
FUNDAMENTALS OF SEMICONDUCTOR MANUFACTURING AND PROCESS CONTROL

Gary S. May, Ph.D.

Georgia Institute of Technology
Atlanta, Georgia

Costas J. Spanos, Ph.D.

University of California at Berkeley
Berkeley, California



 **WILEY-INTERSCIENCE**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2006 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

May, Gary S.

Fundamentals of semiconductor manufacturing and process control / Gary S.

May, Costas J. Spanos.

p. cm.

“Wiley-Interscience.”

Includes bibliographical references and index.

ISBN-13: 978-0-471-78406-7 (cloth : alk. paper)

ISBN-10: 0-471-78406-0 (cloth : alk. paper)

1. Semiconductors—Design and construction. 2. Integrated circuits—Design and construction. 3. Process control—Statistical methods. I. Spanos, Costas J. II. Title.

TK7871.85.M379 2006

621.3815'2—dc22

2005028448

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface	xvii
Acknowledgments	xix
 1 Introduction to Semiconductor Manufacturing	 1
Objectives / 1	
Introduction / 1	
1.1. Historical Evolution / 2	
1.1.1. Manufacturing and Quality Control / 3	
1.1.2. Semiconductor Processes / 5	
1.1.3. Integrated Circuit Manufacturing / 7	
1.2. Modern Semiconductor Manufacturing / 8	
1.2.1. Unit Processes / 9	
1.2.2. Process Sequences / 11	
1.2.3. Information Flow / 12	
1.2.4. Process Organization / 14	
1.3. Goals of Manufacturing / 15	
1.3.1. Cost / 15	
1.3.2. Quality / 17	
1.3.3. Variability / 17	
1.3.4. Yield / 17	
1.3.5. Reliability / 18	
1.4. Manufacturing Systems / 18	
1.4.1. Continuous Flow / 19	
1.4.1.1. Batch Processes / 20	
1.4.1.2. Single Workpiece / 20	
1.4.2. Discrete Parts / 21	
1.5. Outline for Remainder of the Book / 21	
Summary / 22	
Problems / 22	
References / 23	

2 Technology Overview**25**

Objectives / 25

Introduction / 25

2.1. Unit Processes / 25

2.1.1. Oxidation / 26

2.1.1.1. *Growth Kinetics* / 272.1.1.2. *Thin Oxide Growth* / 312.1.1.3. *Oxide Quality* / 33

2.1.2. Photolithography / 34

2.1.2.1. *Exposure Tools* / 352.1.2.2. *Masks* / 382.1.2.3. *Photoresist* / 392.1.2.4. *Pattern Transfer* / 412.1.2.5. *E-Beam Lithography* / 432.1.2.6. *X-Ray Lithography* / 45

2.1.3. Etching / 47

2.1.3.1. *Wet Chemical Etching* / 472.1.3.2. *Dry Etching* / 48

2.1.4. Doping / 51

2.1.4.1. *Diffusion* / 522.1.4.2. *Ion Implantation* / 56

2.1.5. Deposition / 58

2.1.5.1. *Physical Vapor Deposition* / 592.1.5.2. *Chemical Vapor Deposition* / 60

2.1.6. Planarization / 61

2.2. Process Integration / 61

2.2.1. Bipolar Technology / 63

2.2.2. CMOS Technology / 66

2.2.2.1. *Basic NMOS Fabrication Sequence* / 672.2.2.2. *CMOS Fabrication Sequence* / 70

2.2.3. BiCMOS Technology / 74

2.2.4. Packaging / 75

2.2.4.1. *Die Separation* / 762.2.4.2. *Package Types* / 772.2.4.3. *Attachment Methods* / 79

Summary / 80

Problems / 80

References / 81

3 Process Monitoring 82

Objectives / 82

Introduction / 82

3.1. Process Flow and Key Measurement Points / 83

3.2. Wafer State Measurements / 84

3.2.1. Blanket Thin Film / 85

3.2.1.1. *Interferometry* / 853.2.1.2. *Ellipsometry* / 883.2.1.3. *Quartz Crystal Monitor* / 913.2.1.4. *Four-Point Probe* / 92

3.2.2. Patterned Thin Film / 93

3.2.2.1. *Profilometry* / 933.2.2.2. *Atomic Force Microscopy* / 933.2.2.3. *Scanning Electron Microscopy* / 953.2.2.4. *Scatterometry* / 963.2.2.5. *Electrical Linewidth Measurement* / 98

3.2.3. Particle/Defect Inspection / 98

3.2.3.1. *Cleanroom Air Monitoring* / 993.2.3.2. *Product Monitoring* / 100

3.2.4. Electrical Testing / 102

3.2.4.1. *Test Structures* / 1023.2.4.2. *Final Test* / 106

3.3. Equipment State Measurements / 107

3.3.1. Thermal Operations / 109

3.3.1.1. *Temperature* / 1093.3.1.2. *Pressure* / 1093.3.1.3. *Gas Flow* / 110

3.3.2. Plasma Operations / 111

3.3.2.1. *Temperature* / 1113.3.2.2. *Pressure* / 1123.3.2.3. *Gas Flow* / 1123.3.2.4. *Residual Gas Analysis* / 1123.3.2.5. *Optical Emission Spectroscopy* / 1143.3.2.6. *Fourier Transform Infrared Spectroscopy* / 1153.3.2.7. *RF Monitors* / 116

3.3.3. Lithography Operations / 116

3.3.4. Implantation / 117

3.3.5. Planarization / 118

Summary / 118

Problems / 119

References / 120

4 Statistical Fundamentals

122

Objectives / 122

Introduction / 122

4.1. Probability Distributions / 123

4.1.1. Discrete Distributions / 124

4.1.1.1. *Hypergeometric* / 124

4.1.1.2. *Binomial* / 125

4.1.1.3. *Poisson* / 127

4.1.1.4. *Pascal* / 128

4.1.2. Continuous Distributions / 128

4.1.2.1. *Normal* / 129

4.1.2.2. *Exponential* / 131

4.1.3. Useful Approximations / 132

4.1.3.1. *Poisson Approximation to the
Binomial* / 132

4.1.3.2. *Normal Approximation to the
Binomial* / 132

4.2. Sampling from a Normal Distribution / 133

4.2.1. Chi-Square Distribution / 134

4.2.2. *t* Distribution / 134

4.2.3. *F* Distribution / 135

4.3. Estimation / 136

4.3.1. Confidence Interval for the Mean with Known
Variance / 137

4.3.2. Confidence Interval for the Mean with Unknown
Variance / 137

4.3.3. Confidence Interval for Variance / 137

4.3.4. Confidence Interval for the Difference between Two
Means, Known Variance / 138

4.3.5. Confidence Interval for the Difference between Two
Means, Unknown Variances / 138

4.3.6. Confidence Interval for the Ratio of Two
Variances / 139

4.4. Hypothesis Testing / 140

4.4.1. Tests on Means with Known Variance / 141

4.4.2. Tests on Means with Unknown Variance / 142

4.4.3. Tests on Variance / 143

Summary /	145
Problems /	145
Reference /	146

5 Yield Modeling **147**

Objectives /	147
Introduction /	147
5.1. Definitions of Yield Components /	148
5.2. Functional Yield Models /	149
5.2.1. Poisson Model /	151
5.2.2. Murphy's Yield Integral /	152
5.2.3. Negative Binomial Model /	154
5.3. Functional Yield Model Components /	156
5.3.1. Defect Density /	156
5.3.2. Critical Area /	157
5.3.3. Global Yield Loss /	158
5.4. Parametric Yield /	159
5.5. Yield Simulation /	161
5.5.1. Functional Yield Simulation /	162
5.5.2. Parametric Yield Simulation /	167
5.6. Design Centering /	171
5.6.1. Acceptability Regions /	172
5.6.2. Parametric Yield Optimization /	173
5.7. Process Introduction and Time-to-Yield /	174
Summary /	176
Problems /	177
References /	180

6 Statistical Process Control **181**

Objectives /	181
Introduction /	181
6.1. Control Chart Basics /	182
6.2. Patterns in Control Charts /	184
6.3. Control Charts for Attributes /	186
6.3.1. Control Chart for Fraction Nonconforming /	187
6.3.1.1. Chart Design /	188
6.3.1.2. Variable Sample Size /	189
6.3.1.3. Operating Characteristic and Average Runlength /	191
6.3.2. Control Chart for Defects /	193
6.3.3. Control Chart for Defect Density /	193

6.4. Control Charts for Variables /	195
6.4.1. Control Charts for $\bar{x}$ and R /	195
6.4.1.1. <i>Rational Subgroups</i> /	199
6.4.1.2. <i>Operating Characteristic and Average Runlength</i> /	200
6.4.2. Control Charts for $\bar{x}$ and s /	202
6.4.3. Process Capability /	204
6.4.4. Modified and Acceptance Charts /	206
6.4.5. Cusum Chart /	208
6.4.5.1. <i>Tabular Cusum Chart</i> /	210
6.4.5.2. <i>Average Runlength</i> /	210
6.4.5.3. <i>Cusum for Variance</i> /	211
6.4.6. Moving-Average Charts /	212
6.4.6.1. <i>Basic Moving-Average Chart</i> /	212
6.4.6.2. <i>Exponentially Weighted Moving-Average Chart</i> /	213
6.5. Multivariate Control /	215
6.5.1. Control of Means /	217
6.5.2. Control of Variability /	220
6.6. SPC with Correlated Process Data /	221
6.6.1. Time-Series Modeling /	221
6.6.2. Model-Based SPC /	223
Summary /	224
Problems /	224
References /	227

7 Statistical Experimental Design

228

Objectives /	228
Introduction /	228
7.1. Comparing Distributions /	229
7.2. Analysis of Variance /	232
7.2.1. Sums of Squares /	232
7.2.2. ANOVA Table /	234
7.2.2.1. <i>Geometric Interpretation</i> /	235
7.2.2.2. <i>ANOVA Diagnostics</i> /	237
7.2.3. Randomized Block Experiments /	240
7.2.3.1. <i>Mathematical Model</i> /	242
7.2.3.2. <i>Diagnostic Checking</i> /	243
7.2.4. Two-Way Designs /	245
7.2.4.1. <i>Analysis</i> /	245
7.2.4.2. <i>Data Transformation</i> /	246

7.3. Factorial Designs /	249
7.3.1. Two-Level Factorials /	250
7.3.1.1. <i>Main Effects</i> /	251
7.3.1.2. <i>Interaction Effects</i> /	251
7.3.1.3. <i>Standard Error</i> /	252
7.3.1.4. <i>Blocking</i> /	254
7.3.2. Fractional Factorials /	256
7.3.2.1. <i>Construction of Fractional Factorials</i> /	256
7.3.2.2. <i>Resolution</i> /	257
7.3.3. Analyzing Factorials /	257
7.3.3.1. <i>The Yates Algorithm</i> /	258
7.3.3.2. <i>Normal Probability Plots</i> /	258
7.3.4. Advanced Designs /	260
7.4. Taguchi Method /	262
7.4.1. Categorizing Process Variables /	263
7.4.2. Signal-to-Noise Ratio /	264
7.4.3. Orthogonal Arrays /	264
7.4.4. Data Analysis /	266
Summary /	269
Problems /	269
References /	271

8 Process Modeling

272

Objectives /	272
Introduction /	272
8.1. Regression Modeling /	273
8.1.1. Single-Parameter Model /	274
8.1.1.1. <i>Residuals</i> /	275
8.1.1.2. <i>Standard Error</i> /	276
8.1.1.3. <i>Analysis of Variance</i> /	276
8.1.2. Two-Parameter Model /	277
8.1.2.1. <i>Analysis of Variance</i> /	279
8.1.2.2. <i>Precision of Estimates</i> /	279
8.1.2.3. <i>Linear Model with Nonzero Intercept</i> /	280
8.1.3. Multivariate Models /	283
8.1.4. Nonlinear Regression /	285
8.1.5. Regression Chart /	287
8.2. Response Surface Methods /	289
8.2.1. Hypothetical Yield Example /	289

8.2.1.1. <i>Diagnostic Checking</i>	/ 292
8.2.1.2. <i>Augmented Model</i>	/ 293
8.2.2. Plasma Etching Example	/ 294
8.2.2.1. <i>Experimental Design</i>	/ 295
8.2.2.2. <i>Experimental Technique</i>	/ 297
8.2.2.3. <i>Analysis</i>	/ 298
8.3. Evolutionary Operation	/ 301
8.4. Principal-Component Analysis	/ 306
8.5. Intelligent Modeling Techniques	/ 310
8.5.1. Neural Networks	/ 310
8.5.2. Fuzzy Logic	/ 314
8.6. Process Optimization	/ 318
8.6.1. Powell's Algorithm	/ 318
8.6.2. Simplex Method	/ 320
8.6.3. Genetic Algorithms	/ 323
8.6.4. Hybrid Methods	/ 325
8.6.5. PECVD Optimization: A Case Study	/ 326
Summary	/ 327
Problems	/ 328
References	/ 331

9 Advanced Process Control

333

Objectives	/ 333
Introduction	/ 333
9.1. Run-by-Run Control with Constant Term Adaptation	/ 335
9.1.1. Single-Variable Methods	/ 335
9.1.1.1. <i>Gradual Drift</i>	/ 337
9.1.1.2. <i>Abrupt Shifts</i>	/ 339
9.1.2. Multivariate Techniques	/ 343
9.1.2.1. <i>Exponentially Weighted Moving-Average (EWMA) Gradual Model</i>	/ 343
9.1.2.2. <i>Predictor–Corrector Control</i>	/ 343
9.1.3. Practical Considerations	/ 346
9.1.3.1. <i>Input Bounds</i>	/ 346
9.1.3.2. <i>Input Resolution</i>	/ 348
9.1.3.3. <i>Input Weights</i>	/ 348
9.1.3.4. <i>Output Weights</i>	/ 350
9.2. Multivariate Control with Complete Model Adaptation	/ 351
9.2.1. Detection of Process Disturbances via Model-Based SPC	/ 352

9.2.1.1. <i>Malfunction Alarms</i>	/ 352
9.2.1.2. <i>Alarms for Feedback Control</i>	/ 353
9.2.2. <i>Full Model Adaptation</i>	/ 354
9.2.3. <i>Automated Recipe Generation</i>	/ 356
9.2.4. <i>Feedforward Control</i>	/ 358
9.3. <i>Supervisory Control</i>	/ 359
9.3.1. <i>Supervisory Control Using Complete Model Adaptation</i>	/ 359
9.3.1.1. <i>Acceptable Input Ranges of Photolithographic Machines</i>	/ 361
9.3.1.2. <i>Experimental Examples</i>	/ 363
9.3.2. <i>Intelligent Supervisory Control</i>	/ 364
Summary	/ 373
Problems	/ 373
References	/ 378

10 Process and Equipment Diagnosis 379

Objectives	/ 379
Introduction	/ 379
10.1. <i>Algorithmic Methods</i>	/ 381
10.1.1. <i>Hippocrates</i>	/ 381
10.1.1.1. <i>Measurement Plan</i>	/ 382
10.1.1.2. <i>Fault Diagnosis</i>	/ 383
10.1.1.3. <i>Example</i>	/ 383
10.1.2. <i>MERLIN</i>	/ 384
10.1.2.1. <i>Knowledge Representation</i>	/ 385
10.1.2.2. <i>Inference Mechanism</i>	/ 387
10.1.2.3. <i>Case Study</i>	/ 390
10.2. <i>Expert Systems</i>	/ 391
10.2.1. <i>PIES</i>	/ 391
10.2.1.1. <i>Knowledge Base</i>	/ 393
10.2.1.2. <i>Diagnostic Reasoning</i>	/ 394
10.2.1.3. <i>Examples</i>	/ 395
10.2.2. <i>PEDX</i>	/ 395
10.2.2.1. <i>Architecture</i>	/ 396
10.2.2.2. <i>Rule-Based Reasoning</i>	/ 397
10.2.2.3. <i>Implementation</i>	/ 398
10.3. <i>Neural Network Approaches</i>	/ 398
10.3.1. <i>Process Control Neural Network</i>	/ 398
10.3.2. <i>Pattern Recognition in CVD Diagnosis</i>	/ 400
10.4. <i>Hybrid Methods</i>	/ 402

10.4.1. Time-Series Diagnosis /	402
10.4.2. Hybrid Expert System /	403
10.4.2.1. Dempster–Shafer Theory /	406
10.4.2.2. Maintenance Diagnosis /	408
10.4.2.3. Online Diagnosis /	409
10.4.2.4. Inline Diagnosis /	413
Summary /	414
Problems /	414
References /	415
 Appendix A: Some Properties of the Error Function	 417
Appendix B: Cumulative Standard Normal Distribution	420
Appendix C: Percentage Points of the χ^2 Distribution	423
Appendix D: Percentage Points of the t Distribution	425
Appendix E: Percentage Points of the F Distribution	427
Appendix F: Factors for Constructing Variables Control Charts	438
 Index	 441

TECHNOLOGY OVERVIEW

OBJECTIVES

- Provide an overview of the critical unit processes in semiconductor manufacturing.
- Describe the integration of such processes into sequences for fabricating specific technology families.

INTRODUCTION

Planar fabrication technology is used extensively for integrated circuit manufacturing. In Section 1.2.1 we briefly described the major steps of a planar process. We provide a more thorough description of these steps, as well as their integration for particular technology families, in this chapter. However, this treatment is in no way intended to be comprehensive. More complete and detailed discussions can be found in several other texts, such as *Fundamentals of Semiconductor Fabrication* [1], for example.

2.1. UNIT PROCESSES

Chapter 1 provided an introduction to the key unit process steps in IC fabrication, including oxidation, photolithography, etching, ion implantation, and metallization. This was accomplished using the description of the process sequence used

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos
Copyright © 2006 John Wiley & Sons, Inc.

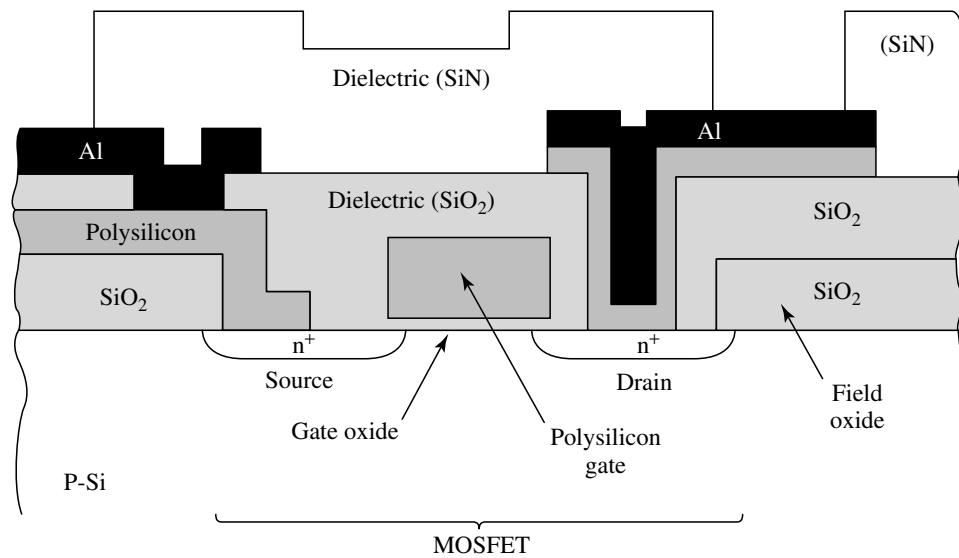


Figure 2.1. Cross section of a MOSFET [1].

to fabricate a p - n junction. Here, we describe each of these steps, as well as planarization, in more detail.

2.1.1. 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. For example, a silicon n -channel MOSFET (Figure 2.1) uses all four groups of films. An important oxide layer is the gate oxide, 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.

Semiconductors can be oxidized by various methods, including thermal oxidation, electrochemical anodization, and plasma-enhanced chemical vapor deposition (PECVD; see Section 2.1.5). Among these, thermal oxidation is the most important for silicon devices. It is a key process in modern silicon IC technology. The basic thermal oxidation apparatus (shown in Figure 2.2) 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. Oxidation temperature is generally in the range of 900–1200°C, and the typical gas flowrate 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, to maintain the oxidation temperature to within $\pm 1^\circ\text{C}$, and to ramp the temperature down when oxidation is completed.

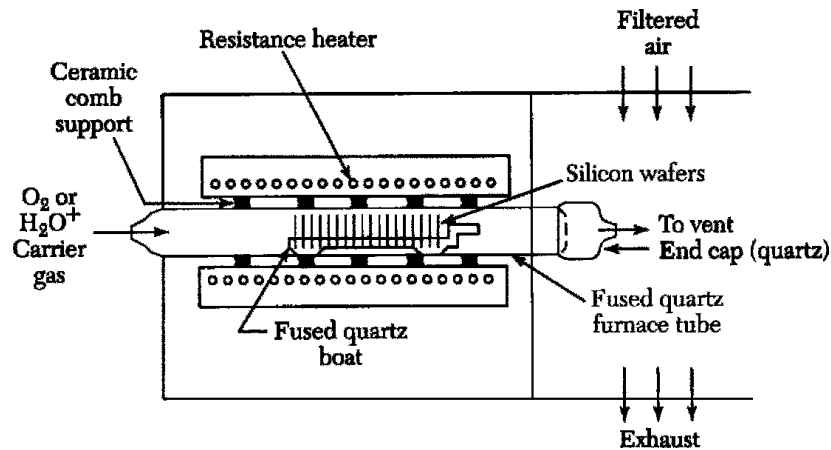
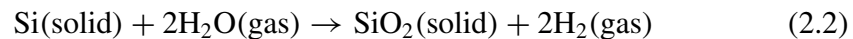


Figure 2.2. Schematic of an oxidation furnace [1].

2.1.1.1. Growth Kinetics

The following chemical reactions describe the thermal oxidation of silicon in oxygen (“dry” oxidation) and water vapor (“wet” oxidation), respectively:



The silicon–silicon dioxide interface moves into the silicon during the oxidation process. This creates a new interface region, with surface contamination on the original silicon ending up on the oxide surface. As a result of the densities and molecular weights of silicon and silicon dioxide, growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Figure 2.3).

The kinetics of silicon oxidation can be described on the basis of the simple model illustrated in Figure 2.4. 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.

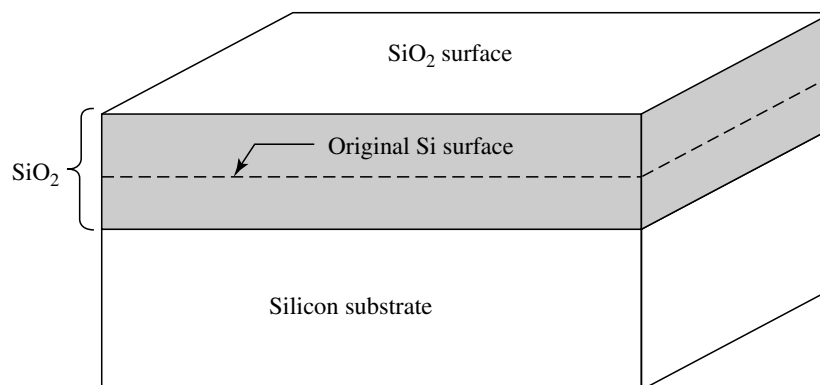


Figure 2.3. Movement of silicon–silicon dioxide interface during oxide growth [1].

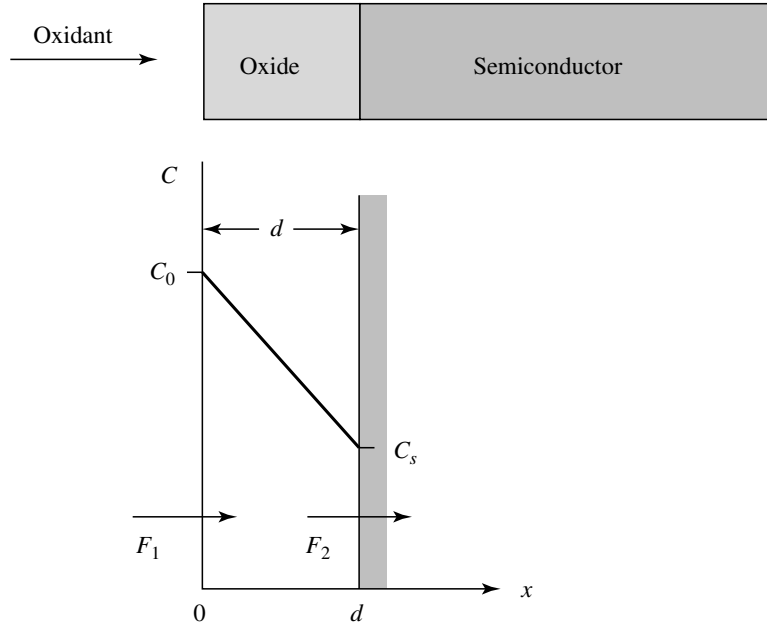


Figure 2.4. Basic model for the thermal oxidation of silicon [1].

At 1000°C and a pressure of 1 atm, the concentration C_0 is $5.2 \times 10^{16} \text{ cm}^{-3}$ for dry oxygen and $3 \times 10^{19} \text{ cm}^{-3}$ 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

$$F_1 = D \frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x} \quad (2.3)$$

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

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

$$F_2 = \kappa C_s \quad (2.4)$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. (2.3) and (2.4) gives

$$F = \frac{DC_0}{x + (D/\kappa)} \quad (2.5)$$

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 one oxygen molecule (O₂) is added to each silicon dioxide molecule, whereas we add two water molecules (H₂O) to each SiO₂ molecule. Therefore, C_1 for oxidation in dry oxygen is $2.2 \times 10^{22} \text{ cm}^{-3}$, and for oxidation in water vapor it is twice

this number ($4.4 \times 10^{22} \text{ cm}^{-3}$). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)} \quad (2.6)$$

This differential equation can be solved 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. (2.6) yields the general relationship for the oxidation of silicon:

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t + \tau) \quad (2.7)$$

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

$$x = \frac{D}{\kappa} \left[\sqrt{1 + \frac{2C_0\kappa^2(t + \tau)}{DC_1}} - 1 \right] \quad (2.8)$$

For small values of t , Eq. (2.8) reduces to

$$x \cong \frac{C_0\kappa}{C_1}(t + \tau) \quad (2.9)$$

and for larger values of t , it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}(t + \tau)} \quad (2.10)$$

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 (2.7) is often written in a more compact form

$$x^2 + Ax = B(t + \tau) \quad (2.11)$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$ and $B/A = \kappa C_0/C_1$. Using this form, Eqs. (2.9) and (2.10) can be written as

$$x = \frac{B}{A}(t + \tau) \quad (2.12)$$

for the linear region and as

$$x^2 = B(t + \tau) \quad (2.13)$$

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 $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 25 nm. Thus, the use of Eq. (2.11) for dry oxidation on bare silicon requires a value for τ that can be generated using this initial thickness. Table 2.1 lists the values of the rate constants for wet oxidation of silicon, and Table 2.2 lists the values for dry oxidation.

The temperature dependence of the linear rate constant B/A is shown in Figure 2.5 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers [1]. 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 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 2.6 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

Table 2.1. Rate constants for wet oxidation of silicon.

Temperature (°C)	A (μm)	B ($\mu\text{m}^2/\text{h}$)	τ (h)
1200	0.05	0.72	0
1100	0.11	0.51	0
1000	0.226	0.287	0
920	0.5	0.203	0

Table 2.2. Rate constants for dry oxidation of silicon.

Temperature (°C)	A (μm)	B ($\mu\text{m}^2/\text{h}$)	τ (h)
1200	0.04	0.045	0.027
1100	0.09	0.027	0.076
1000	0.165	0.0117	0.37
920	0.235	0.0049	1.4
800	0.37	0.0011	9.0
700	—	—	81.0

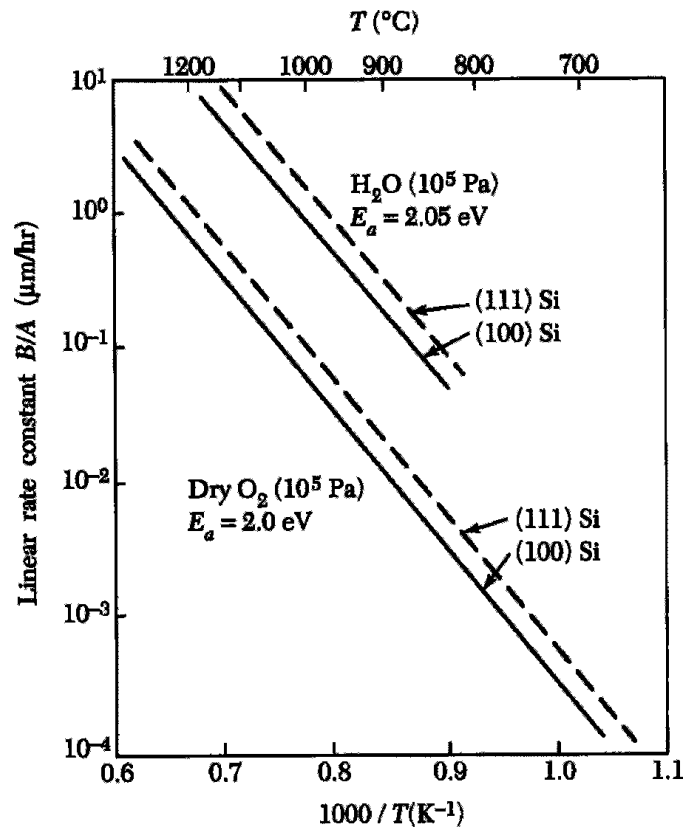


Figure 2.5. Linear rate constant versus temperature [1].

the diffusion process of the oxidizing species through a random network layer of amorphous silica.

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.

2.1.1.2. Thin Oxide Growth

Relatively slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. Approaches to achieve such slower growth rates include growth in dry O_2 at atmospheric pressure and lower temperatures (800–900°C); growth at pressures lower than atmospheric pressure; growth in a 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 chemical vapor deposition (CVD) SiO_2 . However, the mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800–900°C). With this approach, processing

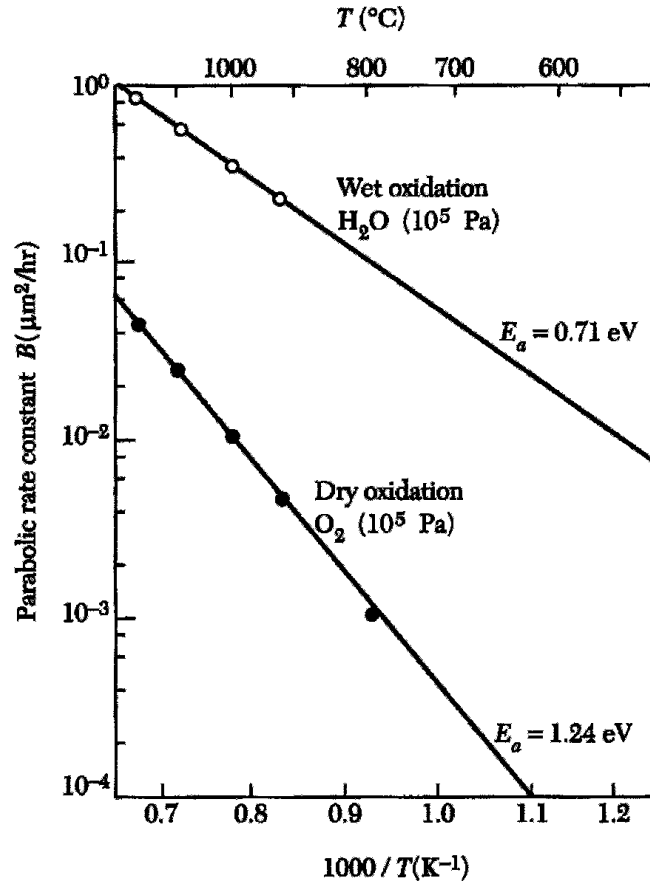


Figure 2.6. Parabolic rate constant versus temperature [1].

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 a rapid early growth that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model given by Eq. (2.11) is not valid for dry oxidation with an oxide thickness ≤ 20 nm. For ultra-large-scale integration, the ability to grow thin (5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important.

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 will be reduced due to 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. (2.11) and obtain

$$x^2 - d_0^2 = Bt \quad (2.14)$$

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.

2.1.1.3. Oxide Quality

Oxides used for masking are usually grown by wet oxidation. A typical growth cycle consists of a dry–wet–dry sequence. 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 (see Section 2.2) 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, shown in Figure 2.7, are interface-trapped charge, fixed-oxide charge, oxide-trapped charge, and mobile ionic charge. Interface-trapped charges (Q_{it}) are due to the SiO_2 –Si interface properties and 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 fixed charge (Q_f) is located within approximately 3 nm of the SiO_2 –Si interface. Generally, Q_f is positive and depends on oxidation and annealing conditions, as well as on the orientation of the silicon substrate. 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. 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

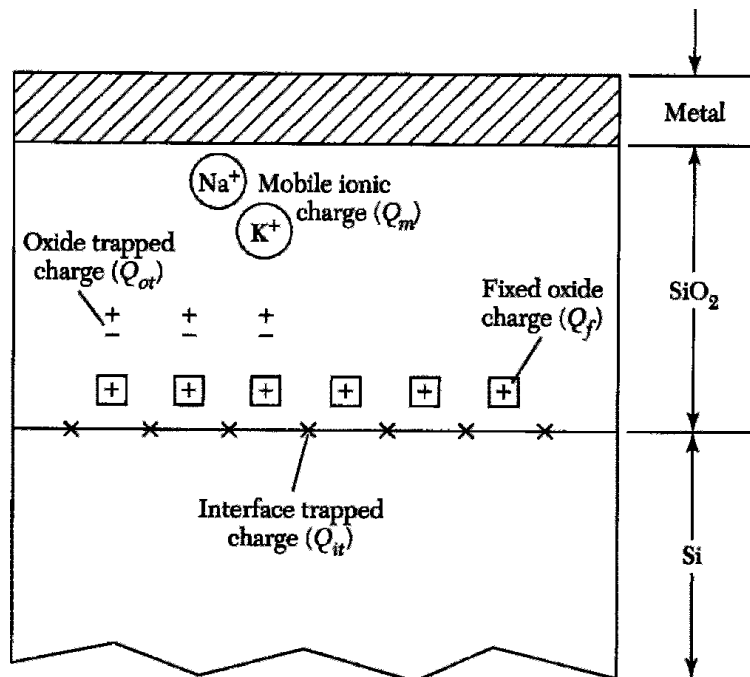


Figure 2.7. Description of charges associated with thermal oxides [1].

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.

2.1.2. 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. 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 of the device. Pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer (see Section 2.1.4).

Photolithography requires a clean processing room. The need for a cleanroom arises because dust particles in the air can settle on semiconductor wafers or lithographic masks and cause defects that result in circuit failure. For example, a dust particle on a semiconductor surface can disrupt the growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into a gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in photolithography. 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 2.8 shows three dust particles on a photomask. Particle 1 may result in the formation of a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause

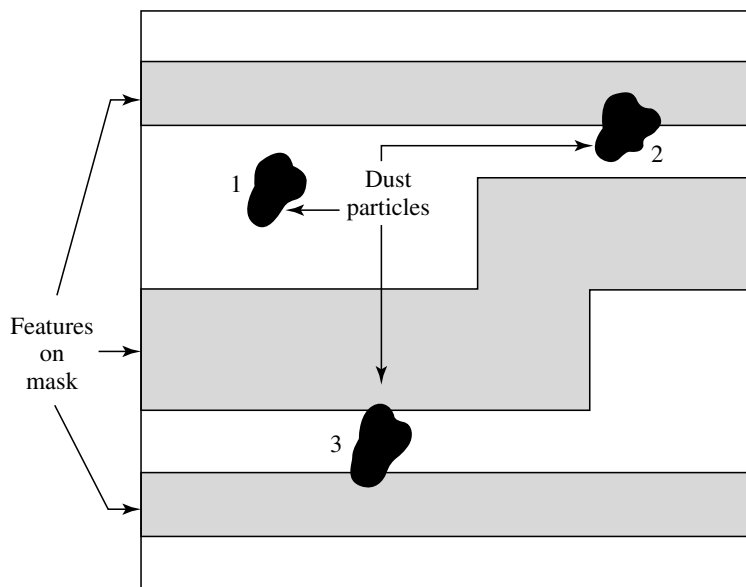


Figure 2.8. Various ways in which particles can interfere with photomask patterns [1].

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 cleanroom, the total number of dust particles per unit volume must be tightly controlled along with the temperature and humidity. There are two systems to define the classes of cleanroom. For the English system, the numerical designation of the class is taken from the maximum allowable number of particles $0.5\text{ }\mu\text{m}$ and larger per cubic foot of air. For the metric system, the class is taken from the logarithm (base 10) of the maximum allowable number of particles $0.5\text{ }\mu\text{m}$ and larger, per cubic meter. For example, a class 100 cleanroom (English system) has a dust count of 100 particles/ft³ with particle diameters of $0.5\text{ }\mu\text{m}$ and larger, whereas a class M 3.5 cleanroom (metric system) has a dust count of $10^{3.5}$ or about 3500 particles/m³ with particle diameters of $0.5\text{ }\mu\text{m}$ or larger.

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

2.1.2.1. Exposure Tools

The pattern transfer process is accomplished by using a lithographic exposure tool. The performance of an exposure tool is determined by 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 unit time for a given mask level.

There are two primary optical exposure methods: shadow printing and projection printing. 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 2.9a shows a basic setup for contact printing where 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 the resist and mask provides a resolution of $\sim 1\text{ }\mu\text{m}$. However, contact printing suffers from one major drawback—a dust particle on the wafer can be embedded into the mask when the mask makes contact with the wafer. The embedded 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 2.9b shows the basic setup, which is similar to contact printing except that there is a small gap ($10\text{--}50\text{ }\mu\text{m}$) between the wafer and the mask during exposure. The small gap, however, 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, resolution is degraded to the $2\text{--}5\text{-}\mu\text{m}$ range.

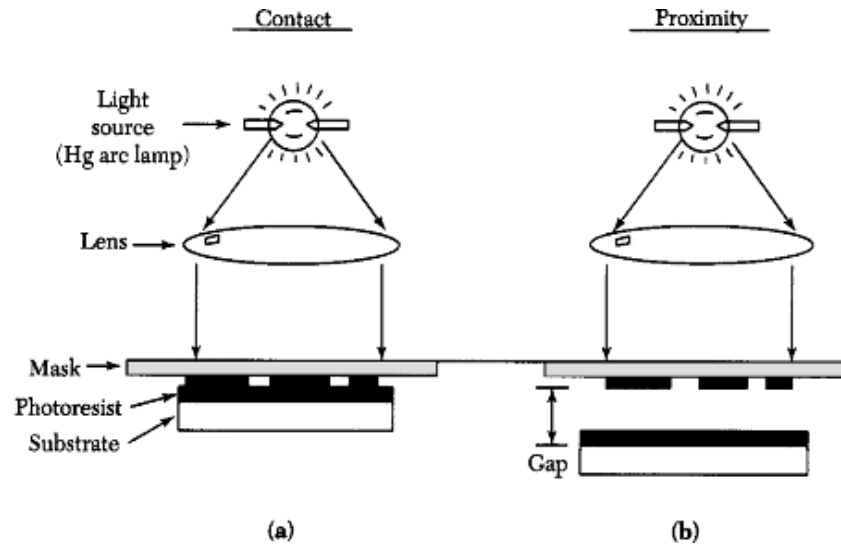


Figure 2.9. Optical shadow printing techniques: (a) contact printing; (b) proximity printing [1].

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

$$CD \cong \sqrt{\lambda g} \quad (2.15)$$

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 $4.5 \mu\text{m}$. If we reduce λ to $0.25 \mu\text{m}$ (a wavelength range of $0.2\text{--}0.3 \mu\text{m}$ is in the deep-UV spectral region) and g to $15 \mu\text{m}$, the CD becomes $2 \mu\text{m}$. 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 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 2.10a shows a 1:1 wafer scan projection system. A narrow, arc-shaped image field $\sim 1 \text{ mm}$ 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.

The small image field can also be stepped over the surface of the wafer by two-dimensional translations of the wafer only, whereas 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 2.10b and 2.10c show the partitioning of the wafer image by *step-and-repeat projection* with a ratio of 1:1 or at a demagnification ratio $M:1$ (e.g., 10:1 for a 10 times reduction on the wafer), respectively. The 1:1 optical systems are easier to design and fabricate than a 10:1 or a 5:1 reduction system, 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.

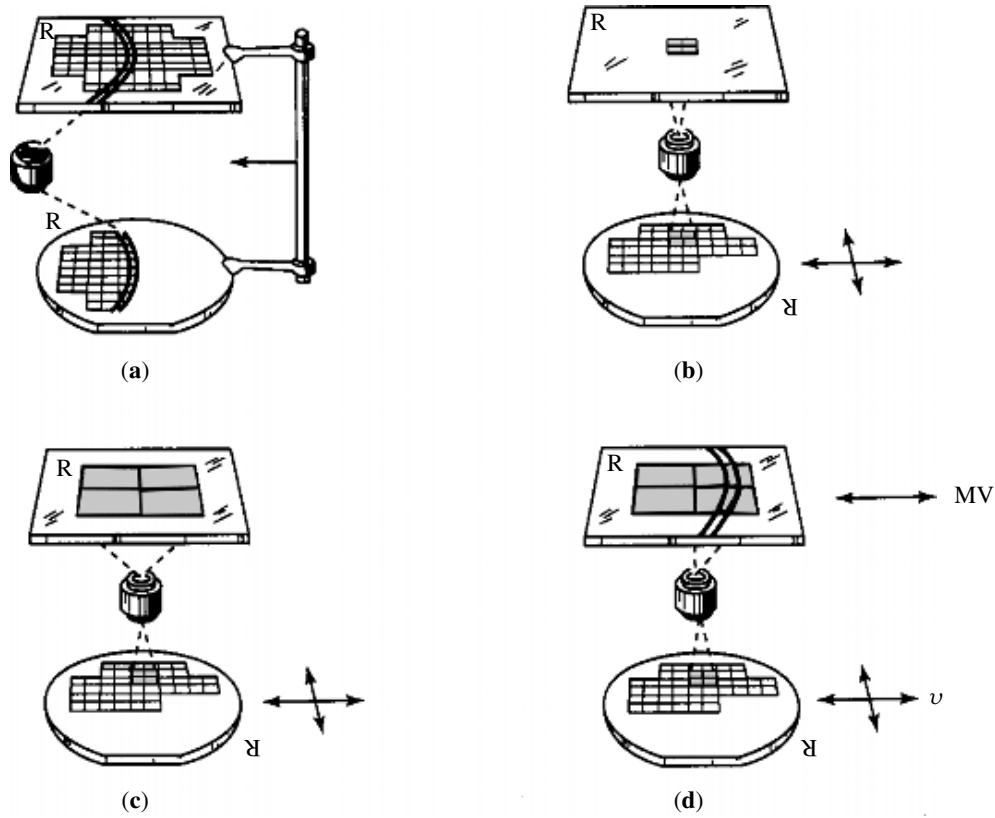


Figure 2.10. Image partitioning techniques for projection printing: (a) annual field wafer scan; (b) 1 : 1 step-and-repeat; (c) $M : 1$ step-and-repeat; and (d) $M : 1$ step-and-scan [1].

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 ICs. When the chip size exceeds the field size of the lens, further partitioning of the image on the reticle is necessary. In Figure 2.10d, the image field on the reticle can be a narrow, arc shape for $M : 1$ step-and-scan projection lithography. For the step-and-scan system, we have two-dimensional translations of the wafer with speed v , and one-dimensional translation of the mask with M times that of the wafer speed.

The resolution of a projection system is given by

$$l_m = k_1 \frac{\lambda}{\text{NA}} \quad (2.16)$$

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

$$\text{NA} = \bar{n} \sin \theta \quad (2.17)$$

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

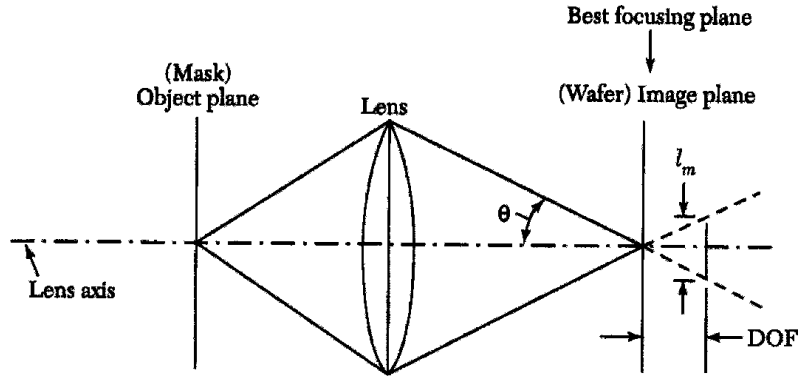


Figure 2.11. Illustration of DoF [1].

wafer, as shown in Figure 2.11. Also shown in the figure is the depth of focus (DoF), which can be expressed as

$$\text{DoF} = \frac{\pm l_m/2}{\tan \theta} \approx \frac{\pm l_m/2}{\sin \theta} = k_2 \frac{\lambda}{(\text{NA})^2} \quad (2.18)$$

where k_2 is another process-dependent factor.

Equation (2.16) indicates that resolution can be improved (i.e., smaller l_m) by either reducing the wavelength, increasing NA, or both. However, Eq. (2.18) 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.

2.1.2.2. Masks

Masks used for semiconductor manufacturing are usually reduction reticles. The first step in maskmaking 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 drive a pattern generator, which is an electron-beam lithographic system (see Section 2.1.2.5) 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 into the underlying chromium layer for the finished mask. The patterns on a mask represent one level of an IC design. The composite layout is broken into mask levels that correspond to the manufacturing process sequence, such as the isolation region on one level, the gate region on another, and so on. Typically, 15–20 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. This 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

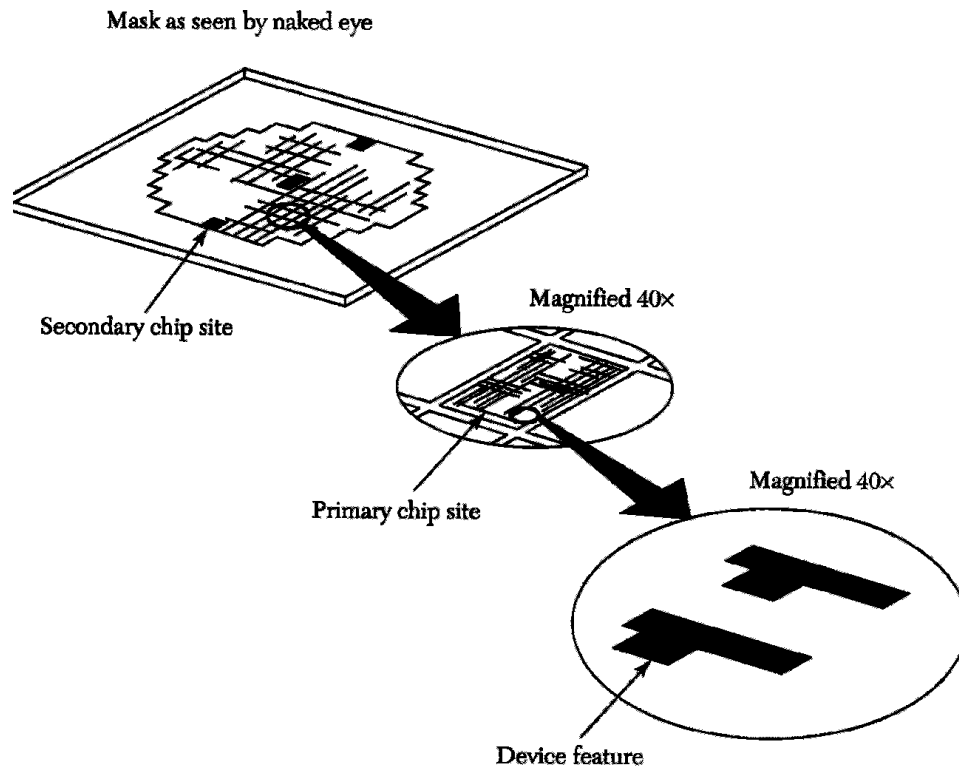


Figure 2.12. A typical IC photomask [1].

wavelengths, and its mechanical strength. Figure 2.12 shows a 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 the 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 5). Inspection and cleaning of masks are important to achieve high yields on large chips. An ultraclean processing area is mandatory for photolithographic processing.

2.1.2.3. Photoresist

Photoresist is a radiation-sensitive compound that can be classified as positive or negative, depending on how they respond to radiation. For *positive resists*, the exposed regions become more soluble and thus more easily removed in the development process. The result is that the patterns formed in the positive resist are the same as those on the mask. 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.

With *negative resists*, exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns. 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 crosslinking reaction. This reaction causes crosslinking of the polymer molecules. The crosslinked 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 2.13a shows a typical exposure response curve and image cross section for a positive resist. The response curve describes the percentage of resist remaining after exposure and development versus the exposure energy. 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

$$\gamma \equiv \left[\ln \left(\frac{E_T}{E_1} \right) \right]^{-1} \quad (2.19)$$

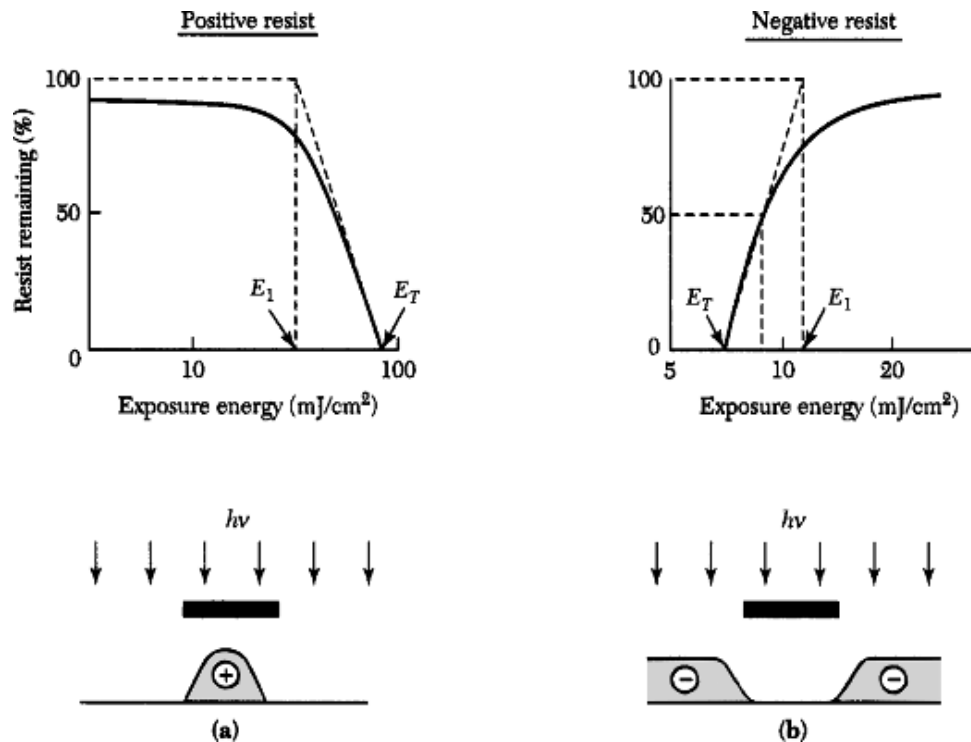


Figure 2.13. Exposure response curve and cross section of resist image after development for (a) positive photoresist and (b) negative photoresist [1].

where E_1 is the energy obtained by drawing the tangent at E_T to reach 100% resist thickness, as shown in Figure 2.13a. 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 2.13a 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 2.13b 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. (2.19), except that E_1 and E_T are interchanged. The image cross section for the negative resist (Figure 2.13b) is also influenced by the diffraction effect.

2.1.2.4. Pattern Transfer

Figure 2.14 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. The wafer is placed in a cleanroom, 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, adhesion promoter is then applied. The most common adhesion promoter for silicon ICs is hexamethylene–disiloxane (HMDS). After the application of this adhesion layer, the wafer is held on a vacuum spindle, and liquidous resist is applied to the center of wafer. The wafer is then rapidly accelerated up to a constant rotational speed, which is maintained for about 30 s. Spin speed is generally in the range of 1000–10,000 rpm [revolutions per minute (r/min)] to coat a uniform film about $0.5\text{--}1\ \mu\text{m}$ thick, as shown in Figure 2.14a. The thickness of photoresist is correlated with its viscosity.

After spinning, the wafer is “soft-baked” (typically at $90\text{--}120^\circ\text{C}$ for 60–120 s) to remove solvent from the photoresist 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 ultraviolet light, as shown in Figure 2.14b. If a positive photoresist is used, the exposed resist is dissolved in the developer, as shown on the left side of Figure 2.14c. Photoresist development is usually done by flooding the wafer with the developer solution. The wafer is then rinsed and dried. After development, “postbaking” at approximately $100\text{--}180^\circ\text{C}$ 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 2.14d. Finally, the resist is stripped (using solvents or plasma oxidation), leaving behind an insulator image that is the same as the opaque image on the mask (left side of Figure 2.14e). For negative photoresist,

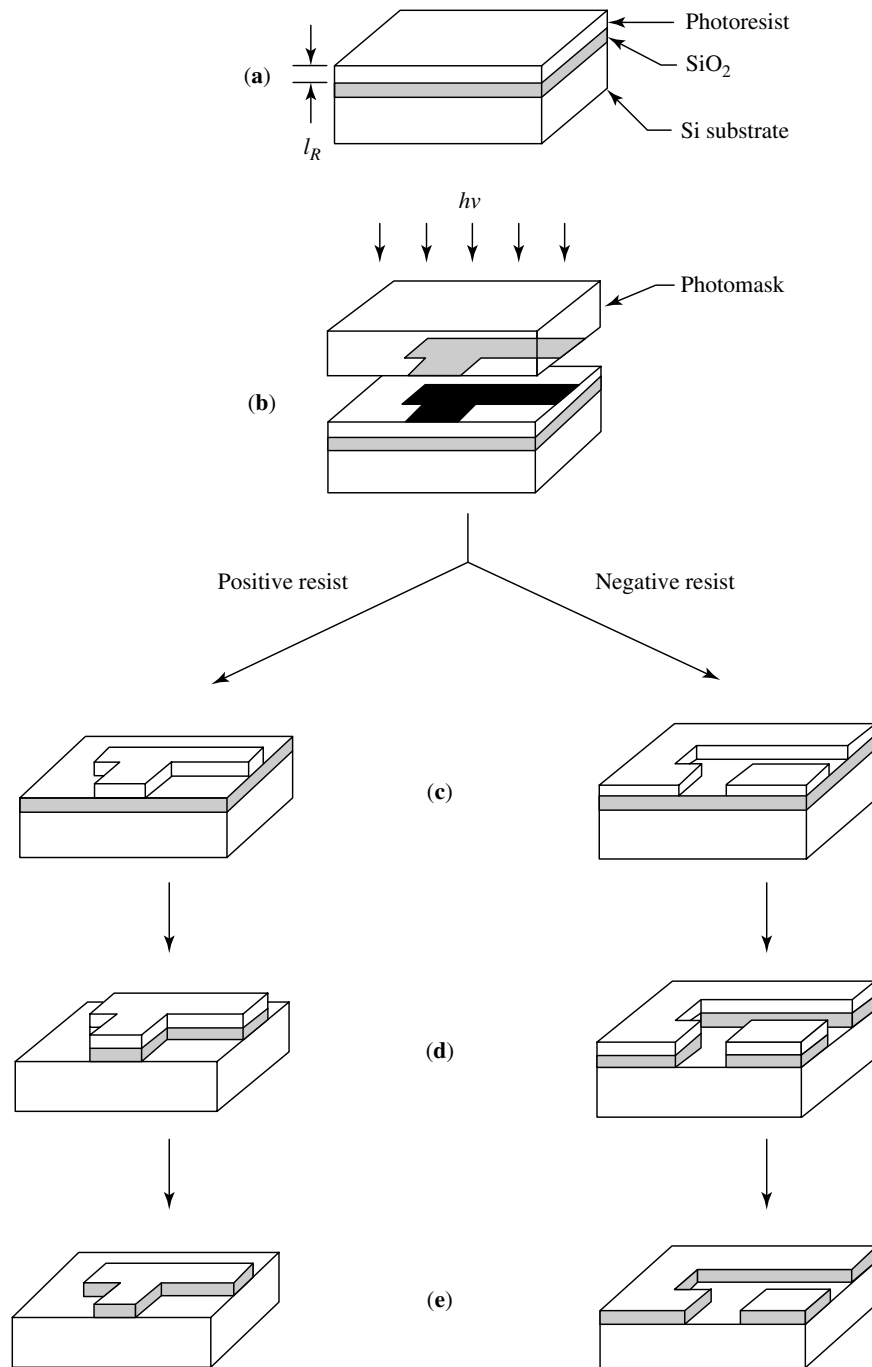


Figure 2.14. Photolithographic pattern transfer process: (a) photoresist application; (b) exposure; (c) development; (d) etching; (e) resist stripping [1].

the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Figure 2.14e) 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* (Section 2.1.3) 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.

2.1.2.5. E-Beam Lithography

Optical lithography is so widely used because it has high throughput, good resolution, low cost, and ease of operation. However, due to deep-submicrometer 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 lifespan, the complexity of mask production and mask inspection cannot be easily resolved. In addition, the cost of the masks is very high. Therefore, we need to find alternatives to optical lithography to process deep-submicrometer or nanometer ICs.

Electron-beam (or *e-beam*) lithography is used primarily to produce photomasks. Relatively few tools are dedicated to direct exposure of the resist by a focused electron beam without a mask. Figure 2.15 shows a schematic of an e-beam lithography system. 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 10–25 nm in diameter. Beam blanking plates that turn the electron beam on and off, and beam deflection coils are computer-controlled and operated at MHz or higher

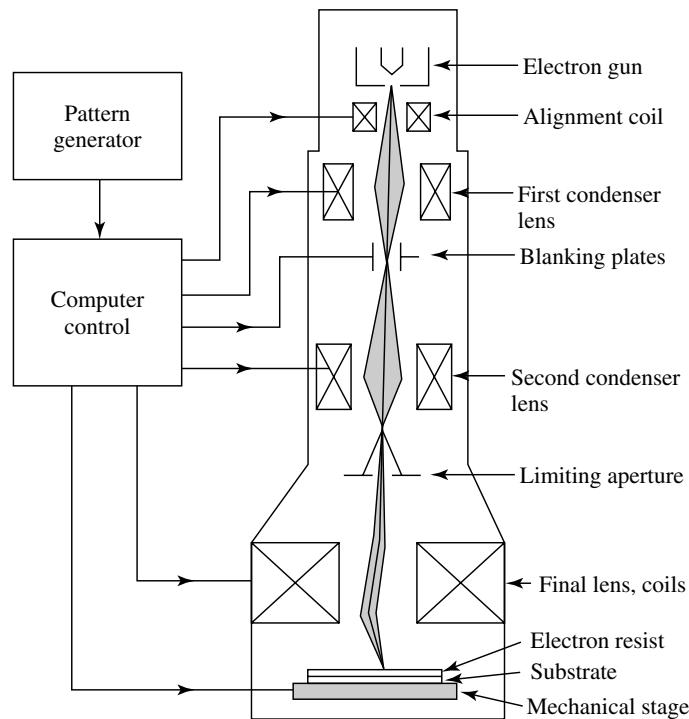


Figure 2.15. E-beam lithography system [1].

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 submicrometer resist geometries, highly automated and precisely controlled operation, depth of focus greater than that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron-beam lithographic machines have low throughput—approximately 10 wafers per hour at less than $0.25\ \mu\text{m}$ 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 two ways to scan the focused electron beam: raster scan and vector scan. In a *raster scan* system, resist patterns are written by a beam that moves through a regular mode, vertically oriented, as shown in Figure 2.16a. 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

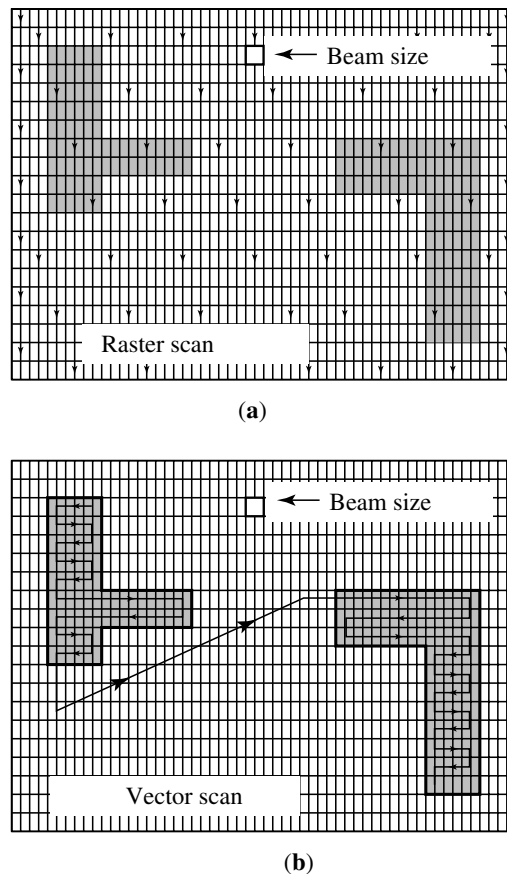


Figure 2.16. (a) Raster scan and (b) vector scan systems [1].

subdivided into individual addresses, and a given pattern must have a minimum incremental interval that is evenly divisible by the beam address size. In the *vector scan* system (Figure 2.16b), 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, which saves time.

Electron resists are polymers. The behavior of an e-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 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 μm or better. For a negative electron resist, the irradiation causes radiation-induced polymer linking. The crosslinking 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. Polyglycidylmethacrylate–coethylacrylate (COP) is a common negative electron resist. COP, like most negative photoresists, also swells during development, so the resolution is limited to about 1 μm .

While resolution is limited by diffraction of light in optical lithography, in e-beam lithography, the resolution is not impacted by diffraction (because the wavelengths associated with electrons of a few keV and higher energies are less than 0.1 nm), but by electron scattering. When electrons penetrate the resist and underlying substrate, they undergo collisions. These collisions lead to energy losses and path changes. The incident electrons spread out as they travel until either all of their energy is lost or they leave the material because of backscattering. Because of backscattering, electrons can irradiate regions several micrometers away from the center of the exposure beam. Since the dose of a resist is the sum of the irradiations from all surrounding areas, the 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.

2.1.2.6. X-Ray Lithography

X-ray lithography (XRL) is a potential candidate to succeed optical lithography for the fabrication of integrated circuits with feature sizes less than 100 nm. XRL

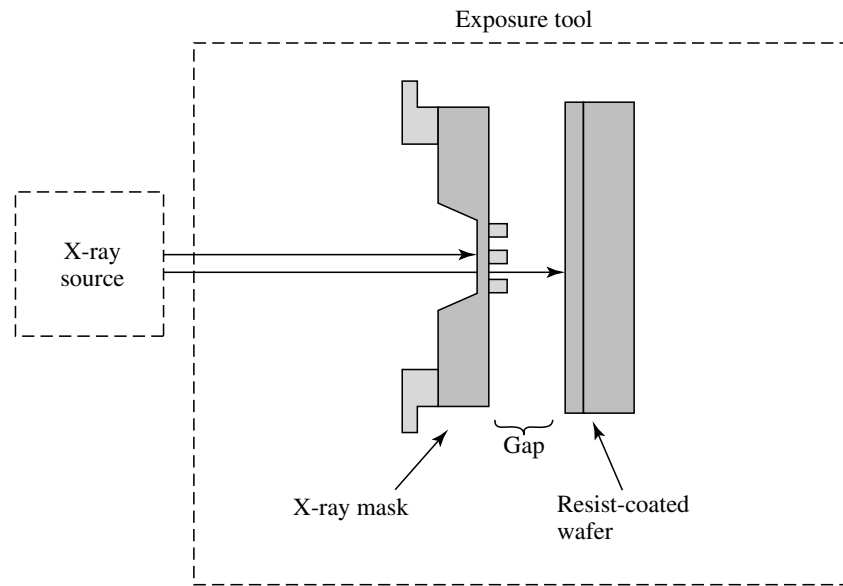


Figure 2.17. Schematic of x-ray lithography system.

uses a shadow printing method similar to optical proximity printing. Figure 2.17 shows a schematic of an XRL system. The X-ray wavelength is about 1 nm, and the printing is through a $1\times$ mask in close proximity (10–40 μm) to the wafer. Since X-ray absorption depends on the atomic number of the material and most materials have low transparency at $\lambda \cong 1$ nm, the mask substrate must be a thin membrane (1–2 μm thick) made of low-atomic-number material, such as silicon carbide or silicon. The pattern itself is defined in a thin (~ 0.5 μm), 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 an optical 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 a vacuum, which is separated from the helium by a thin vacuum window (usually of beryllium). The mask substrate will absorb 25–35% of the incident flux and must therefore be cooled. An X-ray resist 1 μm thick will absorb about 10% of the incident flux, and there are no reflections from the substrate to create standing waves, so antireflection coatings are unnecessary.

Electron-beam resists can be used 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 wavelength different from that of 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.

2.1.3. Etching

As discussed in Section 2.1.2, photolithography 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 of the device. Pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer.

2.1.3.1. Wet Chemical Etching

Wet chemical etching is used extensively in semiconductor processing. Prior to thermal oxidation (Section 2.1.1) or epitaxial growth (Section 2.1.5), 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 2.18; 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 will 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

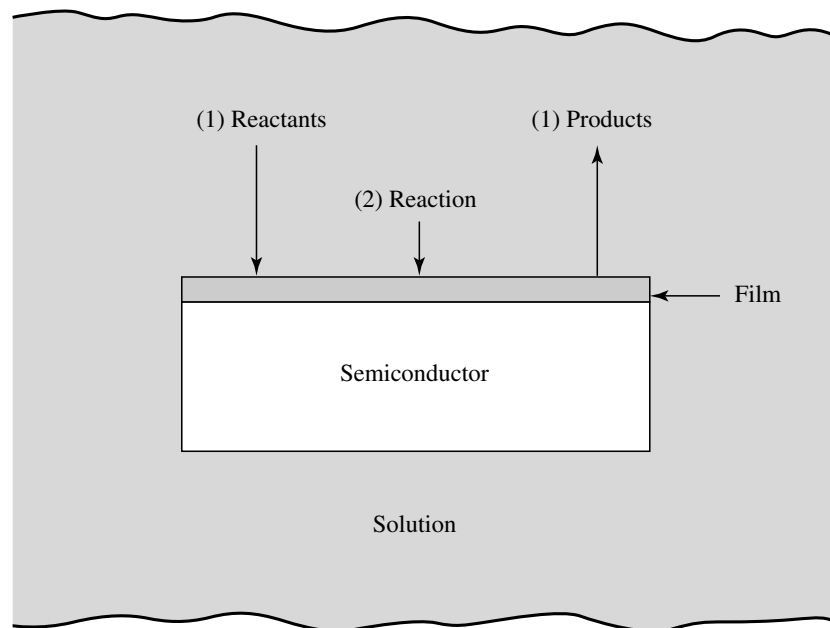


Figure 2.18. Basic mechanisms in wet chemical etching [1].

because it greatly increases the etch rate and uniformity by constantly supplying fresh etchant to the wafer surface.

In 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

$$\text{Etch rate uniformity (\%)} = \frac{(\text{max. etch rate} - \text{min. etch rate})}{\text{max. etch rate} + \text{min. etch rate}} \times 100\% \quad (2.20)$$

2.1.3.2. Dry Etching

In pattern transfer operations, a resist pattern is defined by a photolithographic process to serve as a mask for etching of its underlying layer (Figure 2.19a). 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 etchant, the etch rate

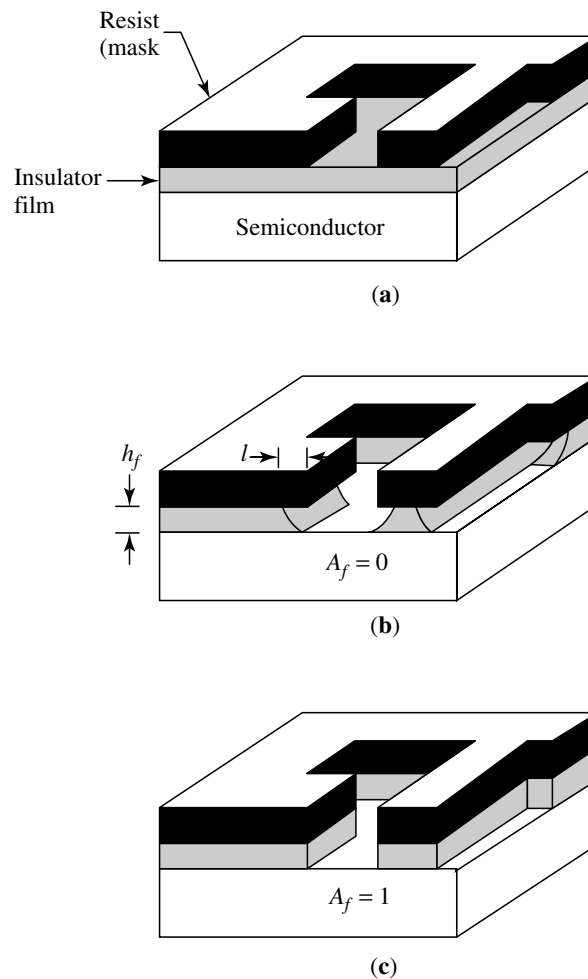


Figure 2.19. Comparison between wet and dry etching: (a) resist pattern; (b) isotropic etching; (c) anisotropic etching [1].

is generally isotropic (i.e., the lateral and vertical etch rates are the same), as illustrated in Figure 2.19b. If h_f 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_f) by

$$A_f \equiv 1 - \frac{l}{h_f} = 1 - \frac{R_l t}{R_v t} = 1 - \frac{R_l}{R_v} \quad (2.21)$$

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 $A_f = 0$.

The major disadvantage of wet 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., $1 \geq A_f > 0$) must be used. In practice, the value of A_f is chosen to be close to unity. Figure 2.19c shows the limiting case where $A_f = 1$, corresponding to $l = 0$ (or $R_l = 0$).

To achieve a high-fidelity transfer of the resist patterns required for ultra-large-scale integration (ULSI) 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.

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

Plasma etching is a process in which a solid film is removed by a chemical reaction with ground-state or excited-state neutral species. The process is often enhanced or induced by energetic ions generated in a gaseous discharge. Plasma etching proceeds in five steps, as illustrated in Figure 2.20: (1) the etchant species is generated in the plasma, (2) the reactant is then transported by diffusion through a stagnant gas layer to the surface, (3) the reactant is adsorbed on the surface, (4) a chemical reaction (along with physical effects such as ion bombardment) follows to form volatile compounds, and (5) the compounds are desorbed from the surface, diffused into the bulk gas, and pumped out by the vacuum system.

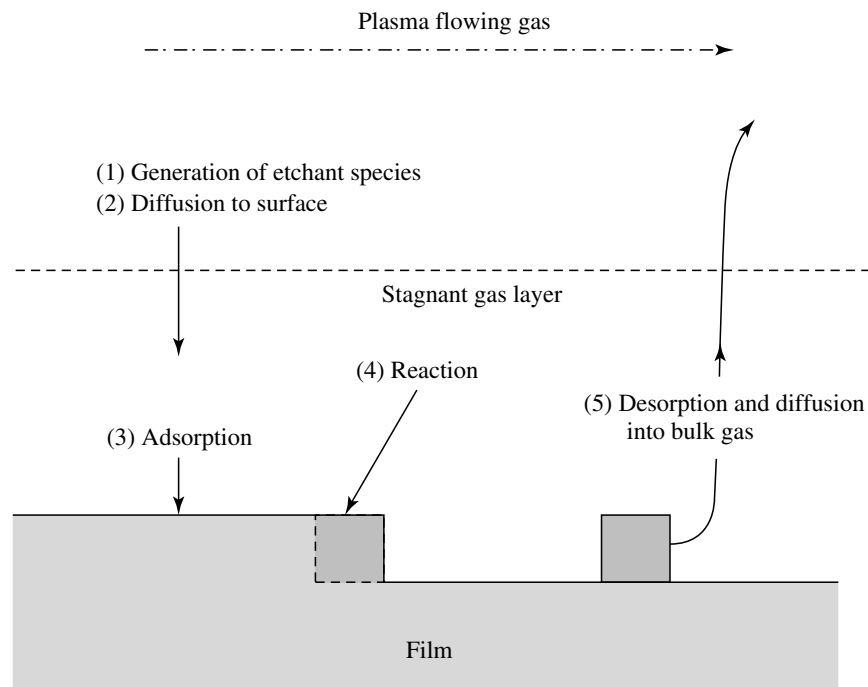


Figure 2.20. Basic steps in dry etching [1].

Plasma etching is based on the generation of plasma in a gas at low pressure. Two basic 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) produces low ion bombardment-induced damage but yields isotropic profiles. Physical etching can yield anisotropic profiles, but it is associated with low etch selectivity and severe bombardment-induced damage. Combinations of chemical and physical etching give anisotropic etch profiles, reasonably good selectivity, and moderate bombardment-induced damage. An example is reactive-ion etching (RIE), which uses a physical method to assist chemical etching or creates reactive ions to participate in chemical etching.

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 *endpoint detector* (see Chapter 3). 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.

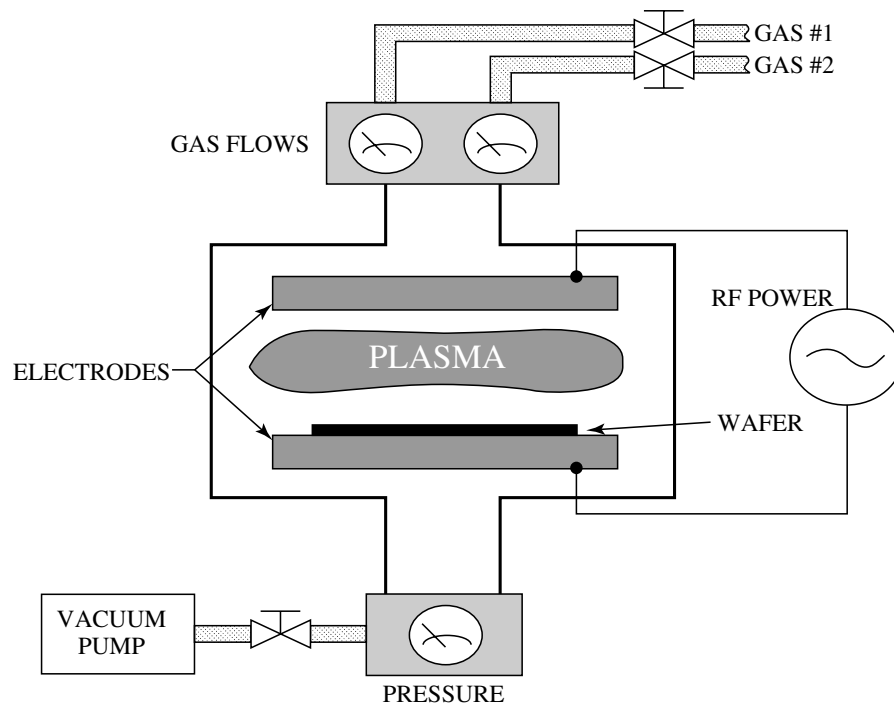


Figure 2.21. Typical reactive-ion etching system.

RIE has been extensively used in the microelectronic industry. In a parallel-plate diode system (Figure 2.21), a radio frequency (RF), 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 by choosing the proper etch chemistry.

2.1.4. Doping

Impurity doping is the introduction of controlled amounts of impurities into semiconductors to change their electrical properties. 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. For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS), whereas ion implantation is used to form a shallow junction (e.g., a source–drain junction of a MOSFET).

Until the early 1970s, impurity doping was performed by diffusion at elevated temperatures, as shown in Figure 2.22a. 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

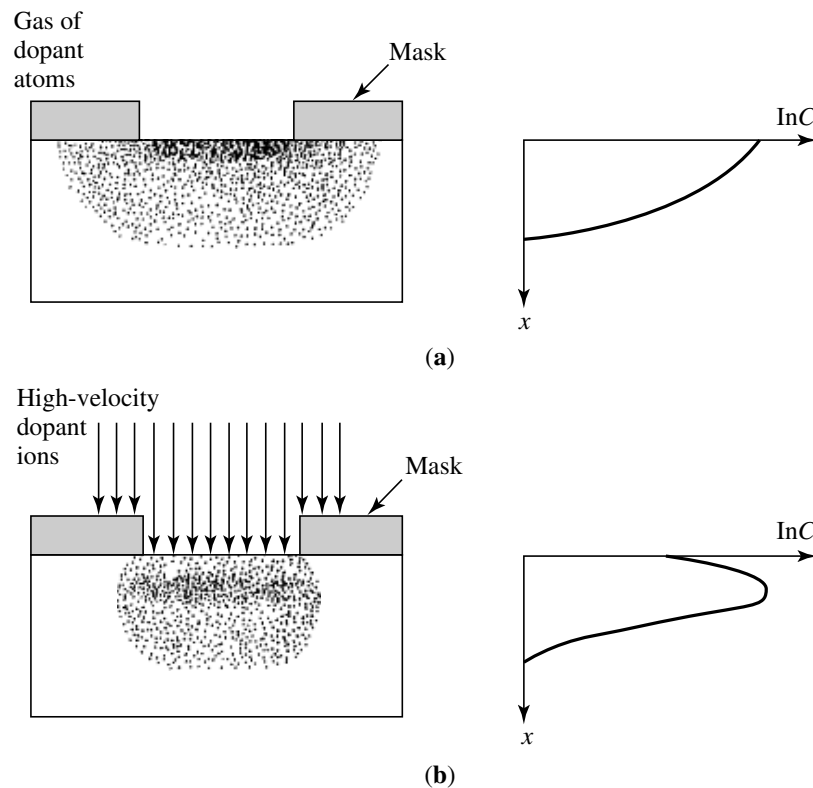


Figure 2.22. (a) Diffusion and (b) ion implantation techniques for impurity doping [1].

monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and the diffusion time. Since the 1970s, doping operations have been performed chiefly by ion implantation, as shown in Figure 2.22b. 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.

2.1.4.1. Diffusion

Diffusion of impurities is accomplished 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 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 dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As₂O₃ for arsenic, and P₂O₅ for phosphorus), liquid sources (BBr₃, AsCl₃, and POCl₃), and gaseous sources (B₂H₆, AsH₃, and PH₃). 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 2.23. This arrangement is similar to that used for thermal

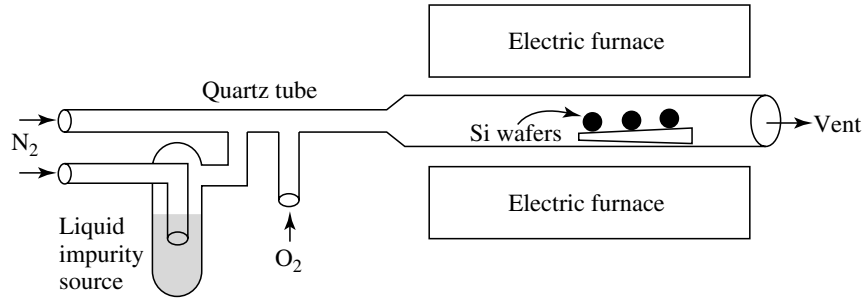
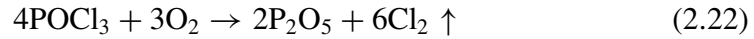
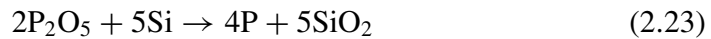


Figure 2.23. Schematic of a diffusion system [1].

oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is



The P_2O_5 is then reduced to phosphorus by silicon according to



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

Diffusion in a semiconductor is the atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. The diffusion process is similar to that of charge carriers (electrons and holes). Let a flux F be defined as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. Then

$$F = -D \frac{\partial C}{\partial x} \quad (2.24)$$

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 dC/dx . 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 Eq. (2.24) is substituted into the one-dimensional continuity equation under the condition that no materials are formed or consumed in the host semiconductor, the result is

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right) \quad (2.25)$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered to be independent of doping concentration, and Eq. (2.25) becomes

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} \quad (2.26)$$

Equation (2.26) is often referred to as *Fick's diffusion equation* or *Fick's law*. Note that the diffusion coefficient varies with temperature. Over the temperature

ranges commonly used in semiconductor manufacturing, the diffusion coefficient can be expressed as

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) \quad (2.27)$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature and E_a is the activation energy in eV. The values of E_a are typically found to be between 0.5 and 2 eV for interstitial diffusion. For vacancy diffusion, E_a is much larger, usually between 3 and 5 eV.

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. There are two important cases to consider: (1) constant-surface-concentration diffusion and (2) 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.

For case 1, the initial condition at $t = 0$ is

$$C(x, 0) = 0 \quad (2.28)$$

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

$$C(0, t) = C_s \quad (2.29a)$$

$$C(\infty, t) = 0 \quad (2.29b)$$

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

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

$$C(x, t) = C_s \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \quad (2.30)$$

where erfc is the complementary error function and $\sqrt{Dt}$ is the diffusion length. The diffusion profile for the constant surface concentration condition is shown in Figure 2.24a, where, 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 are plotted. Note that as the time progresses, the dopant penetrates deeper into the semiconductor.

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

$$Q(t) = \int_0^\infty C(x, t) dx \quad (2.31)$$

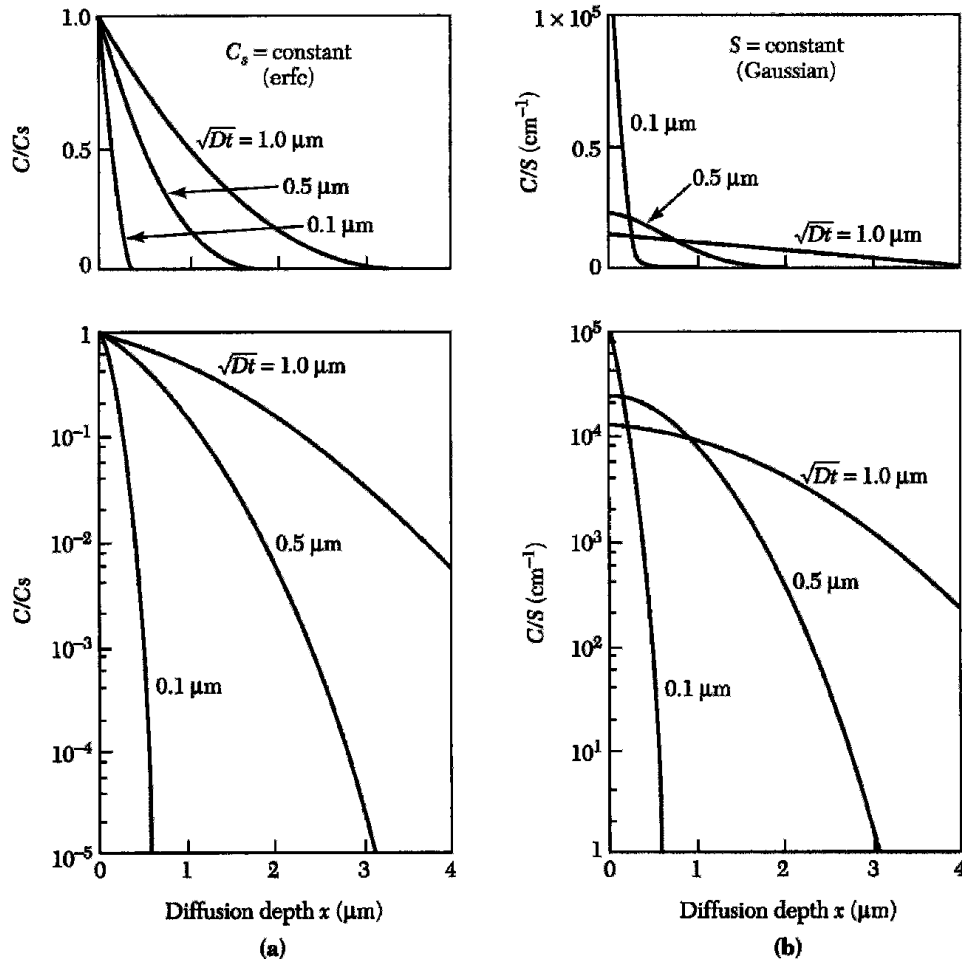


Figure 2.24. Diffusion profiles: (a) normalized erfc versus distance for successive diffusion times; (b) normalized Gaussian function versus distance [1].

Substituting Eq. (2.30) into Eq. (2.31) yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \cong 1.13 C_s \sqrt{Dt} \quad (2.32)$$

The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Figure 2.24a. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \cong C_s \sqrt{Dt}$, which is close to the exact result obtained from Eq. (2.32).

Now consider constant total dopant diffusion. 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. (2.28). The boundary conditions are

$$\int_0^{\infty} C(x, t) dx = S \quad (2.33a)$$

$$C(\infty, t) = 0 \quad (2.33b)$$

where S is the total amount of dopant per unit area. The solution of the diffusion equation that satisfies these conditions is

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \quad (2.34)$$

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 the case, since the surface concentration is given by Eq. (2.34) with $x = 0$:

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \quad (2.35)$$

Figure 2.24b shows the dopant profile for a Gaussian distribution where the normalized concentration (C/S) as a function of the distance for three increasing diffusion lengths is plotted. Note the reduction of the surface concentration as the diffusion time increases.

In IC processing, a two-step diffusion process is commonly used, in which a *predeposition* diffused layer is first formed under the constant-surface-concentration condition (case 1, above). This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under constant total dopant conditions. 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 as negligibly small compared with that of the final profile that results from the drive-in step.

2.1.4.2. Ion Implantation

As discussed above, 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, which is shown in Figure 2.22b. In this process the energetic 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 energy.

Implantation energies are typically 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/cm² for threshold voltage adjustment in MOSFETs to 10^{18} ions/cm² for the formation of buried insulating layer. Note that the dose is expressed as the number of ions implanted into 1 cm² 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 2.25 shows schematically a medium-energy ion implantor. The ion source has a heated filament to break up source gas, such as BF₃ or AsH₃, into

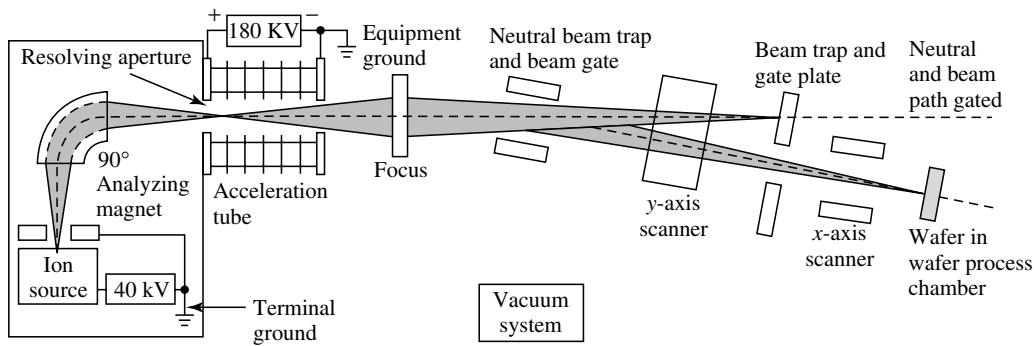


Figure 2.25. Schematic of ion implanter [1].

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 : 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 their energies through collision 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 principal side effect is the disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove these damages.

The total distance that an ion travels in coming to rest is called its *range* (R) and is illustrated in Figure 2.26a. 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* ($\sigma_{\perp}$).

Figure 2.26b shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x - R_p)^2}{2\sigma_p^2}\right] \quad (2.36)$$

where S is the ion dose per unit area. This equation is similar to Eq. (2.34) for constant total dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$

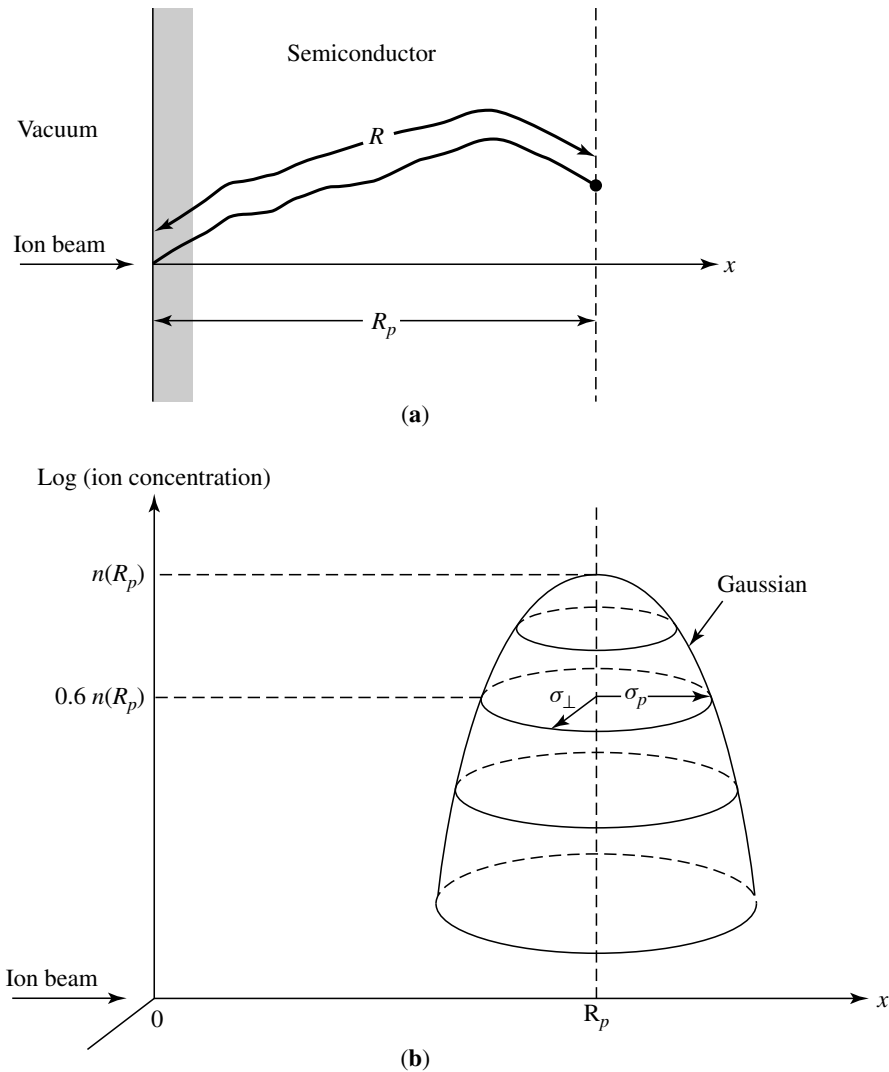


Figure 2.26. (a) Ion range and projected range; (b) two-dimensional distribution of implanted ions [1].

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 $\pm 4.8\sigma_p$.

2.1.5. Deposition

Many different types of thin films are used to manufacture integrated circuits, including thermal oxides, dielectric layers, epitaxial layers, polycrystalline silicon, and metal films. This section addresses two of the various techniques for depositing such films: physical vapor deposition and chemical vapor deposition.

2.1.5.1. Physical Vapor Deposition

The most common methods of physical vapor deposition (PVD) of metals are evaporation, electron-beam evaporation, plasma spray deposition, and sputtering. Metals and metal compounds can be deposited by PVD. Evaporation occurs when a source material is heated above its melting point in an evacuated chamber. The evaporated atoms then travel at high velocity in straight-line trajectories. The source can be melted by resistance heating, by radio frequency (RF) heating, or with a focused electron beam (or e-beam). Evaporation and e-beam evaporation were used extensively in earlier generations of integrated circuits, but they have been replaced by sputtering for modern ICs.

In ion-beam sputtering, a source of ions is accelerated toward the target and impinges on its surface. Figure 2.27a shows a standard sputtering system. The sputtered material deposits on a wafer that is placed facing the target. The ion current and energy can be independently adjusted. Since the target and wafer are placed in a chamber that has lower pressure, more target material and less contamination are transferred to the wafer.

One method to increase the deposition rate in sputtering is to use a third electrode that provides more electrons for ionization. Another method is to use a magnetic field, such as in *electron cyclotron resonance* (ECR) systems, to capture and spiral electrons, increasing their ionizing efficiency in the vicinity of the sputtering target. This technique, referred to as *magnetron sputtering*, has found widespread applications for the deposition of aluminum and its alloys at a rate that can approach 1 $\mu\text{m}/\text{min}$.

Long-throw sputtering is another technique used to control the angular distribution. Figure 2.27b shows a long-throw sputtering system. In standard sputtering configurations, there are two primary reasons for a wide angular distribution of incident flux at the surface: (1) the use of a small target to substrate separation d_{ts} and (2) scattering of the flux by the working gas as the flux travels from the target to the substrate. These two factors are linked because a small d_{ts} is needed to achieve good throughput, uniformity, and film properties when there is substantial gas scattering. A solution to this problem is to sputter at very low

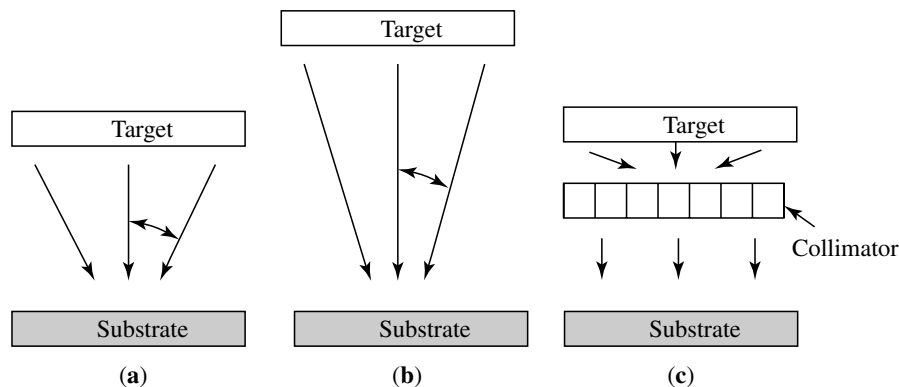


Figure 2.27. (a) Standard sputtering; (b) long-throw sputtering; (c) sputtering through a collimator [1].

pressures, a capability that has been developed using a variety of systems, which can sustain the magnetron plasma under more rarefied conditions. These systems allow for sputtering at working pressures of less than 0.1 Pa. At these pressures, gas scattering is less important, and the target–substrate distance can be greatly increased. From a simple geometric argument, this allows the angular distribution to be greatly narrowed, which permits more deposition at the bottom of high-aspect features such as contact holes.

Contact holes with large aspect ratio are difficult to fill with material, mainly because scattering events cause the top opening of the hole to seal before appreciable material has deposited on its floor. This problem can be overcome by collimating the sputtered atoms by placing an array of collimating tubes just above the wafer to restrict the depositing flux to normal $\pm 5^\circ$. Sputtering with a collimator is shown in Figure 2.27c. Atoms whose trajectory is more than 5° from normal are deposited on the inner surface of the collimators.

2.1.5.2. Chemical Vapor Deposition

Chemical vapor deposition (CVD), also known as *vapor-phase epitaxy* (VPE), is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD). Figure 2.28 shows three common susceptors for epitaxial growth. Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in epitaxial reactors are analogous to crucibles in the

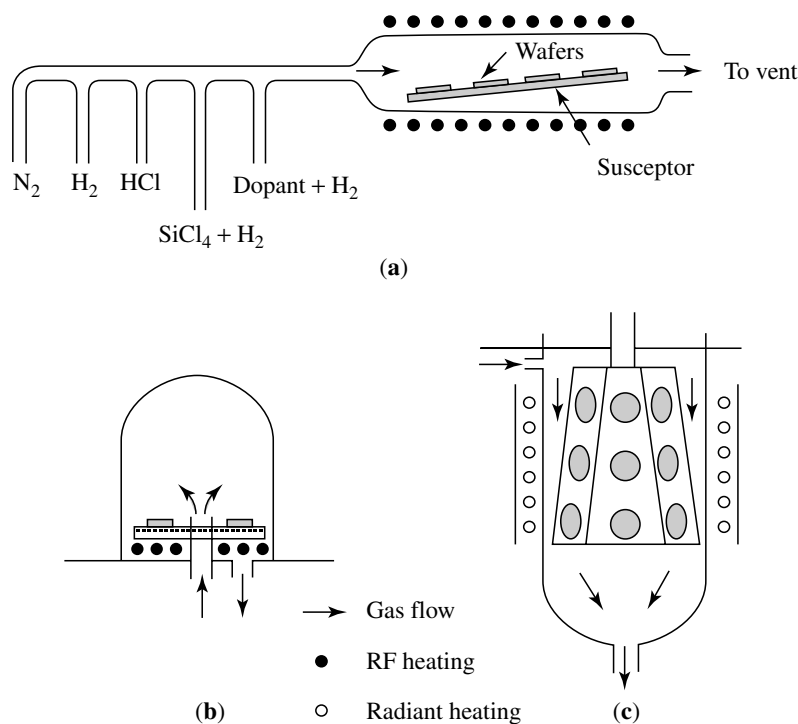


Figure 2.28. Common susceptors for CVD: (a) horizontal; (b) pancake; (c) barrel [1].

crystal growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors, they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (1) the reactants (gases and dopants) are transported to the substrate region; (2) they are transferred to the substrate surface, where they are adsorbed; (3) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer; (4) the gaseous products are desorbed into the main gas stream; and (5) the reaction products are transported out of the reaction chamber.

CVD is attractive for metallization because it offers coatings that are conformal, has good step coverage, and can coat a large number of wafers at a time. The basic CVD setup is the same as that used for deposition of dielectrics and polysilicon (see Figure 1.14). Low-pressure CVD (LPCVD) is capable of producing conformal step coverage over a wide range of topographical profiles, often with lower electrical resistivity than that from PVD. One of the major new applications of CVD metal deposition for integrated circuit production is in the area of refractory metal deposition. For example, tungsten's low electrical resistivity ($5.3 \mu\Omega \cdot \text{cm}$) and refractory nature make it a desirable metal for use in IC fabrication.

2.1.6. Planarization

The development of chemical–mechanical polishing (CMP) has become important for multilevel interconnection technology because it is the only method that allows global planarization (i.e., a flat surface across the whole wafer). It also offers other advantages, including reduced defect density and the avoidance of plasma damage (which would occur in an RIE-based planarization system).

The CMP process consists of moving the sample surface against a pad that carries slurry between the sample surface and the pad. Abrasive particles in the slurry cause mechanical damage on the sample surface, loosening the material for enhanced chemical attack or fracturing off the pieces of surface into a slurry where they dissolve or are swept away. The process is tailored to provide an enhanced material removal rate from high points on surfaces. Mechanical grinding alone may theoretically achieve the desired planarization, but is undesirable because of extensive associated damage to the material surface. There are three main parts of the process: (1) the surface to be polished; (2) the pad, which is the key medium enabling the transfer of mechanical action to the surface being polished; and (3) the slurry, which provides both chemical and mechanical effects. Figure 2.29 shows a typical CMP setup.

2.2. PROCESS INTEGRATION

An integrated circuit is an ensemble of active (e.g., transistors) and passive devices (e.g., resistors, capacitors, and inductors) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern.

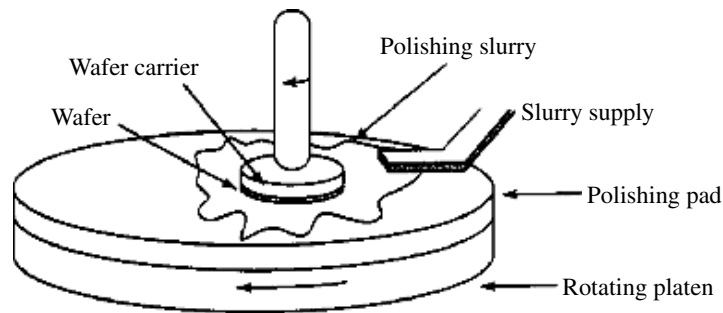


Figure 2.29. CMP schematic [1].

ICs have enormous advantages over discrete devices, including (1) reduction of the interconnection parasitics, (2) full utilization of a semiconductor wafer's area, and (3) drastic reduction in processing cost. In this section, we discuss the manner in which the basic processes described in previous portions of this chapter are combined to fabricate ICs. We consider three major IC technologies associated with two transistor families (viz., bipolar junction transistors and metal–oxide–semiconductor field-effect transistors, or MOSFETs): bipolar, CMOS, and BiCMOS. In addition, we will discuss the packaging of ICs by various techniques.

Figure 2.30 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown

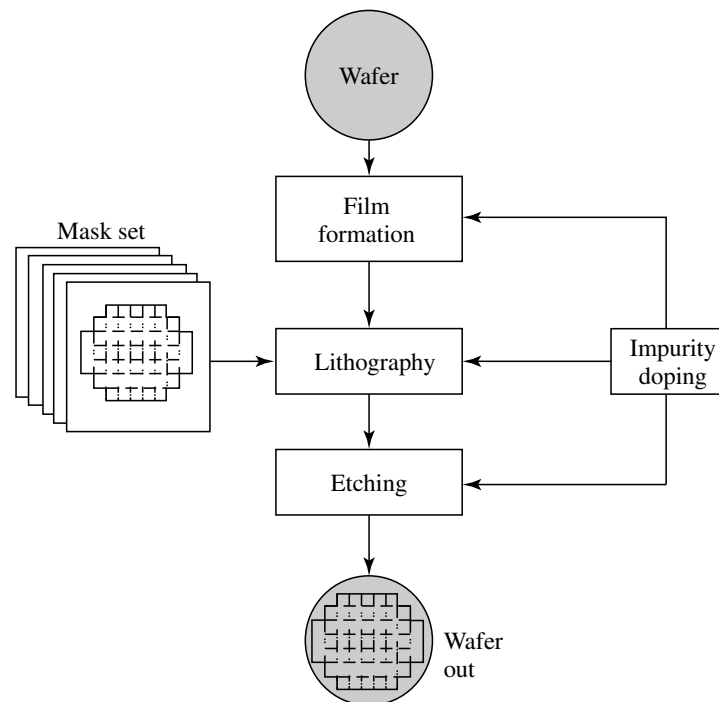


Figure 2.30. Schematic diagram of IC fabrication [1].

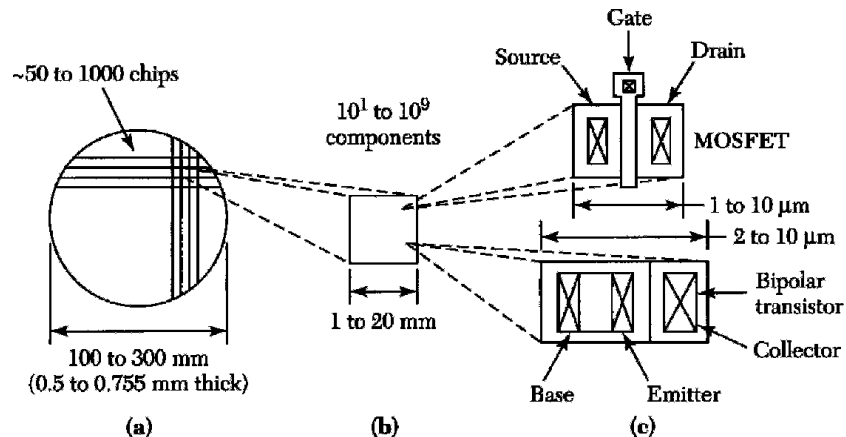


Figure 2.31. (a) Semiconductor wafer; (b) IC chip; (c) MOSFET and bipolar transistor [1].

oxide films (Section 2.1.1), deposited polysilicon, dielectric, and metal films (Section 2.1.5). Film formation is often followed by lithography (Section 2.1.2) or impurity doping (Section 2.1.4). Lithography is generally followed by etching (Section 2.1.3), which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dies), typically between 1 and 20 mm on each side, as shown in Figure 2.31a. The chips are separated by sawing or laser cutting. Figure 2.31b shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Figure 2.31c to give some perspective of the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic applications.

2.2.1. Bipolar Technology

The majority of bipolar transistors used in ICs are of the $n-p-n$ type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with $p-n-p$ types. Figure 2.32 shows a perspective view of an $n-p-n$ bipolar transistor in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+-p junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9, compared with 11.9 for silicon).

For an $n-p-n$ bipolar transistor, the starting material is a p -type, lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 111 \rangle$ - or $\langle 100 \rangle$ -oriented, polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices (see Section 2.2.2). The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide ($0.5\text{--}1 \text{ }\mu\text{m}$) is thermally grown

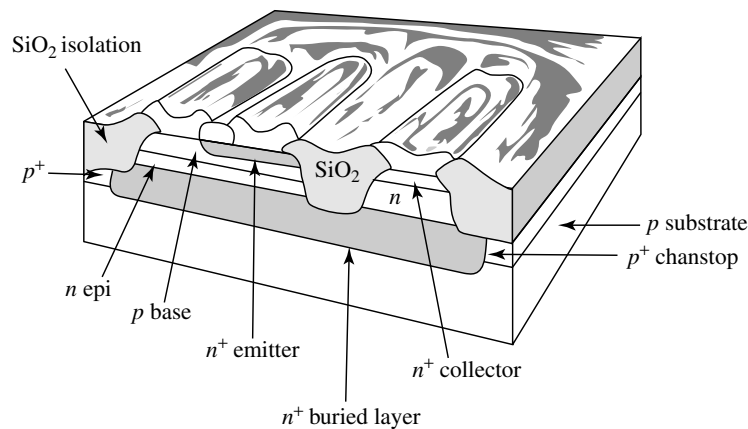


Figure 2.32. Oxide-isolated bipolar transistor [1].

on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~ 30 keV, $\sim 10^{15} \text{ cm}^{-2}$) is implanted into the window region to serve as a predeposit (Figure 2.33a). Next, a high-temperature ($\sim 1100^\circ\text{C}$) drive-in step forms the n^+ -buried layer, which has a typical sheet resistance of $20 \Omega/\square$.

The second step is to deposit an n -type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layers ($\sim 10 \mu\text{m}$) and lower dopings ($\sim 5 \times 10^{15} \text{ cm}^{-3}$), whereas digital circuits (with their lower voltages for switching) require thinner layers ($\sim 3 \mu\text{m}$) and higher dopings ($\sim 2 \times 10^{16} \text{ cm}^{-3}$). Figure 2.33b shows a cross-sectional view of the device after the epitaxial process.

The third step is to form the lateral oxide isolation region. A thin oxide pad (~ 50 nm) is thermally grown on the epitaxial layer, followed by a silicon nitride deposition (~ 100 nm). If nitride is deposited directly onto the silicon without the thin oxide pad, the nitride may cause damage to the silicon surface during subsequent high-temperature steps. Next, the nitride–oxide layers and about half of the epitaxial layer are etched using a photoresist as mask (Figures 2.33c and 2.33d). Boron ions are then implanted into the exposed silicon areas (Figure 2.33d).

The photoresist is removed, and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called *local oxidation of silicon* (LOCOS). Figure 2.34a shows the cross section of the isolation oxide after removal of the nitride layer. Because of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+ layer. This is called the p^+ -channel stop (or *chanstop*), because the high concentration of p -type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

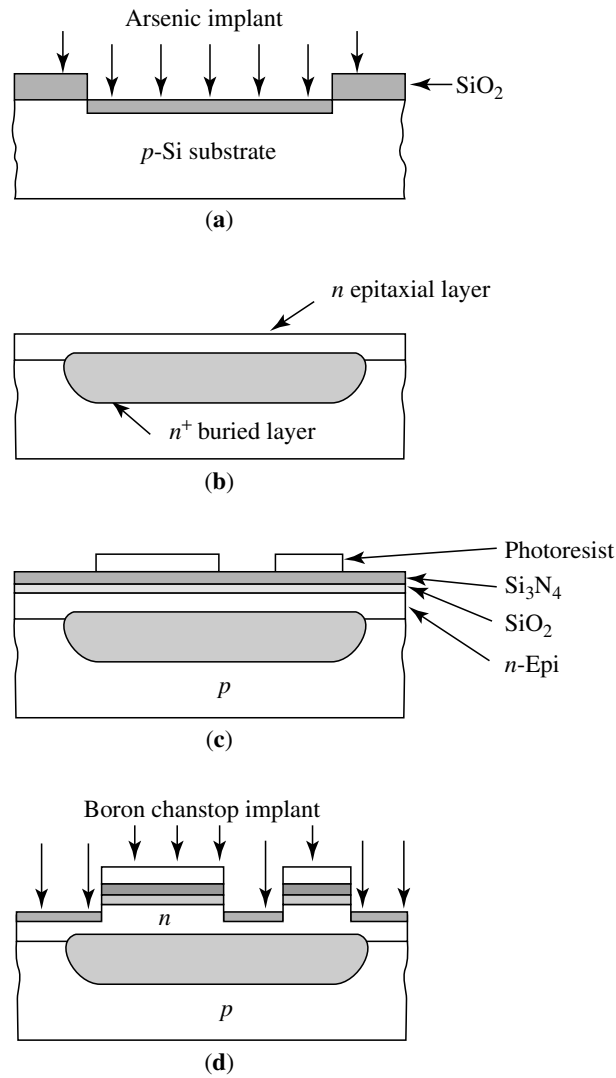


Figure 2.33. Cross-sectional views of bipolar transistor fabrication: (a) buried-layer implantation; (b) epitaxial layer; (c) photoresist mask; (d) channel-stop layer [1].

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device. Then, boron ions ($\sim 10^{12} \text{ cm}^{-2}$) are implanted to form the base regions, as shown in Figure 2.34b. Another lithographic process removes all the thin pad oxide except for a small area near the center of the base region (Figure 2.34c). The fifth step is to form the emitter region. As shown in Figure 2.34d, the base contact area is protected by a photoresist mask. Then, a low-energy, high-arsenic-dose ($\sim 10^{16} \text{ cm}^{-2}$) implantation forms the n^+ -emitter and n^+ -collector contact regions. The photoresist is removed, and a final metalization step forms the contacts to the base, emitter, and collector, as shown in Figure 2.32.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the

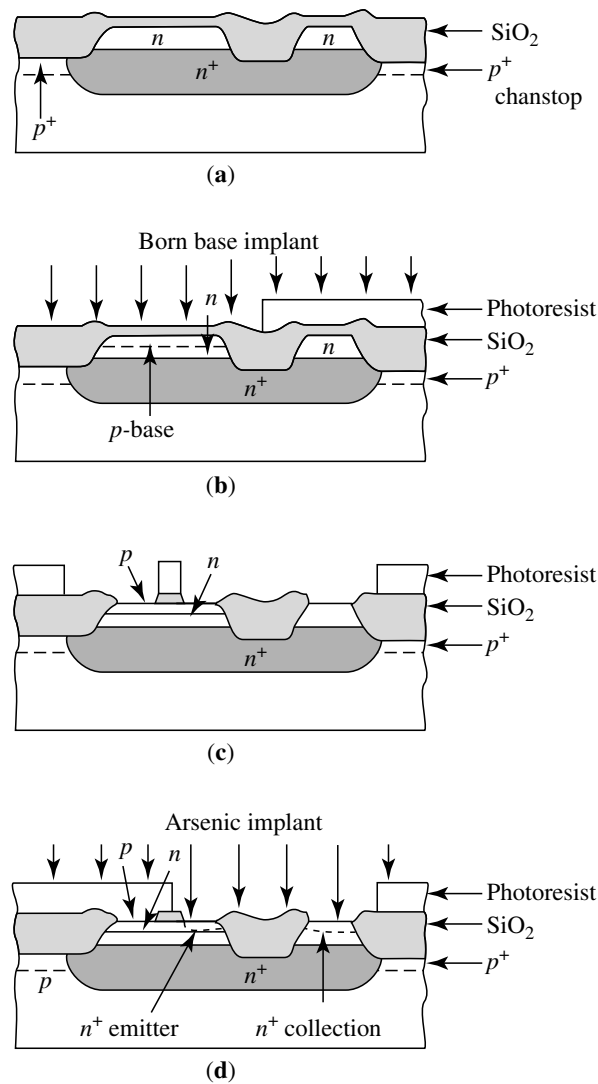


Figure 2.34. Cross-sectional views of bipolar transistor fabrication: (a) oxide isolation; (b) base implantation; (c) removal of thin oxide; (d) emitter–collector implant [1].

operations generally will render the wafer useless. The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Figure 2.35. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for limited-source diffusion. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16} \text{ cm}^{-3}$) for a representative switching transistor.

2.2.2. CMOS Technology

The MOSFET is the dominant device used in modern integrated circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is *complementary MOSFET* (CMOS) technology, in

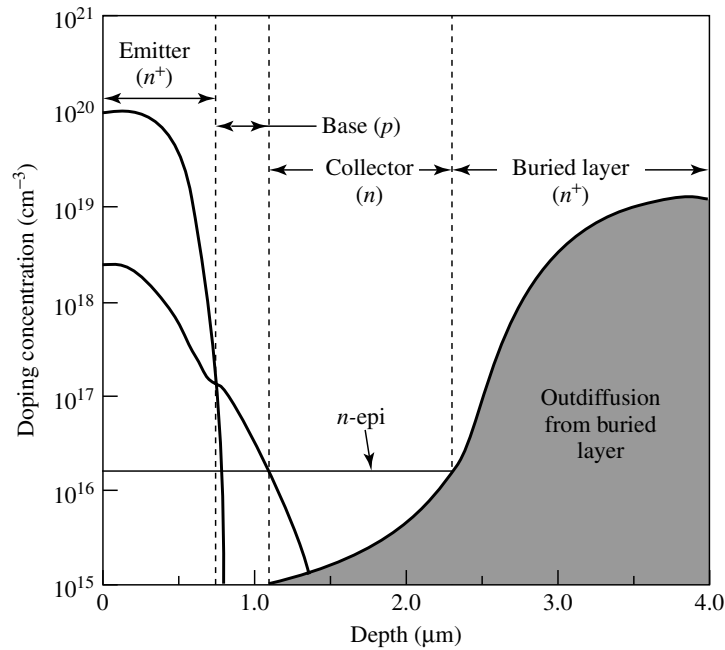


Figure 2.35. *n-p-n* bipolar transistor doping profile [1].

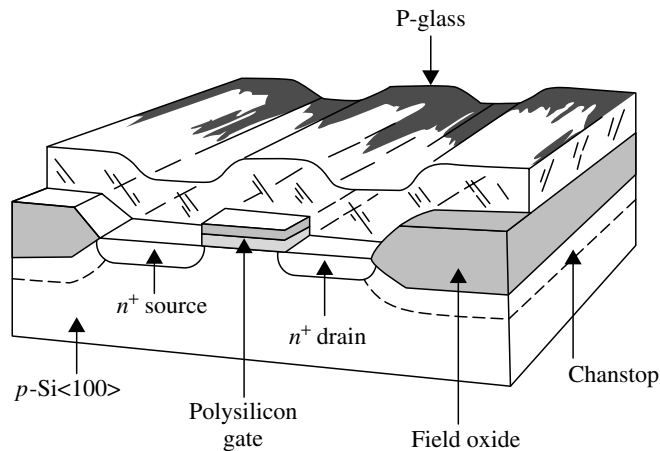


Figure 2.36. *n*-channel MOSFET [1].

which both *n*-channel and *p*-channel devices (NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particularly attractive because it has the lowest power consumption of all IC technology. Figure 2.36 shows a perspective view of an *n*-channel MOSFET prior to final metallization. The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions.

2.2.2.1. Basic NMOS Fabrication Sequence

In an NMOS process, the starting material is a *p*-type, lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 100 \rangle$ -oriented, polished silicon wafer. The first step is to form

the oxide isolation region using LOCOS technology. The process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide (~ 35 nm) is thermally grown, followed by a silicon nitride (~ 150 nm) deposition (Figure 2.37a). The active-device area is defined by a photoresist mask and a boron chanstop layer and is then implanted through the composite nitride–oxide layer (Figure 2.37b). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the *field oxide*), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically $0.5\text{--}1\text{ }\mu\text{m}$.

The second step is to grow the gate oxide and to adjust the threshold voltage. The composite nitride–oxide layer over the active-device area is removed, and

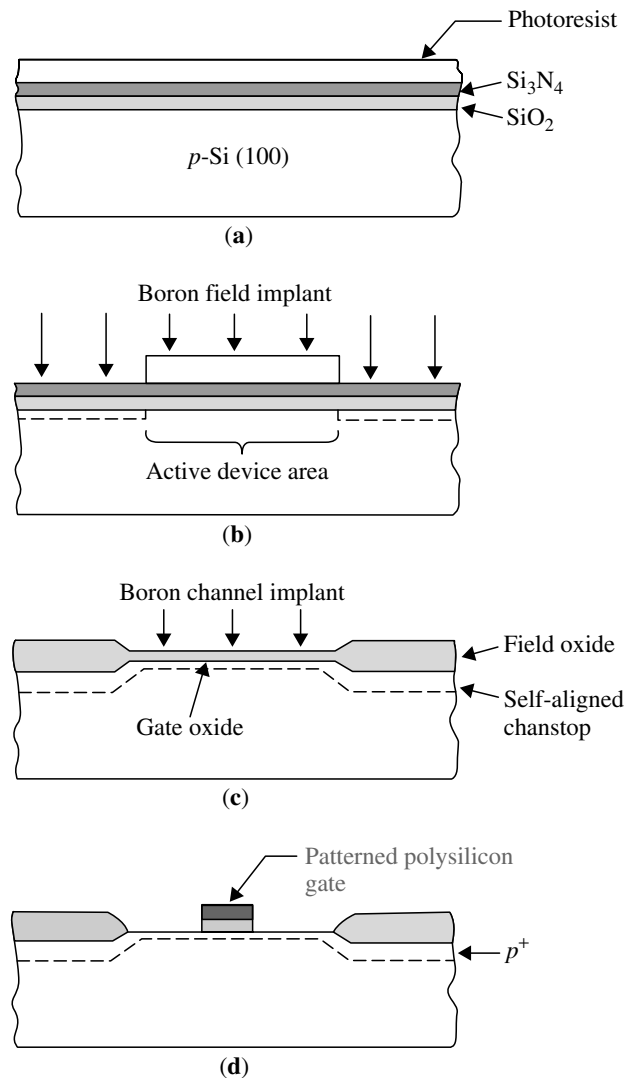


Figure 2.37. NMOS fabrication sequence: (a) formation of SiO_2 , Si_3N_4 , and photoresist layers; (b) boron implant; (c) field oxide; (d) gate [1].

a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode n -channel device, boron ions are implanted in the channel region, as shown in Figure 2.37c, to increase the threshold voltage to a predetermined value (e.g., +0.5 V). For a depletion-mode n -channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., -0.5 V).

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20–30 $\Omega/\square$. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm . For smaller devices, polycide, a composite layer of metal silicide and polysilicon such as W-polycide, can be used as the gate materials to reduce the sheet resistance to about 1 $\Omega/\square$.

The fourth step is to form the source and drain. After the gate is patterned (Figure 2.37d), it serves as a mask for the arsenic implantation (~ 30 keV, $\sim 5 \times 10^{15} \text{ cm}^{-2}$) to form the source and drain (Figure 2.38a), which are self-aligned

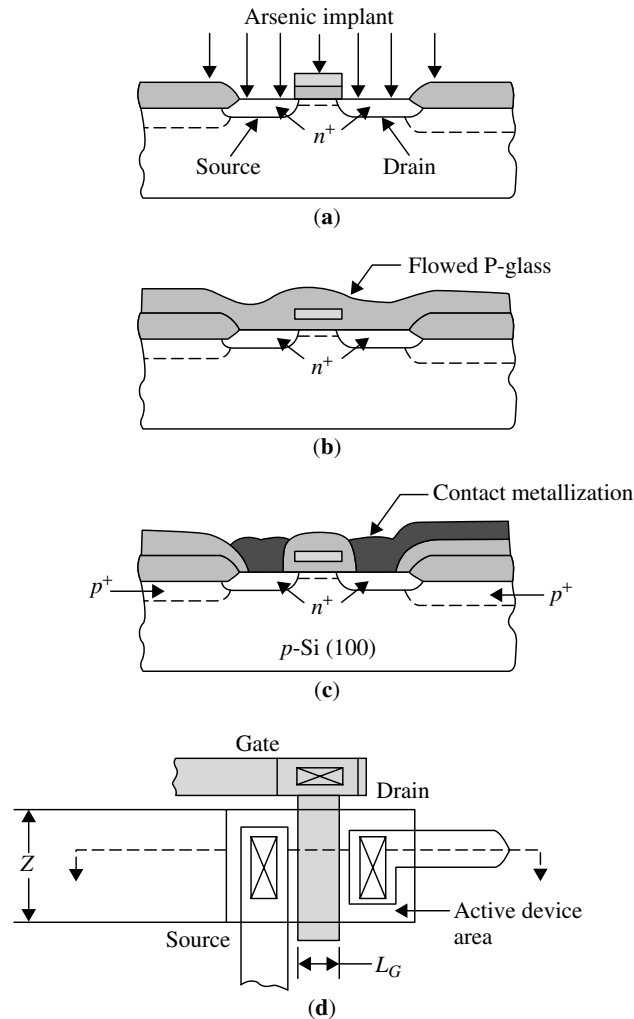


Figure 2.38. NMOS fabrication sequence: (a) source and drain; (b) P-glass deposition; (c) MOSFET cross section; (d) MOSFET top view [1].

with respect to the gate. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\sigma_{\perp}$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

The last step is metallization. P-glass is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Figure 2.38b). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-sectional view of the completed MOSFET is shown in Figure 2.38c, and the corresponding top view is shown in Figure 2.38d. The gate contact is usually made outside the active-device area to avoid possible damage to the thin-gate oxide.

2.2.2.2. CMOS Fabrication Sequence

The MOS process forms the foundation for CMOS technology. Figure 2.39a shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. For the CMOS inverter, in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the inverter. Thus, the average power dissipation is on the order of nanowatts. Low power consumption is the most attractive feature of the CMOS circuit.

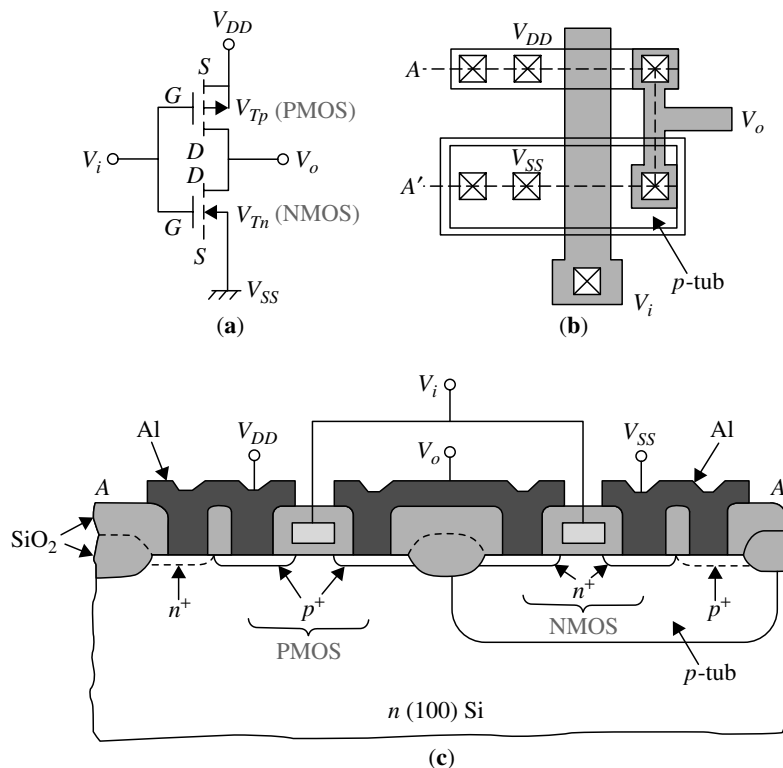


Figure 2.39. CMOS inverter: (a) circuit diagram; (b) layout; and (c) cross section [1].

Figure 2.39b shows a layout of the CMOS inverter, and Figure 2.39c shows the device cross section along the $A-A'$ line. In the processing, a p tub (also called a p well) is first implanted and subsequently driven into the n substrate. The p -type dopant concentration must be high enough to overcompensate the background doping of the n substrate. The subsequent processes for the n -channel MOSFET in the p tub are identical to those described previously. For the p -channel MOSFET, $^{11}\text{B}^+$ or $^{49}(\text{BF}_2)^+$ ions are implanted into the n substrate to form the source and drain regions. A channel implant of $^{75}\text{As}^+$ ions may be used to adjust the threshold voltage and a n^+ chanstop is formed underneath the field oxide around the p -channel device. Because of the p tub and the additional steps needed to make the p -channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, there is a trade-off between the complexity of processing and a reduction in power consumption.

Instead of the p tub described above, an alternate approach is to use an n tub formed in p -type substrate, as shown in Figure 2.40a. In this case, the n -type dopant concentration must be high enough to overcompensate for the background doping of the p substrate (i.e., $N_D > N_A$). In both the p -tub and the n -tub approaches, the channel mobility will be degraded because mobility is determined by the total dopant concentration ($N_A + N_D$). A more recent approach using two separated tubs implanted into a lightly doped substrate is shown in Figure 2.40b. This structure is called a “twin tub.” Because no overcompensation is needed in either of the twin tubs, higher channel mobility can be obtained.

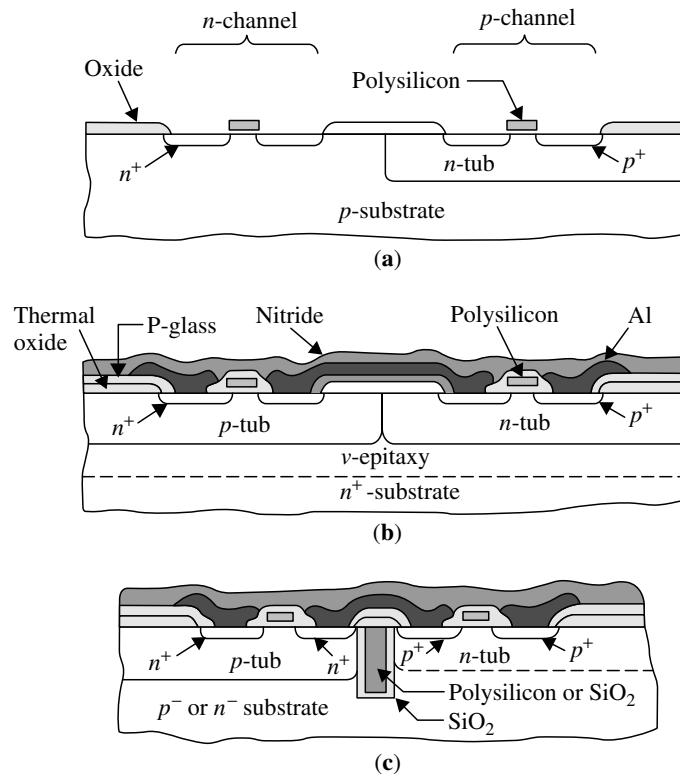


Figure 2.40. Various CMOS structures: (a) n -tub; (b) twin-tub; (c) refilled trench [1].

All CMOS circuits have the potential for a problem called *latchup* that is associated with parasitic bipolar transistors. These parasitic devices consist of the *npn* transistor formed by the NMOS source–drain regions, *p* tub, and *n*-type substrate, as well as the *pn*p transistor formed by the PMOS source-drain regions, *n*-type substrate, and *p* tub. Under appropriate conditions, the collector of the *pn*p device supplies base current to the *n*p*n* and vice versa in a positive feedback arrangement. This latchup current can have serious negative repercussions in a CMOS circuit.

An effective processing technique to eliminate latchup is to use deep-trench isolation, as shown in Figure 2.40c. In this technique, a trench with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for trench isolation and some related CMOS processes are now considered.

Well Formation. The well of a CMOS circuit can be a single well, a twin well, or a retrograde well. The twin-well process exhibits some disadvantages. For example, it needs high-temperature processing (above 1050°C) and a long diffusion time (longer than 8 h) to achieve the required well depth of 2–3 μm . In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used (i.e., implanting the ion to the desired depth instead of diffusion from the surface). The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a *retrograde well*.

The advantage of high-energy implantation is that it can form the well under low-temperature and short-time conditions. Hence, it can reduce the lateral diffusion and increase the device density. The retrograde well offers some additional advantages over the conventional well: (1) because of high doping near the bottom, the well resistivity is lower than that of the conventional well, and latchup can be minimized; (2) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time; and (3) higher well doping in the bottom can reduce the chance of punchthrough from the drain to the source.

Isolation. The conventional MOS isolation process has some disadvantages that make it unsuitable for deep-submicrometer ($\leq 0.25\text{-}\mu\text{m}$) fabrication. The high-temperature oxidation of silicon and long oxidation time result in the encroachment of the chanstop implantation (usually boron for *n*-MOSFET) to the active region and cause a threshold voltage shift. The area of the active region is reduced because of the lateral oxidation. In addition, the field oxide thickness in submicrometer-isolation spacings is significantly less than the thickness of field oxide grown in wider spacings. Trench isolation technology can avoid these problems.

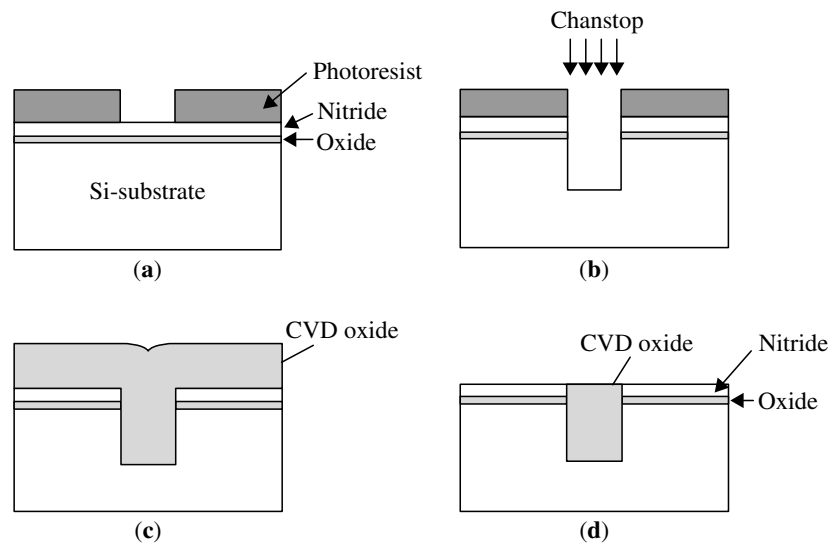


Figure 2.41. Shallow-trench isolation: (a) patterning on nitride–oxide films; (b) dry etching and chanstop implantation; (c) CVD oxide to refill; (d) surface after CMP [1].

An example is *shallow trench* (depth less than $1\text{ }\mu\text{m}$) isolation, shown in Figure 2.41. After patterning (Figure 2.41a), the trench area is etched (Figure 2.41b) and then refilled with oxide (Figure 2.41c). Before refilling, a channel stop implantation can be performed. Since the oxide has overfilled the trench, the oxide on the nitride should be removed. Chemical–mechanical polishing is used to remove the oxide on the nitride and to get a flat surface (Figure 2.41d). Because of its high resistance to polishing, the nitride acts as a stop layer for the CMP process. After the polishing, the nitride layer and the oxide layer can be removed by H_3PO_4 and HF , respectively. This initial planarization step at the beginning is helpful for the subsequent polysilicon patterning and planarizations of the multilevel interconnection processes.

Gate Engineering. If n^+ -polysilicon is used for both PMOS and NMOS gates, the threshold voltage for PMOS has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, as shown in Figure 2.42a. The buried-type PMOS suffers serious short-channel effects as the device size shrinks below $0.25\text{ }\mu\text{m}$. The most noticeable phenomena for short-channel effects are threshold voltage rolloff, drain-induced barrier lowering, and the large leakage current at the OFF state. To alleviate these problems, the n^+ -polysilicon can be changed to p^+ -polysilicon for the PMOS devices. Due to the workfunction difference (1.0 eV from n^+ - to p^+ -polysilicon), a surface p -type channel device can be achieved without the boron V_T adjustment implantation. Hence, as the technology shrinks to $0.25\text{ }\mu\text{m}$ and less, dual-gate structures are required: p^+ -polysilicon gate for PMOS and n^+ -polysilicon for NMOS (Figure 2.42b).

To form the p^+ -polysilicon gate, ion implantation of BF_2 is commonly used. However, boron penetrates easily from the polysilicon through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration

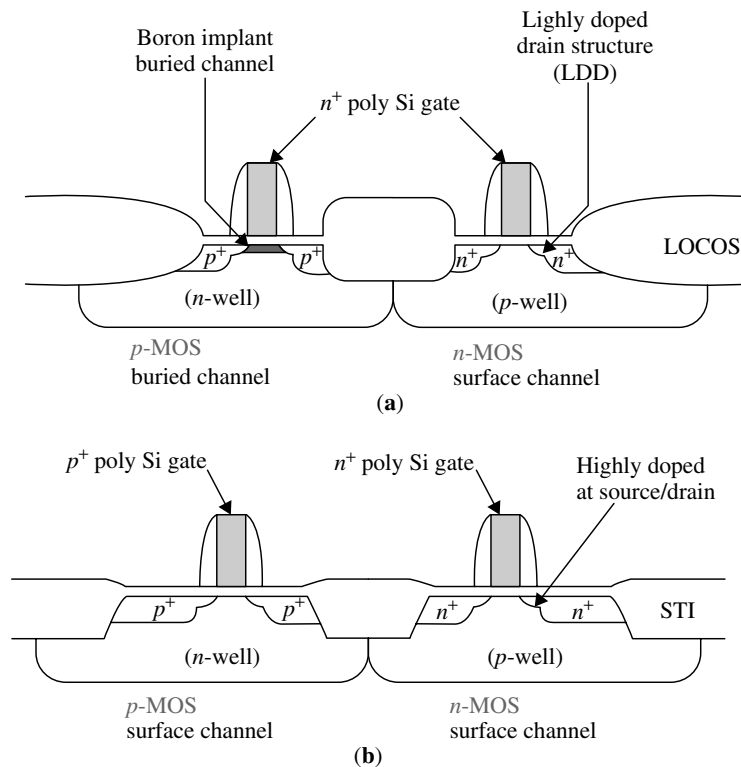


Figure 2.42. (a) Conventional CMOS structure with a single polysilicon gate; (b) advanced CMOS structure with dual polysilicon gates [1].

is enhanced in the presence of a F atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrified oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and the creation of a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

2.2.3. BiCMOS Technology

BiCMOS is a technology that combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip that has the advantages of both CMOS and bipolar devices. We know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar technology shows advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a higher speed than CMOS, better performance in analog circuits than CMOS, a lower power dissipation than bipolar, and a higher component density than bipolar.

BiCMOS has been widely used in many applications. Early on, it was used in static random access memory (SRAM) circuits. Currently, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator

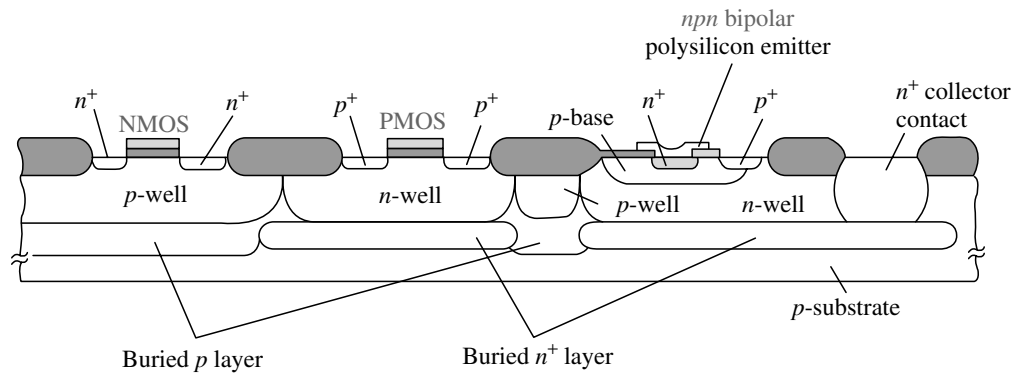


Figure 2.43. BiCMOS device structure [1].

applications in wireless communication equipment. Most BiCMOS processes are based on standard CMOS process with some modifications, such as adding masks for bipolar transistor fabrication. The example shown in Figure 2.43 is for a high-performance BiCMOS process based on the twin-well CMOS approach.

The initial material is a p -type silicon substrate. An n^+ -buried layer is formed to reduce collector resistance. The buried p layer is formed by ion implantation to increase the doping level and prevent punchthrough. A lightly doped n -epi layer is grown on the wafer, and a twin-well process for the CMOS is performed. To achieve high performance for the bipolar transistor, four additional masks are needed: the buried n^+ mask, the collector deep n^+ mask, the base p mask, and the polyemitter mask. The p^+ region for base contact can be formed with the p^+ implant in the source–drain implantation of the PMOS, and the n^+ emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time compared with a standard CMOS process are the main drawbacks of BiCMOS.

2.2.4. Packaging

Before finished ICs can be put to their intended use in various commercial electronic systems and products (such as computers, cellular phones, and digital cameras), several other key processes must take place. These include both *electrical testing* and *packaging*. Testing, which is discussed in detail in Chapter 3, is clearly necessary to ensure high-quality products. The term *packaging* refers to the set of technologies and processes that connect ICs with electronic systems. A useful analogy is to consider an electronic product as the human body. Like the body, these products have “brains,” which are analogous to ICs. Electronic packaging provides the “nervous system,” as well as the “skeletal system.” The package is responsible for interconnecting, powering, cooling, and protecting the IC.

Overall, electronic systems consist of several levels of packaging, each with distinctive types of interconnection devices. Figure 2.44 depicts this packaging hierarchy. Level 0 consists of on-chip interconnections. Chip-to-printed circuit board or chip-to-module connections constitute level 2, and board-to-board

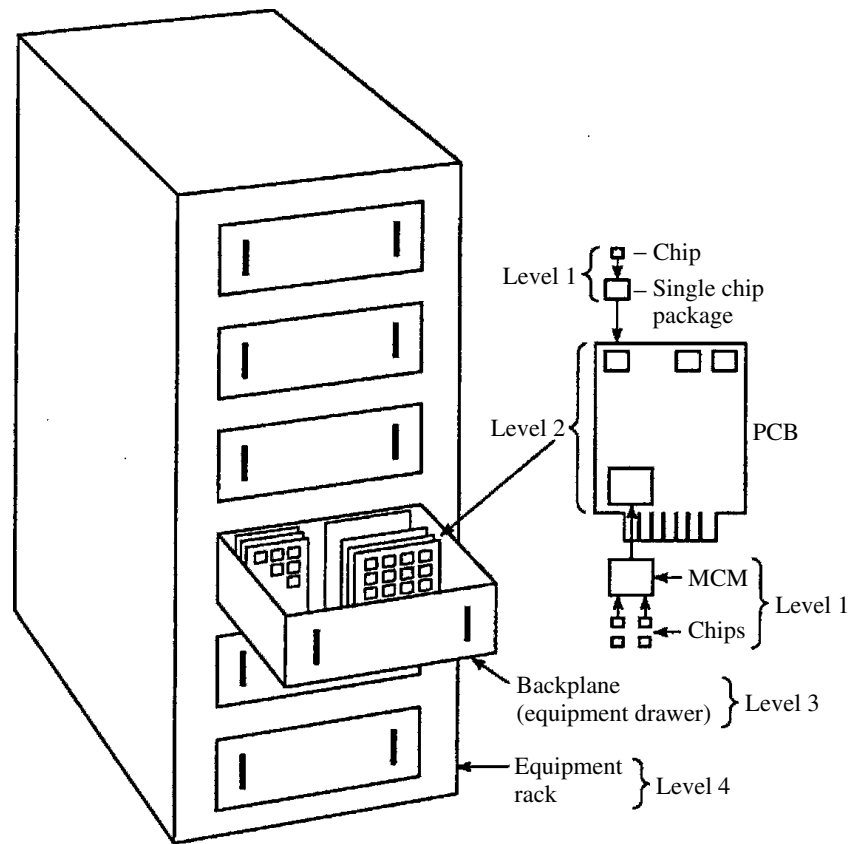


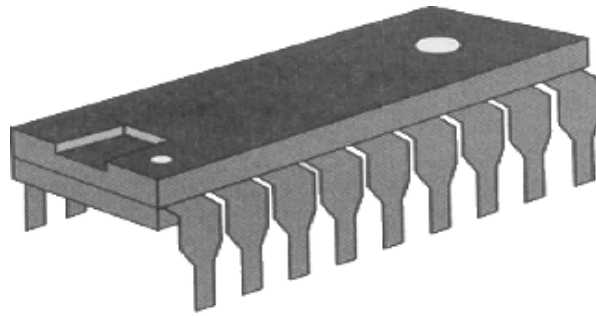
Figure 2.44. Electronic packaging hierarchy [2].

interconnections make up level 3. Levels 4 and 5 consist of connections between subassemblies and between systems (such as computer to printer), respectively.

2.2.4.1. Die Separation

After functional testing, individual ICs (or dies) must be separated from the substrate. This is the first step in the packaging process. In a common method that has been used for many years, the substrate wafer is mounted on a holder and scribed in both the x and y directions using a diamond scribe. This is done along scribe borders of $75\text{--}250\text{ }\mu\text{m}$ in width that are formed around the periphery of the dies during fabrication. These borders are aligned with the crystal planes of this substrate if possible. After scribing, the wafer is removed from the holder and placed upside-down on a soft support. A roller is then used to apply pressure, fracturing the wafer along the scribe lines. This must be accomplished with minimal damage to the individual die.

More modern die separation processes use a diamond saw, rather than a diamond scribe. In this procedure, the wafer is attached to an adhesive sheet of mylar film. The saw is then used to either scribe the wafer or to cut completely through it. After separation, the dies are removed from the mylar. The separated dice are then ready to be placed into packages.



Plastic dual-in-line-package

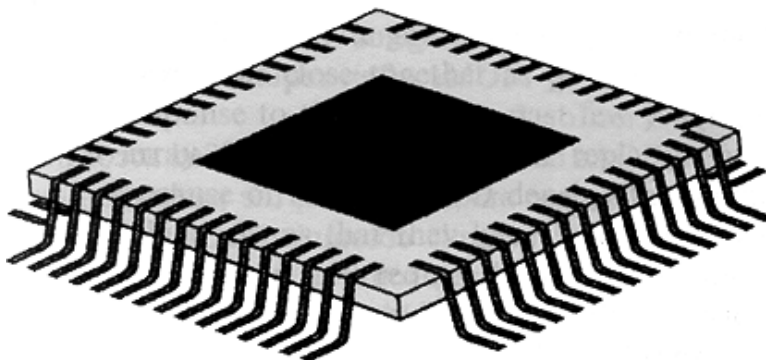
Figure 2.45. Dual-inline package [2].

2.2.4.2. Package Types

There are a number of approaches to the packaging of single ICs. The *dual-inline package* (DIP) (Figure 2.45), is the package most people envision when they think of integrated circuits. The DIP was developed in the 1960s, quickly became the primary package for ICs, and has long dominated the electronics packaging market. The DIP can be made of plastic or ceramic; the latter is called the *CerDIP*. The CerDIP consists of a DIP constructed of two pieces of sandwiched ceramic with leads protruding from between the ceramic plates.

In the 1970s and 1980s, *surface-mount packages* were developed in response to a need for higher-density interconnect than the DIP approach could provide. In contrast to DIPs, the leads of a surface-mounted package do not penetrate the printed circuit board (PCB) on which it is mounted. This means that the package can be mounted on both sides of the board, thereby allowing higher density. One example of such a package is the quad flatpack (QFP) (Figure 2.46), which has leads on all four sides to further increase the number of input/output (I/O) connections.

More recently, the need for rapidly increasing numbers of I/O connections has led to the development of pin-grid array (PGA) and ball-grid array (BGA) packages (Figures 2.47 and 2.48, respectively). PGAs have an I/O density of about 600, and BGAs can have densities greater than 1000, as compared to

**Figure 2.46.** Quad flatpack [2].

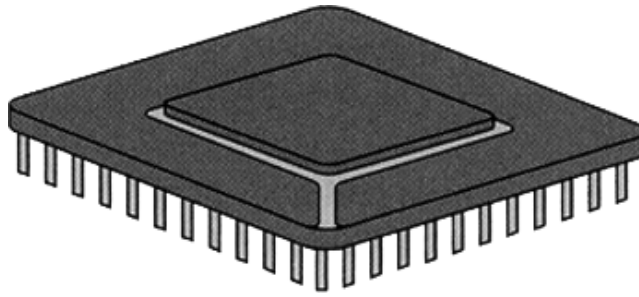


Figure 2.47. Pin-grid array [2].

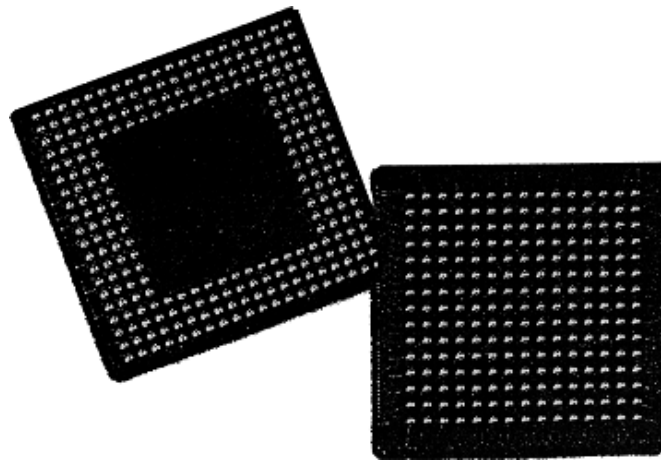


Figure 2.48. Ball-grid array [2].

~200 for QFPs. BGAs can be identified by the solder bumps on the bottom of the package. With QFPs, as the spacing between leads becomes tighter, the manufacturing yield decreases rapidly. The BGA allows higher density and takes up less space than the QFP, but its manufacturing process is inherently more expensive.

The most recent development in packaging is the chip-scale package (CSP), which is shown in Figure 2.49. CSPs, defined as packages no larger than 20% greater than the size of the IC die itself, often take the form of miniaturized ball-grid arrays. They are designed to be *flipchip-mounted* (see Section 2.2.4.3) using conventional equipment and solder reflow. CSPs are typically manufactured in a process that creates external power and signal I/O contacts and encapsulates the finished silicon die prior to dicing the wafer. Essentially, CSPs provide an interconnection framework for ICs so that before dicing, each die has all the functions (i.e., external electrical contacts, encapsulation of the finished silicon) of a conventional, fully packaged IC. Two essential features of this approach are that the leads and interposer layer (an added layer on the IC used to provide electrical functionality and mechanical stability) are flexible enough so that the packaged device is compliant with the test fixture for full testing and burning, and the package can accommodate the vertical nonplanarity and thermal expansion and contraction of the underlying printed circuit board during assembly and operation.

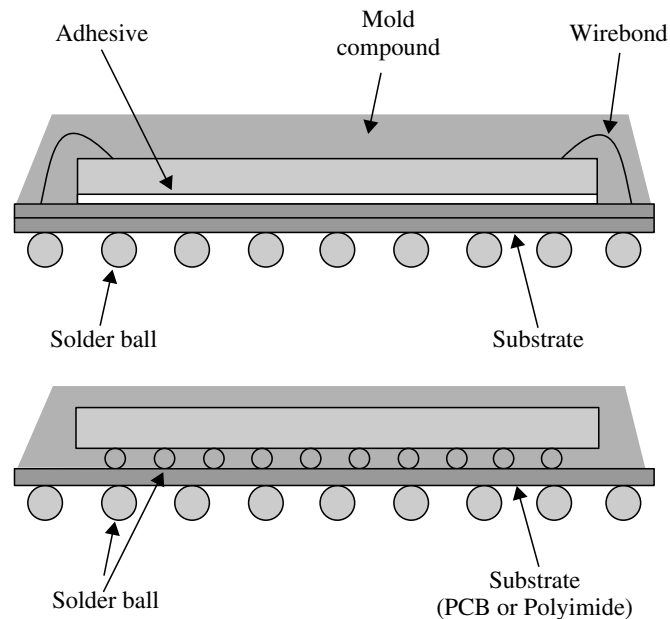


Figure 2.49. Two typical chip-scale packages [3].

2.2.4.3. Attachment Methods

An IC must be mounted and bonded to a package, and that package must be attached to a printed circuit board before the IC can be used in an electronic system. Methods of attaching ICs to PCBs are referred to as *level 1 packaging*. The technique used to bond a bare die to a package has a significant effect on the ultimate electrical, mechanical and thermal properties of electronic system being manufactured. Chip-to-package interconnection is generally accomplished by either *wire bonding*, *tape-automated bonding (TAB)*, or *flipchip bonding* (see Figure 2.50).

Wire bonding is the oldest attachment method and is still the dominant technique for chips with fewer than 200 I/O connections. Wire bonding requires connecting gold or aluminum wires between chip bonding pads and contact points on the package. ICs are first attached to the substrate using a thermally conductive adhesive with their bonding pads facing upward. The Au or Al wires are then attached between the pads and substrate using *ultrasonic*, *thermosonic*, or *thermocompression* bonding [1]. Although automated, this process is still time-consuming since each wire must be attached individually.

Tape-automated bonding (TAB) was developed in the early 1970s and is often used to bond packages to PCBs. In TAB, chips are first mounted on a flexible polymer tape (usually polyimide) containing repeated copper interconnection patterns. The copper leads are defined by lithography and etching, and the lead pattern can contain hundreds of connections. After the IC pads have been aligned to metal interconnection stripes on the tape, attachment takes place by thermocompression. Gold bumps are formed on either side of the die or tape and are used to bond the die to the leads on the tape.

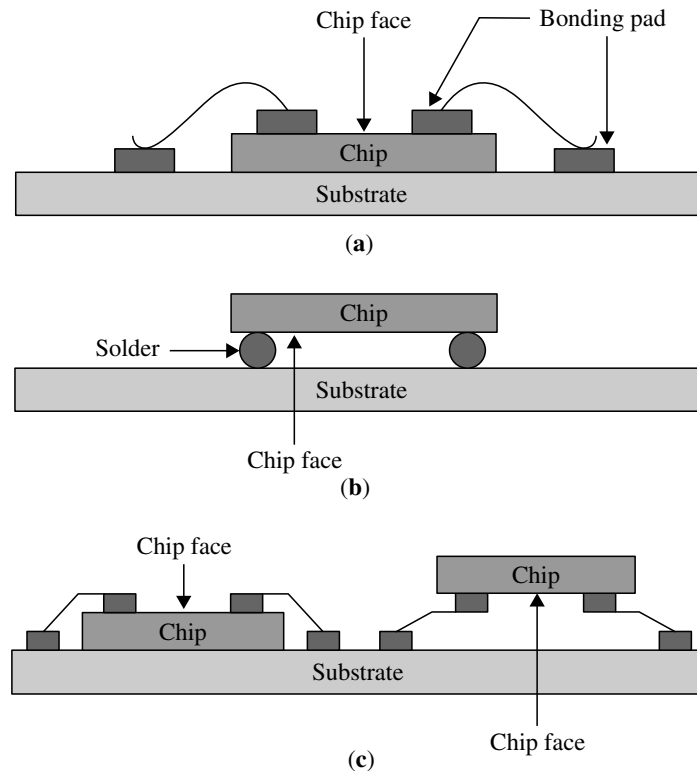


Figure 2.50. (a) Wire; (b) flipchip; (c) tape-automated bonding [2].

Flipchip bonding is a direct interconnection approach in which the IC is mounted upside-down onto a module or printed circuit board. Electrical connections are made via solder bumps (or solderless materials such as epoxies or conductive adhesives) located over the surface of the chip. Since bumps can be located anywhere on the chip, flipchip bonding ensures that the interconnect distance between the chip and package is minimized. The I/O density is limited only by the minimum distance between adjacent bond pads.

SUMMARY

In this chapter, we have provided an overview of the critical unit processes in IC fabrication and described the integration of these unit processes into sequences for fabricating and packaging ICs. In the next chapter, we will discuss how these processes are monitored to facilitate quality control.

PROBLEMS

- 2.1.** Assuming that a silicon oxide layer of thickness x is grown by thermal oxidation, show that the thickness of silicon being consumed is $0.44x$. 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.

- 2.2. A silicon sample is oxidized in dry O₂ at 1200°C for one 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 O₂ at 1200°C?
- 2.3. Find the parameter γ for the photoresists shown in Figure 2.13.
- 2.4. Calculate the Al average etch rate and etch rate uniformity on a 200-mm-diameter silicon wafer, assuming that the etch rates at the center, left, right, top, and bottom of the wafer are 750, 812, 765, 743, and 798 nm/min, respectively.
- 2.5. The electron densities in RIE and HDP systems range within 10⁹–10¹⁰ and 10¹¹–10¹² cm⁻³, respectively. Assuming that 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.
- 2.6. For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at 10¹⁹ cm⁻³ and the diffusion time is 1 h. If the diffusion coefficient of boron at 1000°C is 2 × 10¹⁴ cm²/s, find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches 10¹⁵ cm⁻³.
- 2.7. Arsenic was predeposited by arsine gas, and the resulting total amount of dopant per unit area was 1 × 10¹⁴ 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.
- 2.8. Assume 100-keV boron implants on a 200-mm silicon wafer at a dose of 5 × 10¹⁴ ions/cm². The projected range and project straggle are 0.31 and 0.07 μm, respectively. Calculate the peak concentration and the required ion-beam current for 1 min of implantation.
- 2.9. In a CMP process, the oxide removal rate and the removal rate of a layer underneath the oxide (called a *stop layer*) are $1r$ and $0.1r$, respectively. To remove 1 μm of oxide and a 0.01-μm stop layer, the total removal time is 5.5 min. Find the oxide removal rate.

REFERENCES

1. G. May and S. Sze, *Fundamentals of Semiconductor Fabrication*, Wiley, New York, 2003.
2. W. Brown, ed., *Advanced Electronic Packaging*, IEEE Press, New York, 1999.
3. R. Tummala, ed., *Fundamentals of Microsystems Packaging*, McGraw-Hill, New York, 2001.