateral implantation.³ However, the lateral penetration from the mask edge (on the order of $\sigma_{\perp}$) is considerably smaller than that from the thermal diffusion process discussed in Section 6.3.

7.1.2 Ion Stopping

There are two stopping mechanisms by which an energetic ion, on entering a semiconductor substrate (also called the *target*), can be brought to rest. The first is by transferring its energy to the target nuclei. This causes deflection of the incident ion and also dislodges many target nuclei from their original lattice sites. If E is the energy of the ion at any point x along its path, we can define a nuclear stopping power, $S_n(E) \equiv (dE/dx)_n$, to characterize this process. The second stopping mechanism is by the interaction of the incident ion with the cloud of electrons surrounding the target's atoms. The ion loses energy in collisions with electrons through Coulombic interaction. The electrons can be excited to higher energy levels (excitation), or they can be ejected from the atom (ionization). We can define an electronic stopping power, $S_e(E) \equiv (dE/dx)_e$, to characterize this process.

The average rate of energy loss with distance is given by a superposition of the two stopping mechanisms:

$$\frac{dE}{dx} = S_n(E) + S_e(E) \quad (2)$$

If the total distance traveled by the ion before coming to rest is R , then

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)} \quad (3)$$

where E_0 is the initial ion energy. The quantity R has been defined previously as the range.

We can visualize the nuclear stopping process by considering the elastic collision between an incoming hard sphere (energy E_0 and mass M_1) and a target hard sphere (initial energy zero and mass M_2), as illustrated in Figure 7.4. When the spheres collide, momentum is transferred along the centers of the spheres. The deflection angle (θ) and the velocities, v_1 and v_2 , can be obtained from the requirements for conservation of momentum and energy. The maximum energy loss is in a head-on collision. For this case, the energy loss by the incident particle M_1 , or the energy transferred to M_2 , is

$$\frac{1}{2} M_2 v_2^2 = \left[\frac{4M_1 M_2}{(M_1 + M_2)^2} \right] E_0 \quad (4)$$

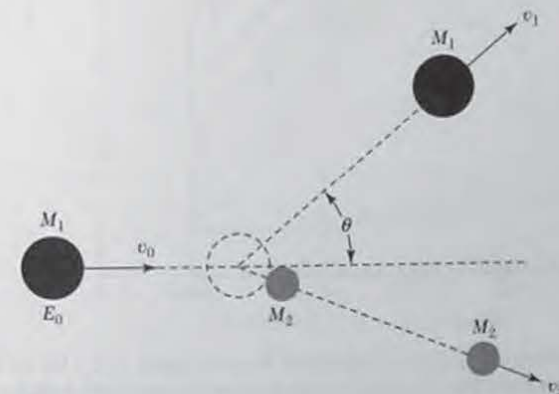


Figure 7.4 Collision of hard spheres.

Since M_2 is usually of the same order of magnitude as M_1 , a large amount of energy can be transferred in a nuclear stopping process.

Detailed calculations show that the nuclear stopping power increases linearly with energy at low energies (similar to Eq. 4), and $S_n(E)$ reaches a maximum at some intermediate energy. At high energies, $S_n(E)$ becomes smaller because fast particles may not have sufficient interaction time with the target atoms to achieve effective energy transfer. The calculated values of $S_n(E)$ for arsenic, phosphorus, and boron in silicon at various energies are shown in Figure 7.5 (solid lines, where the superscript indicates the atomic weight).⁴ Note that heavier atoms, such as arsenic, have a larger nuclear stopping power, that is, larger energy loss per unit distance.

The electronic stopping power is found to be proportional to the velocity of the incident ion, or

$$S_e(E) = k_e \sqrt{E} \quad (5)$$

where the coefficient k_e is a relatively weak function of atomic mass and atomic number. The value of k_e is approximately 10^{-7} (eV)^{1/2}/cm for silicon and 3×10^{-7} (eV)^{1/2}/cm for gallium arsenide. The electronic stopping power in silicon is plotted in Figure 7.5 (dotted line). Also shown in the figure are the crossover energies, at which $S_e(E) = S_n(E)$. For boron, which has a relatively low ion mass compared with the target silicon atom, the crossover energy is only 10 keV. This means that over most of the implantation energy range of 1 keV to 1 MeV, the main energy loss mechanism is due to electronic stopping. On the other hand, for arsenic, with its relatively high ion mass, the crossover energy is 700 keV. Thus, nuclear stopping dominates over most of the energy range. For phosphorus,

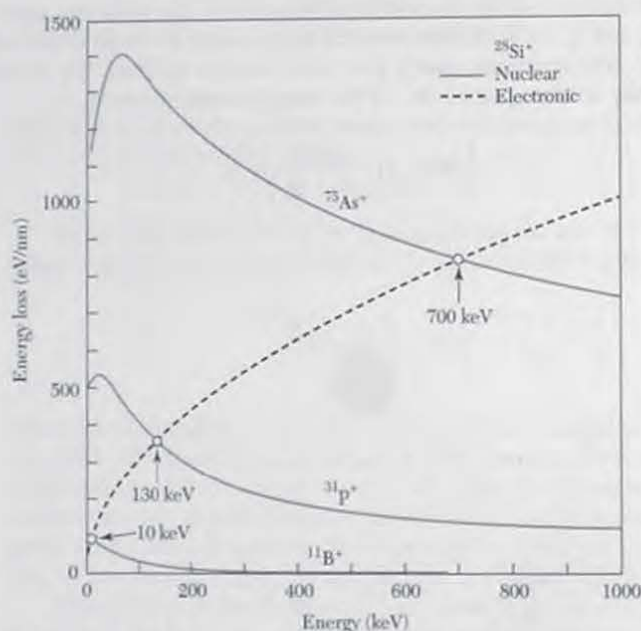


Figure 7.5 Nuclear stopping power, $S_n(E)$, and electronic stopping power, $S_e(E)$, for As, P, and B in Si. The points of intersection of the curves correspond to the energy at which nuclear and electronic stopping are equal.⁴

the crossover energy is 130 keV. For an E_0 less than 130 keV, nuclear stopping will dominate; for higher energies, electronic stopping will take over.

Once $S_n(E)$ and $S_e(E)$ are known, we can calculate the range from Eq. 3. This in turn can give us the projected range and projected straggle with the help of the following approximate equations:¹

$$R_p \approx \frac{R}{1 + (M_2/3M_1)} \quad (6)$$

$$\sigma_p \approx \frac{2}{3} \left[\frac{\sqrt{M_1 M_2}}{M_1 + M_2} \right] R_p \quad (7)$$

Figure 7.6a shows the projected range (R_p), the projected straggle (σ_p), and the lateral straggle (σ_L) for arsenic, boron, and phosphorus in silicon.⁵ As expected, the larger the energy loss, the smaller the range. Also, the projected range and straggles increase with ion energy. For a given element at a specific incident energy, σ_p and σ_L are comparable and usually within $\pm 20\%$. Figure 7.6b shows the corresponding values for hydrogen, zinc, and tellurium in gallium arsenide.³ If we compare Figures 7.6a and b, we see that most of the popular dopants (except hydrogen) have larger projected ranges in silicon than they have in gallium arsenide.

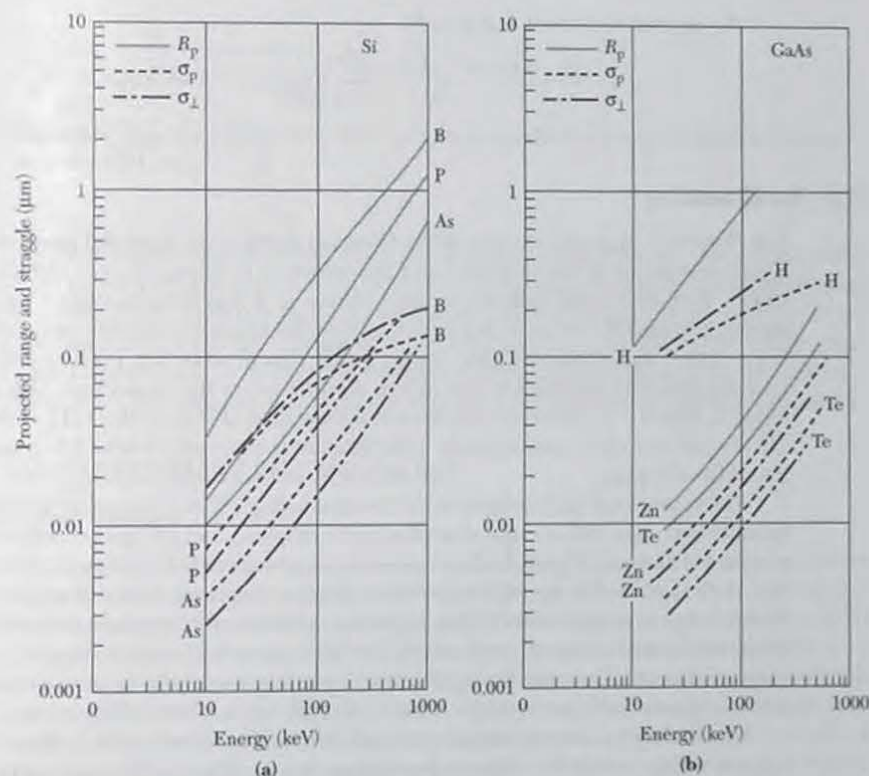


Figure 7.6 Projected range, projected straggle, and lateral straggle for (a) B, P, and As in silicon,⁵ and (b) H, Zn, and Te in gallium arsenide.³

EXAMPLE 1

Assume 100-keV boron implants on a 200-mm silicon wafer at a dose of 5×10^{14} ions/cm². Calculate the peak concentration and the required ion beam current for 1 minute of implantation.

SOLUTION From Figure 7.6a, we obtain 0.31 and 0.07 μm for the projected range and projected straggle, respectively.

From Eq. 1,

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x-R_p)^2}{2\sigma_p^2}\right]$$

$$\frac{dn}{dx} = -\frac{S}{\sqrt{2\pi}\sigma_p} \frac{2(x-R_p)}{2\sigma_p^2} \exp\left[-\frac{(x-R_p)^2}{2\sigma_p^2}\right] = 0$$

The peak concentration is at $x = R_p$, and $n(x) = 2.85 \times 10^{19}$ ions/cm³.

The total number of implanted ions, Q , is given by

$$Q = 5 \times 10^{14} \times \pi \times \left(\frac{20}{2}\right)^2 = 1.57 \times 10^{17} \text{ ions/cm}^2$$

The required ion current, I , is given by

$$I = \frac{qQ}{t} = \frac{1.6 \times 10^{-19} \times 1.57 \times 10^{17}}{60} = 4.19 \times 10^{-4} \text{ A} = 0.42 \text{ mA}$$

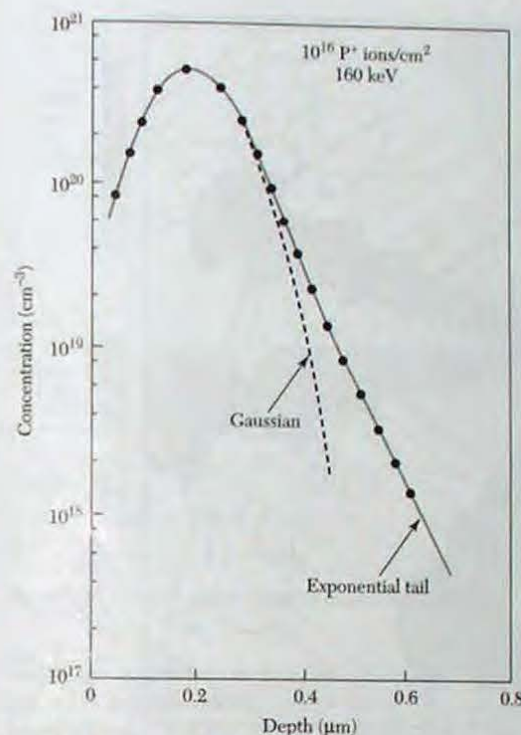


Figure 7.7 Impurity profile obtained in a purposely misoriented target. Ion beam is incident 7° from the $\langle 111 \rangle$ axis.²

7.1.3 Ion Channeling

The projected range and straggle of the Gaussian distribution discussed previously give a good description of the implanted ions in amorphous or fine-grain polycrystalline substrates. Both silicon and gallium arsenide behave as if they were amorphous semiconductors, provided the ion beam is misoriented from the low-index crystallographic direction (e.g., $\langle 111 \rangle$). In this situation, the doping profile described by Eq. 1 is followed closely near the peak and extended to one or two decades below the peak value. This is illustrated in Figure 7.7.² However, even for a misorientation of 7° from the $\langle 111 \rangle$ axis, there still is a tail that varies exponentially with distance as $\exp(-x/\lambda)$, where λ is typically on the order of 0.1 μm .

The exponential tail is related to the *ion channeling* effect. Channeling occurs when incident ions align with a major crystallographic direction and are guided between rows of atoms in a crystal. Figure 7.8 illustrates a diamond lattice viewed along a $\langle 110 \rangle$ direction.⁶ Ions implanted in the $\langle 110 \rangle$ direction will follow trajectories that will not bring them close enough to a target atom to lose significant amounts of energy in nuclear collisions. Thus, for channeled ions, the only energy loss mechanism is electronic stopping, and the range of channeled ions can be significantly larger than it would be in an amorphous target. Ion channeling is particularly critical for low-energy implants and heavy ions.

Channeling can be minimized by several techniques: a blocking amorphous surface layer, misorientation of the wafer, and creating a damaged layer in the wafer surface. The usual blocking amorphous layer is simply a thin layer of grown silicon dioxide (Fig. 7.9a). The layer randomizes the direction of the ion beam so that the ions enter the wafer at

different angles and not directly down the crystal channels. Misorientation of the wafers 5° to 10° off the major plane also has the effect of preventing the ions from entering the channels (Fig. 7.9b). With this method, most implantation machines tilt the wafer by 7° and then apply a 22° twist from the flat to prevent channeling. Pre-damaging the wafer surface with a heavy silicon or germanium implant creates a randomizing layer in the wafer surface (Fig. 7.9c). This method, however, increases the use of the expensive ion implanter.

7.2 IMPLANT DAMAGE AND ANNEALING

7.2.1 Implant Damage

When energetic ions enter a semiconductor substrate, they lose their energy in a series of nuclear and electronic collisions and finally come to rest. The electronic energy loss can be accounted for in terms of electronic excitations to higher energy levels or in terms of the generation of electron-hole pairs. However, electronic collisions do not displace semiconductor atoms from their lattice positions. Only nuclear collisions can transfer sufficient energy to the lattice so that host atoms are displaced, resulting in implant damage (also called *lattice disorder*).⁷ These displaced atoms may possess large fractions of the incident energy, and they in turn cause cascades of secondary displacements of nearby atoms to form a tree of disorder along the ion path. When the displaced atoms per unit volume approach the atomic density of the semiconductor, the material becomes amorphous.

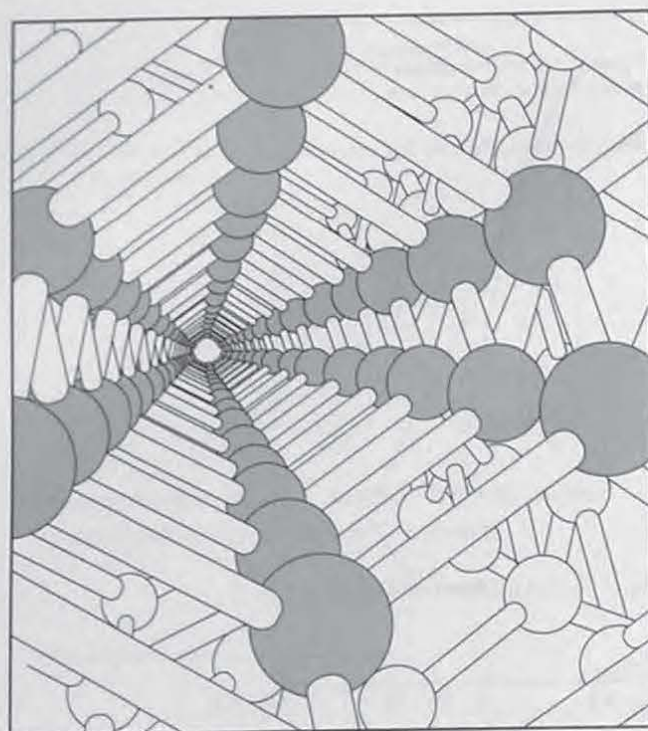


Figure 7.8 Model for a diamond structure, viewed along a $\langle 110 \rangle$ axis.⁶

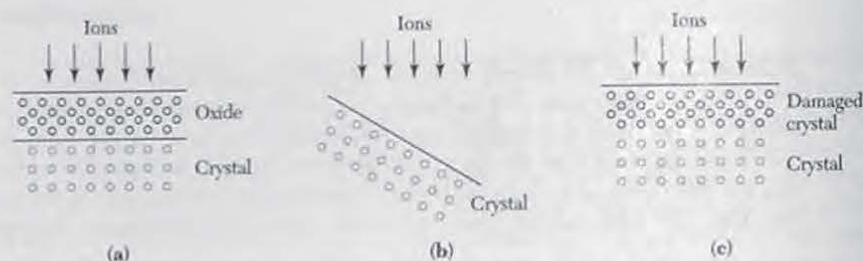


Figure 7.9 Minimizing channeling. (a) Implantation through an amorphous oxide layer. (b) Misorientation of the beam direction to all crystal axes. (c) Pre-damage on the crystal surface.

The tree of disorder for light ions is quite different from that for heavy ions. Much of the energy loss for light ions (e.g., $^{11}\text{B}^+$ in silicon) is due to electronic collisions (see Fig. 7.5), which do not cause lattice damage. The ions lose their energies as they penetrate deeper into the substrate. Eventually, the ion energy is reduced below the crossover energy (10 keV for boron) where nuclear stopping becomes dominant. Therefore, most of the lattice disorder occurs near the final ion position. This is illustrated in Figure 7.10a.

We can estimate the damage by considering a 100-keV boron ion. Its projected range is $0.31\text{ }\mu\text{m}$ (Fig. 6a), and its initial nuclear energy loss is only $3\text{ eV}/\text{\AA}$ (Fig. 7.5). Since the spacing between lattice planes in silicon is about $2.5\text{ }\text{\AA}$, this means that the boron ion will lose 7.5 eV at each lattice plane because of nuclear stopping. The energy required

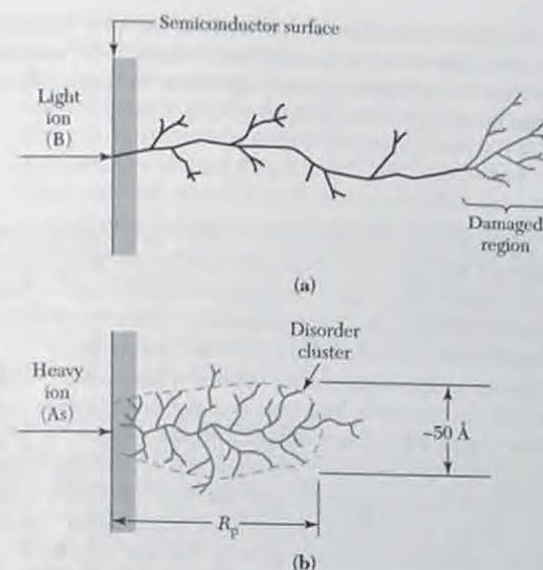


Figure 7.10 Implantation disorder caused by (a) light ions and (b) heavy ions.

to displace a silicon atom from its lattice position is about 15 eV . Therefore, the incident boron ion does not release enough energy from nuclear stopping to displace a silicon atom when it first enters the silicon substrate. When the ion energy is reduced to about 50 keV (at a depth of $1500\text{ }\text{\AA}$), the energy loss due to nuclear stopping increases to 15 eV for each lattice plane (i.e., $6\text{ eV}/\text{\AA}$), sufficient to create a lattice disorder. Assuming that 1 atom is displaced per lattice plane for the remaining ion range, we have 600 lattice atoms displaced (i.e., $1500\text{ }\text{\AA}/2.5\text{ }\text{\AA}$). If each displaced atom moves roughly $25\text{ }\text{\AA}$ from its original position, the damage volume is given by $V_D \approx \pi(25\text{ }\text{\AA})^2(1500\text{ }\text{\AA}) = 3 \times 10^{-18}\text{ cm}^3$. The damage density is $600/V_D \approx 2 \times 10^{20}\text{ cm}^{-3}$, which is only 0.4% of the atoms. Thus, very high doses of light ions are needed to create an amorphous layer.

For heavy ions, the energy loss is primarily due to nuclear collisions; therefore, we expect substantial damage. Consider a 100-keV arsenic ion with a projected range of $0.06\text{ }\mu\text{m}$, or 60 nm . The average nuclear energy loss over the entire energy range is about $1320\text{ eV}/\text{nm}$ (Fig. 7.5). This means that the arsenic ion loses about 330 eV for each lattice plane on the average. Most of the energy is given to one primary silicon atom. Each primary atom will subsequently cause 22 displaced target atoms (i.e., $330\text{ eV}/15\text{ eV}$). The total number of displaced atoms is 5280. Assuming a range of 2.5 nm for the displaced atoms, the damage volume is $V_D \approx \pi(2.5\text{ nm})^2(60\text{ nm}) = 10^{-18}\text{ cm}^3$. The damage density is then $5280/V_D \approx 5 \times 10^{21}\text{ cm}^{-3}$, or about 10% of the total number of atoms in V_D . As a result of the heavy-ion implantation, the material becomes essentially amorphous. Figure 7.10b illustrates the situation in which the damage forms a disordered cluster over the entire projected range.

To estimate the dose required to convert a crystalline material to an amorphous form, we can use the criterion that the energy density is of the same order of magnitude as that needed for melting the material (i.e., $10^{21}\text{ keV}/\text{cm}^3$). For 100-keV arsenic ions, the dose required to make amorphous silicon is then

$$S = \frac{(10^{21}\text{ keV}/\text{cm}^3)R_p}{E_0} = 6 \times 10^{13}\text{ ions}/\text{cm}^2 \quad (8)$$

For 100-keV boron ions, the dose required is 3×10^{14} ions/cm² because R_p for boron is five times larger than for arsenic. However, in practice, higher doses ($>10^{16}$ ions/cm²) are required for boron implantation into a target at room temperature because of the nonuniform distribution of the damage along the ion path.

7.2.2 Annealing

Because of the damaged region and the disorder cluster that result from ion implantation, semiconductor parameters such as mobility and lifetime are severely degraded. In addition, most of the ions as implanted are not located in substitutional sites. To activate the implanted ions and to restore mobility and other material parameters, we must anneal the semiconductor at an appropriate combination of time and temperature.

Conventional annealing uses an open-tube, batch furnace system similar to that used for thermal oxidation. This process requires a long time and high temperature to remove the implant damage. However, conventional annealing may cause substantial dopant diffusion and cannot meet requirements for shallow junctions and narrow doping profiles. *Rapid thermal annealing* (RTA) is an annealing process that employs a variety of energy sources with a wide range of times, from 100 seconds down to nanoseconds—all short compared with conventional annealing. RTA can activate dopants fully with minimal redistribution.

Conventional Annealing of Boron and Phosphorus

Annealing characteristics depend on the dopant type and the dose involved. Figure 7.11 shows the annealing behaviors of boron and phosphorus implantation into silicon substrates.⁵ The substrate is held at room temperature (T_s) during implantation. At a given ion dose, the annealing temperature is defined as the temperature at which 90% of the

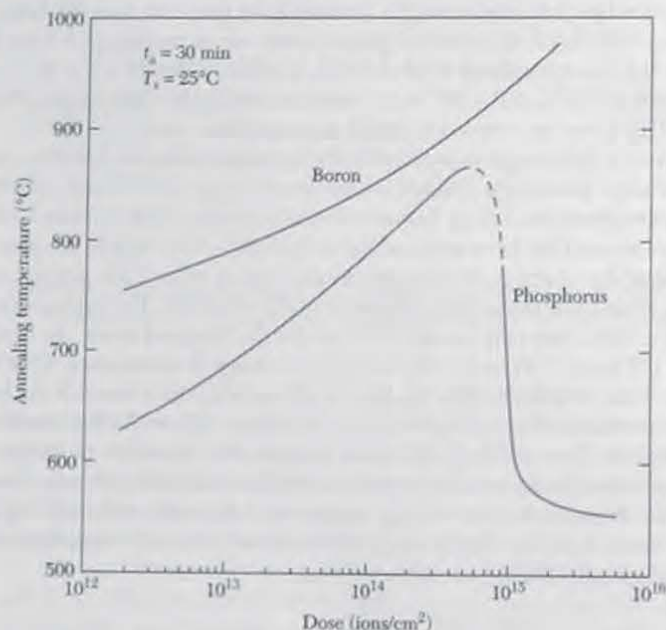


Figure 7.11 Annealing temperature versus dose for 90% activation of boron and phosphorus.⁵

implanted ions are activated by a 30-minute annealing in a conventional annealing furnace. For boron implantation, higher annealing temperatures are needed for higher doses. For phosphorus at lower doses, the annealing behavior is similar to that for boron. However, when the dose is greater than 10^{15} cm⁻², the annealing temperature drops to about 600°C. This phenomenon is related to the solid-phase epitaxy process (see Chapter 8). At phosphorus doses greater than 6×10^{14} cm⁻², the silicon surface layer becomes amorphous. The single-crystal semiconductor underneath the amorphous layer serves as a seeding area for recrystallization of the amorphous layer. The epitaxial growth rate along the $\langle 100 \rangle$ direction is 10 nm/min at 550°C and 50 nm/min at 600°C, with an activation energy at 2.4 eV. Therefore, a 100- to 500-nm amorphous layer can be recrystallized in a few minutes. During the solid-phase epitaxial process, the impurity dopant atoms are incorporated into the lattice sites along with the host atoms. Thus, full activation can be obtained at relatively low temperatures.

Rapid Thermal Annealing

The machine for RTA with transient lamp heating is shown in Figure 7.12. The temperature measured from the heated wafer is usually from 600°C to 1100°C.⁶ A wafer is heated quickly under atmospheric conditions or at low pressure under isothermal conditions. Typical lamps in an RTA system are tungsten filaments or arc lamps. The processing chamber is made of either quartz, silicon carbide, stainless steel, or aluminum and has quartz windows through which the optical radiation passes to illuminate the wafer. The wafer holder is often made of quartz and contacts the wafer in a minimum number of places. A measurement system is placed in a control loop to set wafer temperature. The RTA system interfaces with a gas-handling system and a computer that controls system operation. Typically, wafer temperature in an RTA system is measured with a noncontact optical pyrometer that determines temperature from radiated infrared energy.

Table 7.1 compares conventional furnace and RTA technology. To achieve short processing times using RTA, trade-offs must be made in temperature and process uniformity, temperature measurement and control, and wafer stress and throughput. In addition, there are concerns about the introduction of electrically active wafer defects during the

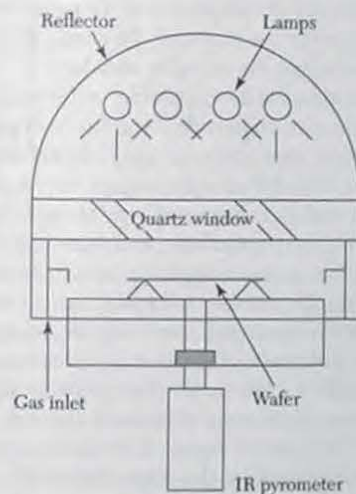


Figure 7.12 Rapid thermal annealing system that is optically heated.

TABLE 7.1 Technology Comparison

Determinant	Conventional Furnace	Rapid Thermal Annealing
Process	Batch	Single-wafer
Furnace	Hot-wall	Cold-wall
Heating rate	Low	High
Cycle time	High	Low
Temperature monitor	Furnace	Wafer
Thermal budget	High	Low
Particle problem	Yes	Minimal
Uniformity and repeatability	High	Low
Throughput	High	Low

very fast (100–300 °C/s) thermal transients. Rapid heating with temperature gradients in the wafers can cause wafer damage in the form of slip dislocations induced by thermal stress. On the other hand, conventional furnace processing causes significant problems, such as particle generation from the hot walls, limited ambient control in an open system, and a large thermal mass that restricts controlled heating times to tens of minutes. In fact, requirements on contamination, process control, and cost of manufacturing floor space have resulted in a paradigm shift to the RTA process.

7.3 IMPLANTATION-RELATED PROCESSES

This section considers a few implantation-related processes, such as multiple implantation, masking, tilt-angle implantation, high-energy implantation, and high-current implantation.

7.3.1 Multiple Implantation and Masking

In many applications, doping profiles other than the simple Gaussian distribution are required. One such case is the preimplantation of silicon with an inert ion to make the silicon surface region amorphous. This technique allows close control of the doping profile and permits nearly 100% dopant activation at low temperatures, as discussed previously. In such a case, a deep amorphous region may be required. To obtain this type of region, we must make a series of implants at varying ion energies and doses.

Multiple implantation can also be used to form a flat doping profile, as shown in Figure 7.13. Here, four boron implants into silicon are used to provide a composite doping profile.⁹ The measured carrier concentration and that predicted using range theory are shown in the figure. Other doping profiles, unavailable from diffusion techniques, can be obtained by using various combinations of impurity dose and implantation energy. Multiple implants have been used to preserve stoichiometry during the implantation and annealing of GaAs. This approach, whereby equal amounts of gallium and an *n*-type dopant (or arsenic and a *p*-type dopant) are implanted prior to annealing, has resulted in higher carrier activation.

To form *p-n* junctions in selected areas of the semiconductor substrate, an appropriate mask should be used for the implantation. Because implantation is a low-temperature process, a large variety of the masking materials can be used. The minimum thickness of the masking material required to stop a given percentage of incident ions can be estimated from the range parameters for ions. The inset of Figure 7.14 shows a profile of an implant in a masking material. The dose implanted in the region beyond a depth *d* (shown shaded) is given by integration of Eq. 1 as

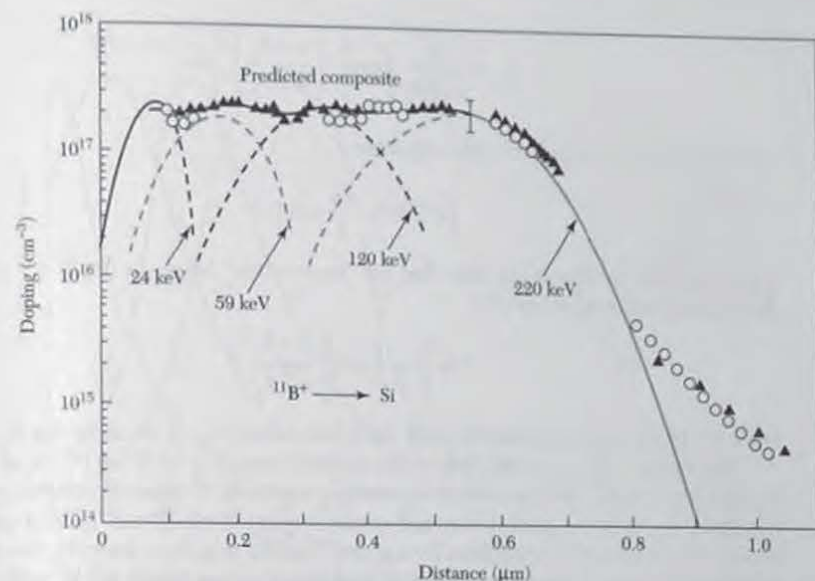


Figure 7.13 Composite doping profile using multiple implants.⁹

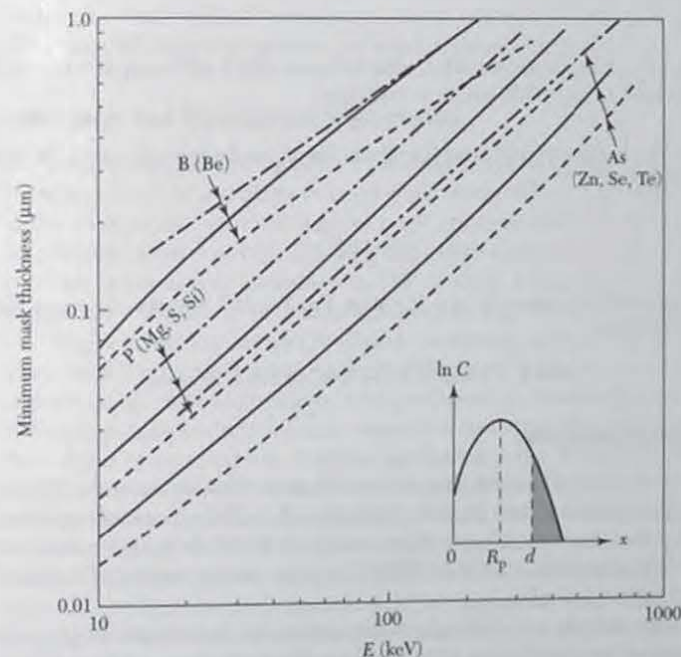


Figure 7.14 Minimum thickness of SiO₂ (—), Si₃N₄ (---), and photoresist (— · —) to produce a masking effectiveness of 99.99%.^{5,10}

$$S_d = \frac{S}{\sqrt{2\pi}\sigma_p} \int_d^{\infty} \exp\left[-\left(\frac{x-R_p}{\sqrt{2}\sigma_p}\right)^2\right] dx \quad (9)$$

From Table 6.1, we can derive the expression

$$\int_x^{\infty} e^{-y^2} dy = \frac{\sqrt{\pi}}{2} \operatorname{erfc}(x) \quad (10)$$

Therefore, the fraction of the dose that has “transmitted” beyond a depth d is given by the transmission coefficient (T):

$$T \equiv \frac{S_d}{S} = \frac{1}{2} \operatorname{erfc}\left(\frac{d-R_p}{\sqrt{2}\sigma_p}\right) \quad (11)$$

Once T is given, we can obtain the mask thickness d from Eq. 11 for any given R_p and σ_p .

The values of d to stop 99.99% of the incident ions ($T = 10^{-4}$) are shown in Figure 7.14 for SiO_2 , Si_3N_4 , and photoresist as masking materials.^{5,10} Mask thicknesses given in this figure are for boron, phosphorus, and arsenic implanted into silicon. These mask thicknesses also can be used as guidelines for impurity masking in gallium arsenide. The dopants are shown in the parentheses. Since both R_p and σ_p vary approximately linearly with energy, the minimum thickness of the masking material also increases linearly with energy. In certain applications, instead of totally stopping the beam, masks can be used as attenuators that can provide an amorphous surface layer to the incident ion beam to minimize the channeling effect.

EXAMPLE 2

When boron ions are implanted at 200 keV, what thickness of SiO_2 will be required to mask 99.996% of the implanted ions ($R_p = 0.53 \mu\text{m}$, $\sigma_p = 0.093 \mu\text{m}$)?

SOLUTION The complementary error function in Eq. 11 can be approximated if the argument is large (see Table 6.1):

$$T \approx \frac{1}{2\sqrt{\pi}} \frac{e^{-u^2}}{u}$$

where the parameter u is given by $(d - R_p)/\sqrt{2}\sigma_p$. For $T = 10^{-4}$, we solve the above equation to obtain $u = 2.8$. Thus,

$$d = R_p + 3.96\sigma_p = 0.53 + 3.96 \times 0.093 = 0.898 \mu\text{m}$$

7.3.2 Tilt-Angle Ion Implantation

In scaling devices to submicron dimensions, it is important also to scale dopant profiles vertically. We need to produce junction depths less than 100 nm, including diffusion during dopant activation and subsequent processing steps. Modern device structures, such as the lightly doped drain (LDD) MOSFET, require precise control of dopant distributions vertically and laterally.

It is the ion velocity perpendicular to the surface that determines the projected range of an implanted ion distribution. If the wafer is tilted at a large angle to the ion beam, then the effective ion energy is greatly reduced. Figure 7.15 illustrates this for 60-keV arsenic ions as a function of the tilt angle, showing that it is possible to achieve extremely shallow distributions using a high tilt angle (86°). In tilt-angle ion implantation, we should

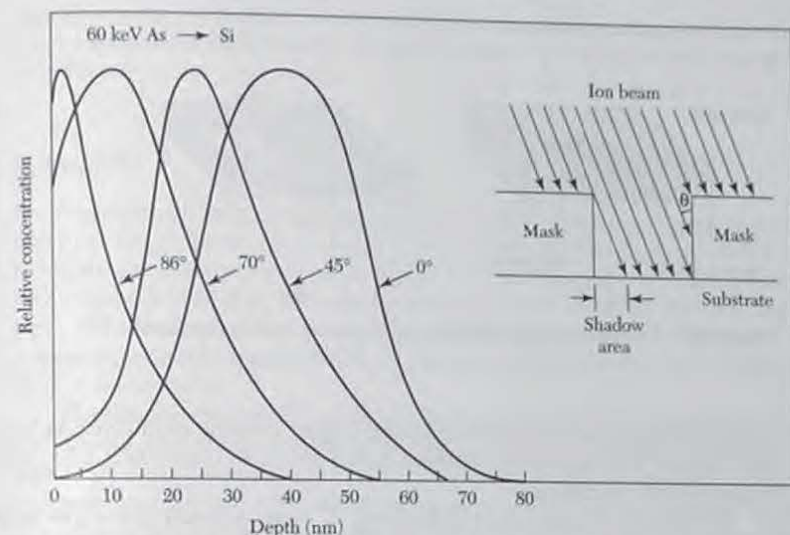


Figure 7.15 60-keV arsenic implanted into silicon, as a function of beam tilt angle. Inset shows the shadow area for tilt-angle ion implantation.

consider the shadow effect (inset in Fig. 7.15) for the patterned wafer. A lower tilt angle leads to a small shadow area. For example, if the height of the patterned mask is $0.5 \mu\text{m}$, with vertical sidewall, a 7° incident ion beam will induce a 61-nm shadow. This shadow effect may introduce an unexpected series resistance in the device.

7.3.3 High-Energy and High-Current Implantation

High-energy implantors, capable of energies as high as 1.5 to 5 MeV, are available and have been used for a number of novel applications. The majority of these depend on the ability to dope the semiconductor to many micrometers in depth, without the need for long diffusion times at high temperatures. High-energy implantors can also be used to produce low-resistivity buried layers. For example, a buried layer 1.5 to $3 \mu\text{m}$ below the surface for a CMOS device can be achieved by high-energy implantation.

High-current implantors (10–20 mA), operating in the 25- to 30-keV range, are routinely used for the predeposition step in diffusion technology because the amount of total dopant can be controlled precisely. After predeposition, the dopant impurities can be driven in by a high-temperature diffusion step at the same time that implant damage at the surface region is annealed out. Another application is the threshold voltage adjustment in MOS devices. A precisely controlled amount of dopant (e.g., boron) is implanted through the gate oxide to the channel region¹¹ (Fig. 7.16a). Because the projected ranges of boron in silicon and silicon oxide are comparable, if we choose a suitable incident energy, the ions will penetrate just the thin gate oxide, not the thicker field oxide. The threshold voltage will vary approximately linearly with the implanted dose. After boron implantation, polysilicon can be deposited and patterned to form the gate electrode of the MOSFET. The thin oxide surrounding the gate electrode is removed, and the source and drain regions are formed, as shown in Figure 7.16b, by another high-dose arsenic implantation.

High-current implantors with energies in the 150- to 200-keV range are now available. A major use for these machines is to form high-quality silicon films, which are