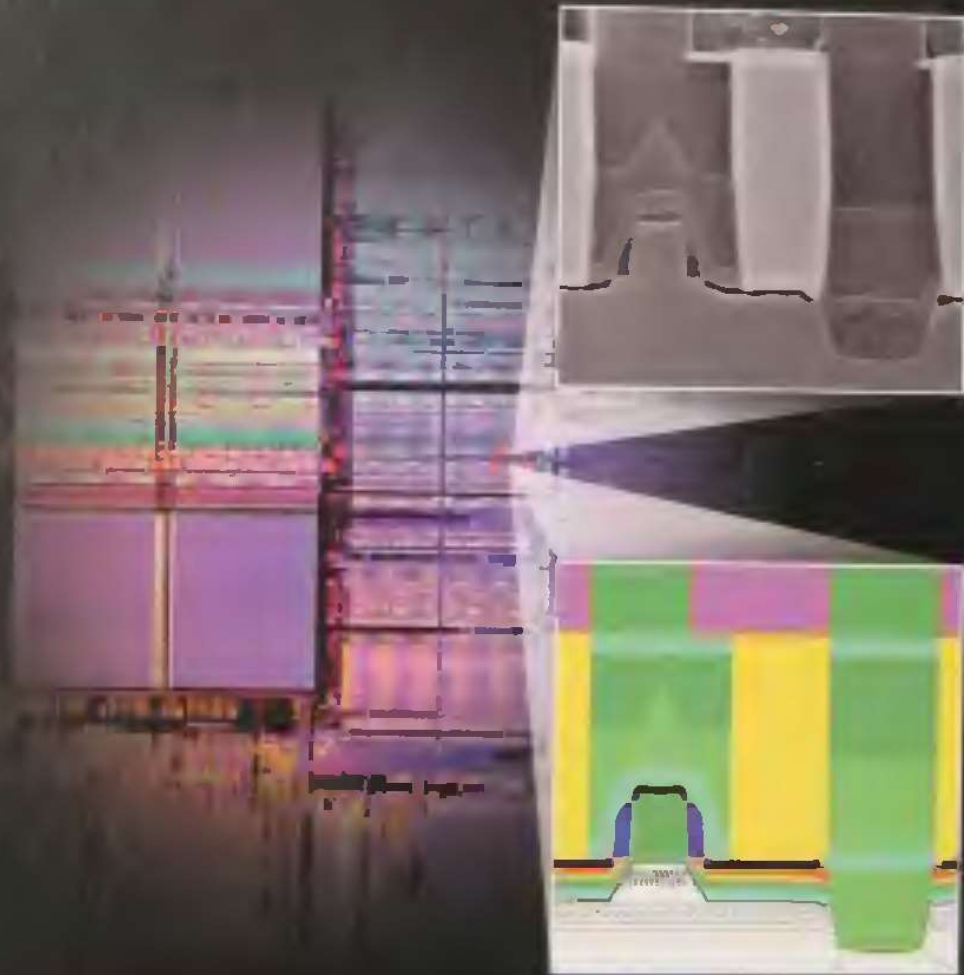


SILICON VLSI TECHNOLOGY

Fundamentals, Practice and Modeling



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

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

YMTC
Exhibit 2028
Micron v. YMTC
PGR2025-00021

Ex. 2028-001

Library of Congress Cataloging-in-Publication Data
Silicon VLSI technology

p. cm.
ISBN 0-13-085037-3
1. Integrated circuits—Very large scale integration—Design and construction. 2. Silicon. 3. Silicon oxide films. 4. Metal oxide semiconductors. 5. Silicon technology.
TK7874.75.S54 2000
621.39'5—dc21

99-42745
CIP

Publisher: Tom Robbins
Associate Editor: Alice Dworkin
Editorial/Production Supervision: Rose Kernan
Vice President and Editorial Director, ECS: Marcia Horton
Vice President of Production and Manufacturing: David W. Riccardi
Executive Managing Editor: Vince O'Brien
Marketing Manager: Danny Hoyt
Managing Editor: David A. George
Manufacturing Buyer: Pat Brown
Manufacturing Manager: Trudy Piscioti
Art Director: Jayne Conte
Cover Design: Bruce Kenselaar
Editorial Assistant: Jesse Power
Copy Editor: Martha Williams
Composition: D&G Limited, LLC



©2000 by Prentice Hall, Inc.
Upper Saddle River, New Jersey 07458

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

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

Printed in the United States of America

10 9 8 7 6 5

ISBN 0-13-085037-3

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

Contents

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

Chapter 3	Crystal Growth, Wafer Fabrication and Basic Properties of Silicon Wafers	93
3.1	Introduction	93
3.2	Historical Development and Basic Concepts	93
3.2.1	Crystal Structure	94
3.2.2	Defects in Crystals	97
3.2.3	Raw Materials and Purification	101
3.2.4	Czochralski and Float-Zone Crystal Growth Methods	102
3.2.5	Wafer Preparation and Specification	105
3.3	Manufacturing Methods and Equipment	109
3.4	Measurement Methods	111
3.4.1	Electrical Measurements	111
3.4.1.1	Hot Point Probe	112
3.4.1.2	Sheet Resistance	113
3.4.1.3	Hall Effect Measurements	115
3.4.2	Physical Measurements	117
3.4.2.1	Defect Etches	117
3.4.2.2	Fourier Transform Infrared Spectroscopy (FTIR)	118
3.4.2.3	Electron Microscopy	119
3.5	Models and Simulation	121
3.5.1	Czochralski Crystal Growth	122
3.5.2	Dopant Incorporation during CZ Crystal Growth	125
3.5.3	Zone Refining and FZ Growth	128
3.5.4	Point Defects	131
3.5.5	Oxygen in Silicon	138
3.5.6	Carbon in Silicon	142
3.5.7	Simulation	143
3.6	Limits and Future Trends in Technologies and Models	144
3.7	Summary of Key Ideas	146
3.8	References	147
3.9	Problems	148
Chapter 4	Semiconductor Manufacturing—Clean Rooms, Wafer Cleaning, and Gettering	151
4.1	Introduction	151
4.2	Historical Development and Basic Concepts	154
4.2.1	Level 1 Contamination Reduction: Clean Factories	157
4.2.2	Level 2 Contamination Reduction: Wafer Cleaning	159
4.2.3	Level 3 Contamination Reduction: Gettering	161
4.3	Manufacturing Methods and Equipment	165
4.3.1	Level 1 Contamination Reduction: Clean Factories	165
4.3.2	Level 2 Contamination Reduction: Wafer Cleaning	166
4.3.3	Level 3 Contamination Reduction: Gettering	167
4.4	Measurement Methods	169
4.4.1	Level 1 Contamination Reduction: Clean Factories	169

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

Chapter 5 Lithography 201

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

5.5.7	Advanced Mask Engineering	271
5.6	Limits and Future Trends in Technologies and Models	272
5.6.1	Electron Beam Lithography	273
5.6.2	X-ray Lithography	275
5.6.3	Advanced Mask Engineering	277
5.6.4	New Resists	278
5.7	Summary of Key Ideas	281
5.8	References	281
5.9	Problems	283

Chapter 6 Thermal Oxidation and the Si/SiO₂ Interface 287

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

Chapter 7 Dopant Diffusion 371

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

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

Chapter 8 Ion Implantation 451

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

8.3.1	High-Energy Implants	466
8.3.2	Ultralow Energy Implants	468
8.3.3	Ion Beam Heating	469
8.4	Measurement Methods	470
8.5	Models and Simulations	471
8.5.1	Nuclear Stopping	473
8.5.2	Nonlocal Electronic Stopping	474
8.5.3	Local Electronic Stopping	475
8.5.4	Total Stopping Powers	476
8.5.5	Damage Production	479
8.5.6	Damage Annealing	482
8.5.7	Solid-Phase Epitaxy	484
8.5.8	Dopant Activation	486
8.5.9	Transient-Enhanced Diffusion	488
8.5.10	Atomic-Level Understanding of TED	497
8.5.11	Effects on Devices	499
8.6	Limits and Future Trends in Technologies and Models	500
8.7	Summary of Key Ideas	500
8.8	References	502
8.9	Problems	

Chapter 9 Thin Film Deposition 509

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

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

Chapter 10 Etching 609

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

Chapter 11 Back-End Technology 681

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

11.2.3	Dielectrics	707
11.3	Manufacturing Methods and Equipment	715
11.3.1	Silicided Gates and Source/Drain Regions	716
11.3.2	First-level Dielectric Processing	718
11.3.3	Contact Formation	719
11.3.4	Global Interconnects	721
11.3.5	IMD Deposition and Planarization	723
11.3.6	Via Formation	724
11.3.7	Final Steps	725
11.4	Measurement Methods	725
11.4.1	Morphological Measurements	726
11.4.2	Electrical Measurements	726
11.4.3	Chemical and Structural Measurements	732
11.4.4	Mechanical Measurements	734
11.5	Models and Simulation	737
11.5.1	Silicide Formation	738
11.5.2	Chemical-Mechanical Polishing	744
11.5.3	Reflow	746
11.5.4	Grain Growth	753
11.5.5	Diffusion in Polycrystalline Materials	762
11.5.6	Electromigration	765
11.6	Limits and Future Trends in Technologies and Models	776
11.7	Summary of Key Ideas	780
11.8	References	781
11.9	Problems	784

Appendices

787

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

Index

805

Thermal Oxidation and the Si/SiO₂ Interface

6

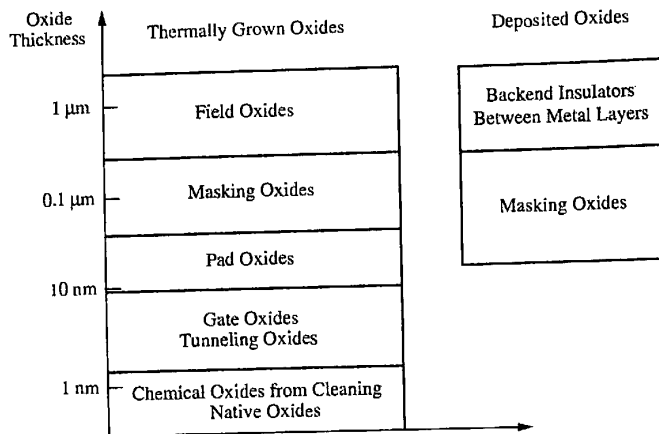
6.1 Introduction

Silicon is unique among semiconductor materials in that its surface can be easily passivated with an oxide layer. The interface between Si and SiO₂ is perhaps the most carefully studied of all material interfaces and its electrical and mechanical properties as well as those of the oxide layer itself are almost ideal. SiO₂ layers are easily grown thermally on silicon or deposited on many substrates. They adhere well, they block the diffusion of dopants and many other unwanted impurities, they are resistant to most of the chemicals used in silicon processing and yet can be easily patterned and etched with specific chemicals or dry etched with plasmas, they are excellent insulators, and they have stable and reproducible bulk properties. The interface that forms between Si and SiO₂ has very few mechanical or electrical defects and is stable over time. These properties make MOS structures easy to build in silicon and they imply that silicon devices of all types are generally reliable and stable. Virtually all other semiconductor/insulator combinations suffer from one or more problems that significantly limit their applicability.

We saw a number of applications of SiO₂ layers in the CMOS technology example in Chapter 2. These included use as the gate dielectric layer in MOS devices, as a mask against implantation, as an isolation region laterally between adjacent devices, and as an insulator between metal layers in back-end processing. These and some other uses are illustrated in Figure 6-1.

The SIA NTRS provides a roadmap for SiO₂ and other insulating layers as part of its general technology requirements. Some of the key issues are summarized in Table 6-1 [6.1]. Most of these requirements relate to oxides < 10 nm since these are the critical gate insulators and tunneling oxides in modern MOS structures. The last two rows in the table relate to thicker insulators used primarily for masking and in back-end processing. These insulating layers are normally deposited rather than thermally grown and will be discussed in more detail in Chapters 9 and 11.

Even at room temperature, silicon exposed to an oxygen or air ambient will form a thin native oxide layer on its surface. This oxide rapidly covers the surface to a thickness of 0.5–1 nm (5–10 Å). Growth then slows down and effectively stops after a few hours, with a final thickness on the order of 1–2 nm. Both the growth rate and the final native oxide thickness depend on the surface preparation and in particular on the presence or absence of chemical residues from cleaning procedures. For example, the SC-1 and SC-2 cleaning procedures we discussed in Chapter 4 create chemical oxides on the silicon surface which are typically 1–2 nm thick.

**Figure 6-1** Uses of SiO₂ in silicon technology.**Table 6-1** Future projections for silicon technology taken from the SIA NTRS [6.1]

Year of First DRAM Shipment	1997	1999	2003	2006	2009	2012
Minimum Feature Size (nm)	250	180	130	100	70	50
DRAM Bits/Chip	256M	1G	4G	16G	64G	256G
Minimum Supply Voltage (volts)	1.8–2.5	1.5–1.8	1.2–1.5	0.9–1.2	0.6–0.9	0.5–0.6
Gate Oxide T _{ox} Equivalent (nm)	4–5	3–4	2–3	1.5–2	<1.5	<1.0
Thickness Control (% 3σ)	± 4	± 4	± 4–6	± 4–8	± 4–8	± 4–8
Equivalent Maximum E-field (MV cm ⁻¹)	4–5	5	5	>5	>5	>5
Gate Oxide Leakage (DRAM) (pA μm ⁻²)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Tunnel Oxide (nm)	8.5	8	7.5	7	6.5	6
Maximum Wiring Levels	6	6–7	7	7–8	8–9	9
Dielectric Constant, K for Intermetal Insulator	3.0–4.1	2.5–3.0	1.5–2.0	1.5–2.0	<1.5	<1.5

The most critical application of insulators in CMOS technology is as the gate insulator. As shown in Table 6-1, these layers are typically 3–5 nm thick in current state-of-the-art technology and are projected to be < 1 nm in another 10–15 years. Even today, these layers are only about 25–50 atomic layers thick. It is quite remarkable that VLSI chips can be manufactured today with tens of millions of such oxides, each defect free and capable of reliably sustaining electric fields within a factor of about two of the material limit of SiO₂ which is about 10–15 MV cm⁻¹. Notice also in Table 6-1 that gate insulators must be controlled in thickness to essentially single atomic distances. Continued scaling will make it very difficult for SiO₂ films to meet all these requirements. As a result, it is very likely that the very thin gate insulators projected for the fu-

ture will not be pure SiO_2 films. Oxynitride films (SiO_2 with nitrogen incorporated through a variety of processes) are the most likely near-term solution.

With films as thin as those projected in the NTRS, the term “insulator” begins to lose some of its meaning because such thin films will conduct finite amounts of current due to quantum mechanical tunneling of electrons or holes. This process occurs because of the wave nature of electrons and basically allows electrons to pass through an otherwise impenetrable barrier such as SiO_2 , if the barrier is thin enough. The probability of such an event happening decreases exponentially with oxide thickness, so tunneling oxides must be quite thin for significant currents to flow. If a polysilicon gate is deposited on top of such an oxide and electrically isolated by surrounding it with SiO_2 on all sides, then the electrons which tunnel through the thin oxide underneath the polysilicon can be trapped in the poly layer, allowing it to take on a negative charge. This can be used as the basis for a nonvolatile memory element. In fact, oxide layers between 4 and 10 nm are often used in electrically programmable memories and are referred to as tunneling oxides in these applications.

In gate dielectrics, tunneling is undesirable and in some applications like DRAM pass transistors is a major problem since it directly affects data retention. The specifications on leakage current shown in Table 6-1 for DRAMs are likely not achievable in SiO_2 layers thinner than about 4 nm and so it is likely that thicker gate dielectrics will be used in these applications. For more general-purpose logic applications, higher levels of gate leakage current can be tolerated.

Many of the other applications listed in Figure 6-1 were discussed in detail in Chapter 2. Thermal oxides in the range of 10–30 nm are often used under Si_3N_4 layers as stress relief or “pad” oxides during LOCOS type processes. We saw several examples of this application in the CMOS process in Chapter 2. Thicker thermal oxides often serve as masks for ion implantation or gas-phase doping steps. Here it is usually only important that the oxide layer be thick enough to block dopant penetration into the underlying silicon substrate. We will consider these applications more carefully in Chapters 7 and 8. Thick oxides may also serve as “field” or isolation oxides in many technologies. In this case they provide lateral isolation between adjacent components on the wafer surface.

It is also quite possible to deposit SiO_2 layers using CVD or LPCVD techniques. For back-end applications, after the wafers have metal on them, deposition is the only option. In fact for any application in which the oxide must be placed on top of underlying films, deposition is generally the only option since silicon may not be available from the underlying films for SiO_2 thermal growth. Deposition usually involves a much smaller “thermal budget” compared to thermal oxidation and hence is preferred even in front-end processes whenever it is important to limit the temperature cycle the wafers are subjected to. Deposited oxides are usually not used for layers thinner than about 10 nm because control of the deposition process is not as good as the thermal oxidation process. The interface between a deposited oxide and the underlying silicon is not as perfect electrically as that formed by a thermal oxide either, so thermal oxides are usually used for gate dielectrics and other critical oxides in device structures. It is possible to anneal a deposited oxide/Si interface, however, to produce electrical properties close to those of a thermally grown oxide. The simplest way to “anneal” such an interface is simply to grow a thin thermal oxide underneath the deposited oxide. This is straight-

forward to do because new SiO₂ always grows at the Si/SiO₂ interface, as we will see in the next section.

Finally, note that the NTRS in Table 6-1 specifies a dielectric constant less than 3.9 (SiO₂) for intermetal dielectrics in technology generations in the future. This is required in order to improve the speed of interconnects (RC delay) and implies that deposited films will not be SiO₂ in back-end processes in the future. We will discuss so-called “low K” deposited dielectrics in Chapter 11. In contrast to back-end “low K” dielectrics, gate oxides are likely to use “high K” dielectrics in the future. Thus the equivalent electrical gate oxide thicknesses in Table 6-1 may be achieved with higher K materials physically thicker than the values in the table.

This chapter will focus primarily on the thermal oxidation process and the properties of the Si/SiO₂ interface. Deposited oxides and other deposited thin films are described in detail in Chapter 9.

6.2 Historical Development and Basic Concepts

A number of experiments done in the past 30 years [6.2-6.5] have conclusively shown that when silicon is oxidized, the oxidation process occurs at the Si/SiO₂ interface as illustrated in Figure 6-2. Because of this, a new interface is constantly forming and moving inward down into the silicon substrate. Silicon oxidation therefore occurs by the inward diffusion of the oxidant, rather than the outward diffusion of silicon. Conceptually, the simplest way to demonstrate this is to grow an SiO₂ layer with one oxygen isotope and then continue the growth with a second isotope (O¹⁶ and O¹⁸, for example). Profiling through the resulting composite oxide with a mass sensitive technique can then determine whether the second oxide grows at the top or bottom surface. This and other related experiments have consistently shown that the new oxide grows at the Si/SiO₂ interface.

Conceptually one can think of this process in the following way. The Si atoms in the substrate are bonded to other Si atoms. These bonds must be broken, oxygen atoms inserted between the silicon atoms, and finally Si-O bonds formed. The process involves a volume expansion because of the room needed for the oxygen atoms. As illustrated in

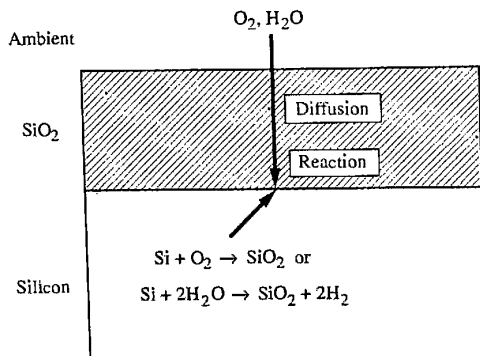


Figure 6-2 Basic process for the oxidation of silicon. The chemical reaction takes place at the Si/SiO₂ interface.

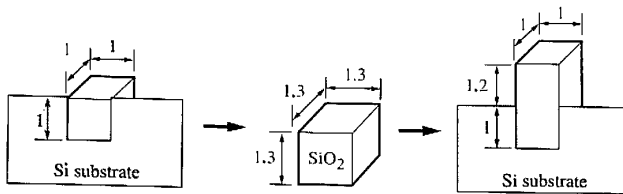


Figure 6-3 Volume expansion that occurs during silicon oxidation. A unit volume of silicon on the left is transformed into SiO_2 . In the middle the volume expansion is unconstrained; on the right the substrate restricts the expansion to one dimension.

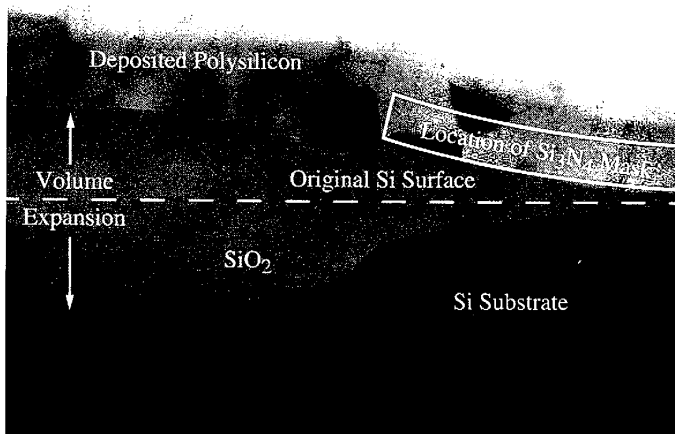


Figure 6-4 Scanning electron micrograph of a simple LOCOS structure. The silicon oxidation was masked on the right side by a Si_3N_4 layer. A polysilicon layer has been deposited on top of the structure, after the Si_3N_4 layer was stripped, to provide contrast in the picture. Note that the grown oxide on the left side occupies approximately twice the volume of the silicon consumed in the oxidation. Courtesy of J. Bravman, Stanford University.

the center of Figure 6-3, the oxide would like to expand by 30% in all three dimensions to accommodate the oxygen atoms. However because the silicon substrate prevents expansion in the two lateral directions, the only option is for the oxide to expand upward as shown on the right of Figure 6-3. The volume is thus accommodated by a 2.2 times upward expansion of the oxide compared to the volume of the silicon oxidized.

Figure 6-4 illustrates this in a real two-dimensional structure. This example is a SEM cross section of a simple LOCOS process like the one we described in Chapter 2. On planar surfaces, most of the volume expansion is accommodated simply by the oxide growing up above the original silicon surface as in Figure 6-3. On shaped surfaces, which are common especially in advanced silicon device structures, the expansion may

not be so easily accommodated. LOCOS results in a nonplanar Si/SiO₂ interface between the nitride masked and oxidized regions and an obvious transition region between the two. This transition region takes on a characteristic “bird’s beak” shape because of lateral oxidation under the nitride mask.

Oxide layers grown on silicon are amorphous. This might seem surprising initially because we are oxidizing a single-crystal substrate and SiO₂ exists in both crystalline and amorphous phases. However, there are no crystalline forms of SiO₂ whose lattice size closely matches the silicon substrate. Any attempt to grow single-crystal SiO₂ on silicon would therefore result in very large stresses. Such stresses could easily be large enough to generate crystallographic defects in the silicon, which would significantly degrade circuit yield and device performance. Even though the oxide that grows is amorphous, it does have short-range order, which is illustrated in Figure 6-5. SiO₄ tetrahedra are the basic units from which SiO₂ forms. These tetrahedra bond together by sharing oxygen atoms as illustrated on the right of the figure. Such shared atoms are called bridging oxygen atoms. In the amorphous forms of SiO₂ there may also be some non-bridging oxygen atoms present. These phases are often referred to as fused silica. Crystalline forms of SiO₂ such as quartz contain only bridging oxygen bonds. The various crystalline and amorphous forms of SiO₂ arise because of the ability of the bridging oxygen bonds to rotate, allowing the position of one tetrahedron to move with respect to its neighbors. This same rotation is what allows the material to lose long-range order and hence become amorphous. At normal processing temperatures, the thermodynamically stable form of SiO₂ is one of the crystalline forms. These forms are rarely observed in IC structures, however, because the time required for amorphous SiO₂ to rearrange itself into a crystalline form is very long compared to normal diffusion, oxidation, or anneal cycles. Occasionally fused silica furnace tubes that are used in oxidation or diffusion furnaces will crystallize (a process called devitrification) after many months or years at high temperature.

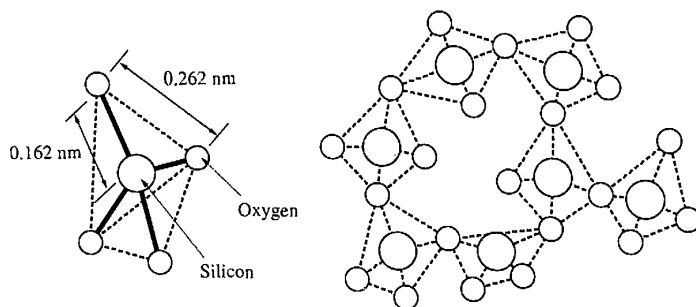


Figure 6-5 Structure of fused silica glass. The silicon atoms are in the center of each of the tetrahedra. The O-O distance is 0.262 nm; and the Si-O distance is 0.162 nm. The Si-Si bond distance depends on the particular form of SiO₂, but is about 0.31 nm. The six-membered ring structure of SiO₂ is shown conceptually on the right.

The oxide layers that grow on Si are normally in compressive stress. At the Si/SiO₂ interface, the growing oxide can expand only upward, not laterally since it must attach to the silicon substrate. This results in compressive stresses that can be as large as 5×10^9 dynes cm⁻². At temperatures above about 1000°C, the oxide can relieve some of this stress by viscous flow; at lower temperatures, the oxide viscosity is too high to allow much relief to take place. In addition to these intrinsic stresses which can be present after an oxide is grown, there is also a large difference in the thermal expansion coefficients of Si and SiO₂. As the wafers are cooled down after an oxide is grown, this results in an additional compressive stress in the oxide that can be as high as a few $\times 10^9$ dynes cm⁻². These two effects place the silicon substrate in tension; however, the substrate is many times thicker than the oxide layer and as a result, the tensile stresses are scaled down in proportion to the thickness ratio. These effects can easily be observed even on flat silicon wafers. After an oxidation, SiO₂ is normally present on both the top and back surface of the wafer and the stresses from the two oxide films therefore balance. If the oxide is chemically stripped from one surface, however, a distinct bow will be observed in the wafer because of the stresses. In fact measurement of this "wafer curvature" is one way to determine the stress levels in SiO₂ or other thin films on silicon substrates.

Figure 3-15 showed a high-resolution Transmission Electron Micrograph (TEM) of the Si/SiO₂ interface. The atomic resolution in this image clearly shows the silicon crystal structure, the amorphous SiO₂ layer, and the abruptness of the interface. A number of recent studies using TEM and other techniques have shown that the Si/SiO₂ interface is normally characterized by a transition region which is very narrow (perhaps only one atomic distance). The interface is usually quite flat, although there are steps of one atomic distance which occur occasionally. The abruptness of the interface depends to some extent on the process conditions used to grow the oxide and the interface is a little rougher for oxides which are rapidly grown (H₂O versus O₂ oxidation or high pressure oxidation, for example). The interface is also a little rougher for oxides grown at low temperatures.

The growth kinetics of SiO₂ layers on silicon have been studied for more than three decades and continue to be studied today. Early studies variously reported the growth rate to have a linear, parabolic, logarithmic, inverse logarithmic, or power-law dependence on time. The first widely accepted model of the growth kinetics was due to Deal and Grove in 1965 [6.6]. This work showed that over a wide range of conditions, the growth followed a linear parabolic law which could be explained on a solid theoretical basis. Later work has generally focused on modeling growth conditions for which the Deal-Grove model is inadequate. Examples include very thin oxides, oxides grown in mixed ambients, oxides grown on 2D or 3D silicon surfaces, and oxides grown on heavily doped substrates. We will discuss these and other models later in this chapter.

There has also been for many years a considerable controversy over which oxidant species diffuse through the SiO₂ to grow new layers of oxide. In the gas phase above the wafer, molecular species (O₂ or H₂O) are present. However, these species could dissociate or even ionize to form O, O⁻, O₂⁻, or many other species once they enter the SiO₂. Some early work [6.2] using an electric field suggested that a charged species was the dominant diffusant. However, later work largely discounted these results and today

most workers believe that neutral O₂ and H₂O and/or OH are the dominant species involved in the oxidation process.

The electrical properties of the Si/SiO₂ interface have been studied intensively for more than 40 years. A great deal has been learned about these properties, how they depend on process conditions, and how to control them. To first order, the interface is perfect. The densities of defects that are present in modern devices are on the order of 10^9 – 10^{11} cm⁻² compared to a silicon surface atom density of about 10^{15} cm⁻². Most of the defects that do exist are usually attributed to incompletely oxidized Si atoms or Si atoms with dangling or unsatisfied bonds. However, only about 1 atom in 10^5 has such a defect.

In 1980, in an effort to unify research in this field, Deal [6.7] suggested the nomenclature shown in Figure 6-6 to represent the various types of electrical defects that are found experimentally at the Si/SiO₂ interface and in SiO₂ layers. There are four basic types of defects or charges that exist. The first, Q_f , is known as the fixed oxide charge. Experimentally we find that a sheet of positive charge (usually 10^9 – 10^{11} cm⁻²) exists in the oxide, very close to the interface. It seems to be located within 2 nm of the interface (perhaps closer) and is likely associated with the transition from Si to SiO₂. Most physical explanations for Q_f suggest that it is due to incompletely oxidized Si atoms that have a net positive charge. Q_f is called the fixed oxide charge because its charge state does not change during normal device operation. It is positive and invariant under normal conditions.

The second type of charged defect present at the Si/SiO₂ interface is Q_{it} , the interface trapped charge. The physical origin of these charges is often suggested to be similar to Q_f . That is, Q_{it} is likely due to some type of incompletely oxidized silicon atom with unsatisfied or dangling bonds located in the oxide, but very close to the interface. However there is a very important difference between Q_{it} and Q_f . The charge associated with Q_f is fixed and positive. The charge associated with Q_{it} may be positive, neutral, or negative and in fact may change during normal device operation because of the capture of holes or electrons. Hence the name trap associated with Q_{it} . These traps behave very much

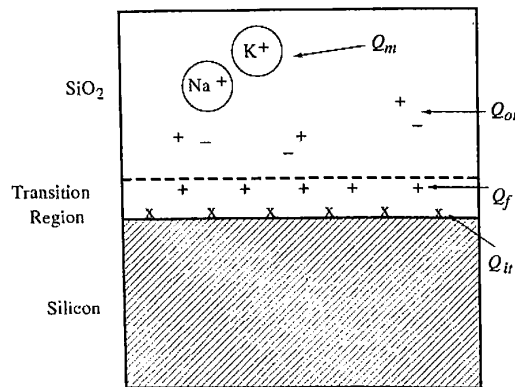


Figure 6-6 Charges associated with the SiO₂/Si system. (After Deal [6.7]).

like the bulk deep-level traps we discussed in Chapters 1 and 4 (Figure 1–27). Energy levels associated with Q_{it} exist throughout the forbidden band, although usually there are more traps at energy levels near the conduction and valence band edges than there are in the middle of the bandgap. Oxidizing a silicon surface usually results in a density of Q_{it} on the order of 10^9 – 10^{11} $\text{cm}^{-2}\text{eV}^{-1}$, about the same density as is found for Q_f . In fact it is usually the case that a process that results in a high value of Q_f will also result in a high density of Q_{it} . This correlation is one of the experimental results that suggests a common origin for the two charges.

The other two types of charges shown in Figure 6–6 are usually less important today. Q_m is the mobile oxide charge that may be located anywhere in the oxide, and was a serious problem in the 1960s. At that time, it was often the case that MOS structures were unstable after fabrication. We discussed this briefly in Chapter 1 and again in Chapter 4 and observed there that this instability in transistor threshold voltage was traced to the presence of mobile ions like Na^+ and K^+ in the gate oxides— Q_m . With proper attention to cleanliness in wafer fabrication facilities, this problem largely disappeared by the 1970s, although even today it occasionally becomes a problem in manufacturing facilities since ppm concentrations of these mobile ions can cause measurable instabilities. In the example we considered in Chapter 4, we saw that the shift in V_{TH} in MOS devices caused by Na^+ or K^+ is inversely proportional to C_{ox} [Eq. (4.1)]. Thus as gate oxide thickness is scaled down, larger Q_m values can be tolerated.

Q_{ot} , or oxide trapped charge, which also may be located anywhere in the oxide, is the final experimentally observed charge type in the Si/SiO₂ system. These defects are likely broken Si–O bonds in the bulk of the oxide, well away from the Si/SiO₂ interface. Such bonds can be broken by ionizing radiation or by some of the process steps used in manufacturing ICs today. Plasma etching, for example, exposes oxides to energetic ions, electrons and other neutral species. Ion implantation is often done through an oxide layer. These and other processes can damage oxides resulting in traps in the bulk oxide— Q_{ot} . Such traps are normally repaired by a high-temperature anneal before device fabrication is complete which allows the broken bonds to repair themselves. If they exist in the oxide because they are not fully annealed, or because the device is exposed to ionizing radiation, these traps can capture holes or electrons that may be injected into the oxide during device operation, resulting in trapped charge. Hence the name Q_{ot} .

Q_{ot} has taken on increased importance in recent years because of the high electric fields present in scaled devices. These higher fields result in more energetic or “hot” carriers that can achieve energies high enough to be injected into the gate oxides of modern MOS devices. If oxide traps are present, or if they are created by the energetic carriers themselves, charge trapping can occur. This results in device threshold shifts with time [see Eq. (4.1)] and reliability concerns. A common measure of the quality of a gate insulator is Q_{BD} or the amount of charge that can be passed through the oxide before failure (breakdown) occurs. Charge trapping in SiO₂ insulators is also a major issue in programmable devices like EPROMs in which current is purposely passed through the gate oxide as part of the write operation.

All four of the charges in Figure 6–6 can have deleterious effects on device operation. As a result, great care is normally taken during fabrication to choose process sequences that will minimize these charges. Generally this is accomplished by high-temperature in-

ert anneals in Ar or N₂ toward the end of the process flow, and by a final moderate temperature ($\approx 400^\circ\text{C}$) anneal in H₂ or forming gas (N₂/H₂) at the end of the process. We will discuss such anneals and charges more carefully in Section 6.5.13 later in this chapter.

A final point with regard to SiO₂ layers on Si is important. One of the earliest and most important uses of SiO₂ layers was for masking against impurity diffusion. Oxides are still widely used for this purpose today and this use depends fundamentally on the fact that most impurities have much lower diffusivities in SiO₂ than they do in Si. It is reasonably simple to calculate the thickness of an oxide layer needed to successfully mask against a diffusion or ion implantation operation. However, we will defer discussion of these points until Chapters 7 and 8 where we discuss diffusion and ion implantation in detail.

The appendix contains a number of useful and important properties of three insulators commonly used in silicon technology. SiO_xN_y is an insulator usually formed by exposing an SiO₂ layer to an NH₃ or other nitrogen containing ambient to convert the oxide into an oxynitride. The properties of both oxynitrides and Si₃N₄ vary somewhat with deposition and/or nitridation conditions and will be discussed in more detail in Section 6.5.11 and at the end of the chapter when future trends are discussed.

6.3 Manufacturing Methods and Equipment

Oxidation systems are among the simplest types of semiconductor processing equipment. Conceptually all that is needed, as illustrated in Figure 6-7, is an oven capable of temperatures from 600–1200°C and a simple gas distribution system capable of introducing O₂ or H₂O. In practice, such systems are far more complex because of the need for uniformity, reproducibility, and cleanliness in the process. Modern furnaces are capable of handling up to several hundred 8" wafers, with a temperature uniformity of $\pm 0.5^\circ\text{C}$.

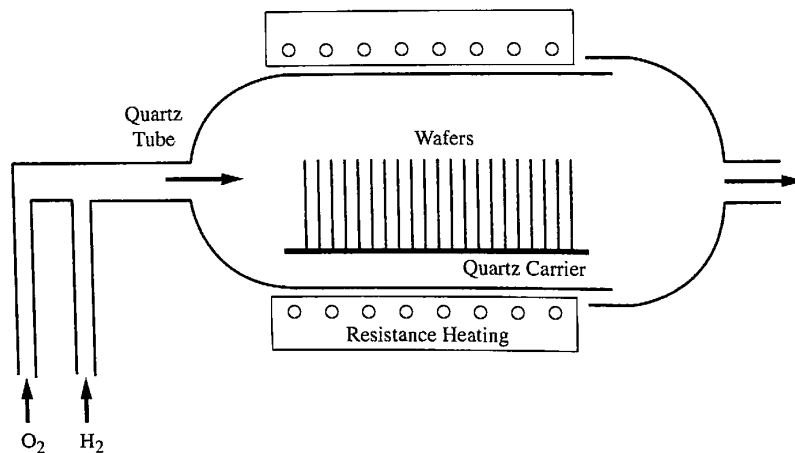


Figure 6-7 Conceptual silicon oxidation system.

Most such systems in use today are horizontally oriented with the oxidant species introduced at the back end and the wafers at the front. Many recently developed furnaces are vertically oriented since these take up less floor space in a manufacturing facility. Liquid sources of O_2 and H_2 are commonly used, with the holding tanks located outside the manufacturing facility. Gases from these liquid sources are carried to the furnaces through high-purity stainless steel lines. For O_2 oxidations, the oxygen is injected directly into the furnace. For H_2O oxidations, O_2 and H_2 are burned in the back end of the furnace to produce H_2O . Additional source lines to the furnace may allow for gaseous HCl or TCA (trichloroethane), both of which provide a Cl source. The Cl may be used to clean the furnace tube prior to oxidation or may actually be used during the oxidation as a small percentage additive. Cl serves the useful purpose of reacting with many unwanted metal ions and producing gas phase byproducts that can be exhausted from the furnace. (Recall our discussion in Chapter 4 of cleaning procedures. HCl is a component of the SC-2 cleaning solution.) If the HCl or TCA are used during the oxidation, some Cl is incorporated in the growing SiO_2 layer. This provides some degree of protection from trace amounts of ionic contaminants such as Na^+ and also changes the growth kinetics. We will discuss these issues more fully later in Section 6.5.6.

The wafers are normally loaded into the furnace on carriers or “boats” which hold 10–50 wafers. These boats are carried into the furnace by automatic loading systems that are usually cantilever arms that move slowly into the furnace hot zone. Early horizontal oxidation systems used sleds, sometimes with quartz wheels that were pushed into the furnace hot zone. However, such systems invariably stir up particles from the furnace system that can deposit on the wafers causing yield loss. Today cantilever systems carry the wafers into the furnace without actually touching the furnace walls. In the case of vertical furnaces, elevators are used to carry the wafers up into the furnace hot zone.

The furnace temperature control system is simple in concept. The furnace is normally divided in three to five zones for temperature control, with the wafers located in the center section during an oxidation. The outer zones are designed to help compensate for heat losses out the ends of the tube, so that a long central section with uniform temperature can be maintained. Thermocouples monitor the temperature in each of the sections and a controller delivers power to the resistive heating elements in each section to maintain a flat temperature profile in the central zone. In order to avoid large thermal gradients across the wafers, which can induce crystallographic defects, the wafers are normally loaded at a moderate temperature (about $800^\circ C$) and the furnace is then ramped up to the oxidation temperature after the wafers are in the central hot zone. The furnace controller is usually designed to control the ramp up and ramp down rates so that thermal stress on the wafers is minimized. Ramp rates on the order of $1^\circ C \text{ sec}^{-1}$ are typical.

At $1000^\circ C$, oxidation rates in H_2O are on the order of 0.1 nm sec^{-1} and approximately double for a $100^\circ C$ temperature rise. As a result, temperature control on the order of $\pm 0.5^\circ C$ and time control on the order of seconds are required to produce consistent oxide thicknesses from batch to batch.

The vast majority of oxidation furnaces today operate at atmospheric pressure since these are the simplest to operate. However, we will see in Section 6.5.3 that oxide growth rates are proportional to oxidant pressure (to first order), so it is possible to grow a given oxide thickness in less time at high pressure. Such systems exist today and

are used in applications in which a thick thermal oxide must be grown with minimum thermal budget (Dt). Pressures lower than atmospheric are simple to obtain by diluting the oxidant gas with an inert gas such as Ar or N₂. In this case, the oxidation rate is slowed down, approximately in proportion to the oxidant partial pressure.

The effort to continue decreasing the size of VLSI devices has in recent years demanded very short high-temperature steps. In addition, gate oxides in modern MOS devices are < 10 nm thick, requiring short, well-controlled growth cycles. These issues have resulted in the investigation of Rapid Thermal Oxidation (RTO) techniques for growing very thin SiO₂ layers. Typically RTO systems use a lamp heated chamber which can heat a wafer up to oxidation temperature at a rate of 100°C sec⁻¹, hold the wafer there while the oxide is grown, and then cool it back down to room temperature, again in a few seconds. Usually such systems are single-wafer machines. When reports of RTO oxide growth kinetics first began to appear in the literature, there were suggestions that the growth kinetics were different from those observed in oxidation furnaces and that new models would be needed to explain the kinetics. This might be the case, for example, if transient effects, which are negligible in a furnace which takes 15 minutes to heat the wafers, become important when the system time constants are measured in seconds. However, it now appears that the conventional models may in fact be satisfactory. One of the very difficult problems in an RTO system is in knowing exactly what the wafer temperature really is. These systems usually support the wafer on a small thermal mass in order to be able to heat the wafer rapidly. This makes it very difficult to use thermocouples for temperature measurement as is done in a furnace. Other techniques such as using a pyrometer are often used, but these methods have their own set of problems. The result is that some of the early RTO experiments which suggested "new" kinetics may have done so because of uncertainties or variations in the wafer temperature. Additional work in the future should clarify whether or not transient effects will require additions to the standard oxidation models.

Some oxidation furnaces today have a "fast ramp" capability ($\approx 10^\circ\text{C sec}^{-1}$) which provides an intermediate capability between conventional furnaces and RTO systems. These new furnaces provide very good temperature control through the use of conventional thermocouples for temperature measurement, along with the fast ramp rates consistent with shorter Dt cycles and better control of thin oxides.

6.4 Measurement Methods

Experimental measurements of the results of an oxidation procedure are an essential part of any manufacturing process. The parameters of interest for an oxide (or other dielectric) layer are usually thickness, dielectric constant, index of refraction, dielectric strength, and defect density. The uniformity of all these parameters across a wafer and from wafer to wafer are also important.

The available measurement techniques can be broadly grouped into three classes. The first of these involves physical measurements that are often destructive. The second involves optical techniques that can measure some, but not all, of the parameters of interest. They are normally nondestructive. The third class of measurements involves electrical measurements. These are also usually nondestructive and are potentially the

most powerful because they directly measure the parameters that are of interest in electrical devices.

6.4.1 Physical Measurements

In one of the simplest physical measurements, the dielectric is etched away in some regions using a chemical etchant that attacks the thin film but not the underlying substrate (HF for an SiO_2 layer). A small needle stylus is then moved across the surface step corresponding to the etched edge of the film. Mechanically or electrically amplifying the small signal produces a measurement of the step height. Instruments of this type have been commercially available for many years and have a resolution below 10 nm. The recent development of the scanning tunneling microscope, the Atomic Force Microscope (AFM), and derivative instruments has pushed the resolution of these “stylus” type tools to atomic dimensions. Figure 1–3 showed an example of the capability of these instruments. Other physical techniques for measuring film thickness include cross sectional SEM images similar to Figure 6–4 that can image film thickness, or higher-resolution TEM cross-sectional images such as that shown in Figure 3–15. All of these techniques require sample preparation and as a result are not well suited to in-process measurements on a manufacturing line. They also only provide information on film thickness.

6.4.2 Optical Measurements

Optical techniques are very widely used to measure film thickness and a number of instruments are commercially available. Many depend on the measurement of reflected light from the sample as illustrated in Figure 6–8. If monochromatic light of wavelength λ is incident on the sample surface at an angle ϕ , some light will be directly reflected. If layer x_0 is transparent, some light will also reflect from the lower interface. For some values of λ , the two reflected waves will be in phase and will add; for other values of λ ,

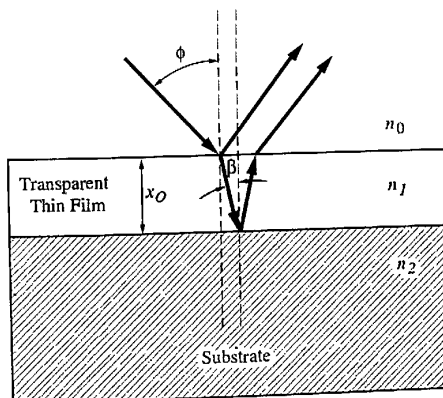


Figure 6–8 Light reflection from a sample with a transparent thin film on its surface. n_0 is the index of refraction of air (1.0), n_1 is the index of refraction of the thin film and n_2 is the index of refraction of the substrate. ϕ is the angle of the incident light, β is the angle of the reflecting light at the bottom interface.

destructive interference will occur. The result is that the intensity of the reflected light will go through minima and maxima as λ is varied. These effects are exactly the same as those that cause standing waves during photoresist exposure. (Recall Figures 5-23 and 5-24.) The maxima and minima occur at values of λ given by

$$\lambda_{min,max} = \frac{2n_1x_0\cos\beta}{m} \quad (6.1)$$

where

$$\beta = \sin^{-1} \left[\frac{n_0 \sin \phi}{n_1} \right] \quad (6.2)$$

and $m = 1, 2, 3, \dots$ for maxima and $1/2, 3/2, 5/2, \dots$ for minima. These are simply generalized versions of the simple $\lambda/2n_1$ periodicity we used for the standing waves in photoresist in Chapter 5 for vertically incident light. Commercially available instruments use spectrophotometers to sweep the incident wavelength with ϕ constant. They then measure dielectric film thicknesses by fitting the resulting reflected intensity to Eq. (6.1). This technique works reliably for film thicknesses greater than a few tens of nm. For thinner films, it is difficult to detect the first minimum unless very short wavelength light is used. In using this technique, the index of refraction of the dielectric film must be known since it shows up directly in Eq. (6.1).

There are many cases of interest in silicon technology in which dielectric films thinner than a few tens of nanometers need to be measured, or in which the film dielectric constant is not known precisely. Examples include gate oxides in MOS structures that are now less than 10 nm in thickness and Si₃N₄ or SiO_xN_y films in which n_1 may vary with processing conditions. For these applications, ellipsometry is a better method to measure film properties. Conceptually, ellipsometry works in a similar fashion to the reflectance technique described above. However, polarized light is used in an ellipsometer and the change in the polarization when the light is reflected from the dielectric/substrate interface is measured. In general the change in polarization depends on the properties of both the film and the substrate. However in the simple case when the optical properties of the substrate are known and the film is transparent at the wavelengths being used, the change in polarization of the reflected light depends only on the film thickness and index of refraction. Commercially available ellipsometers can easily measure films down to 1 nm in thickness and accurately determine the index of refraction. Usually these instruments are computer controlled so that the film properties are calculated automatically.

By far the simplest optical technique to evaluate dielectric film thickness is the use of color charts. These charts were originally described by Pliskin and Conrad [6.8] and are based on the idea that if white light is used to illuminate a thin transparent dielectric film on a reflecting substrate (as in Figure 6-8), destructive interference will occur for some wavelengths in the reflected light, giving the reflected light a characteristic color. This allows a simple observation of the color of the oxide or nitride layer to provide an estimate of the film thickness. Tables in the appendix provide these colors for SiO₂ or Si₃N₄ films on reflecting substrates (usually silicon). The charts assume that the

films are being viewed perpendicularly under daylight fluorescent lighting. The colors repeat about every 300 nm for SiO_2 and about every 200 nm for Si_3N_4 . For this reason, standards are often used in conjunction with these charts. If the sample under investigation is viewed side by side with a standard of the same color and the observer then views the two at an angle, the colors of both will change as the viewing angle changes. The colors on both will change in the same way, only if the thicknesses are actually the same. This allows one to distinguish where on the color chart a particular sample fits. The eye is remarkably good at distinguishing colors and an estimate of film thickness within about 10–20 nm can be made with these charts, particularly if standards are available as well. The ranges covered by these charts are those normally encountered in silicon technology. Note however, that oxides thinner than about 50 nm do not have any characteristic color so this method is not useful in this thickness range.

6.4.3 Electrical Measurements — The MOS Capacitor

The third category of measurement techniques are those that involve electrical measurements. These are potentially the most powerful methods because they measure parameters that are of direct interest to semiconductor devices—capacitances, electrical charges, and so on. By far the most dominant electrical technique is the Capacitance-Voltage or CV method. This technique is very widely used and provides a large amount of information about dielectric films and the interfaces that such films make with underlying semiconductors. We will describe how this technique works and the information that it provides in the following paragraphs. We will apply the CV technique in this chapter to help understand the properties of the Si/SiO_2 interface. In later chapters (7 and 8 in particular), we will apply it to measurement of doping profiles in semiconductors.

The basic structure used to make CV measurements is the MOS capacitor, consisting of a semiconductor substrate, a dielectric layer and a conducting electrode. We briefly considered this structure in Chapter 4 (Figures 4–15 and 4–19) in connection with measurements of dielectric breakdown and carrier lifetime. Since the primary use of the MOS capacitor is in the characterization and study of insulators and the Si/SiO_2 interface, we will discuss this structure in more detail here.

Generally the MOS system is (polysilicon or Al) SiO_2 -Si although any conductor/insulator/semiconductor combination can be used. The basic measurement is illustrated in Figure 6–9 where we consider the specific case of a metal/ SiO_2 /N silicon structure. This figure shows three regions of operation of the MOS capacitor. Positive voltages on the metal electrode correspond to accumulation [Figure 6–9(a)], negative voltages first cause depletion [Figure 6–9(b)] and then inversion [Figure 6–9(c)] in the substrate. We will define these terms more carefully shortly. The metal part of the structure is often called the gate because that is the role it plays in a MOS transistor that has the same structure. It is important to note that the gate voltage V_G is a DC voltage and it is assumed in the following discussion to be applied for a long enough time for the system to come to a final steady state condition. We will return to this point later when we discuss high-frequency and low-frequency CV curves as well as other applications of the technique.

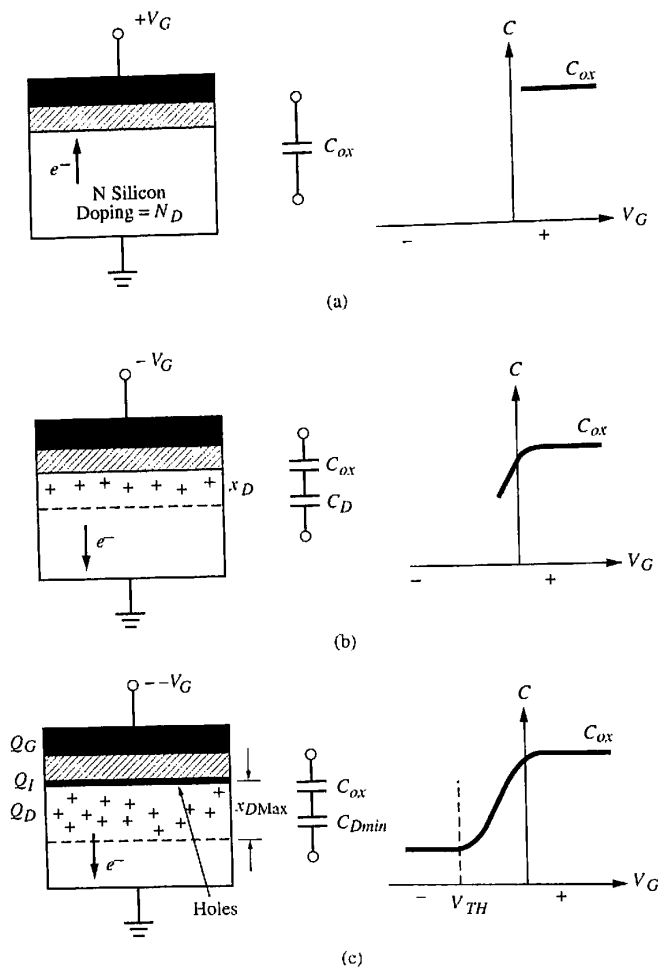


Figure 6-9 MOS capacitor structure and resulting CV plot. (a) corresponds to accumulation, (b) to depletion, and (c) to inversion.

We make the further assumption that the insulator and the insulator/semiconductor interface (SiO₂/Si in this case) have only small numbers of charges or defects. In other words, the four types of charges described in Figure 6-6 are assumed to be relatively small in density. If this is the case, electric field lines that start on charges on the gate will extend through the insulator and into the substrate where there are charges of opposite sign available to terminate the field lines. If there were large numbers of charges or defects in the insulator or at the insulator/semiconductor interface, the electric field lines could terminate on these charges before they reached the underlying semiconductor.

The result would be that applying voltages to the gate would have little effect on the semiconductor. Depletion and inversion could not be controllably achieved and the structure would not be useful in MOS transistors. This is in fact what happens in many insulator/semiconductor systems, but the $\text{SiO}_2\text{-Si}$ system is close to electrically perfect.

Consider first the case with a DC voltage of $+V_G$ on the gate [Figure 6-9(a)]. Since the substrate is N type, positive gate voltages will attract the majority carrier electrons to the silicon surface. If we now measure the small signal capacitance of the structure by superimposing a small AC signal on V_G , the capacitance measured will be simply the oxide capacitance C_{ox} . The silicon substrate will act simply as a resistor in series with C_{ox} since no depletion region will form for positive gate voltages. The measurement extracts only the capacitive part of the impedance, which will be independent of V_G , as shown in the right-hand part of Figure 6-9(a). Frequencies in the range of 100 kHz to 1 MHz are often used for this measurement.

Consider now the application of negative DC gate voltages in part (b) of the figure. Negative voltages will repel the majority carrier electrons from the surface creating a depleted region. Any negative charge we place on the gate must be balanced by a corresponding positive charge in the substrate to maintain charge neutrality. The positive charge is provided by the substrate donor atoms that have a net positive charge when the mobile electrons are forced away from the surface. Conceptually what happens is that the electrons forced away from the semiconductor surface region flow out the bottom of the substrate, through the power supply connected to the gate and then on to the gate where they provide the negative gate charge. As a result,

$$|Q_G| = |Q_D| = N_D x_D \quad (6.3)$$

where Q_G and Q_D are in units of number of charges cm^{-2} , and N_D is the doping in the substrate which is assumed to be uniform. A depletion region will form to a depth x_D . The concept here is very similar to the PN junction depletion region we discussed in Chapter 1. The depletion region will have a capacitance per unit area associated with it given by

$$C_D = \frac{\epsilon_S}{x_D} \quad (6.4)$$

where ϵ_S is the permittivity of silicon. However, x_D is a function of the gate voltage, increasing as V_G increases. The overall capacitance that is measured for the structure is now C_{ox} in series with a varying C_D . The plot on the right of Figure 6-9(b) reflects this, with a capacitance that decreases as V_G becomes more negative because x_D grows as V_G becomes more negative.

For larger values of negative DC gate voltage, the silicon surface will actually “invert” from N type to P type [Figure 6-9(c)]. The negative voltage on the gate will attract the minority carrier holes in the substrate to the surface and if enough of them are present there, they can form an inversion layer of P-type carriers. The gate voltage at which this occurs is called the threshold voltage and is exactly the same voltage that we discussed in earlier chapters in connection with turning on a MOS transistor. Once the inversion layer forms, x_D stops expanding and reaches a maximum value $x_{D\text{Max}}$.

For all regions of operation of the capacitor, the gate charge must be balanced by charge in the substrate. That is

$$Q_G = N_D x_D + Q_I \quad (6.5)$$

where Q_I is the charge density (number of charges per cm²) in the inversion layer. Q_I is negligible during the depletion condition so that x_D expands during depletion to balance Q_G . Once the inversion layer forms, however, additional gate charge is balanced by Q_I rather than Q_D so that x_D reaches a maximum value. This means that the CV curve will reach a minimum as shown in Figure 6-9(c).

In order to understand why additional Q_G is balanced by Q_I and not by Q_D once inversion occurs, we need to return to the concept of band diagrams that we introduced in Chapter 1. Figure 6-10 shows the band diagram for the MOS capacitor under conditions of accumulation, depletion, and inversion. The oxide prevents any current flow through the structure so the semiconductor is in equilibrium and we can draw the Fermi

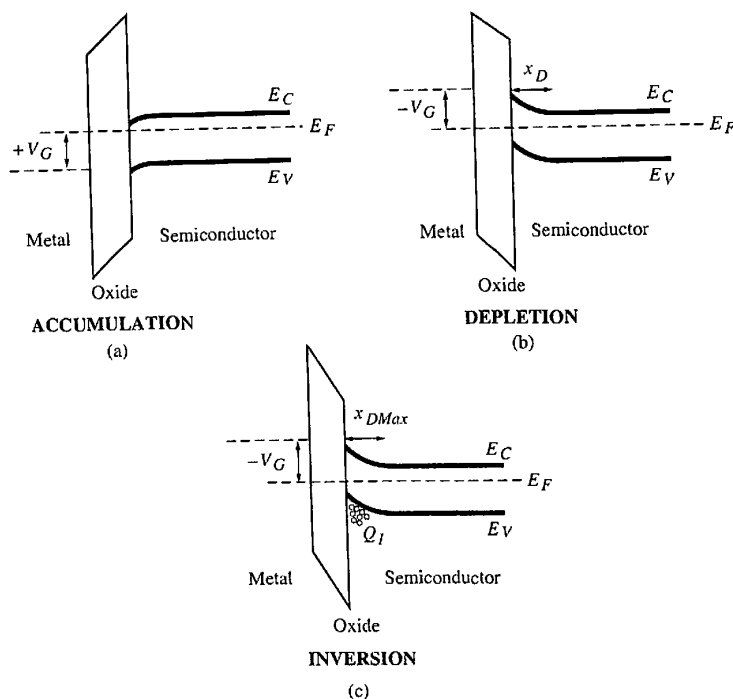


Figure 6-10 Band diagrams for the MOS capacitor in (a) accumulation, (b) depletion, and (c) inversion.

level as flat in the silicon. The Fermi level in the metal is separated from E_F in the silicon by the amount of the applied gate voltage. In Figure 6-10(a), a positive DC gate voltage is applied (recall that positive is downward in these diagrams by convention). This will cause the bands in the semiconductor to bend downward in the near surface region. This places E_F closer to E_C at the surface than is the case in the bulk far away from the surface. As we saw in Chapter 1, this means that the electron population is therefore higher at the surface than it is in the bulk. This accumulates electrons at the surface, the same conclusion we reached when discussing Figure 6-9(a).

In Figure 6-10(b), a negative DC gate voltage has been applied, causing the bands in the semiconductor to bend upward near the surface. This creates the depletion region that we discussed in Figure 6-9(b) since over the distance x_D , E_F is away from both E_C and E_V . This implies that both n and p are small, that is that the mobile carrier concentrations are much less than N_D . In this region, Q_D is given by Eq. (6.3).

Figure 6-10(c) shows the situation in inversion. Here the DC gate voltage is negative enough so that the bands at the surface are bent far enough to place E_F near E_V . This means that the hole population is significant at the surface, that is, that inversion has occurred and the surface is P type. We can now appreciate why additional Q_G is balanced by Q_I rather than by Q_D in inversion. Further increases in the negative voltage on the gate will try to bend the bands further in the semiconductor and move E_F closer to E_V . However, we saw in Chapter 1 that moving E_F nearer to E_V will result in an exponential increase in p . [See Eqs. (1.10) or (1.14) for example.] Because of this, only a very slight movement of E_F with respect to E_V is needed to provide enough Q_I to balance additional Q_G . Referring again to Eq. (6.5), we see that increasing Q_G after inversion is reached, will result in a very small increase in x_D and hence in Q_D , and that most of the Q_G increase will be balanced by an increase in Q_I . We therefore speak of x_D as being "pinned" at x_{DMax} once inversion occurs.

There are a few other basic concepts that we should understand about CV measurements before we discuss how they can be used to evaluate insulators and insulator/semiconductor interfaces. The CV curves we have described to this point are called high-frequency (HF) CV curves. This simply means that the small AC signal used to measure the MOS capacitance is a moderately high-frequency signal (typically 100 kHz to 1 MHz). We need to consider why this frequency is important.

Consider again the inversion condition in Figure 6-9(c) or Figure 6-10(c). Suppose that we have applied the DC gate voltage $-V_G$ for a long enough time for the system to come to equilibrium. (We will define how long this is shortly.) In equilibrium, the inversion layer is present and x_D is "pinned" at x_{DMax} . Suppose we now apply a small HF signal to measure the structure capacitance. The AC signal is superimposed on $-V_G$ and hence modulates Q_G very slightly. This change in Q_G must be balanced by a change in either Q_D or Q_I . In equilibrium, Q_I would change as we saw above. However, high frequency in this context means that Q_G is being modulated faster than Q_I can respond. If this is true, then the only option the system has is to modulate Q_D in order to provide the balancing charge. If this happens, then the charge balancing that occurs is between the gate and the bottom of the depletion region. The bottom edge of the depletion regions moves up and down along with the gate voltage, providing the ΔQ_D needed to

balance the ΔQ_G . Hence the capacitance that is measured is between those two plates (C_{ox} in series with C_D). For any $-V_G$ in the inversion region, x_D is at x_{DMax} and so C_D is at C_{DMin} . This is why the HF capacitance measurement produces a minimum capacitance for negative V_G that is independent of V_G .

Now why is it that Q_i cannot change as fast as the high frequency signal on the gate? The answer is simply that there is no source of additional holes available in the substrate to provide the ΔQ_i . There are very few holes present in an N-type semiconductor. The only mechanisms usually available to create or remove holes are thermal generation and recombination, respectively, and these processes are quite slow in materials like silicon which have long carrier lifetimes. As a result Q_i cannot change rapidly and the HF CV curve measures a constant minimum value in inversion.

Now suppose that we slow down the AC signal on the gate. If we reduce the frequency to a low enough value (typically less than 1 Hz for silicon), then generation and recombination processes can keep up with the AC signal and Q_i can follow the changes in Q_G . What do we now measure for the CV curve? The answer, not surprisingly, is that we now measure the Low-Frequency (LF) CV curve. If Q_i follows Q_G , then the capacitance measured in the structure will be just the oxide capacitance because the top and bottom plates of the capacitor where Q is changing are on the two edges of the oxide. Hence we should expect that the CV curve would come back up to C_{ox} in the inversion region. This is in fact what happens and the resulting curve is the LF CV curve. Figure 6-11 illustrates the HF and LF curves. Note that the point at which these ideal HF or LF CV curves reach their minimum corresponds to the onset of inversion and so is the threshold voltage for the MOS structure.

Figure 6-11 also illustrates the deep depletion condition that can be observed in MOS capacitors. We discussed this case qualitatively in Chapter 4 (Section 4.4.3) in connection with measuring the carrier generation lifetime in semiconductors. Deep depletion results if the DC bias voltage on the capacitor is swept too rapidly into the

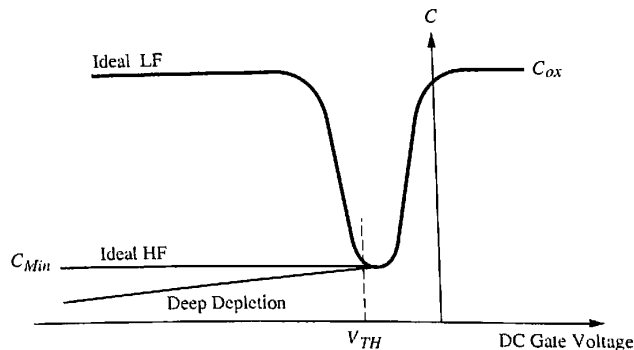


Figure 6-11 Ideal MOS CV curves at high and low frequencies. Also shown is the deep depletion CV curve.

inversion region. If this is done, the holes needed for the inversion layer cannot be generated rapidly enough to follow the applied “DC” bias. If Q_i cannot change rapidly enough, changes in Q_G can only be balanced by Q_D . Thus a deeper depletion region than would be present in equilibrium forms. A smaller capacitance than the equilibrium value is measured, which gradually relaxes back to C_{Min} as hole generation takes place.

Eq. (1.20) gave the net rate of recombination or generation from the Shockley-Read-Hall model. In a depletion region, n and p are negligible and U reduces to

$$U = -\frac{n_i}{\tau_G} \quad (6.6)$$

Assuming that this generation rate holds throughout the volume of the depletion region, then the current density corresponding to U is simply

$$J_{gen} = \frac{qn_i W}{\tau_G} \quad (6.7)$$

This current will flow to the surface, providing the holes for the inversion layer. In order to avoid deep depletion, the rate at which the “DC” bias voltage is swept on the MOS capacitor must be less than

$$\frac{dV}{dt} \leq \frac{J_{gen}}{C} \approx \frac{J_{gen}}{C_{OX}} = \frac{qn_i W}{\tau_G C_{OX}} \quad (6.8)$$

In practice, this means that “DC” sweep rates less than about 0.1 V sec^{-1} must be used.

Example In an experimental structure in Figure 6–12, a phosphorus N^+ region is formed by ion implantation using a 50-nm SiO_2 mask. A metal electrode is then

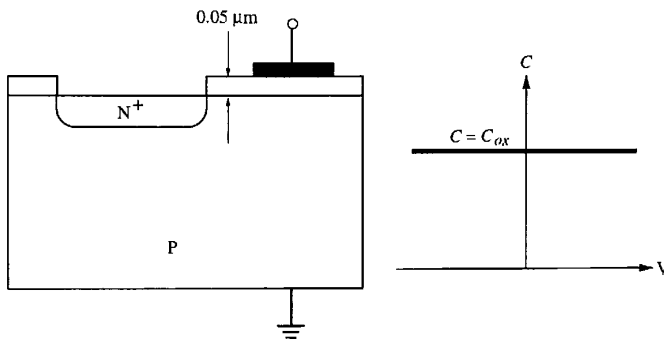


Figure 6–12 MOS capacitor structure with “anomalous” CV measurement.

formed as shown and a CV measurement is made in the region outside the N⁺ region. The measurement gives $C = C_{ox}$ for all values of applied voltage as shown. Explain what might have gone wrong in this experiment. That is explain why no surface inversion is observed in the CV measurement.

Answer While a number of explanations are possible, the basic observation here is that the gate voltage appears to have essentially no effect on the underlying silicon. This can happen if there are charges at or very close to the Si/SiO₂ interface which terminate the field lines before they penetrate into the silicon. This could be caused by a very poor Si/SiO₂ interface, but there is nothing in the process flow to suggest that this should have happened (unless there is damage from the implant that was not properly annealed). A more likely explanation is that the 0.05- μm oxide was not thick enough to mask the implant and as a result the N⁺ implant penetrated everywhere. If this happened, there would be plenty of donor atoms available to terminate the gate field very close to the surface. Thus as $+V_G$ is applied, the material tries to start depleting, but since the doping is so high close to the surface, the depletion region x_d is extremely shallow. Thus the capacitance associated with x_d is very large and the total capacitance (C_o in series with C_D), is essentially just C_o . We will discuss implantation and masking issues more carefully in Chapter 8.

We saw in Section 6.2 that a number of charges are generally associated with insulator/semiconductor interfaces. These charges cause shifts and distortions in CV curves because they require gate charge Q_G to balance them. Figure 6-13 illustrates these effects on the “ideal” HF CV curve. Q_f is a fixed positive charge that is present in the oxide.

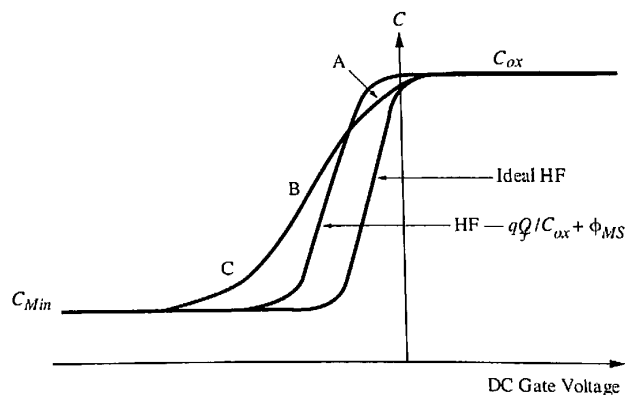


Figure 6-13 HF MOS CV curves illustrating some of the nonidealities that can be present in actual experimental structures. A, B, and C illustrate the effects of interface states with different energy levels in the silicon bandgap.

near the SiO_2/Si interface. Such a positive charge will induce a corresponding negative image charge in the silicon, effectively making the silicon surface more N type and therefore harder to invert to P type. This shows up as a lateral shift in the CV curve, with the magnitude of the shift simply given by qQ_f/C_{ox} . Measurement of this shift thus allows calculation of the magnitude of Q_f . An additional lateral shift occurs in general because of a difference in the work functions of the metal and the semiconductor ϕ_{MS} . The work function is by definition the energy required to remove an electron in the material at energy E_F to a position just outside the material. This is a measurable property of a material and is generally known for any given experiment. The lateral offset in the CV curve due to ϕ_{MS} is thus known. These two terms, ϕ_{MS} and qQ_f/C_{ox} , together account for an increase in the gate voltage required to cause inversion and are often lumped together $\phi_{MS} - qQ_f/C_{ox}$ and called the flat band voltage V_{FB} since if a gate voltage of this magnitude is applied, the bands in the semiconductor will be flat. These terms also show up in the turn-on voltage of a MOS transistor [see Eq. (4.1)].

If mobile charges (Q_m) or oxide trapped charge (Q_{ot}) are present, they produce a similar effect to Q_f . That is, they cause a lateral shift in the CV curve. The only real difference is that these charges are usually not located directly at the $\text{SiO}_2\text{-Si}$ interface and so the magnitude of the shift they cause is reduced in proportion to their distance away from the interface. (A positive charge located in the central portion of the oxide will induce a balancing negative image charge partially on the gate and partially in the silicon.)

The "distorted" CV curve shown in Figure 6-13 illustrates what happens when interface traps Q_{it} are present. We saw in Section 6.2 that these traps can have energy levels throughout the forbidden band. As applied gate voltage causes the semiconductor surface to move from accumulation to depletion to inversion, sweeping out a CV curve, E_F at the semiconductor surface will move from one band edge to the other (E_C to E_V in the N type example we have been considering). As a result, the Q_{it} traps will fill and empty as E_F moves through their energy levels. As this happens, the CV curve will distort from its ideal shape. The form of the distortion depends on the density and energy levels of the Q_{it} traps. In Figure 6-13 the region labeled A corresponds to interface states near E_C , B corresponds to states near the middle of the bandgap, and C corresponds to states near E_V . The assumption is often made that the interface states cannot follow the HF signal used to measure the capacitance. That is, they do not charge and discharge as the HF signal changes. But these traps will respond to the DC signal applied on the gate since we assume that this signal is applied for a long enough period of time to reach equilibrium. Thus Q_G is partially balanced by Q_{it} . The result is a stretching out of the CV curve along the horizontal (DC voltage) axis.

A number of methods have been developed to experimentally extract the interface state density from CV measurements on MOS capacitors. Many of these are described in detail in references like [6.9, 6.10]. The most commonly used method is the "quasi-static" or LF method that is illustrated in Figure 6-14. In this method HF and LF CV curves are superimposed from measurements on the same MOS capacitor. It is assumed that interface states cannot respond to the HF measurement and that they do respond to the low-frequency measurement. In practice, this means that the HF must be high enough that the traps cannot charge and discharge at the measurement frequency, and

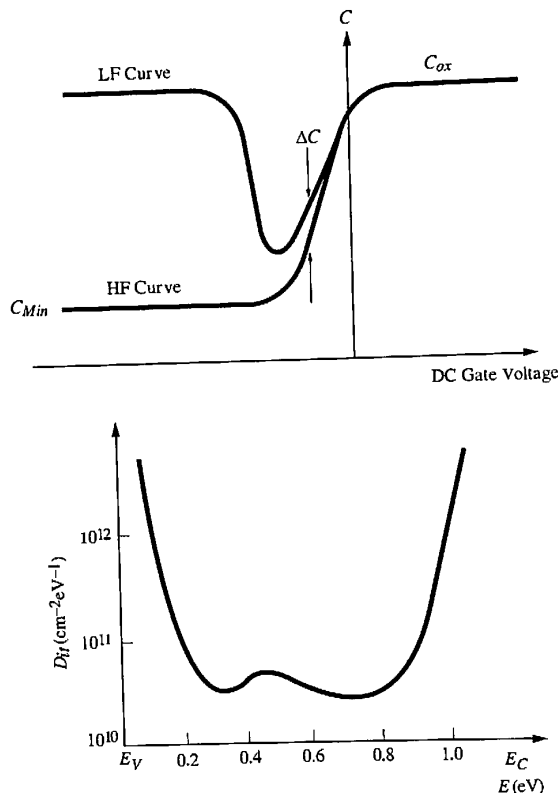


Figure 6-14 Typical experimental HF and LF CV curves used to extract the interface state density D_{it} as a function of energy in the bandgap. The D_{it} plot is typical of extracted data.

the LF must be low enough to allow the traps to charge and discharge as the LF signal varies. If this is the case, then the traps will behave as an additional capacitance for the low frequency measurement and ΔC in Figure 6-14 will be proportional to the trap density at the DC potential corresponding to the measurement point. This information can be used to generate a plot like that shown on the bottom in Figure 6-14. D_{it} is the interface state density at a particular position in the bandgap. Q_{it} has units of coulombs cm^{-2} , while D_{it} has units of traps $\text{cm}^{-2}\text{eV}^{-1}$. The U shape qualitatively shown in the figure is typical of what is generally observed for the Si-SiO₂ interface.

Another CV measurement that is sometimes useful characterizes Q_m , the mobile charge density in the insulator. Typically these ions are Na⁺ or K⁺, both of which are mobile in SiO₂, especially at slightly elevated temperatures. The measurement method

is called Bias Temperature Stressing or simply BTS and involves making an initial HF CV measurement at room temperature. The MOS capacitor is then heated to $\approx 200^\circ\text{C}$ with an applied DC bias on the gate. At this temperature Na^+ and K^+ are very mobile and will easily drift upwards or downwards in the SiO_2 . The capacitor is then cooled to room temperature (still under bias), and a second HF CV curve is measured. If drifting of Q_m has occurred, a lateral shift in the HF CV curve will be measured, like the lateral shift in Figure 6-13 due to Q_f . The extent of the shift is proportional to Q_m . Sometimes a second BTS procedure with opposite polarity bias is used to drift the Q_m to the other SiO_2 interface so that the lateral difference between the two CV curves corresponds to the total Q_m present regardless of where it was originally located in the insulator.

Another simple measurement of insulator properties that is possible with the MOS capacitor is the I-V characteristic of the insulator. For thin oxides ($< 10\text{ nm}$), this measurement can directly measure tunneling currents which are used in some types of memory devices. For thicker oxides, these currents are too small to measure easily, but a destructive measurement of the dielectric breakdown strength of the oxide is sometimes made simply by ramping the voltage across the oxide until breakdown occurs. (See Figure 4-15.) For high-quality SiO_2 layers, this normally occurs at about $10\text{--}15\text{ MV cm}^{-1}$. If a large number of capacitors are measured on a wafer, a histogram plot of the measured breakdown fields can provide information on defect densities in the oxides.

We briefly mentioned earlier Q_{BD} or the charge to breakdown that can be passed through a thin SiO_2 layer before breakdown occurs. The process of oxide degradation as charge passes through it is the trapping of charge at defect sites in the oxide. This process can also be monitored using a MOS capacitor. If a DC current is forced through a MOS capacitor for a period of time and CV measurements were made before and after this process, then trapped charge will show up as Q_{ot} . Thus the CV curves before and after will show a horizontal shift whose magnitude is proportional to Q_{ot} as illustrated in Figure 6-13. A series of such experiments produce a plot of Q_{ot} versus the time the current is forced through the oxide. Forcing current through an insulator can also create damage at the Si/SiO_2 interface. If this happens, such damage is also evident in CV curves through the distortion of the curve Q_{it} creates (also shown in Figure 6-13).

Finally, in most modern MOS structures, the doping is not constant with depth as we have assumed in our discussion to this point. In fact this provides another powerful opportunity for CV measurements. Since the depletion depth x_D depends on doping, a CV measurement can also be used to extract N_D or N_A versus depth. This is a powerful doping profiling capability, which we will discuss in Chapter 7.

The MOS capacitor can thus provide a tremendous amount of information about insulators and the Si/SiO_2 interface. This simple structure has had such a profound influence on the semiconductor industry that entire books have been written on it [6.10].

In summary then, we see that a variety of measurement techniques are available to determine insulator properties. The thickness of a deposited or grown insulating layer is obtainable through physical, optical, or electrical measurements. The interface properties between the insulator and underlying semiconductor are directly measurable through CV methods. And defect densities are available through destructive breakdown tests on large numbers of capacitors. All of these techniques have been widely applied to

both thermally grown oxide/Si interfaces, and to more general insulator/semiconductor structures. We will find many examples of such measurements throughout this text. A final point, although it may be obvious, is that although all the examples we have shown in this section have been for N-type substrates. P-type substrates produce similar CV plots except that the horizontal axis is flipped (mirror imaged).

6.5 Models and Simulation

A substantial amount of effort over more than 25 years has been devoted to understanding and modeling silicon oxidation kinetics. We will focus in this section on a hierarchy of models that have been developed to explain and predict oxidation kinetics and the properties of the Si/SiO₂ interface. The first general models date back to the work of Deal and Grove in the early 1960s [6.6]. This work resulted in the linear parabolic model that is still used today to model the planar oxidation of silicon. This model cannot fully explain the oxidation of shaped surfaces, which are often encountered in modern VLSI devices, and it also cannot fully explain the oxidation kinetics in mixed ambients or for very thin oxides. Finally, it cannot explain why oxidation affects other process phenomena like dopant diffusion, often at great distances away from the oxidizing surface. However, it provides a useful starting point for our discussion and much of the more recent work on modeling builds on Deal and Grove's pioneering work.

Many of the models we will describe in this chapter have been implemented in process simulation programs. We will illustrate the use of these simulation tools as we describe the models. The simulator we will use is the Stanford University program SUPREM IV that has been implemented in commercially available versions like TSUPREM IV [6.11] and ATHENA [6.12]. These programs implement many (but not all) of the models we will discuss. SUPREM IV is a 2D simulator, which is generally available and used fairly widely in the semiconductor industry. The simulation examples in this chapter were run on commercially available versions of this program.

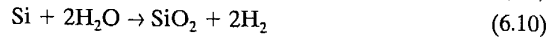
It is important to note that many of the oxidation models we will describe were developed specifically to deal with one aspect of the problem—for example, the oxidation of heavily doped substrates. The fabrication of real device structures involves many processes acting simultaneously. The only real way to model such technologies is to use a computer program that integrates all the individual models and allows them to operate simultaneously during a high-temperature step. This is the real power of simulation programs. However, the very fact that such programs integrate individual models, perhaps even models that were developed independently of each other, means that the interactions between the models may not have been fully tested experimentally. In fact, the simulators really become a means of testing the generality of the individual models. At the end of the chapter we will show some examples of more complete oxidation process simulation—examples that integrate many models simultaneously.

6.5.1 First-Order Planar Growth Kinetics—The Linear Parabolic Model

The central ideas behind the Deal-Grove or linear parabolic model are illustrated in Figure 6-15. (Note the similarity of this model to the photoresist development model we discussed in connection with Figure 5-48.) We assume that an oxide of some thickness x_o is already present on the silicon surface. We also assume that the structure is one dimensional so that this model will only apply to oxide films grown on flat substrates. The oxide grows by indiffusion of the oxidizing species to the oxide/silicon interface, where a simple chemical reaction such as



or



takes place. We initially assume that three sequential steps are necessary for this to happen, although it will turn out that only two of them are important. Figure 6-15 illustrates three fluxes associated with the process. The first represents the transport of the oxidant in the gas phase to the oxide surface. We can describe this flux as

$$F_1 = h_G(C_G - C_S) \quad (6.11)$$

where F_1 is the flux in molecules $\text{cm}^{-2} \text{sec}^{-1}$, $(C_G - C_S)$ is the concentration difference between the main gas flow and the oxide surface, and h_G is the mass transfer coefficient in cm sec^{-1} . The process represented by F_1 is the gas phase diffusion of the oxidant species through the boundary or stagnant layer that always forms adjacent to a solid object placed in a gas which is flowing over its surface. In the case of oxidation, this diffusion process is very fast compared to the other steps that must occur and hence F_1 will turn out to be unimportant in determining the overall growth kinetics. Other processing techniques such as epitaxy and CVD that also involve reactant transport from the

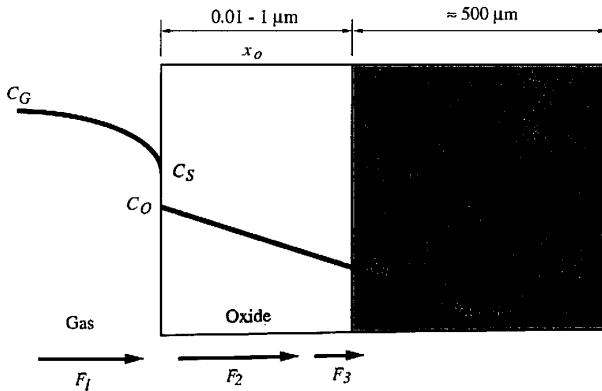


Figure 6-15 Oxidant flux from the gas phase to the silicon surface during thermal oxidation. The bold line represents the O_2 or H_2O concentration.

gas phase to the wafer surface, may be limited by this transport process. We will discuss these technologies in Chapter 9.

We can relate the concentration of the oxidant just inside the oxide surface C_O to the pressure in the gas phase adjacent to the surface, through Henry's law:

$$C_O = HP_S \quad (6.12)$$

This law states that the equilibrium concentration of a gas species dissolved in a solid is proportional to the partial pressure of that species at the solid surface. It holds for molecular species, so we are assuming that O₂ or H₂O are absorbed into the oxide. P_S is not usually a known experimental parameter, so it is convenient to write the oxidant concentration in the oxide in terms of P_G the bulk gas pressure, which is known.

$$C^* = HP_G \quad (6.13)$$

We define C^* to be the oxidant concentration in the oxide that would be in equilibrium with P_G . Since it will turn out that F_1 is not the rate limiting step in the oxidation process, this implies that $C^* \approx C_O$, and $P_G \approx P_S$. C^* therefore really represents the solubility of the oxidant in the SiO₂.

From the ideal gas law, we have that

$$C_G = \frac{P_G}{kT} \quad \text{and} \quad C_S = \frac{P_S}{kT} \quad (6.14)$$

Substitution of Eqs. (6.12)–(6.14) into (6.11), now leads directly to the result that

$$F_1 = h(C^* - C_O) \quad (6.15)$$

where $h = h_c/HkT$. Experimentally we find that wide changes in gas flow rates in oxidation furnaces, changes in the spacings between wafers on the boat or carrier in the furnace and changes in wafer orientation (standing up or lying down in the furnace) make little difference in oxidation rates. These results imply that h is very large, or that only a small difference between C^* and C_O is required to provide the necessary oxidant flux.

Flux F_2 in Figure 6–15 represents the diffusion of the oxidant through the oxide to the Si/SiO₂ interface. Using Fick's law, we can express this as

$$F_2 = -D \frac{\partial C}{\partial x} = D \left(\frac{C_O - C_I}{x_O} \right) \quad (6.16)$$

where D is the oxidant diffusivity in the oxide, C_O and C_I are the concentrations at the two interfaces, and x_O is the oxide thickness. In writing this expression, we have assumed that the process is in steady state (not changing rapidly with time), and that there is no loss of oxidant as it diffuses through the oxide. Under these conditions, F_2 must be constant through the oxide and hence the derivative can be replaced simply by a constant gradient. The oxidant concentration falls off linearly through the oxide as shown in Figure 6–15. The experimental evidence seems to suggest that O₂ diffuses in molecular

form through the SiO_2 , likely interstitially between the atoms in the oxide. H_2O on the other hand, seems to diffuse in a more complex manner, interacting with the SiO_2 matrix. The effective diffusivities of both O_2 and H_2O are on the same order (about $5 \times 10^3 \mu\text{m}^2 \text{hr}^{-1}$ at 1100°C).

The third part of the oxidation process is the reaction at the Si/SiO_2 interface. We represent this with a third flux given by

$$F_3 = k_s C_I \quad (6.17)$$

The rate at which this reaction takes place should be proportional to the oxidant concentration at the interface C_I . There are also many other factors that are likely involved in the reaction such as Si-Si bond breaking, Si-O bond formation, and possibly O_2 or H_2O dissociation. All of these effects and others are lumped into k_s which is termed the interface reaction rate constant (cm sec^{-1}). We will return to a more detailed discussion of this interface reaction later in this chapter.

Under steady-state conditions, the three fluxes representing the oxidation process must be equal since they occur in series with each other and the overall process will proceed at the rate of the slowest process. Thus $F_1 = F_2 = F_3$. Combining Eqs. (6.15)–(6.17) then leads to

$$C_I = \frac{C^*}{1 + \frac{k_s}{h} + \frac{k_s x_O}{D}} \cong \frac{C^*}{1 + \frac{k_s x_O}{D}} \quad (6.18)$$

$$C_O = \frac{C^* \left(1 + \frac{k_s x_O}{D}\right)}{1 + \frac{k_s}{h} + \frac{k_s x_O}{D}} \cong C^* \quad (6.19)$$

In writing the right-hand expressions in these equations, we have made use of the experimental observation that h is very large. Physically, we can think of the overall process as involving two interface reactions (h and k_s) and a diffusion process. Of the two interface reactions, the process at the oxide surface (gas absorption) occurs very rapidly compared to the chemistry occurring at the Si/SiO_2 and hence h can be neglected compared to k_s .

There are two limiting cases that are of interest in Eqs. (6.18) and (6.19). These are illustrated in Figure 6-16. These cases occur when $k_s x_O/D$ is much less than 1 or much greater than 1. The left-hand side of Figure 6-16 illustrates the case when $k_s x_O/D \ll 1$. In this case $C_I \approx C^*$ and the oxidant profile is essentially flat through the oxide. Physically this means that the diffusion process is not rate limiting or in other words that oxidant is being supplied to the Si/SiO_2 interface at a rate which is fast compared to that required to sustain the chemical reaction occurring there. This condition is often referred to as reaction rate controlled since the interface is determining the rate at which

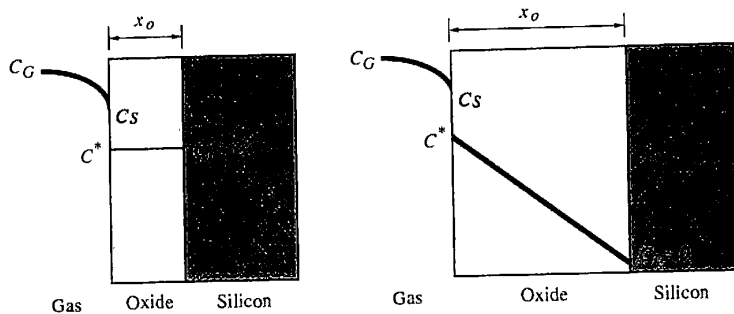


Figure 6-16 Limiting cases in silicon oxidation. The interface reaction is the rate limiting step on the left; the oxidant transport through the SiO₂ rate limits on the right.

the oxide grows. Generally this is the case for thin oxides since $k_s x_o / D \ll 1$ when x_o is small. The oxide thickness at which $k_s x_o / D \approx 1$ varies with temperature since k_s and D change with temperature, but is usually in the range of 50–200 nm.

For oxide thicknesses above this value, the growth tends more and more toward the limiting case shown on the right of Figure 6-16. In this case, $k_s x_o / D \gg 1$ and the oxidant profile in the oxide becomes linear, with $C_i \approx 0$. In this case, the oxidant is reacting at the interface as fast as it arrives and the overall growth rate is limited by the diffusion process. This is often referred to as the diffusion controlled regime.

Combining Eqs. (6.17) and (6.18), we have that

$$\frac{dx_o}{dt} = \frac{F}{N_1} = \frac{k_s C^*}{N_1 \left[1 + \frac{k_s}{h} + \frac{k_s x_o}{D} \right]} \quad (6.20)$$

where N_1 is the number of oxidant molecules incorporated per unit volume of oxide grown. $N_1 = 2.2 \times 10^{22} \text{ cm}^{-3}$ for O₂ oxidation and twice this value for H₂O oxidation. Integrating this equation from an initial oxide thickness x_i to a final thickness x_o leads us to our final result describing the oxide growth kinetics:

$$N_1 \int_{x_i}^{x_o} \left[1 + \frac{k_s}{h} + \frac{k_s x_o}{D} \right] dx_o = k_s C^* \int_0^t dt \quad (6.21)$$

$$\therefore \frac{x_o^2 - x_i^2}{B} + \frac{x_o - x_i}{B/A} = t \quad (6.22)$$

where

$$B = \frac{2DC^*}{N_1} \quad (6.23)$$

and

$$\frac{B}{A} = \frac{C^*}{N_1 \left(\frac{1}{k_s} + \frac{1}{h} \right)} = \frac{C^* k_s}{N_1} \quad (6.24)$$

B and B/A are often termed the parabolic and linear rate constants respectively because of the x^2 and x terms in which they appear. Physically, they represent the contributions of fluxes F_2 (oxidant diffusion) and F_3 (interface reaction), respectively.

It is sometimes convenient to rewrite the linear parabolic growth law in the following form:

$$\frac{x_o^2}{B} + \frac{x_o}{B/A} = t + \tau \quad (6.25)$$

where

$$\tau = \frac{x_i^2 + Ax_i}{B} \quad (6.26)$$

In these expressions, x_i or τ account for any oxide present at the start of the oxidation. They can also be used to provide a better fit to data in the “anomalous” thin oxide regime in dry O_2 as we will see later. Solving the parabolic equation leads to the following explicit expression for oxide thickness in terms of growth time:

$$x_o = \frac{A}{2} \left\{ \sqrt{1 + \frac{t + \tau}{A^2/4B}} - 1 \right\} \quad (6.27)$$

There are two limiting forms of the linear parabolic growth law that may be seen directly from Eq. (6.25). These occur when one of the two terms dominates, leading to

$$x_o \cong \frac{B}{A}(t + \tau) \quad \text{or} \quad x_o^2 \cong B(t + \tau) \quad (6.28)$$

The reasons for referring to B/A and B as the linear and parabolic rate constants are clear when the growth law is expressed in these forms. The linear term will dominate for small x values, the parabolic term for larger x values. SiO_2 growth on a bare silicon wafer therefore usually starts out with a linear x versus t characteristic, which becomes parabolic as the oxide thickens.

Eqs. (6.23) and (6.24) suggest that we should be able to calculate values for B and B/A and therefore test the predictions of the linear parabolic model against experiment. In fact, B and B/A are normally determined experimentally by extracting them from growth data. The reason for taking this approach is simply that we usually do not know all the parameters in Eqs. (6.23) and (6.24). k_s in particular contains a lot of “hidden” physics associated with the interface reaction. What we can do, however, is compare experimental values of B and B/A with the form of Eqs. (6.23) and (6.24) to test the reasonableness of the linear parabolic model.

Figure 6-17 suggests how the oxidation rate constants can be extracted from experimental data. By replotting oxide thickness versus time data in the form shown on the

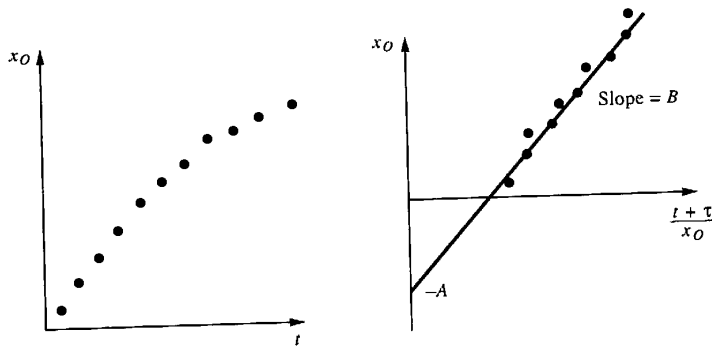


Figure 6-17 Extraction of rate constants from oxide thickness versus time experimental data.

right in the figure, B is directly extracted from the slope of the line and A and hence B/A , is extracted from the intercept. This approach has been widely used to determine values for B and B/A for a wide range of experimental conditions. If the experimental data do fall on a straight line when plotted in the form of Figure 6-17, then the data are well described by a linear parabolic growth law. In fact, there are many growth conditions that do not result in data that lie on a straight line, some of which we will discuss in later sections of this chapter. In these cases, we can conclude that the data are not well modeled by the simple linear parabolic model we have described thus far.

When oxidations are performed on flat unpatterned surfaces, on lightly doped substrates, in simple O₂ or H₂O ambients and when the oxide thickness is larger than about 20 nm, the growth kinetics are usually well described by the linear parabolic law. Under such conditions, values for B and B/A can be readily extracted. Experimentally, we find that both B and B/A are well described by Arrhenius expressions of the form

$$B = C_1 \exp(-E_1/kT) \quad (6.29)$$

$$\frac{B}{A} = C_2 \exp(-E_2/kT) \quad (6.30)$$

In these expressions, E_1 and E_2 are the activation energies associated with the physical processes that B and B/A represent; C_1 and C_2 are the preexponential constants.

It is often the case that considerable physical significance can be attached to activation energies that are measured experimentally when Arrhenius relationships are followed. Table 6-2 below lists the experimental values for the parameters in Eqs. (6.29) and (6.30). The corresponding values for B and B/A are also plotted in Figure 6-18. In Table 6-2, the middle oxidation ambient ("wet O₂") corresponds to an oxidation system in which O₂ is bubbled through H₂O at 95°C. The result is an oxidation ambient which is mostly H₂O, but also contains some O₂. These kinds of systems were common some years ago, although most H₂O oxidation systems today directly react H₂ and O₂ to produce H₂O in the back of the furnace.

Consider first the parabolic rate constant B . We find experimentally, that the activation energy E_1 is quite different for O₂ and H₂O ambients. Eq. (6.23) suggests that the

Table 6-2 Rate constants describing (111) silicon oxidation kinetics at 1 Atm total pressure. For the corresponding values for (100) silicon, all C_2 values should be divided by 1.68.

Ambient	B	B/A
Dry O_2	$C_1 = 7.72 \times 10^2 \mu\text{m}^2 \text{hr}^{-1}$ $E_1 = 1.23 \text{ eV}$	$C_2 = 6.23 \times 10^6 \mu\text{m hr}^{-1}$ $E_2 = 2.0 \text{ eV}$
Wet O_2	$C_1 = 2.14 \times 10^2 \mu\text{m}^2 \text{hr}^{-1}$ $E_1 = 0.71 \text{ eV}$	$C_2 = 8.95 \times 10^7 \mu\text{m hr}^{-1}$ $E_2 = 2.05 \text{ eV}$
H_2O	$C_1 = 3.86 \times 10^2 \mu\text{m}^2 \text{hr}^{-1}$ $E_1 = 0.78 \text{ eV}$	$C_2 = 1.63 \times 10^8 \mu\text{m hr}^{-1}$ $E_2 = 2.05 \text{ eV}$

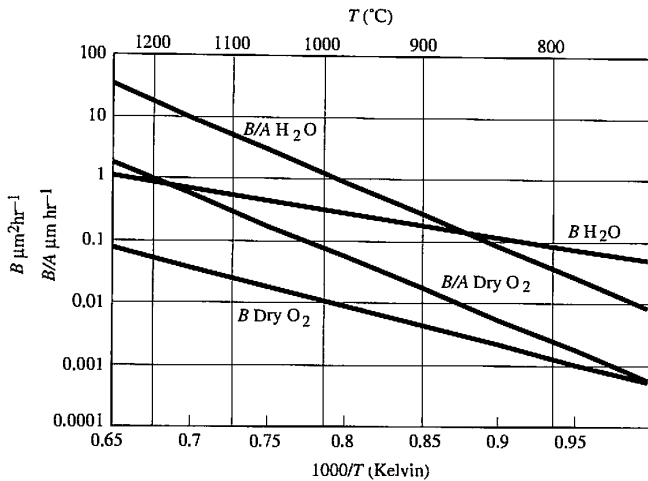


Figure 6-18 B and B/A for O_2 and H_2O oxidation of <111> silicon. Values taken from the parameters in Table 6-2.

physical mechanism responsible for E_1 might be the oxidant diffusion through the SiO_2 . (N_1 is a constant and C^* is not expected to increase exponentially with temperature.) In fact, independent measurements of the diffusion coefficients of O_2 and H_2O in SiO_2 show that these parameters vary with temperature in the same way as Eq. (6.29) and with E_1 values close to those shown in Table 6-2. The clear implication is that B in the linear parabolic model really does represent the oxidant diffusion process.

The E_2 values for B/A in the table are all quite close to 2 eV. Eq. (6.24) suggests that the physical origin of E_2 is likely connected with the interface reaction rate constant k_s . k_s really represents a number of processes occurring at the Si/SiO_2 interface. These may

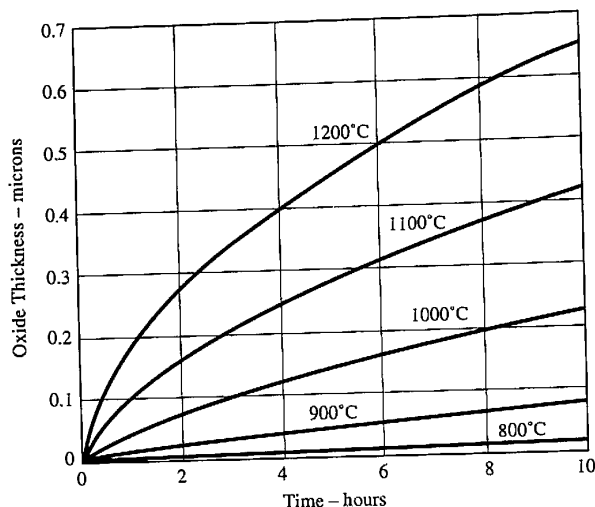


Figure 6-19 Calculated oxidation rates for (100) silicon in dry O₂ based on the Deal-Grove model. Parameter values taken from Table 6-2. The initial fast oxidation for the first ≈ 20 nm is not included ($\tau = 0$).

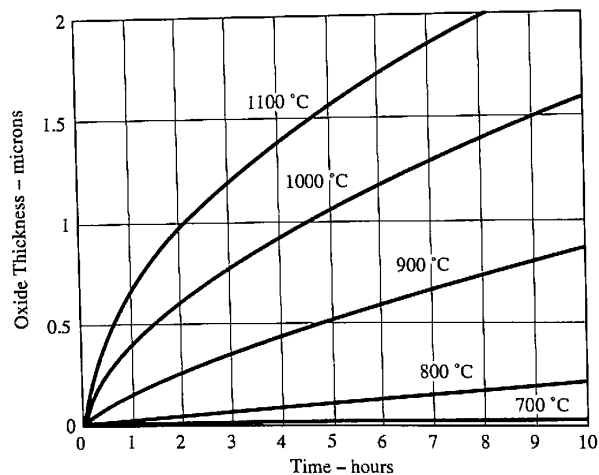


Figure 6-20 Calculated oxidation rates for (100) silicon in H₂O based on the Deal-Grove model. Parameter values taken from Table 6-2.

include oxidant dissociation ($O_2 \rightarrow 2O$), Si-Si bond breaking, and/or Si-O bond formation. Traditionally, the 2-eV activation energy has been associated with the Si-Si bond breaking process because of measurements by Pauling that suggested that the Si-Si bond energy was in the correct range to explain the B/A values. However, the interface reaction is very complex and it is likely that other effects also affect the experimental

B/A values. An additional observation that supports the idea that it is something associated with the Si substrate that determines E_2 , is that E_2 is essentially independent of oxidation ambient. It is also essentially independent of substrate crystal orientation, which suggests that E_2 represents a fundamental part of the oxidation process, not something only associated with the substrate.

Using the values for B and B/A in Table 6-2, Figures 6-19 and 6-20 show the calculated oxide thickness versus time curves predicted by the Deal-Grove model. The general behavior of an initially linear growth rate that becomes parabolic as the oxide grows is apparent in these calculated curves. It is also apparent from the figures that SiO_2 grows much faster in an H_2O ambient than it does in dry O_2 . The principal reason for this is that the oxidant solubility in SiO_2 [C^* in Eqs. (6.23) and (6.24)], is much higher for H_2O than for O_2 . At 1100°C , typical values for C^* are $\approx 5 \times 10^{16} \text{ cm}^{-3}$ for dry O_2 and $\approx 3 \times 10^{19} \text{ cm}^{-3}$ for H_2O . As a result, both rate constants, B and B/A are much larger for H_2O than for O_2 . Dry O_2 oxidations are thus generally useful for producing oxide films up to 100–200 nm. Films thicker than this would normally be grown using H_2O ambients.

Example Local oxidation is a process that is widely used to provide lateral isolation between devices in IC chips. In some cases, it is desirable to end up with a more planar surface than standard LOCOS provides, and so a silicon etch is used prior to the oxidation step as illustrated in Figure 6-21. For the structure shown on the left, with $0.5 \mu\text{m}$ of silicon etched prior to the oxidation, how long must the wafer be oxidized at 1000°C in H_2O to produce the planar oxide shown on the right?

Answer There is a 2.2X volume expansion during thermal oxidation, so growing $1 \mu\text{m}$ of SiO_2 consumes $0.45 \mu\text{m}$ of silicon. Thus the growing oxide will consume an additional thickness of silicon (y in Figure 6-22) in filling up the etched trench.

Thus we need to grow a total thickness of SiO_2 given by

$$\frac{y}{0.45} = y + 0.5 \Rightarrow y = 0.41 \mu\text{m}$$

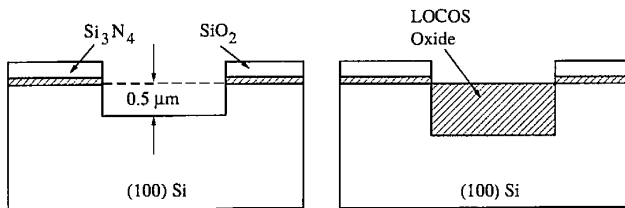


Figure 6-21 Recessed LOCOS process in which the silicon is etched prior to LOCOS to produce a final planar surface. The “bird’s beak” produced at the boundaries of such structures is not illustrated here (see Figure 6-41 later in the chapter).

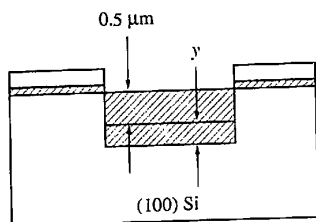


Figure 6-22 Recessed LOCOS geometry showing additional silicon consumed during the oxidation.

Thus we need to grow a total of 0.91 μm of SiO₂. At 1000°C in H₂O we have

$$B = 3.86 \times 10^2 \exp\left(-\frac{0.78 \text{ eV}}{kT}\right) = 0.316 \mu\text{m}^2\text{hr}^{-1}$$

$$\frac{B}{A} = \frac{1.63 \times 10^8}{1.68} \exp\left(-\frac{2.05 \text{ eV}}{kT}\right) = 0.75 \mu\text{m hr}^{-1}$$

$$\therefore t = \frac{x_0^2}{B} + \frac{x_0}{B/A} = \frac{(0.91)^2}{0.316} + \frac{0.91}{0.75} = 3.83 \text{ hrs}$$

6.5.2 Other Models for Planar Oxidation Kinetics

The linear parabolic model described above is the most widely accepted oxidation model. However, it does not explain a number of experimental observations, many of which we will discuss in subsequent sections of this chapter. One fact of particular importance is that experimental measurements of oxide growth rates in dry O₂ are not accurately predicted by the Deal-Grove model for thin oxides less than about 20 nm. Because of this and other problems, many other models have been proposed, each claiming to be an improvement over the linear parabolic approach. However, none of these models has gained widespread acceptance. We will briefly discuss here a few of these alternative models, in order to provide a broader picture of oxidation kinetics.

In 1987, Reisman and colleagues [6.13] suggested that a very simple power law of the form

$$x_O = a(t + t_i)^b \quad \text{or} \quad x_O = a\left(t + \left(\frac{x_i}{a}\right)^{\frac{1}{b}}\right)^b \quad (6.31)$$

where a and b are constants for a given set of process conditions and t_i is the time corresponding to the growth of any existing oxide (x_i) at the start of the process, could fit all published dry O₂ oxidation data. In fitting this equation to experimental data, they extracted b values between 0.25 and 1.0, depending on temperature and oxidant partial pressure. This equation has two fitting parameters (a and b), just as the Deal-Grove

model has two (B and B/A). There are key differences between the models, however. First, Reisman and colleagues claim that Eq. (6.31) can fit data down to an oxide thickness of essentially 0. There is no anomalous thin regime as there is with the Deal-Grove model. Second, the physical bases for the models are quite different. The Deal-Grove model as we have seen is based on the idea of oxidant diffusion and an interface reaction, with each process dominating the growth under different conditions. In a follow-up paper to [6.13], Nicollian and Reisman suggested that the interface reaction actually controlled the oxidation process at all times and that the volume expansion necessary at that interface to accommodate the growing oxide was provided by viscous flow (relaxation) of the oxide layer [6.14]. The time-dependent viscous flow of the oxide, in their model, is used to explain the extracted pressure and temperature dependence of the parameters in their model (b and a).

The model expressed in Eq. (6.31) has not yet been widely implemented in process simulation programs. These programs generally still rely on the Deal-Grove linear parabolic model. However, the physical approach used in [6.13, 6.14] is interesting because the concepts of volume expansion and oxide viscous flow are also the basic ideas used to explain oxidation on nonplanar silicon surfaces as we will see in Section 6.5.7.

Another approach to modeling oxidation kinetics was recently suggested by Han and Helms [6.15]. They showed that a wide body of oxidation data, including the dry O_2 thin regime, could be modeled by an expression of the form

$$\frac{dx_O}{dt} = \frac{B_1}{2x_O + A_1} + \frac{B_2}{2x_O + A_2} \quad (6.32)$$

The two terms represent parallel oxidation processes, perhaps O_2 and O diffusing through the SiO_2 in parallel, or perhaps O_2 and oxygen vacancies. This type of parallel diffusion and reaction was first proposed by Hirabayashi and Iwamura [6.16], to model O_2/HCl oxidations, a topic we will return to later in Section 6.5.6. All of the rate constants in Eq. (6.32) were found to fit Arrhenius expressions of the form

$$B_i = C \exp(-E_A/kT) \quad (6.33)$$

with the values given in the Table 6-3 below for dry O_2 oxidation.

Except for the B_2/A_2 values, all numbers in Table 6-3 apply to either (111) or (100) silicon orientations. Note that no values are given for B_1/A_1 since Han and Helms found that setting this term to ∞ ($A_1 = 0$), produced reasonable agreement with experiment.

Table 6-3 Experimental rate constants for the parallel oxidation model of Han and Helms [6.15]

	B_1	B_1/A_1	B_2	B_2/A_2
C	$3.9 \times 10^5 \mu m^2 \text{ hr}^{-1}$	∞	$1.56 \times 10^4 \mu m^2 \text{ hr}^{-1}$	$4.98 \times 10^6 \mu m \text{ hr}^{-1}$ (111) $1.56 \times 10^6 \mu m \text{ hr}^{-1}$ (100)
E_A	2.2 eV		1.6 eV	1.9 eV

Presumably this implies that whatever process the first term in Eq. (6.32) represents, the interface reaction associated with that process is very fast. Integrating (6.32) produces a closed form expression for oxide thickness versus time :

$$(x_o^2 - x_i^2) + C(x_o - x_i) - G \ln \left(\frac{2Ex_o + F}{2Ex_i + F} \right) = Et \quad (6.34)$$

where $C = \frac{A_1 B_1 + A_2 B_2}{B_1 + B_2}$, $E = B_1 + B_2$, $F = A_1 B_2 + A_2 B_1$, $G = \frac{B_1 B_2 (A_1 - A_2)^2}{2(B_1 + B_2)^2}$

As mentioned above, Han and Helms found that they could fit experimental data satisfactorily with $A_1 = 0$ which simplifies these expressions somewhat. This model has not been applied to H₂O oxidations. However, the simple linear parabolic model works quite well for H₂O, and it is really just a special case of Eq. (6.32) in which only one of the two terms is used.

Another model that has been proposed is due to Ghez and van der Meulen [6.17]. The model was motivated by a series of oxidation experiments done by van der Meulen with various partial pressures of O₂ in the oxidizing ambient. The linear parabolic model predicts that both B and B/A should be linearly proportional to pressure. (This is a consequence of assuming that Henry's law holds during the oxidant adsorption into the SiO₂ and that only a molecular species is involved in the transport through the oxide and reaction at the Si/SiO₂ interface.) van der Meulen's measurements showed that

$$\frac{B}{A} \propto P^n \quad (6.35)$$

where n approached 0.5 at low oxidation temperatures and 1 at high temperatures. (In contrast, n is experimentally equal to 1 for H₂O oxidations, suggesting that the assumptions made in the Deal-Grove model are more accurate in that case.)

To explain their results, Ghez and van der Meulen proposed a model in which molecular O₂ is absorbed from the gas phase and diffuses through the SiO₂ to the silicon surface. At the oxidizing interface, two parallel reactions were proposed, one involving molecular O₂ and the other atomic O. This results in a complicated growth law with no simple analytic formulation. However the atomic O reaction does lead to an n value of 0.5 for the interface reaction; the molecular O₂ reaction leads to an n value of 1. van der Meulen's data can therefore be explained by assuming that the atomic reaction dominates at low temperature and the molecular reaction at high temperatures. We will discuss these results more carefully in Section 6.5.4.

Figures 6-23 and 6-24 compare the predictions of several of these alternative models with the Deal-Grove model. The model parameters for these plots were taken from Tables 6-2 and 6-3 for the Deal-Grove and Han and Helms models, respectively. For the Reisman and colleagues model at 800°C, $a = 0.302$ nm, $b = 0.704$ and $t_0 = 13.1$ min. At 1000°C, $a = 3.02$ nm, $b = 0.701$ and $t_0 = 1.26$ min [6.13].

At 800°C (Figure 6-23), the differences in these models for very thin oxides are apparent. The Han and Helms and Reisman and colleagues models agree fairly well with each other and both match experimental data in the thin regime well. The Deal-Grove

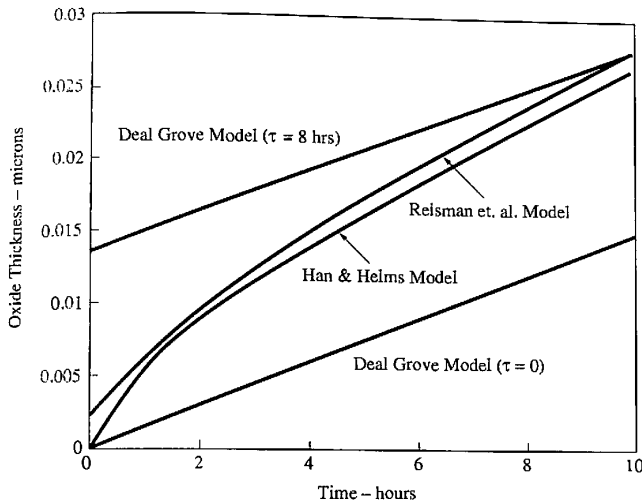


Figure 6-23 Comparison of three oxidation models for 1 atmosphere dry O_2 oxidation at $800^\circ C$.

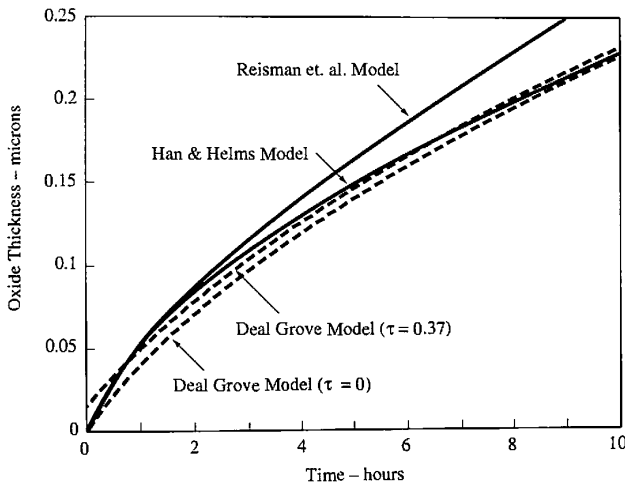


Figure 6-24 Comparison of three oxidation models for 1 atmosphere dry O_2 oxidation at $1000^\circ C$. The dashed lines are the Deal-Grove model.

model does not do a very good job for thin oxides even with τ set at 8 hours (the value Deal and Grove extracted in their original work [6.6]). We will see in Section 6.5.3, that the Deal-Grove model can be “fixed” to do a better job in the dry O_2 thin regime. Note that for oxides thicker than about 20 nm all three models converge to approximately the same result, as long as we use $\tau = 8$ hours in the Deal-Grove model.

At $1000^\circ C$ (Figure 6-24), the differences in the models in the very thin regime are less apparent since much thicker oxides grow at the higher temperature on the same time scale. Including or not including τ in the Deal-Grove model also makes less difference,

for the same reason. For the particular parameters chosen for this plot, the Reisman and colleagues model diverges from the predictions of the other models for thicker oxides.

In summary, many other models for oxidation kinetics have been proposed, all of them motivated by shortcomings in the linear parabolic model. However, none of these models has found the widespread acceptance of the linear parabolic model. Most modeling efforts today use the Deal-Grove model as a starting point and then add to it or modify it as necessary to include effects that the basic model itself either does not include or does not model well. We will consider many of these effects in the sections below.

6.5.3 Thin Oxide SiO₂ Growth Kinetics

The very earliest oxidation studies suggested that there was something unusual about the initial growth of oxides in dry O₂ on bare silicon wafers. Deal and Grove in their original paper describing the linear parabolic model observed that the model did not fit the first 20 nm of growth in dry O₂ and we saw an example of this in Figure 6-23. They attributed the fast growth rate they experimentally observed in this regime to the effect of a space charge region in the oxide, set up by the indiffusing oxygen ions (presumably O₂⁻). If the principal diffusing species really were ionic, then positive charge carriers would also be present to maintain charge neutrality. The two species O₂⁻ and holes (h⁺) would diffuse as a coupled pair. The holes, however, diffuse much faster than the O₂⁻ and would tend to move on ahead. This sets up an electric field due to the charge separation, which acts to keep the two together. The effect of this field is to slow the holes down and to speed up the O₂⁻ ions. This coupled diffusion is observed anytime a charged species diffuses. The distance over which the electric field exists is on the order of the extrinsic Debye length, given by

$$L_D = \sqrt{\frac{kT\epsilon_s}{2q^2C^*}} \quad (6.36)$$

Substituting in numbers for O₂ oxidations gives a Debye length of about 15 nm which agrees with the thickness over which fast growth occurs. For H₂O oxidations, C^{*} is about 10³ times higher, resulting in a Debye length of less than 1 nm even if a charged species were involved in water oxidations. Deal and Grove used this observation to suggest why an anomalous fast initial growth is not observed in H₂O oxidations. The difficulty with this explanation, however, is that the original assumption that O₂⁻ was the species involved in oxidant transport was based on experiments done by Jorgenson in the early 1960s [6.2]. These results have been subsequently largely discounted.

If neutral O₂ and H₂O are the dominant diffusing species, space charge effects cannot be responsible for the anomalous high initial growth rate observed in O₂. This has led to a large number of new proposals over the past 20 years, all attempting to explain thin oxide kinetics. There is still great debate about the physical mechanism(s) responsible; no consensus has emerged. The problem has taken on new importance recently, because MOS gate oxides are now routinely grown with thicknesses in this "anomalous" range. This is an example of industrial application outpacing basic scientific un-

derstanding. From an industrial perspective, it is more important to be able to grow these thin oxide layers reproducibly and uniformly, and with good electrical properties than it is to understand the governing physical principles. However this is sometimes a dangerous approach because lack of physical understanding may mean that manufacturing problems which arise may not be understood and may have to be solved by trial and error experiments. This is the reason why so much effort has been invested in understanding thin oxide growth kinetics.

One of the factors which confuses model development and data interpretation in the thin regime is that chemical cleaning procedures prior to oxidation have been shown to significantly affect the oxide growth kinetics. We discussed wafer cleaning and some of these effects in detail in Chapter 4 (Figure 4-3). Most experimental studies of thin growth kinetics have not carefully controlled this parameter with the result that data on oxide thickness versus time vary from experiment to experiment even when the same temperatures and other parameters have been used. This makes model development very difficult.

Perhaps the most extensive experimental study yet reported was that by Massoud and colleagues [6.18]. They reported that growth in the thin regime could be fitted by an addition to the Deal-Grove model given by

$$\frac{dx_O}{dt} = \frac{B}{2x_O + A} + C \exp\left(-\frac{x_O}{L}\right) \quad (6.37)$$

where

$$C = C^0 \exp\left(-\frac{E_A}{kT}\right)$$

and $C^0 \approx 3.6 \times 10^8 \mu\text{m hr}^{-1}$, $E_A \approx 2.35 \text{ eV}$, and $L \approx 7 \text{ nm}$. All these numbers apply to either (111) or (100) oriented silicon substrates. The first term in Eq. (6.37) is, of course, the Deal-Grove model. The second term represents an additional oxidation mechanism that is important for the first few tens of nm but which decreases exponentially with a decay length of about 7 nm. With this addition to the Deal-Grove model, the predictions of this model also match the data well over the whole thickness range. In Figure 6-24, for example, Eq. (6.37) provides an x_O versus t curve that matches the Han and Helms and Reisman and colleagues models. There are thus several ways of modeling the dry O_2 thin regime. Process simulation programs like SUPREM IV have generally implemented Eq. (6.37) rather than the alternatives given by Eqs. (6.31) and (6.32) because models for many other oxidation effects have also been tied to the basic Deal-Grove formulation as we will see later.

Many physical mechanisms have been suggested for the thin regime. Deal and Grove proposed O_2 coupled diffusion with holes as was discussed above. Other suggestions have included considering the role of thermionic emission of electrons from the silicon into the SiO_2 [6.19], micropores or channels in the oxide which allow parallel oxidant transport when the oxide is thin [6.20], parallel O_2 and O reactions at the Si/SiO_2 interface [6.15], blocking layers near the Si/SiO_2 interface which form as the oxide grows and slow down the rate for thicker layers [6.21], and a silicon surface layer which provides additional sites for the oxidation reaction in its initial phases [6.18]. None of these

mechanisms has been shown to be clearly correct and none of them has gained widespread acceptance. We are left with several empirical expressions that can be used to model the thin regime kinetics, but with no completely satisfactory physical explanation. Additional work in the future may help to clarify some of these issues.

6.5.4 Dependence of Growth Kinetics on Pressure

The linear parabolic model predicts that the oxide growth rate should be directly proportional to oxidant pressure. Eq. (6.13) shows this. If Henry's law holds and the concentration of oxidant just inside the oxide at the gas/SiO₂ interface C^* is proportional to P_G , then both B and B/A are proportional to P_G from Eqs. (6.23) and (6.24). The oxide growth rate should therefore be proportional to P_G .

Experimental measurements have shown that in H₂O oxidations this prediction is correct. Pressures below atmospheric [6.6] and well above atmospheric [6.22] have been investigated. These results suggest that the assumptions made in the linear parabolic model are correct for H₂O oxidation. However, the situation is a little less clear for dry O₂ oxidation. A considerable body of data has consistently shown that in order to model dry O₂ results with a linear parabolic equation, $B \propto P$ and $B/A \propto P^n$ where $0.5 < n < 1$. This pressure dependence is inconsistent with the linear parabolic model. One can only conclude that the linear parabolic model must be incomplete. Within the context of that model, we can infer that since $B \propto P$, then $C^* \propto P$ from Eq. (6.20). If B/A is not linearly proportional to P , then from Eq. (6.21), k_s must depend on P in a nonlinear fashion.

There have been a number of attempts to modify the Deal-Grove model to account for these experimental results. Ghez and van der Meulen (Section 6.5.2) attempted to extend the linear parabolic model to include both O and O₂ reacting at the Si/SiO₂ interface. This model can explain the observed pressure dependence because an O-based reaction would vary as $P^{0.5}$ and an O₂ reaction would vary as $P^{1.0}$. A mixed reaction could give any value between these two limits. However, this model contains many parameters and it has not been fit to a wide range of oxidation data. As a result, it cannot easily be used today to model the complete range of O₂ growth kinetics. More recently, Hu [6.23] attempted to treat the parallel O/O₂ interface reaction in terms of a chemisorption model. Other attempts to account for the pressure dependence of B/A have departed more radically from the Deal-Grove model. These attempts have almost all been based on a modified Si/SiO₂ interface reaction. None of these models has found widespread acceptance.

Lacking a widely accepted first principles model to explain the pressure dependence of O₂ growth kinetics, the best we can do today to model these effects is to simply use the Deal-Grove model with the following corrections:

$$\frac{B}{A} = \left(\frac{B}{A}\right)^i P \quad B = (B)^i P \quad \text{for H}_2\text{O} \quad (6.38)$$

$$\frac{B}{A} = \left(\frac{B}{A}\right)^i P^n \quad B = (B)^i P \quad \text{for O}_2 \quad (6.39)$$

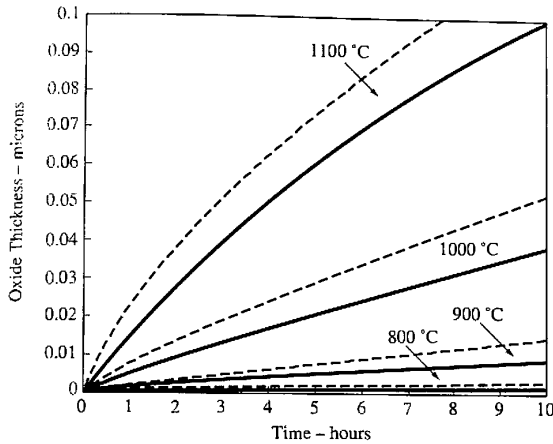


Figure 6-25 Oxidation at 0.1 atmosphere in dry O_2 using the Deal-Grove model. The solid lines correspond to $B \propto P$ and $B/A \propto P$. The dashed lines correspond to $B \propto P$ and $B/A \propto P^{0.8}$.

where $n \approx 0.7-0.8$. The i superscripts refer to the respective values at 1 Atm. A similar $P^{0.8}$ pressure dependence has been noted in the thin regime (< 20 nm). This may be included in Eq. (6.37) by making $C \propto P^{0.8}$. The decay length L in that equation does not seem to have any significant pressure dependence.

Figures 6-25 and 6-26 illustrate the effects of these pressure models on oxidation kinetics, using the Deal-Grove model and Eqs. (6.38) and (6.39), without the added thin oxide model. Note that for pressures below 1 atmosphere, the actual oxide thickness is increased by the sublinear pressure dependence of B/A , while the reverse is true for pressures above 1 atmosphere. Process simulators like SUPREM IV generally implement equations like (6.38) and (6.39) to model these effects.

6.5.5 Dependence of Growth Kinetics on Crystal Orientation

As we saw in Chapter 3, historically two principal crystal orientations have been used in manufacturing ICs. (111) crystals were used in the 1960s almost exclusively. However, as it became clear that the electrical properties of the Si/SiO₂ interface were superior on the (100) crystal surface, a gradual shift to that orientation began, with the result that (100) wafers are almost exclusively used today. Oxidation studies which have investigated orientation effects have usually concentrated on these two types of crystals and it has been shown many times that oxidation growth rates are faster on (111) surfaces than they are on (100) surfaces. In recent years, it has become important to understand orientation effects on oxidation more generally because many structures now use etched trenches and other shaped silicon regions as part of their structure.

Even before the development of the linear parabolic model, it had been observed that crystal orientation affected oxidation rate. Ligenza in 1961 [6.24] suggested that the

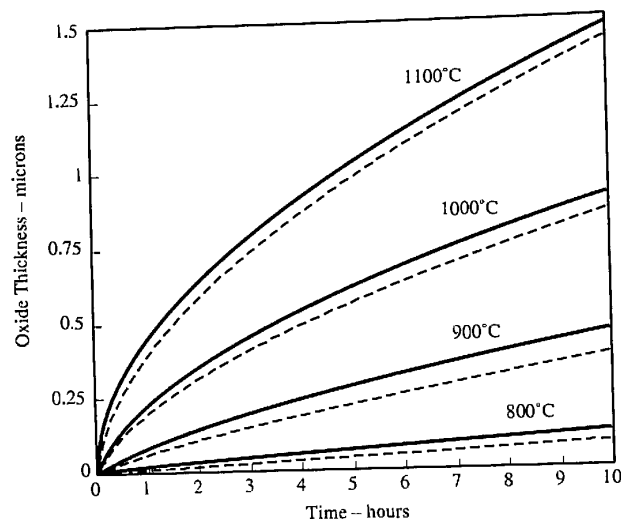


Figure 6-26 Oxidation at 10 atmospheres in dry O₂ using the Deal-Grove model. The solid lines correspond to $B \propto P$ and $B/A \propto P$. The dashed lines correspond to $B \propto P$ and $B/A \propto P^{0.8}$.

effect might be caused by differences in the surface density of silicon atoms on the various crystal faces. He argued that since silicon atoms are required for the oxidation process, crystal planes that have higher densities of atoms should oxidize faster. He further argued that it was not simply the number of silicon atoms cm⁻² that was important, but also the number of bonds since it is necessary for Si-Si bonds to be broken for the oxidation to proceed.

Ligenza calculated the number of "available" bonds cm⁻² on the various silicon surfaces and concluded that oxidation rates in H₂O should be in the order (111) > (100). This is in fact what is observed experimentally in both H₂O and dry O₂. For most oxidation conditions of interest in silicon technology, these two orientations represent the upper and lower bounds for oxidation rates. For very thin oxides grown at very low pressures in dry O₂, there is some evidence that the (110) rate becomes fastest [6.25]. This also appears to be the case at very high pressures and low temperatures in H₂O [6.24].

We can incorporate these effects in the linear parabolic model in the following way. Except perhaps in the region very near the Si/SiO₂ interface, the oxide that grows on silicon is amorphous; that is, it does not incorporate any information about the underlying silicon crystal structure. Because of this, the parabolic rate constant B should not be orientation dependent since B represents oxidant diffusion through the SiO₂. If the oxide structure is unrelated to the underlying substrate, there should be no crystal orientation effect on B . This is in fact what is found experimentally when kinetic data are analyzed in the context of the linear parabolic model and the rate constants B and B/A are extracted. B/A , on the other hand, should be orientation dependent because it involves the reaction at the Si/SiO₂ interface. This reaction surely involves Si atoms and so should be affected by the number of available reaction sites. We can thus incorporate orientation effects as follows:

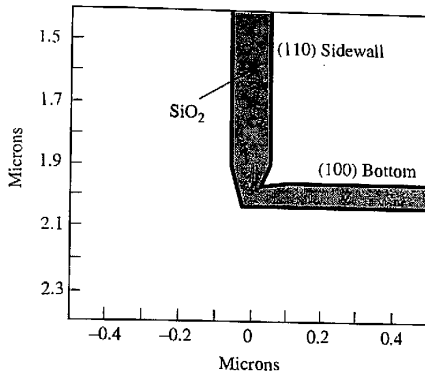


Figure 6-27 Example of an oxidation simulation involving orientation effects using the ATHENA simulator [6.12]. The top and bottom surfaces of the etched trench are (100) surfaces. The trench sidewall is a (110) surface. The oxidation was 30 min at 900°C in H₂O.

$$\left(\frac{B}{A}\right)_{111} = 1.68 \left(\frac{B}{A}\right)_{100} \quad (6.40)$$

All other orientations are normally in between these two extremes and B values are the same for all orientations. The (110) B/A value is often taken as 1.45 times the (100) rate although the amount of data available is not sufficient to be certain of this value.

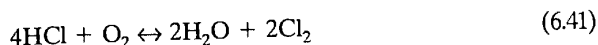
An example of a simulation showing orientation effects is shown in Figure 6-27. A silicon trench is first etched 2 μm deep and then oxidized. Note the thicker oxide along the (110) sidewall surface compared to the (100) wafer surface and trench bottom. Actually the most interesting part of this simulation is at the trench corner, shown on the right, where 2D effects dominate. We will discuss these issues in Section 6.5.7.

Finally, it may be interesting to speculate about why the orientation effect apparently changes for very thin oxides grown at low O₂ partial pressures and also for high pressure steam oxides grown at low temperatures. These two conditions represent the two extremes of oxide growth rates, one extremely slow and the other very fast. In the former case, we have already seen that the growth kinetics are different for thin oxides grown in dry O₂. Since the mechanism is apparently different in this regime, it is perhaps not too surprising that other effects such as the orientation dependence also change. In the case of high-pressure steam oxides, Ligenza observed the (100) surface to oxidize faster than the (111) surface at 800°C and below. These experiments were done at pressures up to 150 atm so the growth rates were extremely high. We will see in Section 6.5.7 on 2D growth kinetics that relaxation of grown-in stresses in SiO₂ layers takes place through viscoelastic flow of the oxide layer. In essence the glass flows as it is grown to relieve the stresses due to volume differences between the oxide and the silicon from which it is grown. At low temperatures this relaxation is more difficult because the oxide cannot flow as easily. It is also more difficult if the oxide is growing very rapidly because there is less time for the oxide to relax before new layers grow. It may be the case in Ligenza's experiments, then, that very high stresses built up in the growing oxides. This may be the origin of the unusual orientation effects that Ligenza observed. The

stresses that build up and the oxide flow to relieve those stresses might well be dependent on the substrate crystal orientation. Perhaps future work will resolve these issues.

6.5.6 Mixed Ambient Growth Kinetics

Situations are sometimes encountered in silicon technology in which mixed ambients are used during an oxidation process. The most common example is a mixture of H₂O and O₂. This occurs when H₂ and O₂ are reacted to form H₂O in modern pyrogenic systems. An excess of O₂ is always present in order to ensure that there is no excess H₂ at the furnace exhaust. The result is a mixture of H₂O (about 95% typically) and O₂ (about 5%) in the furnace. In other cases, small concentrations of HCl are sometimes added to O₂ oxidants because this has been found to help reduce oxide defect densities and to reduce contamination levels. The HCl likely helps because it reacts in the furnace to produce Cl₂ and H₂O according to the reaction



and the Cl₂ readily reacts with trace metals and other contaminants in the furnace to produce volatile chlorides as byproducts. In some cases, other Cl sources are used such as trichloroethane that is a liquid, but the reaction that takes place is similar.

Modeling of mixed ambients can be approached in a straightforward manner, if it is assumed each of the oxidants acts independently. The Han and Helms model we discussed for oxidation kinetics allows for independent parallel processes [6.15, 6.16].

$$\frac{dx_O}{dt} = \frac{B_1}{2x_O + A_1} + \frac{B_2}{2x_O + A_2} \quad (6.32)$$

Choosing B_1 and A_1 to correspond to one of the oxidants (H₂O for example) and B_2 and A_2 to correspond to the other oxidant (O₂ for example) allows calculation of the resulting oxide thickness. Note that each B and B/A value must be corrected for the appropriate partial pressure of the corresponding oxidant, using the relationships described in Section 6.5.4.

This simple approach works well in some cases and poorly in others, presumably because the assumption of independence is better in some cases than in others. H₂O/O₂ mixtures can be described by Eq. (6.32) to first order [6.26] although there is some evidence that the H₂O and O₂ reactions at the oxidizing interface are not completely independent (the B/A values seem to be somewhat different than those suggested by the above approach). H₂O/HCl mixtures can be described by Eq. (6.32) because the HCl in this case seems to only dilute the H₂O and plays no role in the oxidation process. This is likely because the H₂O and HCl do not react significantly in the furnace. As a result, only one term in Eq. (6.32) is needed, with rate constants corresponding to H₂O at the appropriate partial pressure [6.27]. O₂/HCl mixtures are not well described by Eq. (6.32). Clear interactions between the oxidant species occur in this case. The Cl generated by Eq. (6.41) is known to be incorporated into the growing SiO₂ film (mostly near

the Si/SiO₂ interface). B and B/A values extracted from growth kinetic data monotonically increase with HCl percentage, but not in a well understood manner. Process modeling programs like SUPREM IV generally resort to empirical look-up tables to model these processes. The rate constants are simply expressed as

$$\frac{B}{A} = \left(\frac{B}{A}\right)^i \delta \quad \text{and} \quad B = (B)^i \chi \quad (6.42)$$

where δ and χ are functions of the HCl concentration in the furnace ambient and are determined empirically. Other mixed ambient reactions could be dealt with in a similar empirical manner, although the principal ones that have been investigated experimentally are those described above.

6.5.7 2D SiO₂ Growth Kinetics

It has been known for many years that shaped silicon structures oxidize differently than simple flat surfaces. This was not of much technological importance until about 1980 when structures became small enough that 2D and even 3D effects began to be noticeable. We saw a simple example of this in Figure 6-4, which illustrated LOCOS processing, and the 2D growth of oxide layers under the edges of Si₃N₄ layers. Even when they were first observed, the origin of these effects was correctly believed to be the volume expansion that SiO₂ undergoes when it grows. We saw previously that SiO₂ grows upward above the silicon surface as it forms, and that the volume expansion involved is more than 2:1. It is perhaps easy to visualize that oxidation occurring in a confined corner in which the volume expansion is more difficult might be different from oxidation on a flat surface.

A quantitative understanding of the physical mechanisms responsible for these effects came only recently from an elegant set of experiments done by Kao and colleagues [6.28, 6.29]. In those experiments, a number of shaped silicon structures were prepared by anisotropic dry etching of silicon wafers. The structures were then oxidized under a variety of conditions and the resulting oxide thicknesses measured using SEM techniques. The data produced oxide thickness as a function of the radius of curvature of the structure being oxidized. Figures 6-28 to 6-30 illustrate the experimental structure and some of the results.

Several interesting observations can be made from the results of Kao et al. The first is that the retardation is a very significant effect in sharp corners—a factor of two in oxide thickness for nominally 500-nm oxides. Second, the retardation is much more pronounced for low-temperature oxidations than it is for high temperatures. In fact there is virtually no corner effect for oxidations done at 1200°C. Finally, interior (concave) corners show a more pronounced effect than do exterior (convex) corners although both are significantly retarded compared to flat surfaces.

Several physical mechanisms are important in understanding these results.

1. **Crystal orientation:** Shaped surfaces involve many surface orientations. As we saw in Section 6.5.5, orientation affects oxidation rate in the thin oxide or linear regime.

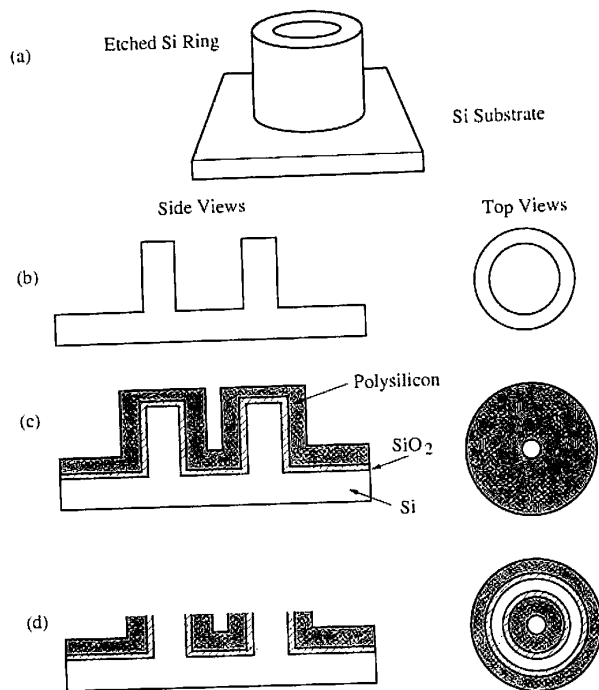


Figure 6-28 Experimental procedure used in the 2D oxidation experiments of Kao [6.28, 6.29]. Silicon wafers were plasma etched to produce a variety of shaped structures including the cylinder illustrated in (a). (b) shows a cross section and top view of the cylinder. In (c) the structure has been oxidized and then coated with a CVD layer of polysilicon. In (d), the structure has been lapped down so that a top view shows the oxide thickness on the sides of the cylinder.

This effect can be modeled in a straightforward manner through Eq. (6.40), although values for the multiplicative constant on B/A for all orientations are needed.

2. **2D Oxidant Diffusion:** In corner regions and other shaped structures, oxidant transport to the Si/SiO₂ interface is a 2D or perhaps even 3D transport problem. Since the oxide is amorphous, the diffusion coefficients of O₂ or H₂O should not be direction dependent, but numerical techniques are generally required to solve the diffusion equation in multiple dimensions.
3. **Stress due to volume expansion:** Oxide layers formed on silicon are under significant compressive stress. This stress is due to the volume expansion that takes place during oxidation and to the difference in thermal expansion coefficients of Si and SiO₂ which causes additional stress when the wafers are cooled from the oxidation temperature

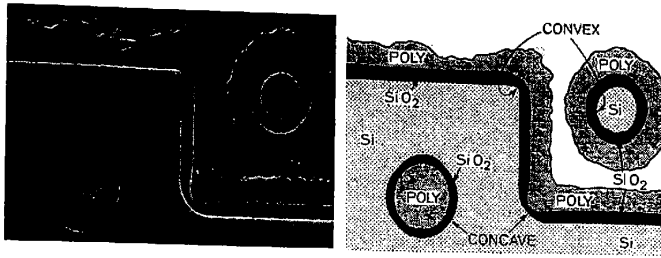


Figure 6-29 Typical experimental result from Kao [6.28, 6.29]. The drawing on the right labels the structure shown experimentally on the left. The oxide is thinner on both concave and convex corners than it is on flat regions. Reprinted with permission from IEEE.

back to room temperature. On shaped silicon surfaces, these stresses can be much larger than they would be on a flat surface simply because the volume expansion is dimensionally confined on shaped structures. Such stresses could in principle affect both oxidant transport through the SiO_2 and the interface reaction at the Si/SiO_2 interface.

To model these stress effects, Kao et al. suggested the following modifications to the parameters usually used in the linear parabolic model.

$$k_s(\text{stress}) = k_s \exp\left(-\frac{\sigma_n V_R}{kT}\right) \exp\left(-\frac{\sigma_t V_T}{kT}\right) \quad (6.43)$$

$$D(\text{stress}) = D \exp\left(-\frac{(P)(V_D)}{kT}\right) \quad (6.44)$$

$$C^*(\text{stress}) = C^* \exp\left(-\frac{(P)(V_S)}{kT}\right) \quad (6.45)$$

where k_s is the normal interface reaction rate, σ_n is the stress normal to the growing interface, σ_t is the stress tangential to the growing interface, D is the normal oxidant diffusivity (at 1 Atm), C^* is the normal oxidant solubility used in Eqs. (6.23) and (6.24) and P is the hydrostatic pressure in the growing oxide. V_R , V_T , V_D , and V_S are considered to

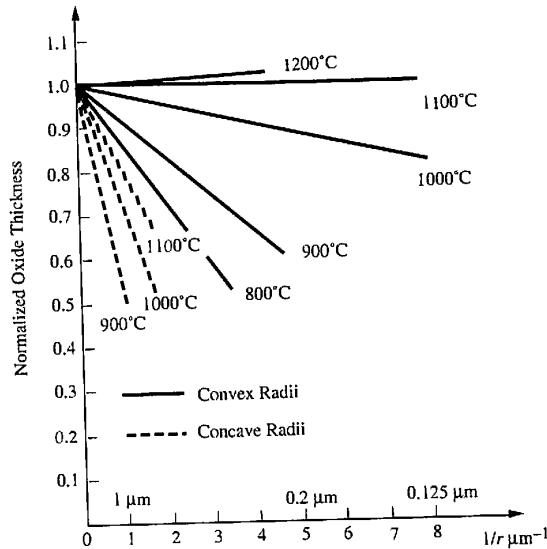


Figure 6-30 Typical experimental data after Kao [6.28, 6.29]. The oxide thickness is plotted versus the radius of curvature of the structure being oxidized. For each temperature, about 500 nm of oxide was grown on a flat surface ($1/r = 0$). These results are for H₂O oxidations. Similar results were found for O₂. Note the retardation of the oxidation for sharp corners (more than a factor of 2 in some cases).

be stress-dependent activation volumes and should be regarded as fitting parameters. They are used in the SUPREM IV implementation of this model.

The final oxide parameter that is needed to calculate growth on shaped surfaces is the oxide viscosity. SiO₂ is a glass and can relax some of the stresses that build up during oxidation by viscoelastic flow. In fact, the stresses that can build up during oxidation are so large, that the oxide viscosity itself needs to be modeled as a function of stress. A relationship of the following form has been found to produce good agreement with experiment [6.30, 6.31]

$$\eta(\text{stress}) = \eta(T) \frac{\sigma_s V_c / 2kT}{\sinh(\sigma_s V_c / 2kT)} \quad (6.46)$$

where $\eta(T)$ is the stress-free, temperature-dependent oxide viscosity, σ_s is the shear stress in the oxide, and V_c is again a fitting parameter.

Much of the interest in recent years in extending oxidation models has been driven by the need to accurately model advanced isolation processes. LOCOS, poly-buffered LOCOS, shallow trench structures, and the like all involve thermal oxidation of shaped structures. As a result, volume expansion during oxidation, stress-dependent oxidation

parameters and oxide viscosity have all taken on first-order importance. Beyond the properties of SiO_2 itself, however, many of the modern isolation structures involve other materials like polysilicon and Si_3N_4 . Thus the viscoelastic properties of these materials have also become important.

The nonlinear viscous nature of oxide films at high temperatures has been parameterized by Rafferty [6.31] as described in Eq. (6.46). Subsequent work determined that thin nitride films are viscous at high temperatures, with the viscosity being dependent on the film stoichiometry [6.32]. Similar work with polysilicon thin films has shown that more complex viscosity behavior can be assigned to polysilicon, possibly dependent on the grain size in the film.

Since the shape of the growing oxide changes with time, $k_s(\text{stress})$, $D(\text{stress})$, and $\eta(\text{stress})$ all change with time during an oxidation. As a result, numerical simulation is required to implement these models. This has been done in SUPREM IV which can successfully predict time-dependent oxide shapes and the stress levels present in both the oxide and the underlying silicon [6.31]. Typical parameters for the SUPREM IV stress-dependent oxidation model are shown in Table 6-4 for wet oxidations [6.30–6.32]. A number of these parameters are fitting parameters (V_R , V_D , V_S , and V_C in particular) and the values should not be regarded as having significant physical meaning. The fact that $V_S \approx 0$ implies that C^* is not highly stress dependent.

An example of the impact of including stress dependent parameters during an oxidation simulation is shown in Figure 6-31. The simulation on the left did not include any stress dependence whereas on the right the types of models for viscosity and stress dependence discussed above were included. Notice that the bird's beak extends much further under the Si_3N_4 mask on the left. Without stress effects, the oxidation rate is not retarded under the masking nitride.

There is a very important implication of the results described in this section. It is quite common in silicon processes to use multiple oxidations during device fabrication. Suppose, for example, that we grow an oxide at 900°C in one step and then further oxidize the silicon in a later step at 1000°C . At the end of the 900°C oxidation, certain stress levels will exist in the SiO_2 . If we then heat the wafer to 1000°C , the stress levels cannot immediately take on the values that would be observed in steady state for a 1000°C oxidation. These stresses would in general be lower because the oxide can flow more easily at high temperatures to relieve them. However, this will take some time,

Table 6-4 Typical parameter values in SUPREM IV stress dependent oxidation model

Parameter	Value (Wet Oxidation)
V_R	0.0125 nm^3
V_D	0.0065 nm^3
V_S	0
V_C	$0.3 \text{ nm}^3 @ 850^\circ\text{C}$
	$0.72 \text{ nm}^3 @ 1050^\circ\text{C}$
V_T	0
$\eta(T) - \text{SiO}_2$	$3.13 \times 10^{10} \exp(2.19 \text{ eV}/kT) \text{ poise}$
$\eta(T) - \text{Si}_3\text{N}_4$	$4.77 \times 10^{10} \exp(1.12 \text{ eV}/kT) \text{ poise}$

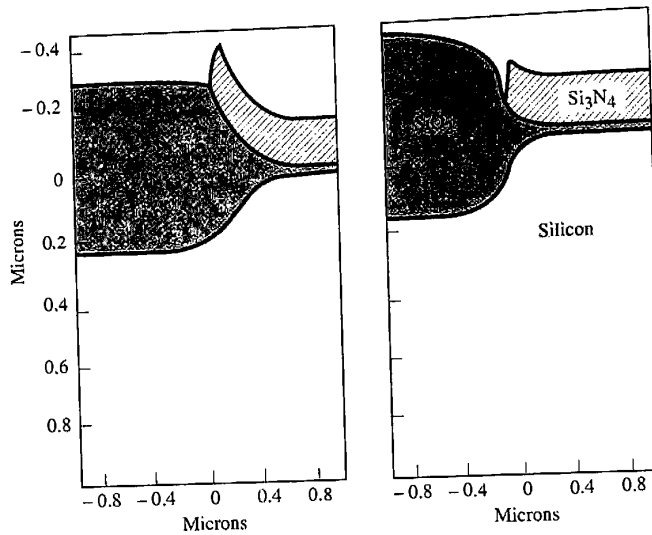


Figure 6-31 Example of an oxidation simulation showing the effects of including stress effects in oxidation using the ATHENA simulator [6.12]. A 20-nm SiO₂ pad oxide is first grown and a 150-nm Si₃N₄ layer is then deposited. The nitride is then etched on the left side of each structure. A 90-minute 1000°C H₂O oxidation was then performed. In the simulation on the left, no stress-dependent parameters were included in the simulation. Stress-dependent parameters were included in the simulation on the right.

and so we should expect a transient at the beginning of the 1000°C oxidation when the oxidation rate constants will not have the values we normally expect for that temperature. These kinds of transients have in fact been observed experimentally and the dominant effect appears to be on B [6.33]. That is, the grown-in stress levels in the oxide change the density of the oxide from that normally expected at the second oxidation temperature; the diffusivity of the oxidant is therefore affected until the oxide relaxes to its “normal” state for the second oxidation. If the first oxidation is at a high temperature and the second at a lower temperature, the second oxide will initially grow faster than expected because the stress levels will be lower than expected at the lower temperature. If the first oxidation is at a lower temperature and the second at a higher temperature, the rate in the second oxidation will be initially slower than expected. These kinds of effects are only beginning to be modeled today in process simulation programs. In essence they imply that wafer history has an effect on the process parameters observed in a given process step.

One of the current trends in the semiconductor industry is toward the use of single-wafer rapid thermal processing type equipment. These systems rapidly ramp wafers from room temperature to process temperatures, typically in a few seconds, using lamp heating. Memory effects like those described in the last paragraph will be exacerbated by such systems because there will be essentially no time during the temperature ramp

for the oxide properties to reach the values characteristic of the processing temperature. If rapid thermal processing is combined with oxidation, the effective oxidation rates may be quite different than expected and will certainly depend on wafer history. This is a largely unstudied area that may be increasingly important in the future.

Finally, two interesting recent results have shown that it is possible to oxidize shaped silicon structures and grow uniform oxide layers on them. This can be done if a means is found to relieve the stresses that build up during the oxidation and hence to eliminate the stress effects on oxidation rate coefficients. It is, of course, possible to do this by oxidizing at very high temperatures since around 1200°C, the oxide can flow easily enough to relieve the stresses. However, such high temperatures are generally not compatible with small geometry devices. Two other ways to achieve essentially the same result involve doing the oxidation in an O_2/NF_3 mixture [6.34], or utilizing a corona discharge above the wafer during the oxidation [6.35]. The physical reasons why these processes apparently relieve the stresses in the growing oxide are not understood at present.

6.5.8 Advanced Point Defect Based Models for Oxidation

The models that we have described to this point have been primarily aimed at explaining the growth kinetics of oxide layers. If that were all that we were required to explain, then the set of models we have considered to this point would be fairly complete. However, there are other experimental results that these models cannot explain, most of which have to do with interactions between oxidation and other process steps like diffusion. One of the earliest indications that more was actually going on during oxidation than the effects we have considered to this point were observations more than 20 years ago that surface oxidations changed the rate of diffusion of dopants in the underlying silicon substrate. This phenomenon is now known as Oxidation Enhanced Diffusion (OED) in cases in which dopant diffusion rates are increased by surface oxidation, or Oxidation Retarded Diffusion (ORD) in cases in which diffusion is slowed down by surface oxidation. We will consider these cases in more detail in Chapter 7. However, we will introduce here some additional ideas about oxidation that will be needed to explain OED, ORD, and other related effects.

All of the oxidation mechanisms we have considered to this point are purely local phenomena. That is, they help to explain why oxides grow at the rates they do, but they offer no suggestions as to why surface oxidation should, for example, perturb dopant diffusion rates tens of microns away from the surface. Clearly some nonlocal phenomenon must be involved. Current thinking regarding such effects centers on the interface reaction itself (B/A or k_s in the Deal-Grove model). We have seen that there is a very large volume expansion that occurs during oxidation. This volume expansion results in very large compressive stresses in the oxide layer and tensile stresses in the silicon in the near surface region. Dobson was among the first to suggest that there might be mechanisms at the atomic level available to the oxidizing interface to help relieve these stresses [6.36]. These mechanisms in essence provide some of the “volume” that is needed for the oxide to grow. Figure 6–32 illustrates these ideas. At a microscopic level, Si-Si bonds are broken at the Si/SiO₂ interface, O atoms are inserted between the Si

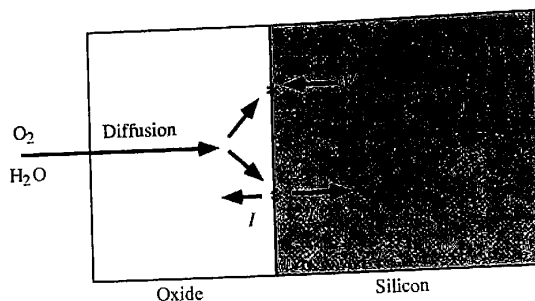
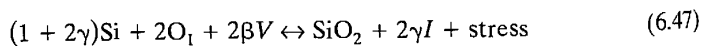


Figure 6-32 Point-defect reactions taking place at the Si/SiO₂ interface during oxidation. Oxidation reactions (*) can consume V or generate I in order to provide the volume needed for the reaction to take place.

atoms, and Si-O bonds are formed. The volume expansion requirement comes from the room needed for the O atoms.

There are two major types of point defects present in crystalline materials like Si—Vacancies (V) or missing Si atoms and Interstitials (I) or excess silicon atoms. We introduced a number of basic ideas about these defects in Chapter 3 where we saw that they can exist in both neutral and charged states and that their concentrations generally increase exponentially with temperature. Volume for the oxidation reaction can be provided at the Si/SiO₂ interface either by consuming vacancies or by generating interstitials. These are the two reactions illustrated in Figure 6-32. We can represent these reactions in the same way we modeled the formation of SiO₂ precipitates in Chapter 3 (Figure 3-25):



The reaction forming SiO₂ consumes vacancies and generates interstitials because the vacancies provide some of the volume needed for the oxidation to proceed, as can removing Si atoms as I . γ is the number of I that contribute to the process per O atom joining the SiO₂, and β is a similar fraction for V . A stress term is also included on the right-hand side because experimentally we know that the point defects do not provide all the necessary volume.

The generation of I is thought to explain the interactions between oxidation and other processes like diffusion (i.e., OED and ORD) because the I which are generated can diffuse far away from the Si/SiO₂ interface and change the diffusivity of dopants which interact with I . We will discuss this further in Chapter 7. The consumption of V as a path to provide the volume required for the oxidation should be favored whenever there are large numbers of V present. This is believed to be the case in heavily doped silicon regions (as we saw in Chapter 3) and this reaction has been used to explain why oxidation proceeds at a much higher rate on heavily doped substrates than it does on lightly doped substrates. We will discuss this more fully in Section 6.5.9.

It is quite clear that the mechanisms proposed in Eq. (6.47) do not provide all the volume required at the oxidizing interface. The fact that grown oxides are under very high compressive stresses is clear evidence for this. In fact, estimates of the number of I which diffuse into the substrate during oxidation, suggest that less than 1 in 10^3 Si atoms does this. The rest are consumed by the oxidation process and end up in the SiO_2 layer. In fact, the number of I actually generated by the oxidation process may be much larger than 1 in 10^3 of the silicon atoms oxidized. Most of them may end up diffusing into the SiO_2 layer where they are oxidized. Nonetheless, the effects that the few point defects that do diffuse into the substrate have on process phenomena in the underlying silicon are quite remarkable.

The coupling between oxidation kinetics and other process phenomena is usually modeled today using reaction diffusion equations and a conceptual framework like that shown in Figure 6-33 [6.37]. An oxidizing interface is characterized by generation and recombination rates for point defects. The balance between G and R determines whether there is a net flux of point defects into or away from the surface. In the case of oxidation, the net result is a flux of interstitials away from the surface and into the bulk. These interstitials interact in the silicon bulk with dopant atoms, producing dopant diffusivities that are proportional to the local concentration of point defects (Chapter 7). This modeling approach has been quite successful in recent years, but it depends on experimentally determined values for G , R , point-defect diffusivities, bulk recombination rates, and many other parameters. The connection to oxidation kinetics is through G and R . For example, in SUPREM IV, G and R for an oxidizing Si/SiO₂ interface are modeled as

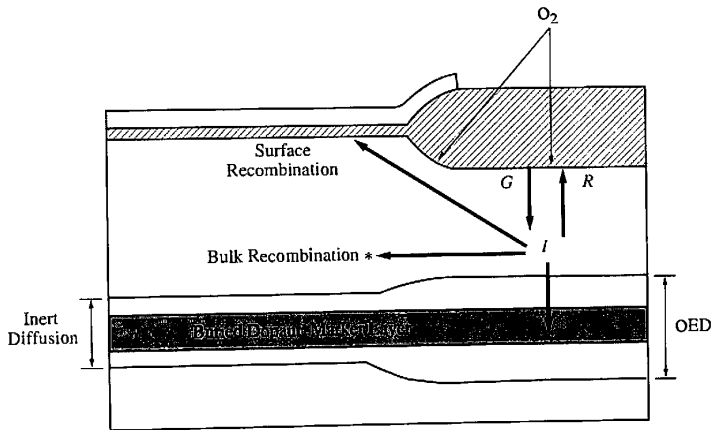


Figure 6-33 Generalized representation of point-defect generation (G), recombination (R , K_B , and K_S) and diffusion processes in silicon. In this example, local oxidation (right side) generates interstitials that diffuse away from the Si/SiO₂ interface, locally enhancing the downward diffusion of the epilayer (OED).

$$G = \Theta \frac{dx_o}{dt} N \quad (6.48)$$

$$R = K_{inert} K_{rat} \left(\frac{dx_o/dt}{B/A} \right)^{K_{pom}} + K_{inert} \quad (6.49)$$

where Θ is the fraction of oxidized silicon atoms that are injected into the substrate as interstitials, dx_o/dt is the oxidation velocity, N is the number of silicon atoms per unit volume, K_{inert} is the recombination velocity at an inert Si/SiO₂ interface, and K_{rat} accounts for the increased recombination rate at an oxidizing interface. Note that Θ is not the same as γ in Eq. (6.47), since Θ represents only the I which are injected into the silicon. All of the K values and Θ are experimentally determined. Other surfaces, such as Si/Si₃N₄ interfaces are modeled with similar expressions but with different parameter values. This modeling approach should be regarded as phenomenological at present because there is no generally accepted underlying physical model.

Models have been proposed to try to explain G and R on a more fundamental level. Hu suggested a regrowth model in which kink sites at the interface act as sites for interstitial recombination [6.38]. Dunham suggested a three-flux model including generation, surface regrowth and segregation of interstitials into the SiO₂ [6.39]. However all of these models must be regarded as speculative at the present time.

Figure 6-34 shows an example of the simulation of these OED effects. The structure is similar to the conceptual drawing in Figure 6-33. After a uniform boron implant, the

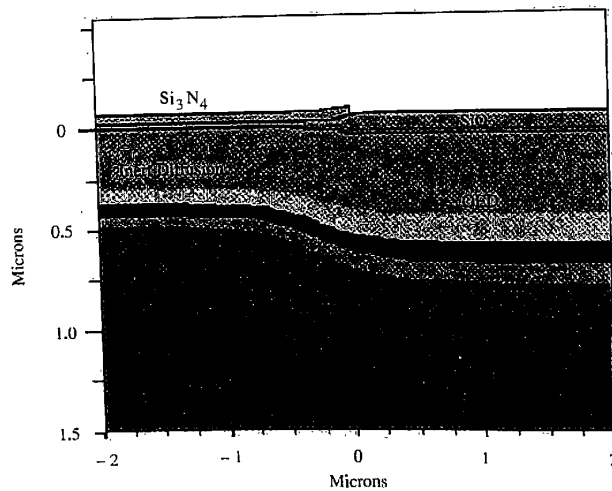


Figure 6-34 Example of the simulation of the OED effect using the ATHENA simulator [6.12]. A uniform $3 \times 10^{14} \text{ cm}^{-2}$, 15-keV implant was performed through a pad oxide, followed by deposition of a nitride masking layer on the left half of the structure. A 60-min H₂O oxidation at 850°C produced the deeper diffusion on the right side.

right half of the structure is oxidized. The I injected by the oxidizing interface enhance the boron diffusivity, producing a deeper junction on the right side. Note the two-dimensional effects associated with the I diffusion away from the surface. The OED effect extends almost a micron under the nitride mask.

6.5.9 Substrate Doping Effects

It has been known for many years that highly doped substrates oxidize more rapidly than do lightly doped wafers. The effect is particularly important at lower temperatures and for thinner oxides. Under these conditions, the difference in oxidation rates between heavily doped and lightly doped regions can be three to four times [6.40]. The difference is more pronounced for N^+ regions than for P^+ regions and is more pronounced for low-temperature compared to high-temperature oxidations. Analysis has indicated that the mechanism associated with the faster oxidation is occurring at the Si/SiO₂ interface. In other words, B/A and not B is affected in the Deal-Grove model. This is perhaps not too surprising because N-type dopants tend to segregate into the silicon during oxidation. As a result, the dopant concentrations in the growing SiO₂ layer are usually fairly small, even if the substrate is heavily doped. As a result, the properties of the grown SiO₂ layer are to first order independent of the substrate doping. Therefore, B , which is primarily determined by the diffusion process of the oxidant through the SiO₂, should not be greatly affected.

While a number of possible physical mechanisms might explain the effect of high doping on B/A , the most widely accepted model relies on the V mechanism illustrated in the top portion of Figure 6-32. The basic idea is as follows. Oxidation requires volume expansion. This volume can be provided by consuming V at the growing Si/SiO₂ interface. We saw in Chapter 3 that the total number of V is much higher in extrinsic, highly doped regions than in intrinsic, lightly doped material. Physically, this is because vacancies can exist in a number of charge states (V^0 , V^+ , V^- and V^{--}) and the charged vacancies have discrete energy levels in the silicon bandgap. As the Fermi level E_F moves away from the middle of the bandgap in extrinsic material, it crosses over the V^- or V^{--} levels in N^+ material or the V^+ level in P^+ material, causing large increases in the charged vacancy populations. In Chapter 3 we expressed these ideas mathematically and found that we could calculate the numbers of vacancies present in the crystal at any given temperature and doping level.

If these vacancies are now able to provide sites for the oxidation reaction, then it should be the case that [6.40]

$$\frac{B}{A} = R_1 + KC_{V^r} \quad (6.50)$$

where R_1 represents all mechanisms other than the vacancy driven process and C_{V^r} is the total vacancy population in all charge states present in the material at the oxidation temperature. Given a temperature and a doping level (i.e., E_F), we can calculate C_{V^r} using the equations developed in Chapter 3.

Eq. (6.50) is usually rewritten in the following form for convenience:

$$\frac{B}{A} = \left(\frac{B}{A}\right)^i \left(1 + 2.62 \times 10^3 \exp\left(-\frac{1.1 \text{ eV}}{kT}\right) \left(\frac{C_{V_i^r}}{C_{V_i^r} - 1}\right)\right) \quad (6.51)$$

$C_{V_i^r}$ is the total vacancy concentration in intrinsic material at the oxidation temperature. The $2.62 \times 10^3 \exp\left(-\frac{1.1 \text{ eV}}{kT}\right)$ factor is an empirically determined parameter that presumably physically represents something to do with the effectiveness of the vacancy reaction compared to other components in B/A . This expression has been quite successful in modeling the oxidation of highly doped regions and is implemented in SUPREM IV.

Several points should be made with regard to this model. The first is that since it is based on V populations and these depend only on T and E_F , it makes no difference if the N^+ regions are formed with arsenic or phosphorus. The same electrically active surface concentration should produce the same E_F and hence the same B/A . Second, the relevant concentration that should be used in calculating E_F and hence $C_{V_i^r}$ is the surface concentration since that is where the oxidation reaction is taking place.

Segregation and pileup effects may increase this concentration above bulk values. We introduced the idea of segregation in Chapter 3 in connection with crystal growth. In that case we were concerned with the distribution of an impurity across the solid/liquid interface of the growing crystal. In the oxidation process, we have a Si/SiO₂ interface where dopant segregation also occurs. N-type dopants tend to segregate into the silicon whereas boron prefers to be in the SiO₂. As a result, the surface concentration for N-type dopants will tend to build up as the oxidation proceeds, and it will tend to decrease during oxidation for P type. Because of this, the B/A values calculated from Eq. (6.51) will change with time during the oxidation. The oxidation rate will therefore change with time. Recent experiments that carefully measured the dopant profiles near the surface clearly showed the time dependence of the oxidation rate and its correlation with the dopant pileup process [6.41]. A significant fraction of the dopant in the piled up layer at the surface may not be electrically active and hence, according to the $C_{V_i^r}$ model, would not affect the oxidation rate.

The model also predicts that P^+ substrates should not show nearly as large an increase in oxidation rate as do N^+ regions, because as we saw in Chapter 3, there are fewer total vacancies present in P^+ regions for a given doping level. This agrees with experiments. Finally, because P-type dopants like boron tend to segregate into the growing oxide layer, rather than staying in the silicon substrate, high boron concentrations in the growing SiO₂ can result. There are indications that this may change the structure of the oxide layer sufficiently to observe changes in oxidant diffusivity through these films. As a result, the parabolic rate constant B may be slightly different (larger usually) than it would be at the same temperature if the oxide were being grown on a lightly doped or N^+ region. This effect has not been characterized well enough to model.

Other models have been proposed to explain the doping dependence of oxidation. Some are based on the idea that high dopant concentrations at the Si/SiO₂ interface

could change the Si-Si bond energy and hence affect B/A . Others are based on the idea that the structure of the Si/SiO₂ interface region is changed by the high dopant concentration, again resulting in a change in B/A in the Deal-Grove model [6.42, 6.43]. At this point in time, none of these models has been sufficiently quantified to use them in oxidation simulators so Eq. (6.50) is the model normally used.

An example of a simulation involving heavily doped oxidation is shown in Figure 6-35. On the left side, a low temperature oxidation (800°C) was simulated whereas a higher temperature (1000°C) was simulated in the right side. In each case, the right side of the structure is heavily doped and the left side is lightly doped. The heavily doped region grows about five times as much oxide as the lightly doped control at 800°C, compared to only about twice as much at 1000°C. The transition region between the lightly doped and heavily doped areas in each simulation can only be properly modeled using 2D diffusion and 2D oxidation models and requires a numerical simulator such as SUPREM IV used in this example.

6.5.10 Polysilicon Oxidation

It is not uncommon in silicon technology for there to be thin films of other materials locally on the wafer surface when an oxidation takes place. The most common examples are thin films of polysilicon, Si₃N₄, and various silicides. We saw several examples of these kinds of processes in Chapter 2. In the next three sections, we briefly consider the effect of oxidation on these thin film materials.

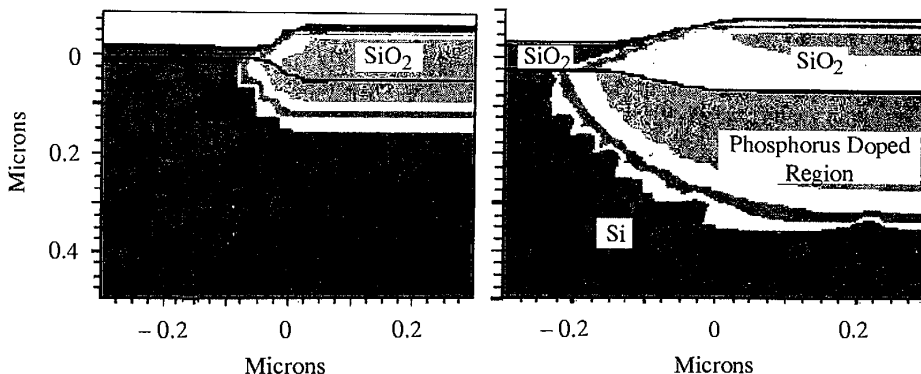


Figure 6-35 Oxidation of heavily doped regions using the ATHENA simulator [6.12]. A $5 \times 10^{15} \text{ cm}^{-2}$ 20-keV phosphorus implant was performed on the right half of the structure. An oxidation was then performed for 30 min in H₂O at 800°C (left) and for 5 min in H₂O at 1000°C (right). The contours correspond to the diffused doping profiles. The solid lines at the top of the structure define the grown SiO₂ layers.

To first order, the same set of models described above for single-crystal silicon can also be used for polysilicon. This is, in fact, the strategy used in most process modeling programs. However, this approach is not always completely satisfactory for two fundamental reasons. First, polysilicon is a polycrystalline material. Second, polysilicon films are thin- or finite-volume structures.

The fact that polysilicon materials are polycrystalline means that a variety of crystal orientations may be present in various grains. The grain structure associated with the polysilicon layers is a strong function of deposition conditions and subsequent heat treatments. In general, deposition at lower temperatures results in smaller grain sizes. In fact depositions below about 600°C tend to produce amorphous rather than polycrystalline films. Annealing at high temperatures generally causes the grains to grow and the number of grains to be reduced. The dominant crystal orientation is often (110). An excellent reference on polysilicon, its properties and applications is [6.44].

There are several implications of the polycrystalline nature of these films for oxidation kinetics. The most obvious is simply that B/A will vary from grain to grain and perhaps with time if the poly grain structure changes during the oxidation. We discussed orientation effects in Section 6.5.5 and saw there that B/A can change by up to 70% depending on the orientation. (110) surfaces oxidize at a rate intermediate between (111) and (100) (the two extremes) and often what is done to model polysilicon oxidation is to simply use an average B/A value characteristic of the (110) orientation. For relatively thin oxides, where B/A is important, we should expect to see variations in oxide thickness across a polysilicon surface because of the varying grain orientations. For thicker oxides, these differences tend to disappear because as we saw in Section 6.5.5, the parabolic rate constant B is not a function of the underlying crystal structure.

Most polysilicon thin films used in silicon technology are heavily doped because they are usually used as local interconnects, gate electrodes in MOS devices, or for bipolar emitter regions. In these situations, modeling any oxidation that takes place will have to include doping effects. This becomes a complicated problem because of the polycrystalline nature of the film, and because of its finite volume. Dopants diffuse very rapidly in polysilicon films, because they tend to migrate along grain boundaries and then into the grains. Grain boundary diffusion is a very rapid process (at least 10^2 times faster than bulk diffusion). As a result, dopant profiles tend to be uniform in polysilicon thin films after a very short time at high temperatures. To first order the dopant chemical concentration is simply given by

$$[\text{Dopant}]_{\text{chem}} = \frac{\text{ImplantDose}}{\text{PolyThickness}} \quad (6.52)$$

when the poly is doped by ion implantation. However, the electrically active concentration may be much lower than this because some of the doping atoms segregate to grain boundaries where they are not electrically active and because some of the free carriers are also trapped at grain boundaries. The net result is that n or p and hence E_F may be quite different from the values one would normally find in single-crystal silicon for the same doping concentration. The result of these effects is that applying models like Eq. (6.51) is often not straightforward in polysilicon. It is often possible to do so, but the correct n or p value must be used to find E_F and hence $C_{V'}.$

The finite volume of polysilicon thin films also has some interesting implications. Many of these issues also apply to silicon on insulator (SOI) structures such as that illustrated in Figure 3–26. Most often, the polysilicon layer is deposited on an underlying SiO₂ layer. During an oxidation, the polysilicon is consumed and the layer will therefore get thinner. Most N-type dopants prefer to stay in silicon or polysilicon rather than segregating into growing SiO₂ layers. This implies that the chemical and probably electrically active dopant concentrations will increase during the oxidation. In fact, very high doping levels may be reached since poly layers are usually quite heavily doped to improve their conductivity. The electrically active dopant concentrations may easily reach maximum solubility levels and in extreme cases, new phases like SiP may be formed. Such phases may result in unusual process problems. For example, SiP is soluble in HF which may be used later as an oxide etch. If the SiP dissolves, “holes” may be left in the polysilicon layer resulting in yield problems. Finally, there are some processes in which the polysilicon layer is fully oxidized. Usually this is done locally as part of a LOCOS type process. Modeling of such process steps is challenging because extreme conditions of dopant concentrations and other effects may be encountered as the poly gets very thin toward the end of the oxidation process. If the polysilicon layer is deposited on an underlying silicon layer (as in a bipolar emitter region, for example), the process physics will be quite different because there is no blocking oxide layer below the poly in that case.

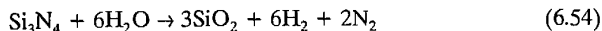
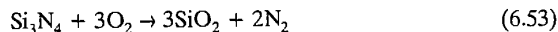
A number of studies of polysilicon oxidation have been done as a function of temperature, poly doping level, and so forth. To first order, these experiments can be explained by the single-crystal models we described earlier and this is the approach normally taken in simulators like SUPREM IV. However, the effects of grain boundaries and the poly finite volume can result in greatly different oxide growth kinetics than those found on single crystal substrates under some conditions. At this point in time, no general-purpose models have been developed, beyond those we have described, to account for these effects.

Figure 6–36 shows an example of a simulation involving polysilicon oxidation using ATHENA. The silicon substrate is (100) orientation and hence has a slower oxidation rate than the poly (both are undoped in this example). The poly is mostly consumed by the oxidation. Notice also the encroachment of the growing oxide under the poly because of 2D oxidation.

6.5.11 Si₃N₄ Growth and Oxidation Kinetics

We have seen several examples in earlier chapters of the use of thin deposited Si₃N₄ films in silicon technology. By far the most common use is as a mask against oxidation in LOCOS type processes. In this application, we are interested in the ability of CVD Si₃N₄ to withstand oxidizing environments for reasonably long times at high temperatures. The use of Si₃N₄ in this way is illustrated in Figure 6–37. The process forms a surface layer of SiO₂ on the nitride.

Conceptually, the following reactions may take place



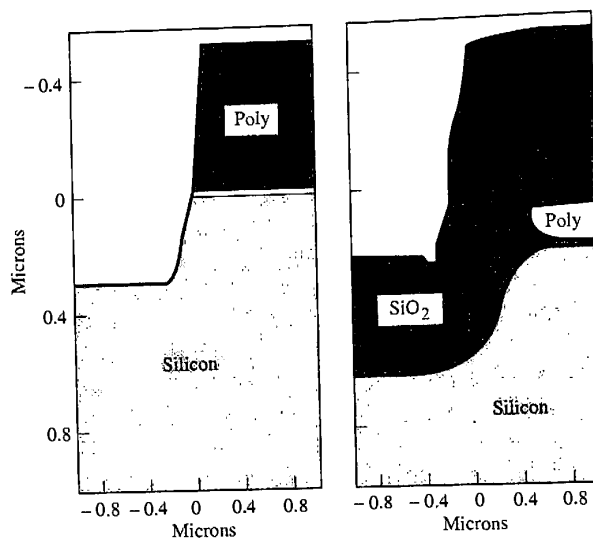


Figure 6-36 Example of an oxidation simulation involving polysilicon using the ATHENA simulator [6.12]. The initial structure (left) has a polysilicon layer on SiO₂ on silicon. The left half of the structure is etched to expose the Si substrate. A 60-min H₂O oxidation at 1000°C is simulated on the right.

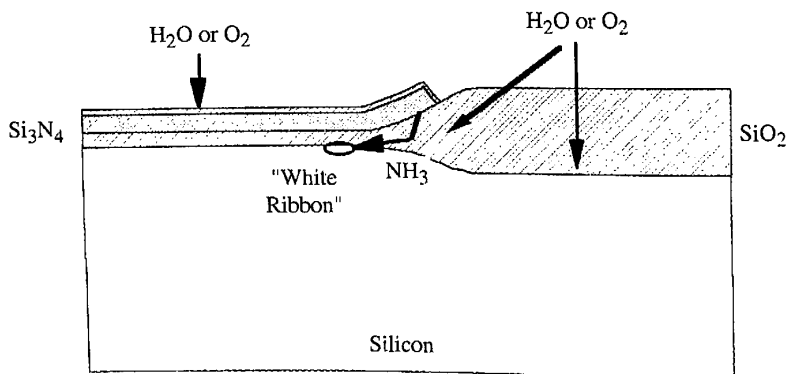


Figure 6-37 LOCOS process illustrating local oxidation of silicon, oxide formation on the Si₃N₄ surface, and "white ribbon" formation due to oxidation of the Si₃N₄.



From a practical point of view, we are usually most interested in the latter two reactions because LOCOS type processes attempt to grow fairly thick field oxides and this usually implies an H_2O oxidation. Physically, the oxidation process should be somewhat similar to silicon oxidation. The H_2O must diffuse through the growing SiO_2 layer and then react at the $\text{SiO}_2/\text{Si}_3\text{N}_4$ interface to form new oxide layers. If this is how the oxide grows, then we might expect a linear parabolic growth law similar to the Deal-Grove model for silicon oxidation. However, the process is apparently not that simple because more complex growth kinetics have been reported [6.45, 6.46]. It has been suggested that the reaction byproducts (N_2 or NH_3) have an effect on the oxidation, either by modifying the interface reaction, or by changing the diffusion process of the indiffusing oxidant as they diffuse out through the oxide.

Lacking a clear physical model, empirical relationships have been developed to allow prediction of Si_3N_4 oxidation kinetics. One such relationship suitable for H_2O oxidations is as follows [6.46]:

$$\Delta x_N = 1.1 \times 10^7 \exp\left(-\frac{1.9 \text{ eV}}{kT}\right) t^{0.7} P \quad (6.56)$$

where Δx_N is the thickness of the nitride consumed in nm, t is the oxidation time in min, and P is the H_2O oxidant pressure in atm. The oxidation process involves a volume expansion just as silicon oxidation does; the resulting oxide thickness is about 1.6 Δx_N . A typical Si_3N_4 film of 80 nm would therefore be completely consumed by a 1050°C oxidation in about 16.6 hours.

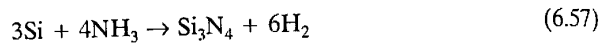
In dry O_2 oxidations, it has been suggested that the reaction that takes place is actually more complicated than Eq. (6.53) would suggest [6.47]. The oxidation may actually take place in two steps with a layer of $\text{Si}_2\text{N}_2\text{O}$ forming on the Si_3N_4 surface and then a layer of SiO_2 forming on top of the $\text{Si}_2\text{N}_2\text{O}$. In any case, the oxidation rate is much slower than in an H_2O ambient and the SiO_2 grown on Si_3N_4 during O_2 oxidations can normally be neglected.

One of the subtleties of LOCOS type processes is also illustrated in Figure 6-37. H_2O diffusing through the growing SiO_2 layer can also react with the underside of the Si_3N_4 layer, generating NH_3 as a byproduct of the oxidation process. [See Eq. (6.55).] It is found experimentally that the NH_3 can then diffuse to the Si/SiO_2 interface and react there to form a thermally grown Si_3N_4 layer. This is sometimes called the “white ribbon” effect or the Kooi effect after the person who first explained the process. The “white ribbon” name came about because of the characteristic appearance in a microscope of the Si_3N_4 layer as a ribbon around the edge of device active regions (just under the edge of the top deposited Si_3N_4 layer). This grown Si_3N_4 layer can sometimes cause problems if it is not removed in later processing because it can slow down or stop later oxidations locally where it is present. Fortunately it is easily removed by chemical etching in hot H_3PO_4 or by oxidizing it after the top Si_3N_4 layer is removed.

The white ribbon effect is a specific example of the fact that it is possible to thermally grow thin Si_3N_4 layers directly on silicon by exposing the silicon to N_2 or NH_3 at high

temperature. This process step is rarely used in VLSI structures because the thickness of Si₃N₄ that can be grown is limited. However the process does have potential for future thin dielectrics in MOS structures and it is of interest as a technique in studying dopant diffusion as we will see in Chapter 7, so we will briefly consider it here.

Most attempts to thermally grow Si₃N₄ layers directly on silicon have used NH₃ rather than N₂ because the growth rates are substantially faster. This may be a result of the hydrogen byproduct of the NH₃ reaction which may react with trace amounts of O₂ in the ambient to prevent competing oxidation processes from taking place although this is speculation at this point. At temperatures below about 900°C, oxidation is thermodynamically favored. Films of Si₃N₄ grown in NH₃ are usually between 2.5 and 7.5 nm thick. The chemical reaction is believed to be



The films appear to grow very rapidly to about half their final thickness and then grow slowly to a final thickness that depends primarily on temperature. The growth kinetics are not fully understood but are generally thought to occur via a surface reaction limited process at the Si/Si₃N₄ interface at first (the fast part of the growth), followed by a slower diffusion limited growth phase. This sounds very much like the Deal-Grove model for oxidation. However, kinetic data for the nitridation process has generally been fit with either a logarithmic or a power law expression. One example is the following empirical growth law [6.48]:

$$x_N = at^b \quad (6.58)$$

$$\text{where} \quad a = 92 \exp\left(-\frac{0.38 \text{ eV}}{kT}\right) \quad \text{and} \quad b = 0.018 \exp\left(\frac{0.24 \text{ eV}}{kT}\right) \quad (6.59)$$

a is in nm and t is in min. The thermally grown Si₃N₄ layers reach thicknesses of 2.5–6 nm in several hours of growth at temperatures of 950–1200°C. A number of recent studies have shown that these growth rates and final thicknesses can be substantially enhanced by growing the Si₃N₄ layers in an N₂ or NH₃ plasma. Such plasmas create highly reactive species, both neutral and ionized, and may also create electric fields to help drive transport processes. It has been possible to grow Si₃N₄ layers 10–30 nm thick with these processes. The growth kinetics have not been modeled as yet.

Commercially available process simulators generally have not yet implemented models for Si₃N₄ oxidation or for the nitridation of silicon. In many applications it is not a bad approximation to assume that Si₃N₄ is not oxidized at all in an O₂ or H₂O ambient and this is by default what simulators like SUPREM IV currently do.

6.5.12 Silicide Oxidation

Beginning in the late 1970s, it became apparent that the resistivity of polysilicon would soon become a limitation on the speed of ICs. In virtually all modern MOS technologies, polysilicon is used both for the gate electrodes and for short-distance interconnects. When it is used as an interconnect, the RC delays associated with charging and discharging the interconnect lines can become appreciable if the line resistance is too

high. Polysilicon resistivities cannot be made much lower than about $400 \mu\Omega\text{cm}$ (or about $10\Omega/\text{sq.}$ for a $0.4 \mu\text{m}$ thick film). This is unacceptable in many applications. As a result, silicides like TiSi_2 and WSi_2 began to be investigated by 1980. We will discuss the principal properties of these films in Chapter 11 when we deal specifically with contacts and interconnects. However, in many cases, silicides are deposited on top of polysilicon to reduce the resistance of poly lines and the composite structure is sometimes thermally oxidized. We did not use thermal oxidation of silicides in the process flow in Chapter 2 (a deposited oxide layer was used on the TiSi_2 instead—see Figure 2–38).

The structure most commonly encountered when silicides are oxidized is shown in Figure 6–38. The silicide/poly part of the structure forms the gate electrode; the lower SiO_2 layer is the gate dielectric and the upper oxide is grown or deposited to insulate the silicide from other metal layers that may be deposited above it. Silicides are generally not used directly on gate oxides because there may be adhesion problems, oxidation problems, gate oxide reliability problems, or simply undesired work functions in the MOS structure.

When such composite structures are thermally oxidized in O_2 or H_2O , a layer of SiO_2 is normally found to grow on the upper surface of the silicide. In principle, either SiO_2 or the metal oxide MO_x could grow. Thermodynamic calculations [6.49] suggest that the MO_x is actually favored on Hf, Zr, and perhaps Ti silicides and the difference between the heats of formation of MO_x and SiO_2 are small on many of the other silicides. Some uncertainty exists in these calculations, but they do suggest that kinetics must play an important role in the system choosing to form SiO_2 rather than MO_x in most cases. MO_x is found to be the preferred oxide on HfSi_2 and on TiSi_2 at temperatures below 900°C . With these few exceptions, SiO_2 appears to grow in all other cases. By kinetics playing an important role, we mean that Si atoms must react more rapidly with the oxidant than do the metal atoms, even though both reactions may be possible thermodynamically.

The oxidation process appears to proceed in a manner very similar to silicon oxidation. O_2 or H_2O diffuse through the SiO_2 layer to the SiO_2 /silicide interface where a reaction similar to that which occurs in silicon oxidation takes place. The silicon atoms needed to form the SiO_2 can be provided either by diffusion upward from the poly, or by M-Si bonds being broken. In the later case, the metal atoms would then diffuse downward to form a new layer of silicide at the silicide/poly interface. In both cases, the silicide layer is conserved. Both possibilities are observed in practice. CrSi_2 , CoSi_2 , NiSi_2 ,

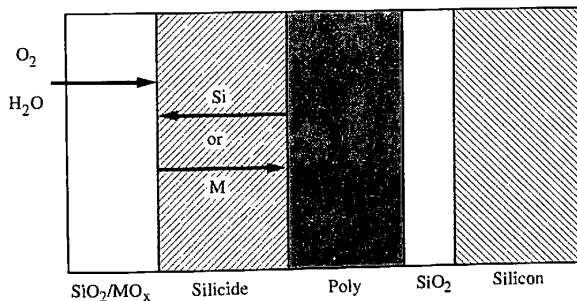


Figure 6–38 Oxidation of a polycide (silicide on poly) structure. M is the metal used to form the silicide.

PdSi, and PtSi all appear to grow oxides primarily by dissociation of the silicide and diffusion of the metal to the silicide/poly interface where new silicide is formed. TiSi₂, WSi₂, TaSi₂, and perhaps others form oxides by the diffusion of Si atoms from the poly layer. In all cases, the growth kinetics are well described by the familiar linear parabolic model

$$\frac{x_o^2 - x_i^2}{B} + \frac{x_o - x_i}{B/A} = t \quad (6.22)$$

B is found to be exactly the same for silicide oxidations as it is for silicon oxidation. This should not be too surprising because the process is the same in the two cases—oxidant diffusion through SiO₂. B/A values extracted from silicide oxidation data are all significantly larger than the values found for silicon oxidation (typically 10–20 times larger) [6.50]. Apparently the processes contributing to B/A in the silicide structure, metal or silicon diffusion, M-Si or Si-Si bond breaking, and so on, are much faster than the corresponding processes in silicon oxidation. In practice, this means that the oxide growth is often simply parabolic ($B/A \rightarrow \infty$).

Difficulties can be encountered in oxidizing silicide structures like that shown in Figure 6–38. For example, if a barrier layer such as a thin native oxide exists between the poly and silicide layers (due to improper cleaning between poly and silicide depositions for example), the reaction at that interface may be slowed down significantly. This can result in voids forming at the poly/silicide interface, apparently because the Si needed to form the growing SiO₂ can be obtained only from localized regions where pinholes exist in the blocking layer. In TiSi₂ in which the metal oxide forms at low temperatures, it is important to ensure that the oxidizing ambient is not turned on until the wafers have reached high temperature in the furnace. There have also been reports of oxidation difficulties in O₂ ambients although these are not uniformly observed. Most data in the literature are for H₂O oxidations. Finally, there are also many reports of problems with gate oxide integrity if the polysilicon layer under the silicide shrinks to below about 200 nm during the oxidation. The poly is consumed during the process, so this constraint may limit the maximum thickness of oxide that can be reliably grown. Commercially available simulation tools have generally implemented models for silicide deposition and for reactions between refractory metals and silicon to form silicides (see Chapter 11), but have not yet implemented oxidation models for silicides, probably because deposited oxides are commonly used in silicide processes.

6.5.13 Si/SiO₂ Interface Charges

Toward the beginning of this chapter, we discussed the four basic kinds of charges that are typically found at semiconductor/insulator interfaces. Figure 6–6 summarizes these charges. Of the four charges illustrated, only two (Q_f and Q_{it}) will be discussed in detail here. Mobile charges in dielectric layers (Q_m), have been largely eliminated from silicon technology because of careful attention to cleanliness during manufacturing. They occasionally are still a problem, but such situations are rare and usually quickly eliminated

by discovering the source of the contamination (chemicals, handling etc.) and eliminating it. Oxide trapped charge (Q_{ot}) is of concern in applications in which the device structure is exposed to ionizing radiation either during manufacturing or during later use of the IC (in space applications, for example). It is also of concern in devices which pass current through oxides as part of normal device operation (EPROMs, for example), and more generally in very small MOS devices because of hot carrier effects which can inject carriers into the gate dielectric. However little quantitative modeling of Q_{ot} has been done to this point. Experiments have determined ways of growing oxides in order to minimize susceptibility to Q_{ot} , largely through proper anneals and choice of growth temperature and ambient. Q_{BD} measurements have also become common for characterizing thin gate insulators. Much of this work on Q_{ot} is, however, empirical at present. Much more work has been aimed at understanding and predicting Q_f and Q_{it} and we will focus on that work here.

Virtually all of the experimental techniques that are available to investigate interface charges are electrical in nature. The problem is easily visualized in connection with the TEM photo in Figure 3-15. Normal densities of Q_f and Q_{it} are around 10^9 – 10^{12} cm⁻² or about 1 charge for every 10^3 – 10^6 silicon atoms at the interface. Physical techniques such as TEM are simply not sensitive enough to “find” such charges. It is quite literally like “looking for a needle in a haystack.” Electrical methods, such as the CV technique described earlier, do have the required sensitivity and are commonly used to measure Q_f and Q_{it} . Such electrical measurements are also the quantities directly of interest from the point of view of devices. However, it is usually not straightforward to relate the electrical results to a physical, atomic-level model.

Many attempts have been made to relate the Si/SiO₂ interface charges to specific atomic level structures. Figure 6-39 illustrates some of these ideas, although the ideas represented by this figure should be regarded as speculative. Most workers today agree that the interface is almost atomically abrupt. If there is a transition region between the two materials, it is not more than one or two atomic distances thick. The bulk of the oxide consists mainly of rings of SiO₄ tetrahedra (see Figure 6-5), each with six silicon atoms in the ring. In the near interface region rings with smaller numbers of tetrahedra may be present. Many of the models for interface charges associate them with incompletely oxidized silicon atoms, often designated as ≡Si· for trivalently bonded silicon atoms. The · is meant to represent a dangling or unsatisfied bond. Both Q_f and Q_{it} have been modeled in this way, although there are clear experimental differences between the two and so there must be some structural difference between the two charges.

The principal electrical difference between Q_f and Q_{it} is that Q_f is a fixed positive charge which does not change during normal device operation, while Q_{it} is in communication with the silicon surface and can trap holes or electrons to become positively or negatively charged depending on where the Fermi level is at the surface. One explanation for this that has been suggested is that the ≡Si· corresponding to Q_f are located physically further away from the interface and thus are unable to trap carriers. Another explanation is that the energy levels corresponding to Q_f are located outside the silicon bandgap whereas those associated with Q_{it} are inside the bandgap. Thus Q_{it} states can be charged and discharged as E_F moves around in the bandgap. Other explanations

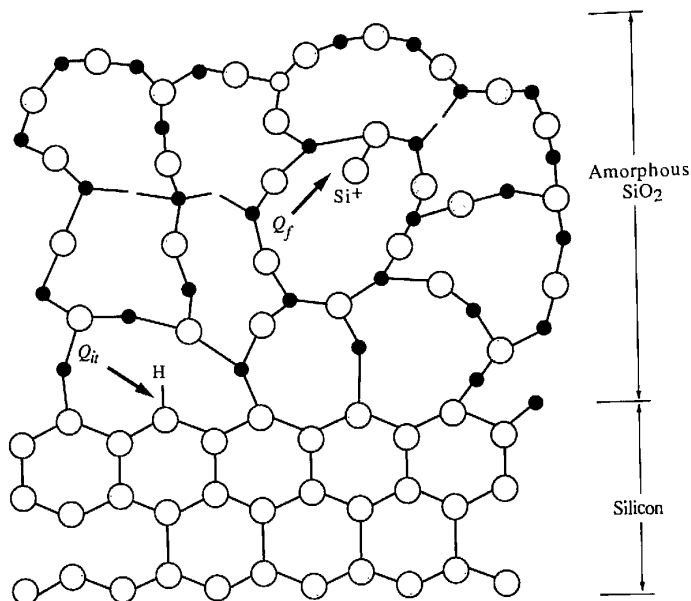


Figure 6-39 Conceptual atomic level picture of the Si/SiO₂ interface showing possible structural origins of Q_f and Q_{it} . Q_f corresponds to the excess Si⁺ atom; Q_{it} is shown as an incompletely bonded Si atom near the interface, passivated in this example with an H bond. The larger circles correspond to Si atoms; the smaller circles to O atoms.

have suggested that the charges are due to stretched or distorted bonds at the interface, that is stress at the interface caused by the size mismatch between Si and SiO₂, and that different bonding arrangements account for the different charge types and the different energy levels in the bandgap found for Q_{it} states. There does appear to be a correlation between the roughness of the Si/SiO₂ interface and the resulting charge densities (the rougher the interface, the higher Q_f and Q_{it}). In addition, there is nearly always a strong correlation between the values of Q_f and Q_{it} that result from a given process, again suggesting a common origin [6.51]. Many other models have also been suggested, but none has been shown to be conclusively correct.

Whatever the physical origins of Q_f and Q_{it} , much is known about how they vary with process conditions. One of the earliest sets of such experimental data was presented by Deal in the form of the “ Q_f triangle” shown in Figure 6-40 [6.52]. The experiments behind this data were done by oxidizing a wafer in a furnace (or annealing it in Ar) at the indicated temperature, and then pulling the wafer from the furnace fairly rapidly (few seconds) to “freeze in” the Q_f corresponding to that condition. Each curve appears to represent equilibrium values for that ambient and that temperature. That is, as the arrows on the figure illustrate, any point on any curve can be reached simply by changing the ambient and temperature in the furnace [6.53].

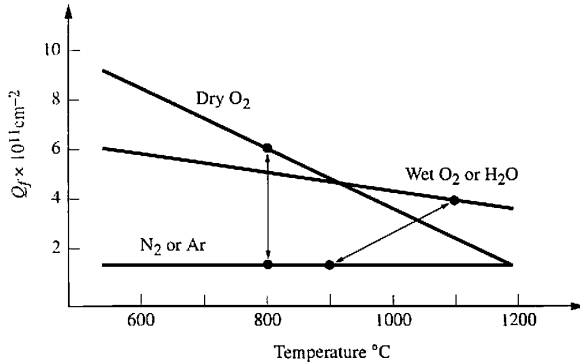


Figure 6-40 Variation of fixed oxide charge Q_f with oxidation temperature and ambient [6.52]. The arrows indicate that changes in temperature or ambient result in equilibrium values of Q_f after a transient time that depends on the temperature. Values shown are for (111) silicon; (100) values are about a factor of 3 lower.

In general, if such a change is made, a transient time will be required to reach the new Q_f value. This time is longer at lower temperatures. These transients and the final value of Q_f achieved after a high temperature Ar anneal can be empirically described by [6.53]

$$Q_f(t) = (Q_f(0) - Q_{fe}) \exp\left(-\frac{t}{\tau}\right) + Q_{fe} \quad (6.60)$$

where

$$\tau = 5.41 \times 10^{-10} \exp\left(\frac{2.42 \text{ eV}}{kT}\right) \text{ min}$$

$$Q_{fe} = 5.24 \times 10^9 \exp\left(\frac{0.36 \text{ eV}}{kT}\right) \text{ cm}^{-2} \quad \text{for (111)}$$

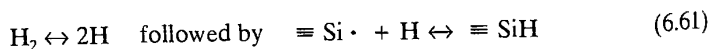
$$Q_{fe} = 5.22 \times 10^7 \exp\left(\frac{0.69 \text{ eV}}{kT}\right) \text{ cm}^{-2} \quad \text{for (100)}$$

Q_f is the fixed oxide charge density per cm^2 and Q_{fe} is the final or equilibrium value in Ar for any process temperature (the bottom curve in Figure 6-40). In almost all cases of practical interest, it is desired to minimize Q_f and so a final high temperature Ar anneal is used toward the end of a device fabrication sequence. The time required to reach Q_{fe} is only a few minutes at 1000°C , but many hours at 700°C . N_2 was the most common anneal ambient until recently because it is less expensive than Ar. However, as control of Q_f has become more important in modern device structures, it has been found that Ar is a better choice. The reason is that Q_f has been found to increase with time in N_2 after the transient described by the above equations is complete. That is, Q_f reduces to Q_{fe} and then begins to increase again in N_2 . The reason likely is that Si/N_2 reactions can take place at the Si/SiO_2 interface whereas such reactions do not take place in Ar. The anneal conditions are thus less critical in terms of time and temperature if Ar is used.

Finally, it should be pointed out that Figure 6-40 indicates that Q_{fe} is essentially independent of temperature. This is historically how the " Q_f triangle" has been presented [6.51]. More recent data on which Eq. (6.60) is based, indicate that Q_{fe} is actually smaller at higher temperatures [6.53].

We mentioned above that the interface states or traps (Q_{it}) can interact with carriers during normal device operation, by capturing or releasing holes and electrons. This distinguishes Q_{it} from Q_f . Another very important distinction between the two is that the Q_{it} states can be passivated through hydrogen annealing at fairly low temperatures (300–500°C) whereas Q_f is unaffected by such an anneal. By passivated we mean that once the Q_{it} traps are bonded with H atoms, they are no longer electrically active and no longer trap carriers. The interface state conceptually illustrated in Figure 6-39 is shown passivated with a bonded H atom. The mechanism of passivation is believed to simply be hydrogen diffusion through the SiO₂ layer to the Si/SiO₂ interface and reaction there to form Si-H bonds. Since the process takes place readily at temperatures metal layers can tolerate, this step is normally done at the very end of the chip fabrication process. If it is done earlier, the Si-H bonds may break at the higher temperatures used for diffusions, implant anneals, and so on, and the hydrogen may then outdiffuse.

Because of the importance of this passivation process, it has been carefully studied over many years in order to understand not only the origins of the interface states themselves, but also how the passivation works. Most of the atomic level models for interface states are based on excess silicon or incompletely bonded silicon atoms with the trivalent silicon defect $\equiv\text{Si}\cdot$ we discussed above being a leading candidate. The annealing or passivation process is usually modeled with molecular hydrogen diffusion to the Si/SiO₂ interface followed by hydrogen dissociation to form atomic hydrogen and then chemical bonding between the $\equiv\text{Si}\cdot$ defect and H. This is often expressed simply as



Both reactions are reversible. The annealing process itself has been shown to follow a power law dependence on time, of the following form [6.54].

$$Q_{it} = \frac{Q_{it}(0)}{(1 + Kt)^\eta} \quad (6.62)$$

where $K = 8.4 \times 10^{11} \exp\left(-\frac{1.21 \text{ eV}}{kT}\right) \text{ sec}^{-1}$ and $\eta \approx 0.55$ for (100) substrates.

The final annealed value is below 10^{10} cm^{-2} at all normal anneal temperatures (350–500°C). Times on the order of 30 minutes are required to reach this minimum at 350°C; shorter times are adequate at higher temperatures. The kinetics appear to be a little different for (111) surfaces [6.54]. A number of detailed physical models have been proposed to account for these kinetics [6.54–6.56].

Several additional points should be made regarding Q_{it} . Usually the diffusion process of H₂ getting from the ambient to the Si/SiO₂ interface is neglected in the annealing kinetics. This is nearly always a reasonable thing to do because

$$D_H \approx 7.2 \times 10^{-5} \exp\left(-\frac{0.58 \text{ eV}}{kT}\right) \text{ cm}^2 \text{ sec}^{-1} \quad (6.63)$$

which implies quite short time delays for diffusion through normal oxide layers. The fact that Q_{it} annealing kinetics are generally independent of oxide thickness further substantiates this argument. However, many modern IC technologies use polysilicon gates and H_2 may have to go under the edges of such films and then diffuse laterally under them. In large area structures such as MOS capacitors, this can involve significant diffusion times [6.56]. Polysilicon gates have the advantage, however, of preventing outdiffusion of H_2 , so that the annealing process is more tolerant of high-temperature exposure than are bare oxide surfaces or Al covered structures.

Al covered oxides actually do not need an external source of H_2 because of the reaction



which occurs at the Al/SiO₂ interface. It is believed that cleaning steps prior to Al deposition leave trace amounts of water on the oxide surface, which accounts for the presence of the OH in the above reaction. This process is sometimes called the "Alneal" reaction. The oxide is denoted "AlO" rather than Al₂O₃ because the composition is not known and likely changes during the annealing process. Finally, it is also possible to implant H ions through the oxide to the vicinity of the Si/SiO₂ interface. This process can provide the H needed for the Q_{it} annealing, although the process is not very common since other techniques using annealing furnaces are simpler.

In summary, modern silicon technologies use a combination of a high temperature Ar (or N₂) anneal at a temperature on the order of 900–1000°C as the last high-temperature step to reduce Q_f to a minimum value. In addition, a lower temperature 400–500°C anneal, usually in "Forming Gas" (10% H₂ in N₂), is used to minimize Q_{it} . The combination of these two process steps reduces both types of charges to about 10¹⁰ cm⁻² (or lower) on (100) surfaces with slightly larger values on (111) surfaces. These values are low enough that they do not have a significant effect on most devices.

Commercially available process simulation programs generally do not contain models to predict Si/SiO₂ interface charges. Since these charges can have significant effects on device electrical performance, device simulators often allow the user to input Q_f and Q_{it} values that are then used in calculating I-V characteristics.

6.5.14 Complete Oxidation Module Simulation

Many modern device structures involve complex geometries and many process steps. The only real way to simulate such processes is with a computer simulation code like SUPREM IV. It is important to note, however, that many of the models we have discussed in the previous sections were developed largely in isolation of each other. For example, a researcher studying thin oxide growth kinetics might propose a new growth law based on experimental measurements. But such models are very often developed based on simple experimental structures, perhaps planar oxidations in this example. When such a new model is implemented in a program like SUPREM, it may well be applied

to nonplanar structures or at temperatures outside the range of the original experiments that led to the model development. New models are also often applied in situations in which many physical effects are occurring simultaneously. For example, the new thin oxide model in this example might be applied to a very heavily doped substrate and as a result, dopants might segregate into the oxide during growth, perhaps changing the growth kinetics in ways that were unanticipated by the researcher doing the model development. Because of all these issues, complete process simulators often provide the most comprehensive test of new models. Figures 6-41 and 6-42 show two examples.

Figure 6-41 shows a simple example of the simulation of a typical isolation process. In this case, the silicon is etched prior to the field oxide growth, so that the final oxide surface is at roughly the same level as the original silicon surface. This produces the well-known "bird's head" shape in the simulation because of lateral oxidation under the masking nitride. Figure 6-41 illustrates one of the powerful features of modern simulators—the ability to "observe" things that are not easily observable experimentally. In this case the time evolution of the structure is shown.

Figure 6-42 shows an example of a more complex isolation process. In this case two nitride depositions (of different thickness) are needed to form the initial structure, along with a silicon etch to form a recessed field oxide. The second (thinner) nitride re-masks the sidewall of the etched silicon, minimizing oxidation in the active device region under the thicker nitride. In the final structure on the top right of the figure, this thin nitride "flap" has lifted up to control the stresses developed during the oxidation. This simulation invoked many of the stress-dependent models that we described earlier

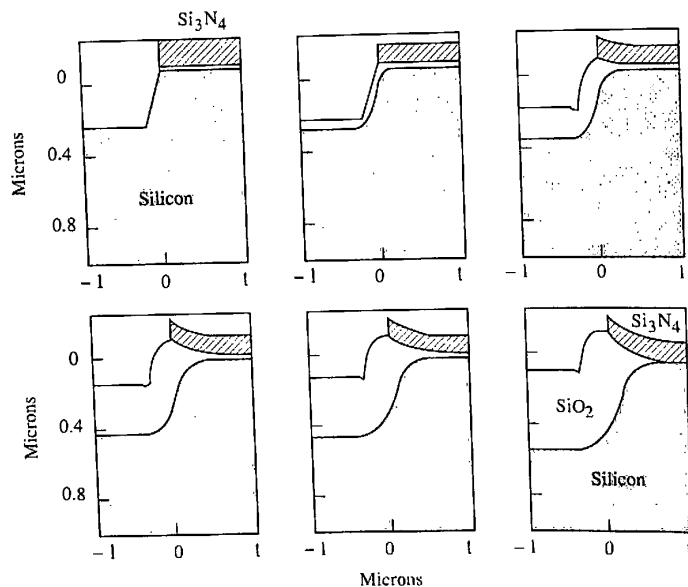


Figure 6-41 Simulation of a recessed LOCOS isolation structure using the ATHENA simulator [6.12]. The initial structure (top left) is formed by depositing a SiO₂/Si₃N₄ structure followed by etching of this stack on the left side. The silicon is then etched to form a recessed oxide and the structure is oxidized for 90 min at 1000°C in H₂O. The time evolution of the bird's head shape during the oxidation is shown in the simulations.

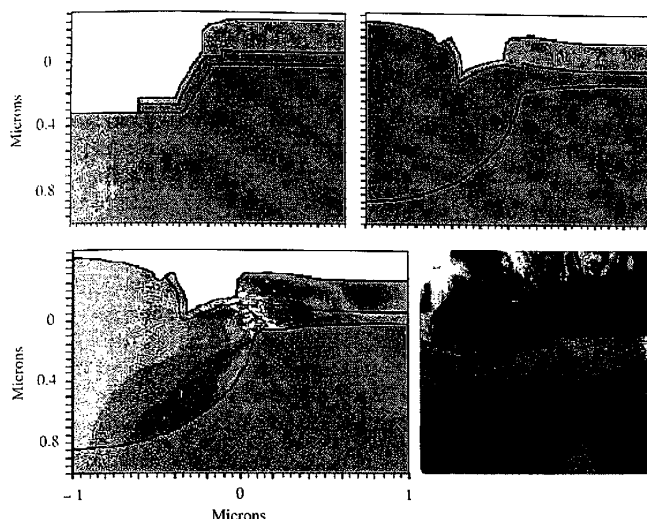


Figure 6-42 Simulation of an advanced isolation structure (the SWAMI process originally developed by Hewlett-Packard), using the ATHENA simulator [6.12]. The structure prior to oxidation is on the top left. This structure is formed by depositing an oxide followed by a thick Si_3N_4 layer, both of which are etched away on the left side. A silicon etch on the left side is then followed by a second oxide and nitride deposition. These layers are then etched away on the far-left side, leaving the thin $\text{SiO}_2/\text{Si}_3\text{N}_4$ stack covering the sidewall of the silicon. A 450-min H_2O oxidation at 1000°C is then performed which results in the structure on the top right. An experimental structure fabricated with a similar process flow is shown on the bottom right. The stress levels in the SiO_2 are shown at the end of the oxidation on the bottom left.

in this chapter. Note that there is reasonable agreement with the experimental structure also shown in the figure. This example again also illustrates one of the unique capabilities of simulators. The calculated stress levels in the SiO_2 are shown in the lower-left portion of the figure. Information like this is often useful in understanding whether a particular process will create defects in the underlying substrate. The stress contours themselves are usually not observable experimentally at all. About all that can be done experimentally is to observe at the end of the process whether or not defects were created in the substrate. Simulation thus provides a powerful design tool in understanding processes and their effects.

6.6 Limits and Future Trends in Technologies and Models

Oxidation is certain to remain an integral part of silicon technology. However, as device structures shrink in size, some changes may occur in how the process is used. There will certainly be more use made of lower-temperature processes in order to control impurity diffusion. Oxides grow more slowly at low temperatures. There are two basic approaches to obtaining relatively thick SiO_2 layers at lower temperatures. The first is to use high pressures during the oxidation, a strategy that is already finding some use in manufacturing. Systems that operate at 10–25 atmospheres are commercially available and are

sometimes used, particularly for LOCOS processes in which thick oxides are required. The second option is to use deposited SiO₂ films. We will discuss deposition in Chapter 9 but relatively simple CVD or LPCVD processes can be used to deposit thick SiO₂ layers. Such processes are increasingly being used. The ideal properties of the Si/SiO₂ interface can be preserved with deposited oxides either by first growing a thin thermal oxide and then depositing SiO₂ on top of it, or by annealing the deposited oxide after deposition. (Such anneals often consist of a short thermal oxidation step which grows a thin thermal oxide underneath the deposited oxide, again producing a virtually ideal interface.)

A second trend that is likely for the future, is the use of "composite" dielectric layers. The SIA NTRS (Table 6-1) calls for extraordinarily thin gate dielectrics in future technology generations. It is unlikely that pure SiO₂ films will be able to meet all the requirements of this roadmap. Very thin layers of SiO₂ conduct excessive currents (due to tunneling) and they are not adequate masks against dopant penetration from polysilicon gates. Boron diffusion through these thin SiO₂ layers is a particular problem. It would also be desirable to have a gate dielectric with a higher dielectric constant than SiO₂ in order to improve the electrical characteristics of MOS transistors or to increase charge storage in DRAM cells.

For these reasons, there has been a considerable amount of work in recent years on composite oxynitride dielectrics. These films were first commonly prepared by exposing thermally grown SiO₂ films to an NH₃ ambient. This results in a dielectric film which is generically called an oxynitride, often labeled as SiO_xN_y, but which usually varies in composition throughout the film. For relatively short NH₃ exposures, chemical profiling through the resulting composite dielectrics has shown that nitrogen rich regions are formed both at the Si/SiO₂ interface and at the top SiO₂/ambient interface. These two "nitride" layers seal the intervening SiO₂ layer and can trap hydrogen in the SiO₂. Because this trapped hydrogen can cause instability problems in the devices, more recent work has focused on using N₂O or NO ambients to create the oxynitride. These ambients do not contain hydrogen. Alternatively, processes that strip off the top "nitride" region and allow the trapped hydrogen to escape, have also been explored. Finally, nitrogen implants into the substrate and subsequent oxidation have also been investigated.

The benefits of the oxynitride dielectric arise because of the stronger Si-N bonds that are formed near the Si/SiO₂ interface. This oxynitride interface can result in more rugged gate dielectrics because the stronger Si-N bonds are more resistant to breaking during hot carrier stressing. Oxynitrides are also more resistant to dopant penetration through the dielectric. Finally, the relative dielectric constant of Si₃N₄ is significantly higher than SiO₂ (≈ 7.5 versus 3.9). While oxynitrides do not achieve values as high as 7.5 because they are composite materials, they usually do improve upon SiO₂. For all these reasons, oxynitrides are likely to find increased use in the future as gate dielectrics.

2D and even 3D effects will become increasingly important in small structures. These include not only predicting the shapes of oxides grown on nonplanar structures, but also accurate prediction of the stresses generated in the oxide and in the silicon substrate. Control of oxide thicknesses at the atomic level will be required for thin dielectrics. The Si-Si bond distances in SiO₂ are a fraction of a nm; this is a significant percentage of the total thickness of films that will be grown in the future for gate dielectrics and other applications. Finally, statistical issues may also become important. For example, an oxide

charge density of 10^{10} cm^{-2} corresponds to one charge in an area of about $0.1 \times 0.1 \mu\text{m}$. As device sizes approach these dimensions, we may find variations from one device to another in the number of such charges that are contained in the device active area.

6.7 Summary of Key Ideas

Thermal oxidation has been a cornerstone of silicon technology since its inception. In fact it is really the ability to easily grow SiO_2 layers that distinguishes silicon technology from all other semiconductors. SiO_2 layers adhere well to silicon, are relatively chemically inert, mask dopant diffusion, block ion implantations, and are easily patterned. However the real strength of SiO_2 lies in the almost perfect electrical interface it makes with silicon. Electrical charges are reproducible and small and the interface is almost atomically abrupt. These features allow junctions to be almost ideally passivated at silicon surfaces and they allow a variety of device structures that critically depend on surface properties to be easily built in silicon (MOS structures in particular). Even devices which are not primarily surface controlled such as bipolar transistors, do not have their electrical characteristics degraded by parasitic surface effects. No other semiconductor/insulator interface has all these desirable properties.

Since SiO_2 is such an important part of silicon technology, the growth kinetics of these layers on silicon have been carefully studied over many years. The basic growth mechanism is oxidant transport through the SiO_2 layer to the Si/SiO_2 interface where a simple chemical reaction produces the new layers of oxide. The growth generally follows a simple linear parabolic law. This basic model of oxidation has been extended in recent years to handle many situations for which the original formulation falls short. These include heavily doped substrates, mixed ambients, 2D effects and thin oxides. The same basic ideas have been applied to other materials like polysilicon and silicides which are common in silicon technology and which are often oxidized. We have not dealt in this chapter in detail with the interaction of oxidation with other process steps such as diffusion. Oxidation enhanced or retarded diffusion and impurity redistribution and segregation during oxidation are important effects. We will discuss them in more detail in Chapter 7 after we describe the basic ideas of dopant diffusion. Finally, many of the models and physical mechanisms described in this chapter have been implemented in process simulators. These simulators today are capable of accurately predicting the oxidation processes in modern VLSI structures. As new understanding develops, particularly with respect to very small structures, new models are incorporated in these simulators so that their capability is constantly improving.

6.8 References

- [6.1]. "National Technology Roadmap for Semiconductors," SIA, 1997
- [6.2]. P. J. Jorgensen, "Effect of an Electric Field on Silicon Oxidation," *J. Chem. Phys.*, vol. 37, p. 874, 1962.
- [6.3]. J. R. Ligenza and W. G. Spitzer, *Phys. and Chem. Solids*, vol. 14, p. 131, 1960.
- [6.4]. E. Rosencher, A. Straboni, S. Rigo, and G. Amsel, "An ^{18}O Study of the Thermal Oxidation of Silicon in Oxygen," *Appl. Phys. Lett.*, vol. 34, p. 254, 1979.

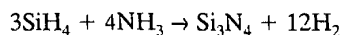
- [6.5]. R. Pretorius, *J. Electrochem. Soc.*, vol 128, p. 107, 1981.
- [6.6]. B. E. Deal and A. S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, vol. 36, p. 3770, 1965.
- [6.7]. B. E. Deal, "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon," *IEEE Trans. Elec. Dev.*, vol. ED-27, p. 606, 1980.
- [6.8]. W. A. Pliskin and E. E. Conrad, "Nondestructive Determination of Thickness and Refractive Index of Transparent Films," *IBM J. Res. and Dev.*, vol. 8, p. 43, 1964.
- [6.9]. D. K. Schroder, *Semiconductor Material and Device Characterization*, John Wiley & Sons, 1990.
- [6.10]. E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*, John Wiley & Sons, 1982.
- [6.11]. TSUPREM IV is the version of SUPREM IV by Avant! Inc. SUPREM IV was originally written at Stanford University by M. E. Law, C. S. Rafferty, and R. W. Dutton.
- [6.12]. ATHENA is the version of SUPREM IV by Silvaco Inc.
- [6.13]. A. Reisman, E. H. Nicollian, K. C. Williams, and C. J. Merz, "The Modeling of Silicon Oxidation from 1×10^{-5} to 20 Atmospheres," *J. Elect. Mat.*, vol. 16, p. 45, 1987.
- [6.14]. E. Nicollian and A. Reisman, "A New Model for the Thermal Oxidation Kinetics of Silicon," *J. Electronic Mat.*, vol. 17, p. 263, 1988.
- [6.15]. C. J. Han and C. R. Helms, "Parallel Oxidation Mechanism for Si Oxidation in Dry O₂," *J. Electrochem. Soc.*, vol. 134, p. 1297, 1987.
- [6.16]. K. Hirabayashi and J. Iwamura, "Kinetics of Thermal Growth of HCl-O₂ Oxides on Silicon," *J. Electrochem. Soc.*, vol. 120, p. 1595, 1973.
- [6.17]. R. Ghez and Y. J. van der Meulen, "Kinetics of Thermal Growth of Ultra-Thin layers of SiO₂ on Silicon, Part I: Experiment and Part II: Theory," *J. Electrochem. Soc.*, vol. 119, pp. 530 and 1100, 1972.
- [6.18]. H. Z. Massoud, J. D. Plummer, and E. A. Irene, "Thermal Oxidation of Silicon in Dry Oxygen: Growth-Rate Enhancement in the Thin Regime I. Experimental Results, II. Physical Mechanisms," *J. Electrochem. Soc.*, vol. 132, pp. 2685 and 2693, 1985.
- [6.19]. E. A. Irene and E. A. Lewis, "Thermionic Emission Model for the Initial Regime of Silicon Oxidation," *Appl. Phys. Lett.*, vol. 51, p. 767, 1987.
- [6.20]. A. G. Revesz, *Phys. Stat. Sol.*, vol. 58, p. 107, 1980.
- [6.21]. W. A. Tiller, *J. Electrochem. Soc.*, vol. 130, p. 501, 1983.
- [6.22]. R. R. Razouk, L. N. Lie, and B. E. Deal, "Kinetics of High Pressure Oxidation of Silicon in Pyrogenic Steam," *J. Electrochem. Soc.*, vol. 128, p. 2214, 1981.
- [6.23]. S. M. Hu, "A New Oxide Growth Law, and the Thermal Oxidation of Silicon," *Appl. Phys. Lett.*, May 1983.
- [6.24]. J. R. Ligenza, "Effect of Crystal Orientation on Oxidation Rates of Silicon in High Pressure Steam," *J. Phys. Chem.*, vol. 65, p. 2011, 1961.
- [6.25]. E. A. Irene, H. Z. Massoud, and E. Tierney, "Silicon Oxidation Studies: Silicon Orientation Effects on Thermal Oxidation," *J. Electrochem. Soc.*, vol. 133, p. 1253, 1986.
- [6.26]. B. E. Deal, D. W. Hess, J. D. Plummer, and C. P. Ho, "Kinetics of the Thermal Oxidation of Silicon in O₂/H₂O and O₂/Cl₂ Mixtures," *J. Electrochem. Soc.*, vol. 125, p. 339, 1978.
- [6.27]. B. E. Deal, "Thermal Oxidation Kinetics of Silicon in Pyrogenic H₂O and 5% HCl/H₂O Mixtures," *J. Electrochem. Soc.*, vol. 125, p. 576, 1978.

- [6.28]. D. B. Kao, J. P. McVittie, W. D. Nix, and K. C. Saraswat, "Two-Dimensional Thermal Oxidation of Silicon—I. Experiments," *IEEE Trans. Elec. Dev.*, vol. ED-34, p. 1008, 1987.
- [6.29]. D. B. Kao, J. P. McVittie, W. D. Nix and K. C. Saraswat, "Two-Dimensional Thermal Oxidation of Silicon—II. Modeling Stress Effects in Wet Oxides," *IEEE Trans. Elec. Dev.*, vol. ED-35, p. 25, 1988.
- [6.30]. P. Sutardja and W. G. Oldham, "Modeling of Stress Effects in Silicon Oxidation," *IEEE Trans. Elec. Dev.*, vol. 36, p. 2415, 1989.
- [6.31]. C. S. Rafferty, "Stress Effects in Silicon Oxidation—Simulation and Experiments." Ph.D. Thesis, Stanford University, 1989.
- [6.32]. P. B. Griffin and C. S. Rafferty, "A Viscous Nitride Model for Nitride/Oxide Isolation Structures," *IEDM Technical Digest*, p. 741, 1990.
- [6.33]. L. M. Landsberger and W. A. Tiller, "Two-Step Oxidation Experiments to Determine Structural and Thermal History Effects in Thermally-Grown SiO₂ Films on Si," *J. Electrochem. Soc.*, vol. 137, p. 2825, 1990.
- [6.34]. K. Imai and K. Yamabe, "Nonplanar Silicon Oxidation in Dry O₂ + NF₃," *Appl. Phys. Lett.*, vol. 56, p. 280, 1990.
- [6.35]. L. M. Landsberger, D. B. Kao, and W. A. Tiller, "Conformal Two-Dimensional SiO₂ Layers on Silicon Grown by Low Temperature Corona Discharge," *J. Electrochem. Soc.*, vol. 135, p. 1766, 1988.
- [6.36]. P. S. Dobson, "The Effect of Oxidation on Anomalous Diffusion in Silicon," *Philosophical Mag.*, vol. 24, p. 567, 1971.
- [6.37]. P. M. Fahey, P. B. Griffin, and J. D. Plummer, "Point Defects and Dopant Diffusion in Silicon," *Reviews of Modern Physics*, vol. 61, p. 289, 1989.
- [6.38]. S. M. Hu, *J. Appl. Physics*, "Formation of Stacking Faults and Enhanced Diffusion in the Oxidation of Silicon," vol. 45, p. 1567, 1974.
- [6.39]. S. Dunham, *J. Appl. Physics*, "Interaction of Silicon Point Defects with Silicon Oxide Films," vol. 71, p. 685, 1992.
- [6.40]. C. P. Ho and J. D. Plummer, "Si/SiO₂ Interface Oxidation Kinetics: a Physical Model for the Influence of High Substrate Doping Levels. I. Theory and II. Comparison with Experiment," *J. Electrochem. Soc.*, vol. 126, pp. 1516 and 1523, 1979.
- [6.41]. E. Biermann, H. H. Berger, P. Linke and B. Muller, "Oxide Growth Enhancement on Highly N-Type Doped Silicon Under Steam oxidation," *J. Electrochem. Soc.*, vol. 143, p. 1434, 1996.
- [6.42]. S. S. Choi, M. Z. Numan, W. K. Chu, and E. A. Irene, "Anomalous Oxidation rate of Silicon Implanted with Very High Doses of Arsenic," *Appl. Phys. Lett.*, vol. 51, p. 1001, 1987.
- [6.43]. F. Lau, L. Mader, C. Mazure, C. Warner, and M. Orlowski, "A Model for Phosphorus Segregation at the Silicon-Silicon Dioxide Interface," *Appl. Phys. A*, vol. 49, p. 671, 1989.
- [6.44]. T. Kamins, *Polycrystalline Silicon for Integrated Circuit Applications*, 2d ed., Kluwer Academic Publishers, 1998.
- [6.45]. T. Enomoto, R. Ando, H. Morita, and H. Nakayama, "Thermal Oxidation Rate of a Si₃N₄ Film and Its Masking Effect Against Oxidation of Silicon," *Jpn. J. Appl. Phys.*, vol. 17, p. 1049, 1978.
- [6.46]. T. I. Kamins, private communication.
- [6.47]. H. Du, R. E Tressler and K. E. Spear, "Thermodynamics of the Si-N-O System and Kinetic Modeling of Oxidation of Si₃N₄," *J. Electrochem. Soc.*, vol. 136, p. 3210, 1989.

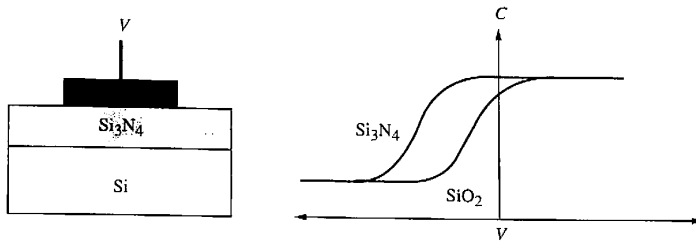
- [6.48]. M. M. Moslehi and K. C. Saraswat, "Thermal Nitridation of Si and SiO₂ for VLSI," *IEEE Trans. Elec. Dev.*, vol. ED-32, p. 106, 1985.
- [6.49]. M. Bartur and M-A. Nicolet, "Thermal Oxidation of Transition Metal Silicides on Si: Summary," *J. Electrochem. Soc.*, vol. 131, p. 371, 1984.
- [6.50]. J. E. E. Baglin, F. M. d'Heurle, and C. S. Petersson, "Interface Effects in the Formation of Silicon Oxide on Metal Silicide Layers Over Silicon Substrates," *J. Appl. Phys.*, vol. 54, p. 1849, 1983.
- [6.51]. R. R. Razouk and B. E. Deal, "Dependence of Interface State Density on Silicon Thermal Oxidation Process Variables," *J. Electrochem. Soc.*, vol. 126, p. 1573, 1979.
- [6.52]. B. E. Deal, "The Current Understanding of Charges in the Thermally Oxidized Silicon Structure," *J. Electrochem. Soc.*, vol. 121, p. 198C, 1974.
- [6.53]. A. I. Akinwande and J. D. Plummer, "Quantitative Modeling of Si/SiO₂ Interface Fixed Charge I. Experimental Results and II. Physical Modeling," *J. Electrochem. Soc.*, vol. 134, pp. 2565 and 2573, 1987.
- [6.54]. M. L. Reed and J. D. Plummer, "Chemistry of Si-SiO₂ Interface Trap Annealing," *J. Appl. Phys.*, vol. 63, p. 5776, 1988.
- [6.55]. K. L. Brower and S. M. Meyers, "Chemical Kinetics of Hydrogen and (111) Si-SiO₂ Interface Defects," *Appl. Phys. Lett.*, vol. 57, p. 162, 1990.
- [6.56]. G. J. Gerardi, E. H. Poindexter, and P. J. Caplan, "Interface Traps and Pb Centers in Oxidized (100) Silicon Wafers," *Appl. Phys. Lett.*, vol. 49, p. 348, 1986.

6.9 Problems

- 6.1. A spherically shaped piece of silicon is cut and polished from a Czochralski single-crystal ingot and oxidized in a thermal oxidation furnace. Upon pulling the silicon sphere from the furnace, the color is observed to vary significantly over the surface. Why?
- 6.2. Using Eqs. (6.1) and (6.2), explain why oxides thinner than about 50 nm do not have any characteristic color associated with them. You do not need to do any calculations. Simply explain the physical reason for the lack of color in thin oxides.
- 6.3. An experimental Metal Insulator Semiconductor (MIS) (I = insulator) structure is fabricated by depositing Si₃N₄ (silicon nitride) on a silicon substrate. The nitride is deposited by directing a jet of silane and ammonia at the surface:

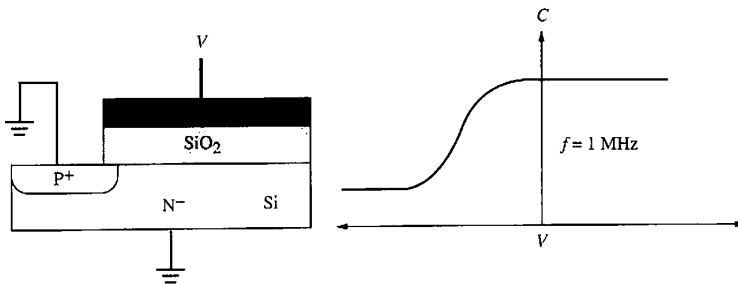


A metal electrode is deposited and a CV plot is made as shown below. A representative CV plot is also shown for an identical structure except with thermally grown SiO₂ as the insulator.

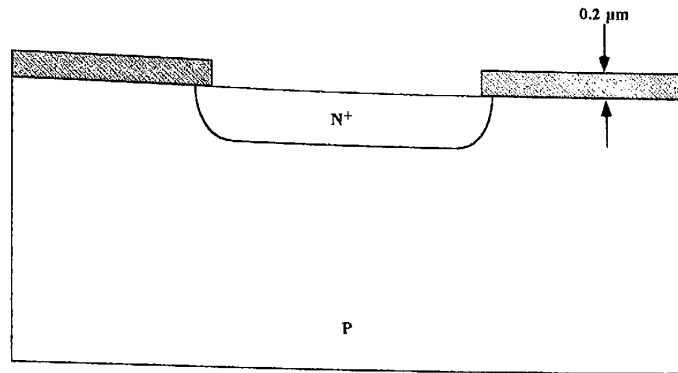


Explain the lateral shift in the CV curve of the Si_3N_4 .

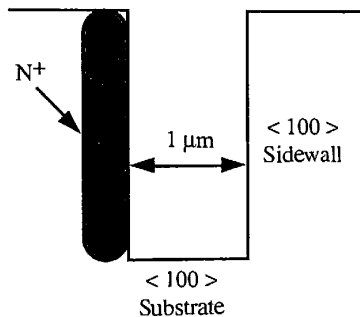
- 6.4. Construct a HF CV plot for a P-type silicon sample, analogous to Figure 6–9. Explain your plot based on the behavior of holes and electrons in the semiconductor in a similar manner to the discussion in the text for Figure 6–9.
- 6.5. A MOS structure is fabricated to make CV measurements as shown below. The CV plot shows the result if the P^+ diffusion is *not* present. Sketch the expected shape of the CV plot with the P^+ diffusion. Explain.



- 6.6. In a small MOS device, there may be a statistical variation in V_T due to differences in Q_F from one device to another. In a 0.13- μm technology minimum device (gate oxide area = $0.1\ \mu\text{m} \times 0.1\ \mu\text{m}$) with a 2.5-nm gate oxide, what would the difference in threshold voltage be for devices with 0 or 1 fixed charge in the gate oxide?
- 6.7. Why is steam oxidation more rapid than dry O₂ oxidation?
- 6.8. Under what conditions is the thermal growth rate of SiO₂ linearly proportional to time?
- 6.9. According to the Deal-Grove model, oxidation kinetics start out linear and become parabolic as the oxidation proceeds. Calculate the oxide thickness at which this transition takes place and plot this versus oxidation temperature.
- 6.10. Does the oxide thickness at which there is a transition from linear to parabolic rates change if we perform an oxidation at a pressure of 20 Atm rather than at 1 Atm.
- 6.11. A MOS device requires a gate oxide of $10\ \text{nm} \pm 0.5\ \text{nm}$. Assume the growth is done at 900°C in dry O₂. Neglect any effect of the anomalous initial growth. Derive a simple expression which gives the sensitivity of the oxide thickness to growth temperature (dx/dT). Evaluate this expression to see how well controlled the furnace T must be in order to obtain $10\ \text{nm} \pm 0.5\ \text{nm}$ at 900°C.
- 6.12. A silicon wafer is covered by an SiO₂ film 0.3 μm thick.
- What is the time required to increase the thickness by 0.5 μm by oxidation in H₂O at 1200°C?
 - Repeat for oxidation in dry O₂ at 1200°C.
- 6.13. Suppose an oxidation process is used in which (100) wafers are oxidized in O₂ for three hours at 1100°C, followed by two hours in H₂O at 900°C, followed by two hours in O₂ at 1200°C. Use Figures 6-19 and 6-20 in the text to estimate the resulting final oxide thickness. Explain how you use these figures to calculate the results of a multi-step oxidation like this.
- 6.14. What is the approximate oxide thickness (use charts) after a 100-minute dry O₂ oxidation followed by a 35-minute H₂O oxidation at 900°C?
- 6.15. The structure shown below is formed by oxidizing a silicon wafer ($x_0 = 200\ \text{nm}$) and then using standard masking and etching techniques to remove the SiO₂ in the center region. An N⁺ doping step is then used to produce the structure shown. The structure is next placed in an oxidation furnace and oxidized at 900°C in H₂O. The oxide will grow faster over the N⁺ region than it will over the lightly doped substrate. Assume that B/A is enhanced by 4X over the N⁺ region. Will the growing oxide over the N⁺ region ever catch up in thickness to the other oxide? If so, when and at what thickness? Use the Deal-Grove model for the oxidation kinetics.

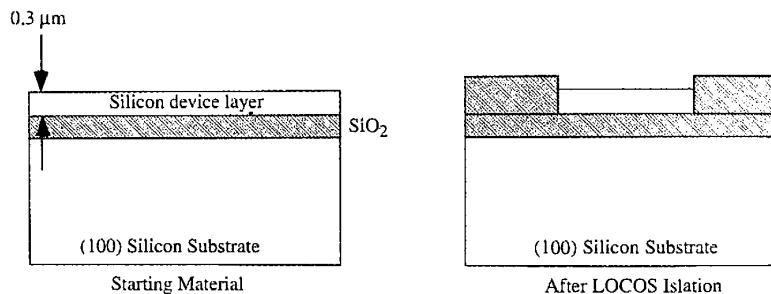


- 6.16. A 1-μm wide trench is etched in a <100> silicon wafer, so that the sides of the trench are <110> planes. An angled implant is performed, doping the sidewall N⁺ and thereby enhancing the linear rate constant by a factor of 4. The structure is then oxidized in steam at 1100°C. At what time during the oxidation will the groove be filled with SiO₂? Assume the appropriate oxidation coefficient scale as [(111 : 110 : 100) = (1.68 : 1.2 : 1.0)].



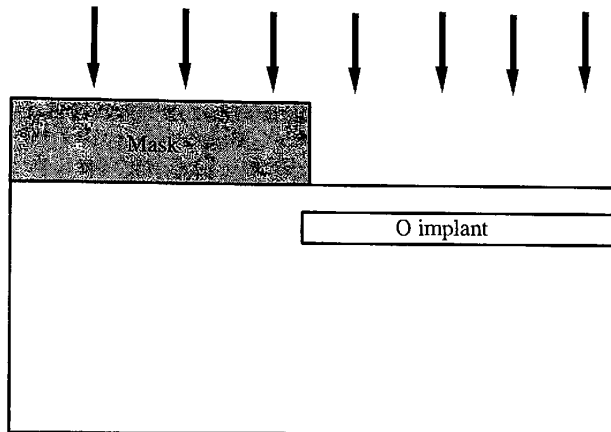
- 6.17. A uniform oxide layer of 0.4-μm thickness is selectively etched to expose the silicon surface in some locations on a wafer surface. A second oxidation at 1000°C in H₂O grows 0.2 μm on the bare silicon.
- Sketch a cross section of the SiO₂ in all locations on the wafer and the position of the Si/SiO₂ interface.
 - Would your picture be the same if the second oxidation grew the 0.2 μm at a different temperature? Explain.

- 6.18. Silicon On Insulator or SOI is a new substrate material that is being considered for future integrated circuits. The structure, shown below, consists of a thin single-crystal silicon layer on an insulating (SiO₂) substrate. The silicon below the SiO₂ provides mechanical support for the structure. One of the reasons this type of material is being considered is because junctions can be diffused completely through the thin silicon layer to the underlying SiO₂. This reduces junction capacitances and produces faster circuits. Isolation is also easy to achieve in this material, because the thin Si layer can be completely oxidized, resulting in devices completely surrounded by SiO₂. A LOCOS process is used to locally oxidize through the silicon as shown on the right below. Assuming the LOCOS oxidation is done in H₂O at 1000°C, how long will it take to oxidize through the 0.3- μ m silicon layer? Calculate a numerical answer using the Deal-Grove model.



- 6.19. As MOS devices are scaled to smaller dimensions, gate oxides must be reduced in thickness.
- As the gate oxide thickness decreases, do MOS devices become more or less sensitive to sodium contamination? Explain.
 - As the gate oxide thickness decreases, what must be done to the substrate doping (or alternatively the channel V_{TH} implant, to maintain the same V_{TH} ?
- 6.20. A SiO₂ layer is thermally grown at 1000°C using a 10–20–30 min dry-wet-dry oxidation cycle. Upon pulling the (100) wafer from the furnace, the oxide color is observed to be tan. Is this right? If not, suggest what might have gone wrong in the experiment.
- 6.21. In the simulation example in Fig. 6.35, the difference in oxidation rate between the heavily doped and lightly doped regions is much more pronounced at low temperatures (the 800°C example) than at high temperatures (the 1000°C example). Explain physically why this is the case using the mechanisms in the Deal-Grove model and the behavior of point defect (V) concentrations versus temperature.

- 6.22. The structure shown below is implanted with oxygen using a $1 \times 10^{18} \text{ cm}^{-2}$ implant at 200 keV. This places the implanted oxygen in a profile centered at $0.35 \text{ } \mu\text{m}$ ($R_p = 0.35 \text{ } \mu\text{m}$) below the surface. The left-hand side is masked from the implant. Following the implant, a high-temperature anneal is performed which forms stoichiometric SiO_2 in a buried layer on the right side. Calculate the structural dimensions on the right side following this anneal (oxide thickness, distance from the surface, and all other important dimensions). You can assume the silicon atomic density is $5 \times 10^{22} \text{ atoms cm}^{-3}$, and the Si lattice planes are 0.25 nm apart. State any other assumptions you make.



- 6.23. As part of an IC process flow, a CVD SiO_2 layer $1.0\text{-}\mu\text{m}$ thick is deposited on a $\langle 100 \rangle$ silicon substrate. This structure is then oxidized at 900°C for 60 minutes in an H_2O ambient. What is the final SiO_2 thickness after this oxidation? Calculate an answer. Do not use the oxidation charts in the text.

9.1 Introduction

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

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

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

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

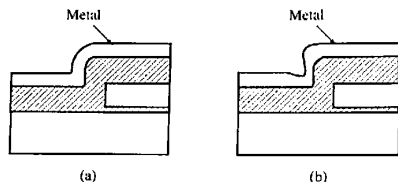


Figure 9-1 Step coverage of metal over nonplanar topography. (a) shows conformal step coverage, with constant thickness on horizontal and vertical surfaces. (b) shows poor step coverage.

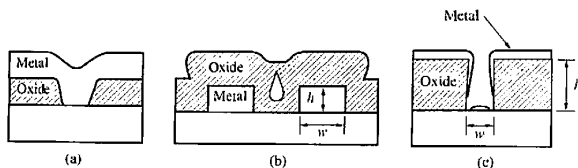


Figure 9-2 Thin film filling issues. (a) shows good metal filling of a via or contact hole in a dielectric layer. (b) shows silicon dioxide dielectric filling the space between metal lines, with poor filling leading to void formation. (c) shows poor filling of the bottom of a via hole with a barrier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

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

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

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

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

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

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

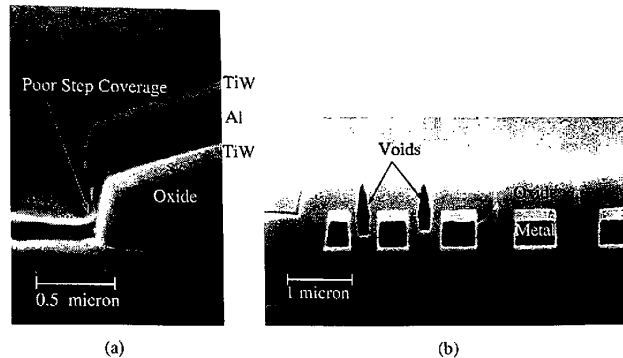


Figure 9-3 SEM images of coverage and filling problems. (a) shows poor step coverage of a TiW/Al/TiW metal stack layer, deposited by sputter deposition, over an oxide step. (b) shows voids in a Chemical Vapor Deposition (CVD) oxide layer for narrow spaces between metal lines. Photos courtesy of VLSI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

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

9.2 Historical Development and Basic Concepts

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

A third category for thin film deposition is the technique of coating a wafer with a liquid film that forms a solid film when heated. Spin-On-Glasses (SOGs) are examples of this. This method, which is used for topography smoothing or “planarization,” will be discussed in Chapter 11. A fourth category includes electrolytic deposition techniques. These have historically been used for printed circuit board fabrication but have recently been extended to the deposition of Cu interconnect layers.

9.2.1 Chemical Vapor Deposition (CVD)

In CVD, gases are introduced into the deposition chamber that react and form the desired film on the surface of the substrate. The silicon substrate will usually have other films deposited and patterned on it which then form a complex topography on which the new film must be deposited. Either a single gas is used that will decompose when heated to supply the necessary component or components for the film, or multiple gases are used which will react to produce the film. In either case, the film-producing reaction should take place on the surface of the substrate as opposed to in the gas stream. Gas stream reactions can create particulates that form above the wafer which can drop down onto the surface and cause the films to have pinholes or low density. Sometimes appropriate reactants are not available in gaseous form, in which case a liquid source may be used. A carrier gas such as hydrogen is bubbled through the liquid source and carries the source vapor to the reaction chamber. Nitrogen or argon may be used as a diluent or carrier gas.

The simplest CVD process uses an atmospheric deposition chamber, and no plasma-enhanced processes are used. This is appropriately called Atmospheric Pressure Chemical Vapor Deposition, or APCVD. The configuration shown in Figure 9-4(a) is a common type that has been used for APCVD, especially for epitaxial silicon deposition, but several other systems in a variety of configurations such as barrel-shaped chambers have been used. The walls of the chamber in this case are not heated—a so-called “cold-wall” reactor. The wafers are heated by using a graphite susceptor which is heated by RF induction. This minimizes deposition on the reactor walls since only the susceptor and wafers are hot. APCVD has been used for many films, especially epitaxial silicon and silicon dioxide—both undoped and doped with boron and phosphorus. However, the increasing use of Low-Pressure CVD, or LPCVD, and Plasma-Enhanced CVD, or PECVD, for reasons to be discussed in the following sections, has made the use of APCVD reactors much less common today.

Figure 9-4(b) shows the low-pressure CVD configuration. In this case the wafers are stacked upright, and the deposition is performed at a reduced pressure. Heating is accomplished using resistive heating elements wrapped around the tube (a “hot-wall” reactor), which heats up the tube and everything inside. This is very similar to the oxidation system described in Chapter 6. Very uniform temperatures can be obtained in this type of system.

There are other types of CVD configurations and techniques, but all are based on the idea of supplying gaseous chemical sources and energy to the surface of the wafers. The energy is usually thermal energy, but one can greatly reduce the amount of thermal energy needed by using a plasma to provide for energetic ions and very reactive free radicals to stimulate the deposition.

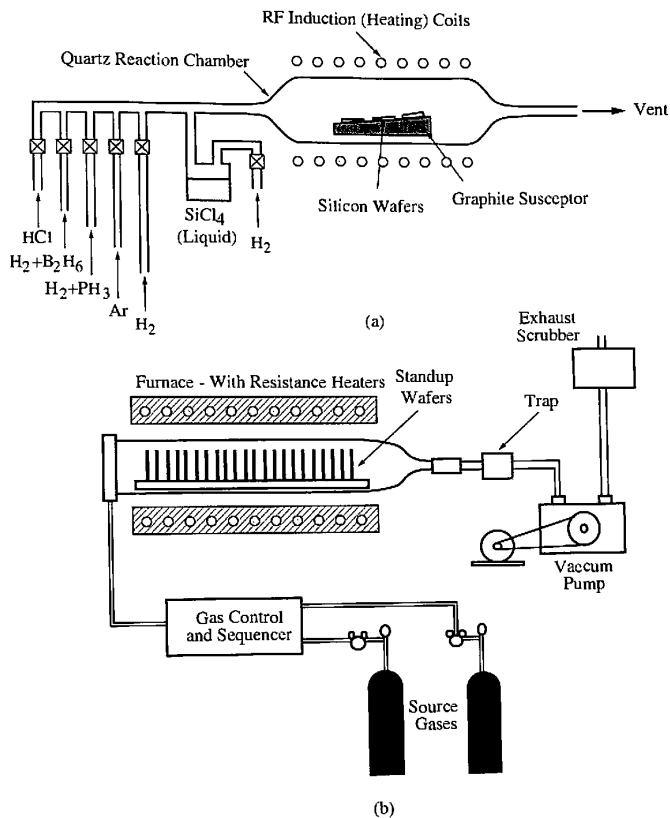


Figure 9-4 Chemical Vapor Deposition (CVD) systems. (a) is an atmospheric cold-wall system used for deposition of epitaxial silicon. (b) is a low-pressure hot-wall system used for deposition of polycrystalline and amorphous films, such as polysilicon and silicon dioxide, respectively.

9.2.1.1 Atmospheric Pressure Chemical Vapor Deposition (APCVD)

Although APCVD systems are not used as commonly today, we will begin with an APCVD system to describe some of the basic concepts of CVD. The steps involved in a CVD process, schematically illustrated in Figure 9-5, are

1. Transport of reactants by forced convection to the deposition region.
2. Transport of reactants by diffusion from the main gas stream through the boundary layer to the wafer surface.

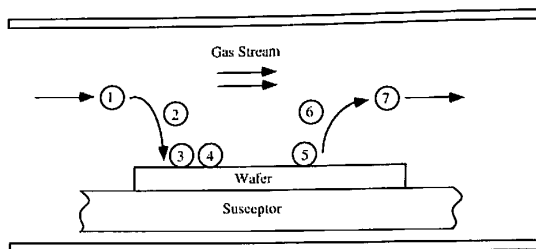


Figure 9-5 Steps involved in a CVD process. Numbered steps are explained in text.

3. Adsorption of reactants on the wafer surface.
4. Surface processes, including chemical decomposition or reaction, surface migration to attachment sites (such as atomic-level ledges and kinks), site incorporation, and other surface reactions.
5. Desorption of byproducts from the surface.
6. Transport of byproducts by diffusion through the boundary layer and back to the main gas stream.
7. Transport of byproducts by forced convection away from the deposition region.

A process not explicitly shown is that of desorption or “reemission,” often just called “emission,” of the reacting species before any reaction can occur. This is related to the sticking probability or coefficient, which will be discussed in more detail later. This process becomes important in coverage and filling issues, particularly in low pressure systems, since the desorbed species may deposit elsewhere such as on the sidewall of a trench or contact hole. In CVD processes the emission and redeposition process may occur quite readily, corresponding to a low sticking coefficient and can lead to better coverage. Surface migration or diffusion can also occur on a more long-range scale on nonflat topographies, due to surface curvature and surface energy reduction considerations for example. This can affect coverage and filling in a similar manner to reemission and redeposition. In the analysis here we are mainly concerned with deposition kinetics of thin films for relatively large-scale depositions on relatively flat surfaces. Local topography and coverage issues will be examined later focusing on low pressure CVD and PVD systems. We will also not consider gas-phase reactions, which sometimes produce intermediary vapor-phase species.

A full treatment of the kinetics of CVD thin film deposition, or growth, would involve all of the steps in Figure 9-5. However, steps 2-5 are usually the most important as far as determining the overall growth rate and thus we consider only those processes here. Step 2 is the mass transfer of reactants through the boundary layer. To simplify matters, we group steps 3-5 and classify them together as surface reactions. In our simple treatment

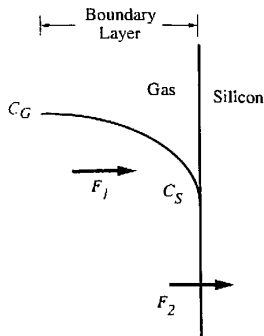


Figure 9-6 Wafer surface region in CVD process showing concentrations and fluxes of reactant species.

we consider the fluxes for the two important processes (mass transfer and surface reactions) and equate them under steady-state conditions. The treatment is very similar to that used for oxidation kinetics in Chapter 6 and for photoresist development in Chapter 5. Figure 9-6 shows the surface region of the depositing films, with the appropriate concentrations and fluxes in the gas-phase boundary layer and at the wafer surface.

The flux of reactant species from the gas phase to the wafer surface through the boundary layer, F_1 (in molecules $\text{cm}^{-2} \text{sec}^{-1}$), can be described by

$$F_1 = h_G(C_G - C_S) \quad (9.2)$$

The $(C_G - C_S)$ term is the difference in concentration of the reactant species (in molecules cm^{-3}) between the main gas flow and the wafer surface, and h_G is the mass transfer coefficient (in cm sec^{-1}). This flux represents the gas-phase diffusion through the stagnant boundary layer that forms between a flowing gas and a solid object. F_2 is the flux of reactant consumed by the reaction at the surface. Assuming first-order reaction kinetics, this can be written as

$$F_2 = k_s C_S \quad (9.3)$$

where k_s is the chemical surface reaction rate (in cm sec^{-1}) and C_S is the concentration of the reacting species at the surface (in molecules cm^{-3}). All of the processes involved with the surface chemical reactions and surface diffusion that occur during the actual film formation, steps 3-5 above, are lumped into this one parameter. This is very similar to the strategy used in the Deal-Grove oxidation model. [Recall Eq. (6.17).]

Assuming steady-state deposition conditions, these two processes that act in series must be equal to each other, and the overall process will proceed at the rate of the slower process. Thus

$$F = F_1 = F_2 \quad (9.4)$$

and equating Eqs. (9.2) and (9.3) leads to

$$C_S = C_G \left(1 + \frac{k_s}{h_G} \right)^{-1} \quad (9.5)$$

The growth rate of the film is now given by

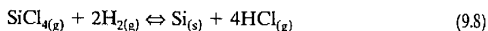
$$v = \frac{F}{N} = \frac{k_s h_G}{k_s + h_G} \frac{C_G}{N} \quad (9.6)$$

where v is the deposition rate or velocity in cm sec^{-1} , and N is the number of atoms incorporated per unit volume in the film, or its density, in cm^{-3} ($5 \times 10^{22} \text{ cm}^{-3}$ for the case of Si deposition).

We now define Y to be the mole fraction of the incorporating species in the gas phase:

$$Y \equiv \frac{C_G}{C_T} \quad (9.7)$$

where C_T is the concentration of all molecules in the gas phase. One example is the case of epitaxial Si deposition by the reaction of SiCl_4 and hydrogen:



Here the incorporating species is Si, C_G is the number of molecules of SiCl_4 per cm^3 in the gas phase and C_T would correspond to the total number of SiCl_4 and H_2 molecules (plus any other species) per cm^3 in the gas phase. Y is also equal to the partial pressure of the incorporating species, P_G , divided by the total pressure in the system such that

$$Y \equiv \frac{C_G}{C_T} = \frac{P_G}{P_{\text{Total}}} = \frac{P_G}{P_G + P_G + \dots} \quad (9.9)$$

where P_G is the partial pressure of SiCl_4 and P_G corresponds to the partial pressure of H_2 . The partial pressures of any other gas species present, such as N_2 or Ar used as carrier or dilutant gases and any byproduct gases, would also have to be added to the total pressure term. The equation for the deposition velocity is now

$$v = \frac{k_s h_G}{k_s + h_G} \frac{C_T Y}{N} \quad (9.10)$$

Example Calculate the deposition rate for a CVD system in which

$$h_G = 1.0 \text{ cm sec}^{-1}$$

$$k_s = 10 \text{ cm sec}^{-1}$$

$$\text{Partial pressure of incorporating species} = P_G = 1 \text{ torr}$$

$$\text{Total pressure} = P_T = 1 \text{ atm} = 760 \text{ torr}$$

Total concentration in gas phase = $C_T = 1 \times 10^{19} \text{ cm}^{-3}$

Density of depositing film = $N = 5 \times 10^{22} \text{ cm}^{-3}$

Answer By Eq. (9.10):

$$\begin{aligned} v &= \frac{k_s h_G}{k_s + h_G} \frac{C_T Y}{N} = \frac{1}{\left(\frac{1}{h_G} + \frac{1}{k_s}\right)} \frac{C_T}{N} Y \\ v &= \frac{1}{\left(\frac{1}{1.0 \text{ cm sec}^{-1}} + \frac{1}{10 \text{ cm sec}^{-1}}\right)} \frac{1 \times 10^{19} \text{ cm}^{-3}}{5 \times 10^{22} \text{ cm}^{-3}} \frac{1}{760} \\ &= 2.4 \times 10^{-7} \text{ cm sec}^{-1} = 0.14 \text{ } \mu\text{m min}^{-1} \end{aligned}$$

From Eq. (9.10) one sees that the deposition velocity is determined by the smaller of k_s or h_G , leading to two limiting cases. If $k_s \ll h_G$, then

$$v \cong \frac{C_T}{N} k_s Y \quad (9.11)$$

This is the surface reaction controlled case. The mass transfer through the gas boundary layer is relatively fast, while the surface reaction is sluggish. From Eq. (9.5) one sees that the surface concentration, C_s , approaches C_G . If $h_G \ll k_s$, then

$$v \cong \frac{C_T}{N} h_G Y \quad (9.12)$$

This is the mass transfer, or gas-phase diffusion, controlled case. Here the surface reaction is quite fast compared to the mass transfer process. C_s approaches zero because the reactant species are used up at the surface as soon as they arrive. In the above example, $h_G \ll k_s$ and so that system is operating in the mass transfer controlled regime.

These two limiting cases are analogous to the oxidation cases in Figure 6–15. The analysis in Chapter 6 showed how thermal oxidation leads to two kinetic regimes, one described by a linear rate constant (interface reaction limited or controlled) and the other described by a parabolic rate constant (solid-state diffusion limited). In the case of CVD deposition, there are also two limiting cases: one reaction controlled and one diffusion controlled. But the growth rate in both regimes is linear with time. That is, both regimes have constant growth velocities, and neither is parabolic. This is because the reaction always occurs at the growing surface, and the diffusion-related process is mass transport through a gas region of constant thickness, not a growing solid region as in the oxidation process. The two CVD regimes are therefore not differentiated by different growth laws. However, the two processes dominate in different temperature regimes.

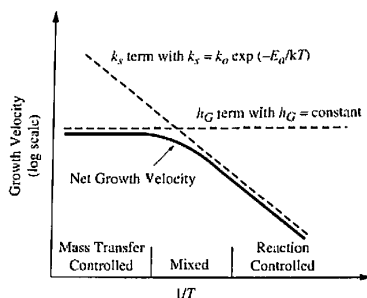


Figure 9-7 Arrhenius plot of growth velocity (or deposition rate) vs. $1/T$ for CVD process. The net growth velocity is the result of the surface reaction and gas-phase mass transfer processes acting in series so that the slower of the two dominates at any temperature.

The temperature dependence of the surface reaction controlled case is Arrhenius just as in the oxidation case. On the other hand, the mass transfer coefficient, h_G , is rather insensitive to temperature. Figure 9-7 plots the growth velocity versus reciprocal temperature. The k_s and h_G components are plotted separately showing their different temperature dependencies. When the two components are combined so that the smaller of the two components dominates (two processes acting in series), as described by Eq. (9.10), the resulting growth velocity behaves as shown in the figure. In practice, the cross-over temperature (where the two processes are equal) and the temperature ranges for the different regimes depend on a variety of factors, such as the specific gases used, the reaction activation energy, wafer configuration, the flow conditions in the reactor, and the pressure.

This basic model for CVD deposition has important implications for deposition conditions and equipment configuration. When the deposition occurs in the mass transfer controlled regime (higher temperatures), the deposition rate is relatively constant with temperature. However because it is limited by mass transfer of species through the gas-phase boundary layer, the flow of the gas over the wafers and transport of the reactants to the wafer surfaces are very important and can place major restrictions on the configuration of the equipment and the placement of the wafers. For example, the wafers must usually be placed face up and edge to edge in the chamber [as in Figure 9-4(a)] to ensure that equal gas flows reach each wafer surface and that equal deposition rates are achieved. By contrast, when the deposition occurs in the surface reaction controlled regime (lower temperatures), the process is very sensitive to temperature. However, the mass transfer through the boundary layer is not as important, leading to fewer restrictions on the gas flow and wafer placement. For example, the wafers may be stacked face-to-face [as in Figure 9-4(b)] without much of a penalty in wafer-to-wafer deposition uniformity.

Figure 9-8 shows typical experimental data for the deposition of silicon at atmospheric pressure. All four silicon sources exhibit similar types of behavior, with surface reaction limited regimes showing Arrhenius behavior, and mass transfer (or diffusion) limited regimes showing little temperature dependence. For all sources, the slopes of the curves in the reaction limited regimes are approximately the same, with an activation

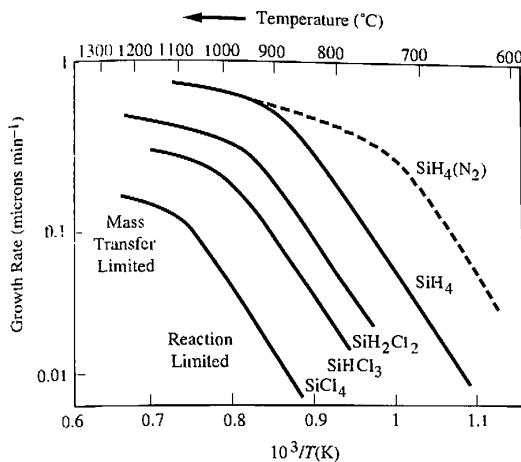


Figure 9-8 Growth or deposition rate versus $1/T$ for silicon deposited by CVD at atmospheric pressure using various silicon source gases. The partial pressure of the reactant gas is 0.8 torr. Hydrogen is used as the carrier or diluent gas for the solid curves. (After [9.1].) Using an inert gas such as N_2 for the diluent shifts the SiH_4 curve to the right, as indicated by the dashed curve and as discussed later in Section 9.3.2.

energy of about 1.6 eV. This suggests that the rate limiting surface reaction process is the same for each of the different sources and may be related to the desorption of hydrogen atoms from the Si surface. The choice of source gas does affect the overall growth velocity. The growth is fastest for a silane (SiH_4) source and slowest for a silicon tetrachloride ($SiCl_4$) source. The growth rate in the mass transfer limited regime is found to be approximately inversely proportional to the square root of the source gas molecular weight. There is also a substrate orientation effect on the growth kinetics. As expected, this is only important at lower temperatures where the surface reaction dominates. The (111) Si surface has a lower growth rate, presumably due to the fewer kink or attachment sites available for atom incorporation.

Epitaxial Si deposition has traditionally been done at higher temperatures to ensure that all the Si atoms being deposited are incorporated into lattice sites in order to obtain a single-crystal thin film. This means that epitaxial Si in these systems is deposited in the mass transfer regime where gas-phase processes are important. On the other hand, polysilicon is usually deposited at lower temperatures and in the surface reaction regime. Here the gas-phase processes are not important (so that stacking of wafers is possible), but the deposition is very temperature sensitive.

Earlier we assumed that the flux of reactant from the main gas flow to the wafer surface could be described simply by

$$F_1 = h_G(C_G - C_s) \quad (9.2)$$

where h_G is a constant. But gas flow in deposition chambers is more complicated than this, and it is worthwhile to take a little deeper look at it. We start with the configuration of Figure 9-4(a) where the wafers lie flat on a susceptor. The simplest model assumes that a completely stagnant boundary layer δ_s exists next to the surface of the wafer. This means that for some distance, away from the surface, the gas is not moving

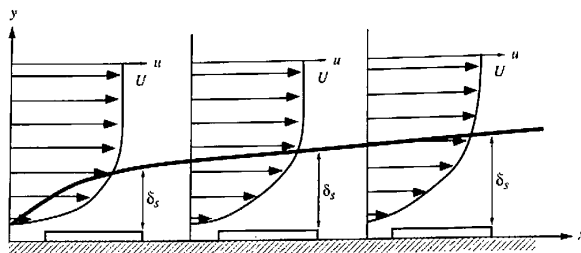


Figure 9-9 Boundary layer velocities along susceptor. δ_s is the thickness of the boundary layer. The boundary layer increases with distance in the direction of the gas flow from Newton's second law. (After [9.2].)

at all. Beyond δ_s , the gas is well mixed and moving at a constant velocity, U , parallel to the substrate surface. Figure 9-9 illustrates this and also the fact that δ_s increases in thickness along the susceptor. The transfer of gas across the boundary layer is assumed to occur via diffusion.

According to Fick's first law, the flux across the boundary layer at steady state is

$$F_1 = -D_G \frac{\partial C}{\partial x} = \frac{D_G}{\delta_s} (C_G - C_s) \quad (9.13)$$

where D_G is the diffusion coefficient of the reacting species across the boundary layer in the gas phase, and C_G and C_s are defined in Figure 9-6. From Eq. (9.2) we see that h_G corresponds to

$$h_G = \frac{D_G}{\delta_s} \quad (9.14)$$

If D_G and δ_s are constant, then h_G is constant. But δ_s is generally not constant with position in the chamber.

Right next to the wafer surface, the gas velocity is zero, and the velocity changes from zero to U within the boundary region. The boundary layer is caused by the frictional force of the gas stream on the stationary susceptor and wafers. If the gas velocity right next to the wafer surface were not zero, an infinitely large frictional force would exist right at the surface which would bring the velocity to zero there. Using Newton's second law ($F = ma$), where F is the friction force, one can determine the position of the boundary layer as a function of x to be [9.2]

$$\delta_s(x) = \left(\frac{\mu x}{\rho U} \right)^{1/2} \quad (9.15)$$

where ρ is the gas density, μ is the gas viscosity, and U is the bulk gas velocity as before. Since δ_s increases along the length of the susceptor, then the effective mass transfer coefficient, h_G , decreases along the length of the susceptor according to Eq. (9.14). Therefore, if the deposition is limited by mass transfer through the boundary layer, then the deposition rate would decrease going from the front of the susceptor to the back.

In addition, source gas depletion occurs down the length of the susceptor. As reacting gases are consumed, their concentrations decrease with distance along the susceptor. Since the growth rate is proportional to the partial pressure of the reacting species, the growth rate will decrease going downstream. While the growth rate decrease due to the increasing boundary layer is important only in the mass transfer regime, this decrease due to reactant depletion occurs in both the mass transfer and surface reaction controlled regimes. This is because Y is important in both regimes [Eqs. (9.11) and (9.12)].

To compensate for both the boundary-layer variation and depletion effects, a change in reactor geometry is needed. The wafer susceptor is tilted, as shown in Figure 9-10. This decreases the cross sectional area of the chamber along the susceptor, which causes the gas velocity to increase. This in turn causes the boundary layer to decrease along the susceptor. This increases the growth rate downstream, compensating for the friction-induced boundary-layer increase as well as the depletion effects. A way to compensate for the depletion effect when operating in the surface reaction limited regime is to impose a 5–25° temperature gradient along the tube or chamber, increasing from front to back. This is often done in LPCVD tubes, which normally operate in the surface reaction limited regime and where depletion is the main problem. For SiO_2 deposition by APCVD, commonly done in the mass transfer limited regime, a different gas flow configuration has been used to help overcome both boundary layer nonuniformities and depletion effects. The wafers still are placed side by side, facing up, but the gas is injected straight down from above. In addition the wafers can be moved horizontally on a conveyor belt. This helps ensure that uniform fluxes of reactant gases are delivered to each wafer.

As we have seen, chemical vapor deposition can be used to deposit compounds, such as SiO_2 and Si_3N_4 , as well as elements. However alloys, such as $\text{Al}(2\% \text{ Cu})$, are much

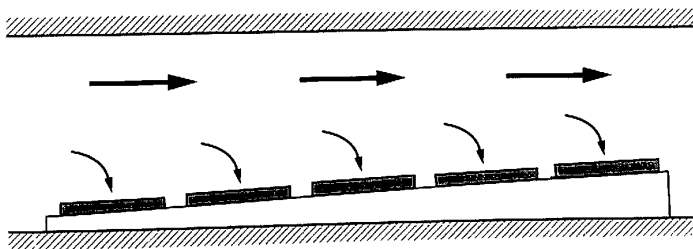


Figure 9-10 The susceptor in a horizontal epitaxial reactor is tilted so that the cross-sectional area of the chamber is decreased, increasing the gas velocity along the susceptor. This compensates for both the boundary layer and depletion effects.

more difficult. The problem is usually that separate reactions are required for each constituent (Al and Cu in our example) and it is often difficult to avoid unwanted reactions between the different reactant species, or precursors. In addition it can be hard to control the exact composition of the alloy, because unlike compounds where limited compositions result (very close to 2:1 oxygen-to-silicon in SiO_2 , for example), a wide range of alloy compositions are possible. Physical vapor deposition methods are usually better suited for alloy deposition.

Thin films deposited by CVD may be doped during the deposition process. This is common for oxides, polysilicon, and epitaxial silicon. A gaseous source of the desired dopant is mixed with the other source gases, often along with hydrogen which acts as a carrier gas and a diluent. Common dopant source gases are arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6). The same steps occur in the incorporation of the dopants as occur in the thin film deposition, including mass transport and surface reactions. The surface reactions include (1) dissociation of the hydride gas, (2) lattice site incorporation, and (3) burying of dopants atoms by the other atoms in the film. It is believed that these are the rate limiting steps for doping of Si during epitaxial growth, and not the mass transport through the boundary layer. Models have been developed to predict the doping concentration in the film as a function of gas partial pressures, temperature, flow rate, and so on. However the models have not been very successful since many factors, especially in regard to the reactor geometry, are not well characterized and not adequately incorporated into the models. In general, for low growth rates the dopant concentration in the growing film C is proportional to the dopant gas-phase concentration (which equals the partial pressure of the dopant species, P_i) at the reactor input. For higher growth rates, C is inversely proportional to v , the deposition velocity.

$$C \propto P_i \text{ for low growth rates} \quad (9.16a)$$

$$C \propto \frac{P_i}{v} \text{ for high growth rates} \quad (9.16b)$$

In practice, the relationship between dopant concentration in a deposited film and the deposition conditions is determined experimentally for each particular reactor.

Unintentional doping of deposited layers can also occur. This is especially common when depositing a lightly doped epitaxial Si film on a highly doped Si substrate—a common occurrence in CMOS technology (Figure 2–20). One way unintentional doping can occur is through normal solid-state diffusion. During the deposition process, dopants from the highly doped substrate can diffuse into the newly deposited, lightly doped region. This is often termed outdiffusion. A large concentration gradient exists and at normal deposition temperatures (800–1100°C), significant diffusion can occur. Another way unintentional doping can occur is when dopant atoms get added to the gas stream and are then subsequently reintroduced into the growing film. This is termed “autodoping.” The dopant may be added to the gas during film deposition (or even during an HCl in situ preclean) by evaporation from the wafer frontside, from the wafer backside or edges, other wafers, or from the susceptor.

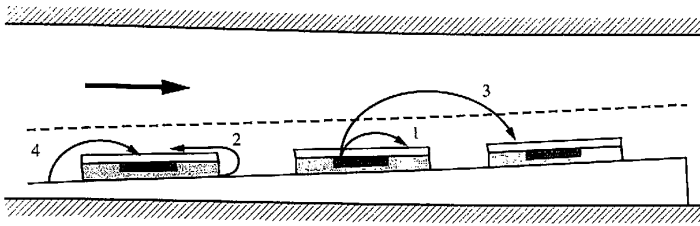


Figure 9-11 Autodoping processes in an epitaxial reactor. Illustrated are evaporation from 1, the wafer frontside; 2, the wafer backside or edges; 3, other wafers; and 4, the susceptor.

Figure 9-11 illustrates these processes. The autodoping effects add to the outdiffusion problem, often making the doping profile in the deposited layer quite complicated. The outdiffusion can be modeled with standard diffusion models. Since the growth of the layer is usually faster than the diffusion from the surface (that is, $vt \gg \sqrt{Dt}$), one can assume this is just diffusion from a doped region into a semi-infinite medium. This is described by a simple error-function profile, as discussed in Chapter 7 and given by Eq. (7.26):

$$C(x,t) = \frac{C}{2} \left[\operatorname{erfc} \frac{x}{2\sqrt{Dt}} \right] \quad (9.17)$$

Here $x = 0$ at the substrate/epitaxial layer interface, and x is positive in the epitaxial layer.

Autodoping from the frontside can usually be modeled with the empirical expression

$$C_{\text{autodoping}} = C_s^* \exp\left(-\frac{x}{L}\right) \quad (9.18)$$

Here C_s^* is the effective substrate surface concentration and L is an experimentally determined parameter, or diffusion length. This decrease in dopant concentration toward the surface arises from the fact that as the film grows in thickness, the dopant must diffuse through more and more film, and less dopant enters the gas phase. Autodoping from the backside, edges, or other sources usually results in a relatively constant level in the deposited film since the source of dopant does not diminish as quickly but is at a much lower level.

Figure 9-12 illustrates these effects in a Si epilayer. The top part of the curve, closest to the substrate, is due to outdiffusion, as described to first order by Eq. (9.17). (In fact, a concentration-dependent diffusivity would model the diffusion profile better than a simple error function.) The next segment, the nearly straight line on the plot, is from frontside autodoping, described by Eq. (9.18). The background portion is due to backside autodoping, and the concentration level is constant. These effects can significantly modify desired doping profiles in epitaxial layers.

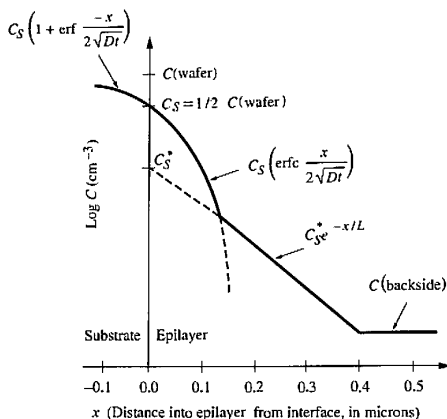


Figure 9-12 Dopant profile in a Si epilayer due to outdiffusion and autodoping. (After [9.3].)

Example Calculate the dopant concentration due to outdiffusion and autodoping in an otherwise undoped epilayer at $0.2 \mu\text{m}$ from the substrate/epilayer interface ($x = 0.2 \mu\text{m}$). The deposition time is 1 hour with a deposition rate of $1 \mu\text{m hour}^{-1}$. The substrate doping, $C(\text{wafer})$, is $1 \times 10^{19} \text{ cm}^{-3}$, C_s^* is $1 \times 10^{18} \text{ cm}^{-3}$, L is $0.25 \mu\text{m}$, D is $6 \times 10^{-15} \text{ cm}^2 \text{ sec}^{-1}$, and the backside doping concentration, $C(\text{backside})$, is $1 \times 10^{15} \text{ cm}^{-3}$.

Answer The dopant concentration profile looks like Figure 9-12, with the concentration at each depth equal to the sum of the contributions from the outdiffusion, frontside autodoping and backside autodoping. For the outdiffusion term, we first check vt versus $\sqrt{Dt}$:

$$vt = (1 \times 10^{-4} \text{ cm hr}^{-1})(1 \text{ hr}) = 1 \times 10^{-4} \text{ cm}$$

$$\sqrt{Dt} = \sqrt{(6 \times 10^{-15} \text{ cm}^2 \text{ sec}^{-1})(3600 \text{ sec})} = 4.6 \times 10^{-6} \text{ cm}$$

Since $vt \gg \sqrt{Dt}$, we can assume the diffusion is occurring into a semi-infinite medium at all times:

$$\therefore C(x,t) = \frac{C}{2} \left[\operatorname{erfc} \frac{x}{2\sqrt{Dt}} \right] + C_s^* \exp\left(-\frac{x}{L}\right) + C(\text{backside})$$

The outdiffusion and autodoping occur throughout the deposition, so $t = 3600$ sec. We wish to know the doping concentration at $x = 0.2 \mu\text{m}$, so

$$\begin{aligned} C(0.2 \text{ micron}, 3600 \text{ sec}) &= \frac{1 \times 10^{19} \text{ cm}^{-3}}{2} \left[\operatorname{erfc} \frac{0.2 \times 10^{-4} \text{ cm}}{2\sqrt{(6 \times 10^{-15} \text{ cm}^2 \text{ sec}^{-1})(3600 \text{ sec})}} \right] \\ &\quad + 1 \times 10^{18} \text{ cm}^{-3} \exp\left(-\frac{0.20 \mu\text{m}}{0.25 \mu\text{m}}\right) + 1 \times 10^{15} \text{ cm}^{-3} \\ &= 0.5 \times 10^{19} \text{ cm}^{-3} [\operatorname{erfc}(2.15)] + 1 \times 10^{18} \text{ cm}^{-3} \cdot 0.45 + 1 \times 10^{15} \text{ cm}^{-3} \\ &= 1.2 \times 10^{16} + 4.5 \times 10^{17} + 1 \times 10^{15} \text{ cm}^{-3} \\ &= 4.6 \times 10^{17} \text{ cm}^{-3} \end{aligned}$$

The major contribution at this depth is from the frontside autodoping.

9.2.1.2 Low-Pressure Chemical Vapor Deposition (LPCVD)

We saw that when operating CVD reactors in the mass transfer limited regime, one has to ensure that equal fluxes of reactant gases must reach every location in order to achieve uniform depositions across the wafers. This generally implies placing the wafers side by side as in Figure 9-4(a). Otherwise the variability in the boundary layer thickness and transport through that layer would result in nonuniform depositions. Operating in the surface reaction limited regime where transport is not important would avoid these problems and allow stacking the wafers as in Figure 9-4(b). This generally provides higher throughput. This can be achieved by going to a lower deposition temperature (Figure 9-7). However as the temperature is lowered, the deposition rate can decrease significantly. In addition the film quality may suffer.

Lowering the total pressure of the gas stream increases the diffusion and extends the reaction controlled regime to higher temperatures. First recall that the CVD film deposition rate is given by

$$v = \frac{k_s h_G}{k_s + h_G} \frac{C_T Y}{N} \quad (9.10)$$

where k_s is the surface reaction coefficient and h_G is the mass transfer coefficient. The smaller of the two terms dominates and controls the overall deposition process. Diffusion through the boundary layer is given by

$$h_G = \frac{D_G}{\delta_s} \quad (9.14)$$

where D_G is the diffusivity of the reacting species through the boundary layer and δ_s is the boundary layer thickness. The key new point is that D_G is inversely proportional to P_{total} ,

$$D_G \propto \frac{1}{P_{\text{total}}} \quad (9.19)$$

since the diffusivity is inversely proportional to the number of collisions that the gas species experience. So decreasing the total pressure, which also means decreasing the total concentration of gases in the gas stream but not necessarily decreasing the concentrations or partial pressures of the reactant gases, causes D_G to increase.

Decreasing P_{total} from 1 atm (760 torr) to 1 torr, a typical LPCVD pressure, increases D_G by 760 times. Decreasing P_{total} also increases the boundary layer thickness, δ_s , but not proportionally. Reducing the pressure to 1 torr only increases δ_s by 3 to 10 times. The net effect on h_G is that it increases by about 100 times. In the earlier APCVD numerical example, if h_G were increased by 100 times, with everything kept the same, the deposition rate would increase to $1.4 \mu\text{m min}^{-1}$ and system operation switches to the surface reaction controlled regime (problem 9.6).

Since h_G is much larger at the lower pressure, mass transport through the boundary layer becomes much less important compared to the surface reactions. Figure 9-13 illustrates this. Now that the deposition velocity is not limited by the mass transfer of reactants through the gas-phase boundary layer, the wafers can be placed much closer together without affecting the deposition uniformity nearly as much. The configuration

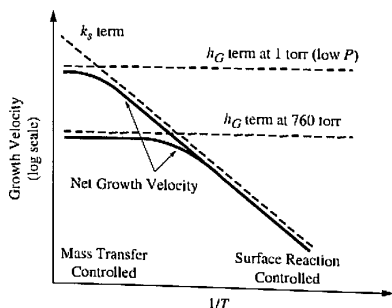


Figure 9-13 Growth velocity vs. $1/T$ for APCVD (760 torr) and LPCVD (1 torr) systems. The lower total pressure (with P_G and C_G remaining fixed) shifts the h_G curve upward, extending the surface reaction regime to higher temperatures.

shown earlier in Figure 9-4(b) is typical for LPCVD reactors. The wafers are stacked vertically. A vacuum pump is needed to reduce and control the pressure, usually in the 0.25–2.0 torr range. Deposition temperatures in LPCVD systems range from 300–900°C. Since these systems are operated in the surface reaction regime, temperature must be controlled typically to $\pm 1^\circ\text{C}$, but this is easily achieved in hot-walled furnaces. Depletion of reactant gas can still be a problem, as discussed earlier, so a 5–25°C temperature gradient (higher temperature downstream) is often used. This speeds up the reaction rate toward the back of the tube to compensate for the depletion. An alternate method to avoid the depletion problem is to use a distributed feed setup in which the reactant gases are injected at regular intervals down the tube.

There are additional advantages to using a low-pressure system. With the lower pressure and much faster diffusion out through the boundary layer to the main gas stream, less autodoping occurs. In addition, little or no carrier or dilutant gas is needed and gas consumption is much lower. Also, at lower pressures there are usually fewer gas-phase reactions and hence fewer particulates form that can deposit on the wafers. There is of course a limit to how much the pressure can be lowered. At some point the reactant gas concentration, C_G , or pressure, P_G , would have to be lowered to reduce the total gas pressure, which would then decrease the deposition rate.

It should be pointed out that even though we describe operating in the surface reaction limited regime at low pressures to achieve uniform deposition, the transport of the depositing material through the gas phase can still be important, especially on a local level near the surface. It is true that the diffusional flow through a boundary layer is no longer the limiting step. One might assume that under such conditions there would always be enough species available to react and deposit at each position along the surface, resulting in perfectly conformal coverage for local topographical features. However, at these low pressures virtually no gas-phase collisions occur in the near-surface region within the topographical features. Therefore line-of-sight transport occurs over this distance, as opposed to more randomly directed diffusional transport that occurs within the surface features at atmospheric pressure. As a result, shadowing can occur where the flux from the gas phase is impeded locally by surface topography features. This shadowing can directly affect step coverage and filling. Reemission processes may not always occur readily enough (or equivalently, the sticking coefficient low enough) or surface diffusion fast enough to allow for constant flux to reach each point on the surface from the gas phase. Thus nonconformal profiles can still result, as we shall see. These issues will be even more important for PVD systems—not only do they operate at very low pressures, but the deposition processes are usually assumed to occur with very little surface diffusion or reemission.

9.2.1.3 Plasma-Enhanced Chemical Vapor Deposition (PECVD)

Sometimes there are restrictions on the temperature that the substrate can be exposed to when depositing a film. The most obvious example is when depositing a silicon dioxide or silicon nitride film when an aluminum metal level is already in place. The melting point of Al is 660°C, and any subsequent processing temperature must be less than about 450°C. APCVD or LPCVD films can be deposited at these lower temperatures,

but the deposition rate is usually quite low because k_s decreases exponentially with temperature. In addition the film quality can be poor, with the films being rather porous and susceptible to moisture absorption. Good film quality is especially important for multilevel metallization schemes where deposited dielectric layers are used to provide electrical and physical barriers between metal interconnect layers. Conformal and void-free films are also important. Therefore a dielectric deposition technique is needed that provides reasonable deposition rates, good film quality, conformality, and low deposition temperatures.

A method developed for these needs is Plasma-Enhanced Chemical Vapor Deposition, or PECVD. In addition to using a thermal source to provide the energy needed for the chemical reactions to occur during the deposition, a plasma source is used. Plasmas are highly ionized gases and will be discussed in more detail later in Section 9.2.2.2. By supplying additional energy from the plasma to the reactant gases, the reactions needed for deposition can occur at temperatures much lower than those needed when only thermal energy is provided. In fact, room temperature depositions are even possible with this method. Usually, though, PECVD depositions of dielectrics are done in the 200–350°C range. In addition to allowing the depositions to occur at lower substrate temperatures, PECVD also allows more easily altering the film's properties (composition, density, stress, etc.) and the tailoring of those properties for particular applications. This is due mainly to the more nonequilibrium nature of the deposition process compared to the nonplasma CVD methods. However, this can also lead to undesirable compositions or properties in the films, such as the incorporation of byproducts or gas molecules into the film.

A typical equipment configuration for PECVD is shown in Figure 9-14. This is similar to the parallel plate plasma etching equipment to be described in Chapter 10. In plasma etching, the energetic species created in the plasma are used to etch materials on

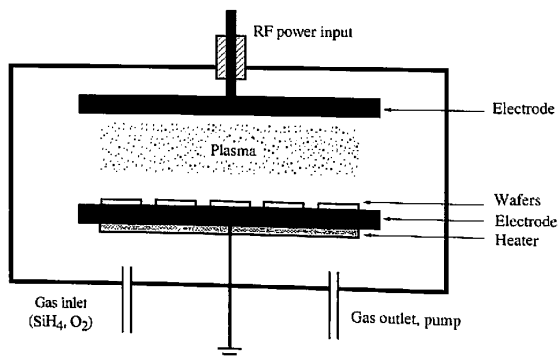


Figure 9-14 Typical PECVD equipment configuration.

the wafer. In plasma-enhanced deposition, the energetic species are used to help form the films on top of wafers. The wafers are placed on the lower plate electrode. Heat can be supplied to the wafers through a heater beneath the electrode. The reactant gases, such as silane and oxygen for the deposition of silicon dioxide, are fed through inlets. A glow discharge, or plasma, is sustained between the lower and upper electrodes. This is formed by applying a high electric field, often at 13.56 MHz, to a low pressure gas (between 50 mtorr and 5 torr), creating ions and free electrons. The plasma is sustained when high-energy electrons strike and ionize atoms and molecules.

In the plasma, interactions with the high-energy electrons cause the reactant gases to dissociate and ionize into a variety of species. These include ionized and excited molecules (or atoms), neutral molecules and neutral and ionized fragments of broken-up molecules, including free radicals. Free radicals are electrically neutral species that have incomplete bonding; that is, they have unpaired electrons. Examples are SiO , SiH_3 , and F (the first two are important in plasma deposition and the last one in plasma etching). These species are extremely reactive, and with other species are absorbed onto the wafer surface, migrate, interact, rearrange, and chemically recombine to form the film. In addition, ion and electron bombardment from the plasma onto the wafer surface can occur. As we will see later in the description of RF sputtering systems, a potential difference will exist between the plasma and the wafer surface in RF plasma systems, with the wafer surface electrically negative with respect to the plasma. This will accelerate positive ions to the wafer surface. Even some of the high-energy electrons can escape the plasma to strike the wafers. This bombardment of ions and electrons transfers even more energy to the species on the surface, breaking chemical bonds for example, and further enhances the different surface processes and reactions. The net result from the plasma-induced fragmentation, free radical generation, and ion bombardment is that the surface processes and deposition can occur at much lower temperatures than in nonplasma systems. The byproducts of the deposition reactions are desorbed and exhausted through the outlet ports.

The reactions and processes that occur in a PECVD system are numerous and complicated and not easy to predict or model. The complexity and nonequilibrium nature of the reactions can lead to nonstoichiometric compositions of the films (Si - or O -rich SiO_2 , for example), as well as the incorporation of byproducts into the films. Incorporation of H_2 , O_2 , or N_2 is common and can result in outgassing, peeling, or cracking of the film during subsequent processing. Film density and stress may also vary, depending on the conditions of the deposition.

PECVD depositions can result in fairly good coverage and filling of non-planar topography in some cases (such as TEOS SiO_2 depositions), better than what might be expected at these low temperatures and often similar to higher temperature depositions in non-plasma systems. This has often been attributed in the past to long-range surface diffusion, enhanced by the more energetic species and by ion bombardment. However, as we will discuss later, for oxide deposition this may be instead due to multiple near-surface emission and redeposition events occurring, characterized by a low effective sticking coefficient for the neutral species even at low temperatures, along with the directional ion component of the deposition to help in filling spaces [9.4, 9.5]. While standard PECVD depositions may be adequate for covering and filling larger dimension

structures, for smaller dimensions and higher aspect ratio spaces (greater than about 1) other methods are required.

9.2.1.4 High-Density Plasma Chemical Vapor Deposition (HDPCVD)

A newly developed version of PECVD utilizes a very high-density plasma and a separate RF bias applied on the substrate. Called HDP (High-Density Plasma) CVD, this technique combines PECVD deposition with bias sputtering to obtain very good filling of narrow gaps. It is used primarily for silicon dioxide depositions. The high-density plasma can be generated by a variety of sources, including Electron Cyclotron Resonance (ECR) and Inductively Coupled Plasma (ICP). This high-density plasma results in a dense CVD silicon dioxide film at low temperatures, typically 150°C down to room temperature, and with a very low chamber pressure in the 1–10 mtorr range. In fact, there is usually no intentional heating in these CVD systems. The ion bombardment supplies enough energy to raise the substrate temperature and cooling of the wafers is often required to keep the temperature below 400°C.

With a separate RF bias applied on the substrate in HDP systems, the angular dependence of ion sputtering is exploited. Sputtering—the knocking off of atoms from a solid by incident ions when the solid is electrically biased—occurs preferentially on sloped surfaces rather than on vertical or horizontal surfaces. This simultaneous sputtering of the film during its deposition can result in planarized and void-free films. This is the same idea behind the bias sputtering deposition method. Bias-sputter deposition and angle-dependent sputtering will be more fully explained in the next section. The strong directed ion component of the HDP deposition process also helps in the filling of holes, and the ion bombardment helps make the film denser.

9.2.2 Physical Vapor Deposition (PVD)

Rather than relying on chemical reactions to produce the reacting species and to form the film as in chemical vapor deposition, physical vapor deposition methods use mainly physical processes to deposit the films. PVD techniques are generally more versatile than CVD methods, allowing for the deposition of almost any material. The constituent species, individual atoms or molecules, are produced by either evaporation of a solid or molten source, or by using energetic gaseous ions in a plasma to knock off, or sputter, the atoms from a source “target.” These atoms or molecules then travel through a vacuum or a very low pressure gas phase, impinge on the wafers, and condense on the surface to form the film. Very few, if any, chemical reactions occur. An exception to this is in reactive sputtering, in which a species is sputtered in the presence of a reactive gas (such as N_2) and a compound is formed and deposited. But in general, physical processes dominate in PVD.

Because of the very low pressures in these systems, very few gas-phase collisions occur. In addition, the surface reactions occur very rapidly and very little rearrangement of atoms usually occurs on the wafer surface. As a result, thickness uniformity, shadowing by surface topography, and step coverage can be very important issues in PVD depositions.

Both PVD techniques, evaporation and sputter deposition, have a long history. Both have been used for over 100 years to coat objects with metal thin films [9.6]. Today sputtering is the dominant PVD technique, but evaporation is still used for some special applications.

9.2.2.1 Evaporation

A schematic of a simple evaporator is shown in Figure 9-15. In this process the source material is heated in a vacuum chamber which has initially been pumped down to less than 10^{-5} torr. Evaporated atoms from the source condense on the surfaces of the wafers. The heater can be of the resistance type, using a tungsten filament which heats up when current passes through it. More popular for microelectronics use is an e-beam heater, in which a high-energy electron beam is focused onto the source material in a crucible using magnetic fields. e-beam heaters can achieve higher temperatures so that a wider range of materials can be evaporated. In addition, the process is cleaner since no metal filaments are used. (Recall our discussion in Chapters 1 and 4 regarding alkali ions as sources of electrical instability in early MOS devices. One source of these contaminants was found to be Al evaporation systems where sodium and potassium were used in the production of the tungsten filaments.) In addition, since only the top of the source in e-beam systems is usually melted, no contamination from the crucible occurs. As a result, purer films are deposited with e-beam evaporation systems. A downside of e-beam systems is that X-rays can be emitted when an e-beam strikes Al. These X-rays can create trapped charges in gate oxides. Annealing of the film is required to remove this damage.

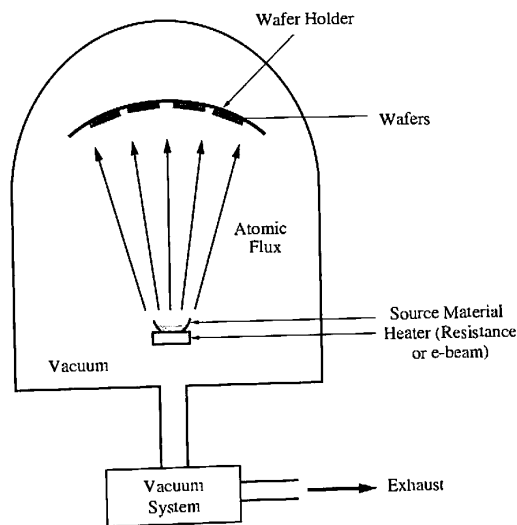


Figure 9-15 Schematic diagram of evaporation equipment.

Since the evaporated atoms are transported through a high-vacuum atmosphere, very little gas-phase scattering occurs. Therefore, unlike CVD systems, the mean free path is very long and the atoms travel in essentially straight lines all the way from the source to the surface of the wafers. This is illustrated in Figure 9–15. Since the source is small, we can first consider it as a point source. For uniform or isotropic emission from a point source, that is, an equal evaporation rate in every direction, the contours of equal flux (atoms per area per time) would be circular in two dimensions or spherical in three dimensions.

For a point source one can derive an expression for the deposition rate or growth velocity, v , of a film deposited on a flat surface as a function of distance along the surface. For this we use the coordinate system of Figure 9–16. We are interested in the amount of material striking a small area on the surface, A_k , per unit time. But first we consider the flux being emitted from the source. If R_{evap} is the evaporation rate from the point source (in either atoms or grams per second), r is the distance from the source to the spot on the surface, and A_k^p is the projected area of A_k facing the source, then the flux F_k^p that strikes A_k^p is simply

$$F_k^p = \frac{R_{\text{evap}}}{\Omega r^2} \quad (9.20)$$

Here Ω is the solid angle over which the source emits evaporated material. For a source emitting in all directions (in three dimensions) this would equal 4π . For a source emitting only upward, this would equal 2π .

Eq. (9.20) describes the flux emitting outward from a point source and it is independent of emission angle θ_i for any radius r . Since the deposition rate on A_k depends on the component of the incoming flux that is normal to the surface A_k , we must then mul-

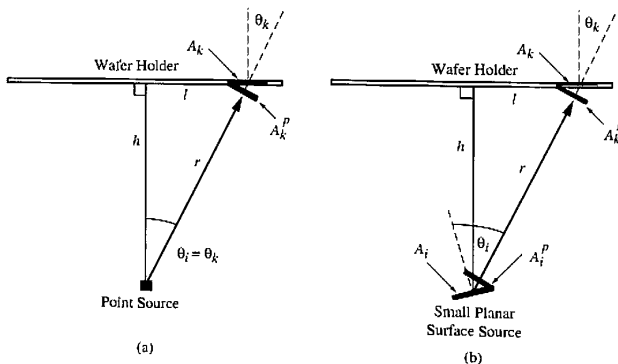


Figure 9–16 Geometries of flux and deposition of small areas on a flat wafer holder for (a), a point source, and (b), a small planar surface source.

tively this by $\cos\theta_k$. Here θ_k is the angle between the surface normal and the direction of the source to a spot on the surface, as indicated in the figure. We designate the normal component of the incoming flux with respect to A_k to be F_k . This corresponds to the amount of material deposited per time per unit surface area. This leads to

$$F_k = \frac{R_{\text{evap}}}{\Omega r^2} \cos\theta_k \quad (9.21)$$

Another way to look at this is that for a larger angle, the material that is coming through the cone to strike the area A_k must be deposited over a larger surface area A_k , and the amount striking A_k per unit area is thereby decreased (by $\cos\theta$). Thus the deposition rate on a surface area will decrease as the area is tilted from an orientation facing the source.

The deposition rate or velocity, v , on area A_k is then just the flux F_k as expressed by Eq. (9.21) divided by the density of the material being deposited, N , so that for a point source:

$$v = \frac{R_{\text{evap}}}{\Omega N r^2} \cos\theta_k \quad (9.22)$$

Eq. (9.22) can be converted to a plot of the relative or normalized deposition rate versus ℓ/h , as is done in Figure 9-17, where ℓ is the distance from the center of the wafer holder and h is the length of the surface normal. Note that $\cos\theta_k = h/r$ and that $r^2 = \ell^2 + h^2$. The thickest deposit is at $\ell = 0$ and leads to v_o (v at $\ell = 0$) being equal to $R_{\text{evap}}/\Omega N h^2$. The nonuniform deposition is obvious. As ℓ increases, not only does r increase, but also the $\cos\theta_k$ term decreases.

Evaporation is usually modeled better as a small surface area source rather than a point source. Evaporation from a small surface area behaves similarly to effusion of gas

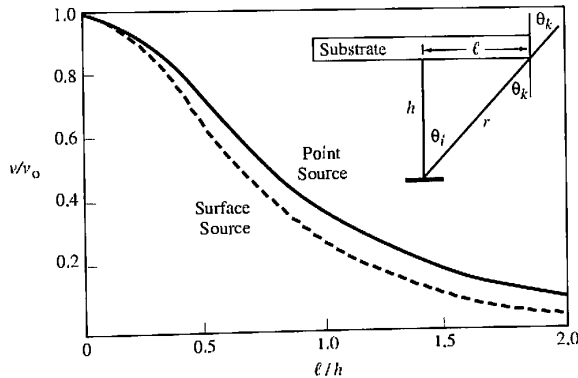


Figure 9-17 Deposition rate of evaporated film as a function of position on substrate for point and surface sources. $\theta_l = \theta_k$ in this configuration for both point and surface sources. (After [9.6].)

out a small opening from a container, as in a Knudsen cell [9.7]. The outward or emitted flux is not uniform in all directions as with an ideal point source but is more directed and dependent on the projected area of the small source area toward the flux direction. Thus the emitted flux is largest in the direction perpendicular to the source surface and less to the sides. The orientation of the 2D surface source to the area being deposited on is defined by the angle θ_i , as shown in Figure 9-16(b), which may be different from θ_k . As θ_i becomes larger, the projected area of the surface source, A_i' , facing the deposition area, becomes smaller and there is less flux in that direction. Hence the emitted flux for any distance r is greatest in the direction normal to the source, $\theta_i = 0$, falling off as a cosine function toward either side. This is different than the $\cos\theta_k$ factor above used to calculate the deposition rate on a flat surface and instead describes how a certain type of source is seen to emit material differently in different directions, following a cosine relation. This emission behavior from a small planar source is known as the cosine distribution law for evaporation from a planar source, or ideal cosine emission.

An extra $2\cos\theta_i$ factor due to this cosine emission now comes into the expressions for the emitted flux [Eq. (9.20)], the normal component of the incoming flux striking A_k , [Eq. (9.21)], and the deposition rate [Eq. (9.22)]. (The factor of 2 normalizes the total evaporation rate with the $\cos\theta_i$ term.) The deposition rate, ν , for a small surface area source becomes

$$\nu = \frac{R_{\text{evap}}}{\pi N r^2} \cos\theta_i \cdot \cos\theta_k \quad (9.23)$$

Eq. (9.23) is valid for any small area emitting material, with ideal cosine emission behavior, to any other small area in any relative orientation to each other, with r the distance between them and the angles as defined above.

While it may not be intuitively obvious that this Knudsen-cell-like behavior, resulting in cosine emission behavior, should apply to evaporation from a small surface, it has been verified experimentally by Knudsen himself [9.8] and others for many materials. (In radiation theory, this type of emission is called “diffuse” emission or more specifically “perfectly diffuse” emission from a surface, and very similar equations express the radiation behavior from a blackbody [9.9].) Cosine emission behavior not only describes source emission in most evaporation systems, but also describes the reemission of deposited material from a surface area in many cases, as we will discuss later. The contours of equal flux for an evaporation system with a small planar source would not be hemispherical as in the point-source case in Figure 9-15 but would fall off more quickly from the center of the wafer holder to each side.

We can now plot the deposition rate versus ℓ/h curve for a small surface source in Figure 9-17, as we did for a point source. We do this for the special case when the small surface source is parallel to the wafer holder. In this particular situation with the surface source facing the wafer holder, θ_i equals θ_k . The deposition curve for the planar surface source case is similar to the point-source case, except that the deposition is even more directed toward the region directly above it. Not only is r increasing and the $\cos\theta_k$ term decreasing as ℓ gets larger, as with the point source, but due to the $\cos\theta_i$ emission from the planar source, even more deposition occurs upward.

To achieve thickness uniformity over many wafers for either type of source, the wafers obviously cannot be placed on a flat holder. Instead, spherical or hemispherical

holders are used, with the wafers all facing inward toward the center. For a point source at the center of the sphere, θ_k and r will be constant giving a uniform deposition rate over the whole wafer holder. (Across each individual wafer, θ_k and r will vary somewhat, giving less than uniform across each wafer, but from wafer to wafer the deposition will be uniform.) It can be shown (problem 9.8) that to obtain uniform deposition around a sphere for a small surface source, the source itself should be placed on the inside surface of the sphere holding the wafers, rather than the sphere's center. Figure 9-18 illustrates these configurations. Since most evaporation sources act more like small surface sources than point sources, it is common practice to place the source at the inside surface of a spherical wafer holder in an evaporation system.

Deviations from ideal point-source or small-area surface source emissions can occur and sometimes even more directed and narrower emission distributions are observed. For small planar surface sources, the flux and deposition rate expressions can be described with a $\cos^n \theta$ term rather than a $\cos \theta$ term to take this nonideal behavior into account. Values of n greater than 1 correspond to angular distributions of evaporant flux from the source being narrower than from ideal sources. This is conceptually illustrated in Figure 9-19 for a planar source, where a value of $n = 1$ leads to ideal cosine emission

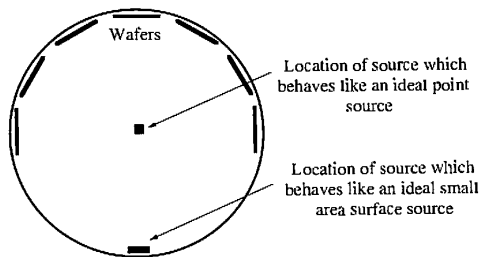


Figure 9-18 Positions of wafers and sources to achieve uniform deposition in evaporation system on all the wafers. Most evaporation sources behave more like a small area surface source and are in that configuration.

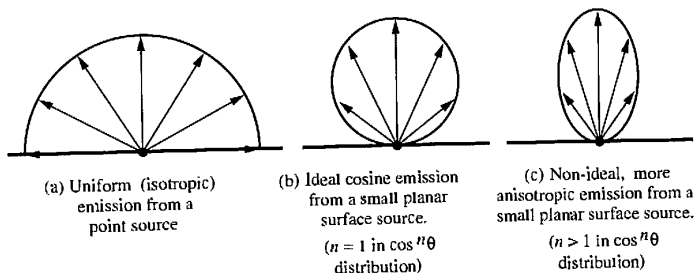


Figure 9-19 Emitted fluxes for point sources and small surface areas sources.

behavior in the emitted angle distribution, and a value of $n > 1$ leads to even more anisotropic behavior.

This nonideal behavior may be due to crucible geometry, the ratio of melt depth to the melt surface area, and other factors. In addition, a large density of evaporated atoms right above the source sometimes causes the gas phase to act more like a viscous medium rather than the preferred near-vacuum condition. (In the latter case, molecular flow, or free molecular flow, is said to occur, with virtually no gas-phase collisions or interactions.) This viscous behavior near the source also causes nonideal, nonuniform behavior. To counter these problems another solution is employed. Not only are hemispherical wafer holders used, but they are rotated. And not only is the entire planetary fixture rotated about the vertical axis, but each group of wafers is also rotated about a second axis. This use of rotating planetaries helps compensate for nonideal deposition behavior and results in more uniform deposition.

The source evaporation rate, R_{evap} , is found to be proportional to the equilibrium vapor pressure (P_e) of the source material at the evaporation temperature. From Langmuir-Knudsen theory, the evaporation rate in g sec⁻¹ is estimated to be

$$R_{\text{evap}} = 5.83 \times 10^{-2} A_s \left(\frac{m}{T} \right)^{1/2} P_e \quad (9.24)$$

where A_s is the area of the source in cm², m is the gram-molecular mass, T is the temperature in Kelvin, and the vapor pressure, P_e , is in torr. The vapor pressure of the material is a very strong function of temperature, and Figure 9-20 shows the vapor pressure of several metals commonly evaporated. A partial vapor pressure of the element of about 1–10 mtorr or more is required to achieve reasonable deposition rates (on the order of 0.1–1 μm per minute), and one can see that a temperature of about

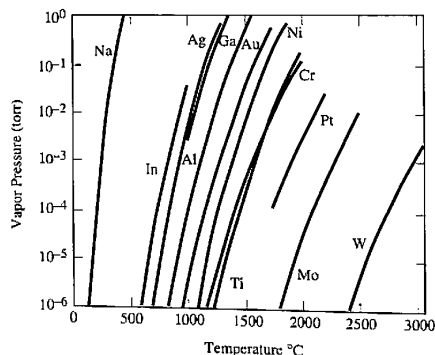


Figure 9-20 Vapor pressure as a function of temperature of commonly evaporated metals. (After [9.10].)

1100°C is needed for the evaporation of aluminum. This means that the aluminum is evaporated from the liquid phase. Tungsten must be heated to a much higher temperature of about 3100°C, where it evaporates or “sublimes” from the solid phase.

Because the vapor pressures of different elements vary so much, evaporating alloys presents a special challenge. Alloying one element with another is common in microelectronics to achieve the desired thin film properties. Alloys of Al with small amounts of Si and Cu for interconnects are examples, as we will see in Chapter 11. But with different vapor pressures at a given temperature (the vapor pressure of Cu is about 5 times lower than that of Al at 1100°C, while that of Si is 10^4 times lower!), the evaporation rates will be different, and the deposited film will have a different composition than the source. When evaporating films from a single, multicomponent source, the differences in evaporation rates of the constituent elements must be taken into account, and the sources must be enriched in the element with the lower evaporation rate. But the source itself will change in composition as one element evaporates faster than the other (called fractionation). This will in turn change the composition of the deposited film, resulting in a deposited film with a composition that varies throughout the thickness of the film. This problem is very similar to the dopant segregation problem during Czochralski crystal growth that we considered in Chapter 3.

The most common solution to this problem is two or more separate e-beams used on separate sources, each containing just one of the elements in the alloy. The e-beam energies, and hence the source temperatures, can be different so that the evaporation rates can be controlled to give the desired deposited film composition. Differences in evaporation rates and changing source compositions are not a problem here. But spatial compositional uniformity when using dual sources is a problem, and it is difficult to obtain constant compositions across the wafer and across the wafer holder.

Compounds, such as SiO_2 , are just as hard, if not harder, to evaporate as alloys. The composition of the vapor phase is usually not the same as the source compound, and chemical reactions and molecular changes often occur as the material is evaporated and deposited. Sputter deposition generally does a much better job in depositing compositionally uniform films of both alloys and compounds. This is partly because the sputter rates of elements are usually much closer to each other than their vapor pressures are.

There are advantages in using evaporation to deposit thin films for microelectronics applications. First, there is little damage caused to the wafer, since the wafers are not subjected to energetic particles. Second, the deposited films are usually very pure. Because the deposition is done in a high vacuum, there are no residual gases or particles to get incorporated in the film.

However, there are also disadvantages to this method which have all but eliminated its use from mainstream silicon fabrication. First, metals with low vapor pressures, such as W, and films of alloys or compounds with precisely controlled composition are difficult to evaporate, as discussed above. Second, there is no *in situ* precleaning method available as there is for sputter deposition. The third problem with evaporation has to do with step coverage. Because of the very low chamber pressure which makes the mean free path of the gaseous species very large, the evaporated deposition species travel essentially in straight lines from the source to the wafer surfaces. This and the fact that the source is small means that the atoms arrive at a limited range of angles, leading

to poor step coverage. Rotating the planetary and wafers does help to widen the range of arrival angles, but shadowing still occurs.

Step coverage also depends on another property or characteristic of the deposition species, the sticking coefficient. This can be even more important than the line-of-sight issues in regards to step coverage and filling ability. The sticking coefficient is the ratio of the number of species that actually stay or “stick” on the surface relative to the number of incident species. We define the sticking coefficient S_c , or more accurately the effective sticking probability, specifically in terms of incident and reacting fluxes, where a deposition species that has “reacted” on the surface is one that properly deposits and stays there:

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \quad (9.25)$$

The idea of a sticking coefficient is important for step coverage since the deposition species that do not stick or react are desorbed and may deposit elsewhere—on the sidewall of a step, a trench, or a contact hole, for example. This is illustrated in Figure 9-21. Therefore, deposition processes with a sticking coefficient close to one, in which the species react and stick locally where they first hit without desorbing, will tend to have poorer step coverage and filling ability. Those with a sticking coefficient much less than one, in which the species bounce around before they stick and deposit, such as is generally the case in CVD depositions, will have better step coverage and better filling. In evaporation systems and other PVD methods, the sticking coefficient of the depositing species is usually very close to one, meaning that they deposit where they hit and there is little or no redepositing of desorbed species to other regions. Surface diffusion is also very slow at low temperatures. In addition, there is no resputtering of material in evaporation systems, as can be in the case in plasma CVD or sputter deposition systems where there is significant ion bombardment. As we will see in more detail later, deposition fluxes that have narrow arrival angle distributions and high sticking coefficients with little or no resputtering or diffusion lead to very poor step coverage. Deposition will occur primarily on horizontal surfaces, not on the sides of features. Heating the substrate during the evaporation process increases the surface diffusion and allows for

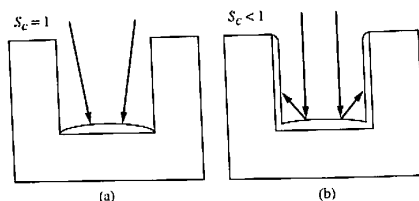


Figure 9-21 In (a) the depositing species have a high sticking coefficient (close to 1) so that they are deposited where they first strike. In (b) the depositing species have a low sticking coefficient (< 1) so that many are reemitted and deposited elsewhere on the topography, such as on the sidewalls.

some coverage on sidewalls. But heating can also affect the film composition and microstructure.

For all these reasons—difficulty in depositing alloys, compounds, and high-vapor pressure elements, lack of an in situ preclean, and poor step coverage—evaporation is rarely used today for standard silicon microelectronics technology. CVD and sputter deposition, described next, are now used instead. Nevertheless, we will make extensive use of the basic concepts of emission distribution laws and deposition rates developed in this section.

9.2.2.2 Sputter Deposition

Because the vacuum requirements for sputter deposition are less severe than for evaporation (in the 1–100 mtorr range for sputtering compared to $<10^{-5}$ torr for evaporation), sputter deposition was the preferred method in a variety of technologies prior to the microelectronics age. However, because the higher pressures in sputtering systems allowed significant contamination, evaporation was generally preferred in the early years of silicon microelectronics technology. But with the availability of ultrahigh purity source gases and the use of an initial low-pressure pumpdown, gas incorporation in sputtered films was found to be greatly reduced and contamination became much less of a problem with sputtered films. Because of the problems with evaporation discussed above, as well as the inability to develop manufacturable CVD methods for some materials, sputter deposition is commonly used today in semiconductor technology.

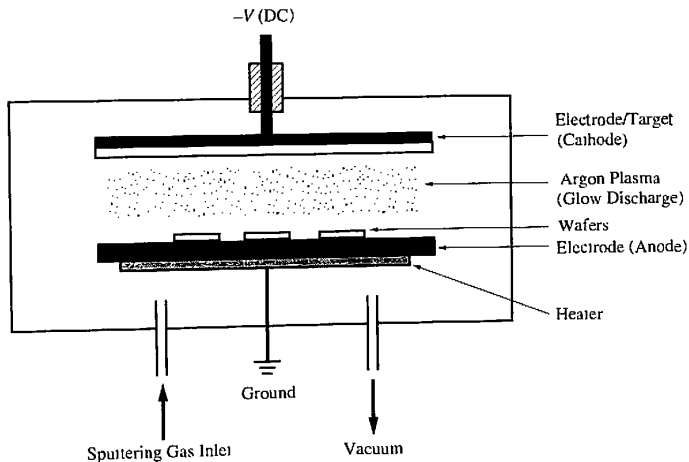


Figure 9-22 Schematic diagram of DC-powered sputter deposition equipment.

DC Sputter Deposition Figure 9-22 illustrates schematically a sputter deposition system. Note the similarity to the PECVD system in Figure 9-14. In the sputter system an inert gas, such as argon, is fed into the sputtering chamber at low pressure. A voltage is applied across the two electrodes and a plasma is created. The plasma contains neutral argon atoms, and roughly equal numbers of positive ions and free electrons, and is a conducting medium. (No free radical species are present here since Ar does not form free radicals in a plasma as O, H, and F species do in plasma deposition and etching systems.) The top electrode, where a negative DC voltage is applied, is actually the source material to be deposited, a plate or "platen" of Al, for example. This electrode becomes the cathode and is called the target. The bottom electrode is a metal platen, which is grounded and becomes the anode. The wafers sit upon the anode electrode. The positive ions in the plasma are accelerated to the negatively biased target, which can be several hundred to several thousand volts negative relative to the plasma. These energetic ions strike the target and dislodge, or sputter, the target atoms. These atoms are then free to travel through the plasma as a vapor and strike the surface of the wafers, where they condense and form the deposited film.

Since the target acts as an electrode in the DC mode of sputter deposition, the target or source material must be conductive. Therefore Al, W, Ti, silicides, and other metals can be sputtered this way. To deposit nonconductors—oxides, nitrides, and lightly doped silicon—another method must be used. We will discuss a technique to do that, the

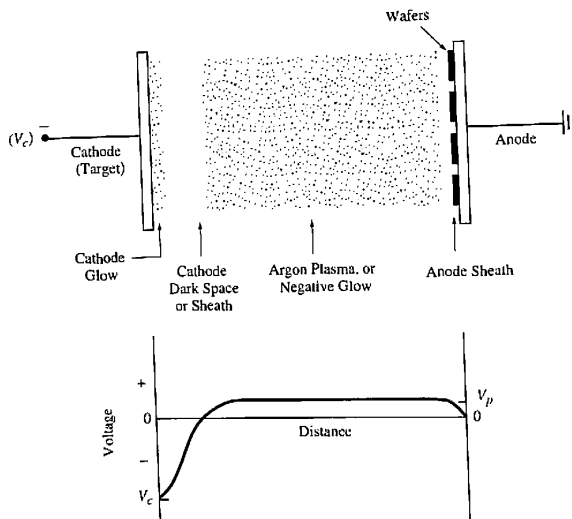


Figure 9-23 Plasma structure and voltage distribution in DC sputter system. (After [9.11].)

RF or AC method, a little later.

The structure and voltage distribution of the plasma in a DC system are shown in Figure 9-23. Most of the space between the electrodes is made up of the argon plasma, called the negative glow due to photon emission from excited Ar atoms. The voltage is fairly constant throughout this region since it has a fairly low resistance. Near the negative electrode, where there is an excess of positive ions and a shortage of electrons, as discussed below, there exists a space charge region that does not glow. This is called either the cathode dark space, the Crookes dark space, or the cathode sheath and is on the order of 0.1 to 10 mm thick. It is across this space charge region that most of the voltage drop in the system occurs. Because of the low density of electrons in the cathode sheath, few collisions between Ar atoms and electrons occur and that region is dark. Right next to the negative electrode is the thin cathode glow region; the light produced in this case is due to the interaction between the incoming ions and the cathode material. In this region incoming argon ions and sputtered target ions are neutralized and secondary electrons created by the ion bombardment are accelerated away from the cathode. These electrons then pass through the cathode sheath and collide with the Ar atoms in the negative glow region, ionizing some of them and sustaining the plasma.

While the electrons are accelerated in one direction across the high electric field in the cathode dark space, the positive argon ions are accelerated in the opposite direction, from the negative glow region to the target. They strike the target surface and sputter off the source atoms, which can then transport to the wafers and form the film.

Additional glow regions and dark spaces can form closer to the anode, such as the Faraday dark space, the positive column region, and the anode sheath [9.11]. But these are not useful in the sputter deposition process, and all but the anode sheath are eliminated or minimized by having the positive electrode close to the negative glow, resulting in the plasma structure and voltage distribution shown in Figure 9-23.

One interesting feature is that the plasma potential, V_p , is positive with respect to the anode (on the order of 10V), as shown in Figure 9-23. In fact, the plasma potential will be positive with respect to any surface within the plasma, with positive space charge regions, or sheaths, formed next to each surface. This is primarily due to the fact that both ions and electrons from the plasma will randomly strike the surface and escape the plasma. But the electrons with their lighter mass and higher mobility will strike the surface more often, causing an imbalance in the ion and electron currents. An electric field develops that opposes this imbalance by opposing the electron flow, raising the voltage (by V_p). Therefore the potential of the plasma will be positive with respect to both the anode and cathode, with a positive space charge region of excess ions and a shortage of electrons forming a sheath at both the anode and the cathode. With the applied negative voltage at the cathode, the voltage distribution shown in Figure 9-23 results, with a net DC current passing through the circuit. In the region near the cathode, this current is predominantly carried by the positive ion flux (which is also responsible for the sputtering). The large voltage drop there accelerates the secondary electrons from the cathode across the sheath and sustains the plasma. In the region near the anode an equal current flows, but is mostly carried by the electrons. Because of the exponential Maxwell-Boltzmann distribution of the electron energy, only a small perturbation of the zero-current plasma potential is needed to supply the net DC current at the anode. Because of the potential drop from the plasma to the anode, some sputtering of the anode

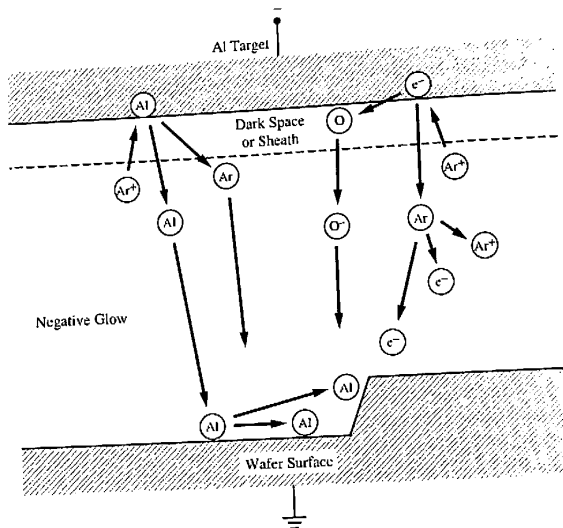


Figure 9-24 Important processes in sputter deposition.

can occur, but very little compared to that at the cathode.

Several of the important processes that occur during sputter deposition are illustrated in Figure 9-24. At the top left, an Ar ion from the negative glow accelerates across the dark space and sputters off an Al atom from the target. (A minimum energy, on the order of 10–20 eV, is needed to sputter an atom, and the sputter yield increases with ion energy.) This Al atom travels through the dark space and negative glow and lands on the wafer surface. If the argon atom is neutralized during gas phase collisions, it can retain enough kinetic energy to also travel to the wafer surface and become incorporated in the growing film. The Al atom now on the wafer surface may stay absorbed and become part of the growing film there. It may also surface migrate or diffuse to another location. Finally, it can leave the surface by either desorbing on its own (though this does not occur very often compared to CVD systems, as we will discuss later), or by being resputtered by another atom hitting the surface. The Al atom could then be redeposited at another location.

In the upper right, an Ar ion strikes the target, and a secondary electron is emitted. The secondary electron can accelerate across the dark space and strike another argon atom, exciting it or ionizing it, and sustaining the plasma. The electron can also ionize an impurity atom, oxygen in this case, which can then be accelerated across the dark space and travel toward the wafer surface. Finally, all the species moving toward the

wafer surface can be embedded in the wafer surface. Even a significant number of positively charged ions can be made to strike the wafer surface with high energies by independently biasing the wafer surface relative to the plasma, as will be described later in the sections on bias sputtering and ionized sputtering. Or the atoms or ions can strike the atoms on the wafer surface, imparting energy to them and enhancing their movements and reactions, and even heating the wafer surface. Thus physical reactions and processes, rather than chemical, dominate in PVD.

In evaporation, the deposition rate was seen to depend on the source material vapor pressure. In sputter deposition, it is the source, or target, sputtering rate that is important. The plasma gas ion strikes the target and physically dislodges the target atom. This is a physical, rather than a chemical, process, and is often explained using billiard ball models. Sputtering occurs when the incoming ion transfers enough energy to the target surface to break the bonds holding the target atom in place. This depends on the energy and mass of the ions, and the target material.

The sputter yields for argon sputtering of different target materials do not vary very much for the different target materials compared to the wide range of vapor pressures important in evaporation. Typically this yield, defined as the number of atoms or molecules ejected from the target per incident ion, is in the range of 0.1 to 3. It does depend on the energy of the incoming ion, determined mainly by the voltage across the sheath, and on the direction of incidence of the ions. An ion coming in at a low, glancing angle will not be as successful at dislodging an atom, with the ion itself being reflected off the surface. At the other extreme, ions arriving normal to the surface will push a surface atom further into the target but not necessarily sputter it off. The sputter yield peaks at some angle less than 90° . The angles that the sputtered atoms leave the surface of the targets are usually found to be diffusely distributed, with an ideal cosine-emitted angle distribution, or close to it, from each position on the target. This means that the sputtered atoms are ejected from each location on the target in a broad range of directions.

The targets used in sputter deposition come in a variety of shapes and sizes. Multiple targets may also be used in these systems for cosputtering of alloys and compounds or for multilayer depositions, where Ti and Al layers may be deposited one after the other in the chamber without breaking vacuum. Achieving good thickness uniformity in the different types of systems is accomplished in different ways: having the target larger than the wafer or wafer holder; having the wafers pass by or spin in front of a long rectangular target; or by using the cylindrical wafer holder around an oblong target (similar to having a spherical holder in front of a point source in evaporation.)

Having a large area target also means that the arrival angles of the atoms at the wafer surface are more widely distributed which generally improves step coverage. Having the wafers rotate or pass by a target is somewhat similar to having a large target in this regard and also helps increase the range of arrival angles in the direction of the rotation.

Let us first discuss how we describe or quantify the "arrival angle distribution." Here we are interested in the flux, as a function of arrival angle, arriving at any small area on the wafer surface. In Figure 9-25 on the left we show fluxes arriving at the surface from all angles with equal magnitude; that is, the number of particles or atoms, per time and per area normal to the flux direction, is equal for all incoming directions. We call this isotropic arrival. We could describe this arrival distribution as being constant since the flux relative to the flux direction is constant, similar to the way we characterized the

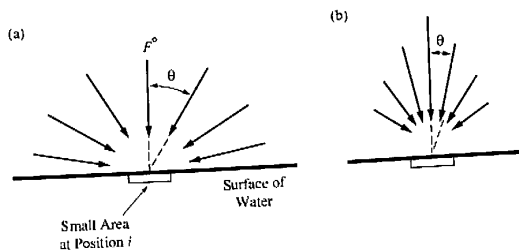


Figure 9-25 Distribution of arrival fluxes for (a) uniform or isotropic arrival distribution; and (b) directed or anisotropic arrival distribution. Arrival angle distribution ($\cos^2\theta$) is defined by arrival flux relative to unit surface area. This flux is equal to the normal component of incoming flux, relative to the vertical direction for a horizontal surface.

emission angle distribution. However, we are going to define the arrival angle distribution a little differently. We are ultimately interested in the deposition on a surface area, so we are concerned with the flux in terms of particles per time per area on the surface, or per unit surface area. This is just the incoming flux, which is constant for isotropic arrival, multiplied by the cosine of the arrival angle θ , where θ is the angle between the incoming flux and the normal direction to the surface. Just as in our treatment of a flux striking and depositing on a flat surface in an evaporation system, the deposition rate depends on the normal component, or the cosine of the angle, of the incoming flux to that surface. Using the cosine function in the arrival angle distribution also allows us to easily describe both isotropic and anisotropic arrival distributions. (This convention may be a little confusing since in our description of the emitted angle distribution for evaporation sources and emission, which uses a different convention, a “cosine distribution” from a small area is not uniform or isotropic as we describe it but is directed toward the surface normal. But to be consistent with normal terminology for emission, as well as with that used to describe arrival angle distributions in our deposition and etch modeling using the SPEEDIE simulator, we will use these two conventions.)

We thus describe the arrival angle distribution in terms of $\cos\theta$, or the normal component of the incoming flux. For isotropic incoming fluxes, the arrival angle distribution by this definition is therefore described as $\cos\theta$, with a maximum at $\theta = 0^\circ$ and decreasing as the angle gets larger, eventually going to 0 at $\theta = 90^\circ$. Another way to look at it is that for isotropic arrival, the atoms that strike the unit surface area from a particular direction, θ , must pass through the projected area of the unit surface area facing the flux direction. As θ increases, the projected area gets smaller. Therefore fewer atoms strike the unit surface area, and the flux relative to the surface area (the atoms per time per unit surface area) decreases. Ultimately, no atoms strike the surface area when $\theta = 90^\circ$, since the flux is parallel to the surface.

For incoming fluxes that are not isotropic, but anisotropic and more directed toward the surface [as shown in part (b) of the figure], a $\cos^2\theta$ distribution can be used. A value of n larger than 1 describes a more directed or anisotropic incoming flux distribution.

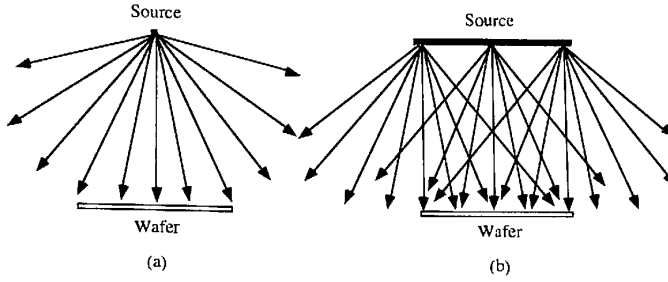


Figure 9-26 Flux lines of atoms from small area and extended sources. In (a) the flux lines from a point or small area source are shown, with a wide emitted angle distribution but a narrow arrival angle distribution. In (b) the flux lines from an extended source are shown. A wider arrival angle distribution is seen with the extended source.

We now return to illustrating the idea that a larger source region, or target area in the case of sputtering, results in a wider arrival angle distribution. In Figure 9-26(a), the flux lines of atoms from either an ideal small area surface source or an ideal point source are shown, in either case emitting in a rather wide angular distribution. While the atoms are emitted from the small source with a wide angular distribution, the distribution of angles arriving at each point or small area on the wafer is very small. This tight or narrow distribution can be described by an arrival distribution of $\cos^n \theta$ with n much greater than 1. Obviously a wide emission or departure distribution of angles does not necessarily mean a wide arrival distribution of angles.

In Figure 9-26(b) an extended source is used, such as a larger target. At each position on the wafer, atoms arrive at a range of angles, as determined by the geometry of the system, and not just from a narrow range of angles. Here n in the $\cos^n \theta$ arrival distribution would be closer to 1, meaning less anisotropic and more isotropic. This can cause more uniform step coverage on steps in the wafer topography, and would also improve thickness uniformity across the substrate.

As we described earlier, the arrival angle distribution for evaporation systems can be made wider by rotating the wafer holder around two axes, effectively making the source wider, and the step coverage can be improved. But there are other reasons that step coverage is usually still better in sputtering systems. Because of the higher pressure in sputtering systems, the atoms experience more gas-phase collisions in sputtering systems. One can estimate the average number of collisions by calculating the mean free path, λ , of a particle in a gas. From kinetic theory, the mean free path of a gas particle is

$$\lambda = \frac{kT}{\sqrt{2}\pi d^2 P} \quad (9.26)$$

where $k = 1.36 \times 10^{-22} \text{ cm}^3 \text{ atm K}^{-1}$, T is the temperature in K, d is the collision diameter of the molecule in cm (approximately $4 \times 10^{-8} \text{ cm}$, or about 0.4 nm, for most molecules of interest), and P is the pressure in atm. In a typical sputtering system with $P =$

5 mtorr, this results in a mean free path on the order of 1 cm, which is a little less than the physical path length the depositing atoms follow.

For evaporation systems, the number of collisions experienced between the source and wafer is virtually zero, while in sputtering systems the average number of collisions experienced is on the order of one to a few. This tends to make the arrival angle distribution somewhat wider. (Increasing the pressure or number of collisions too much to widen the distribution can result in poor film properties, partly due to less energetic species, and contamination problems.) Also, the number of resputtered atoms on the wafer surface is higher for sputtering due to the bombardment of energetic particles. This effectively increases the arrival angle distribution even more. For these reasons, step coverage in sputtering systems is usually better than in evaporation systems. But it is still usually much worse than in CVD depositions, which generally have an isotropic arrival angle distribution and lower sticking coefficients, as we will discuss later. In some cases, though, we actually want to narrow the arrival distribution—for example when filling a deep via with metal in a PVD system. We will see how that can be achieved when we discuss collimated and ionized sputtering systems.

Because the sputtering yields of different elements are much more similar to each other than their vapor pressures are, the deposition of alloys and compounds should be much more controllable using sputtering rather than evaporation. But there is another reason why the compositions of sputter deposited films are more controllable and uniform. Because sputtering occurs just from the top surface of the target and little mixing from the layers below takes place, a steady-state condition soon develops where the atoms are sputtered off at relative rates equal to the target composition, even when the sputtering rate of each type of atom is different. Initially the element with the higher sputtering rate does preferentially sputter off more from the surface. This leaves the top surface enriched, relative to the bulk of the target, with the element with the lower sputtering rate. But now more of the element with the lower sputtering rate is sputtered since the surface region is enriched with that element. A steady state is achieved where the elements leave the surface in a ratio that is the same as the bulk target composition—the difference in sputtering rates is exactly compensated for by the difference in the surface concentration.

The overall target composition thus remains at its original composition, with no depletion or fractionation, since the two elements are being sputtered off in proportion to the overall composition. This does not work for evaporation. Since the evaporation source material is often molten or near molten, the diffusion of atoms is very fast and mixing between the surface and bulk wipes out any concentration gradients. Therefore, the required enriched surface concentration cannot be maintained. But in sputtering, the temperature of the target is much lower—near room temperature—and little diffusion and mixing occur.

Compounds, such as SiO_2 and WSi_2 , may be sputtered as well. In some cases, they are sputtered off as individual atoms, and sometimes they are sputtered as molecules. Both alloys and compounds may be sputtered from two different targets, called cosputtering (as opposed to the use of one "composite" target). For example, WSi_2 can be sputtered by simultaneously using a tungsten target and a silicon target. But it is often difficult to

control the stoichiometry, and composite targets are more commonly used. Even then the composition may vary.

Reactive Sputter Deposition Another way in which compounds can be sputter deposited is by reactive sputter deposition. Here a reactive gas is introduced into the sputtering chamber in addition to the Ar plasma, and the compound is formed by the elements of that gas combining with the sputtered material. For example, TiN can be deposited by sputtering Ti in the presence of nitrogen. The plasma can furnish energy to the N_2 to allow it to dissociate into atomic nitrogen, which then can easily react with the Ti. The reaction usually occurs either on the wafer surface or on the target itself, and not in the plasma. In the case of TiO_2 reactive sputtering, it is found that the oxygen reacts with the Ti on the surface of the Ti target to form the oxide there, and then it is sputtered onto the wafer surface. Controlling the stoichiometry of reactively sputtered films is sometimes difficult and this method is not used very often—sputtering from compound targets is more common. Sometimes a reactive gas such as O_2 or N_2 is used in conjunction with sputtering from a compound target. This could ensure that the composition of a gaseous constituent is maintained in a sputtered film, or it could be used to improve the properties of a film. Sputtering TiW in the presence of oxygen or nitrogen has been found to improve the barrier properties of these films, as discussed in Chapter 11.

RF Sputter Deposition DC sputter deposition is not suitable for insulator deposition. When applying the necessary DC voltage to the insulating target to sustain a plasma, difficulties arise. With a negative DC voltage applied, positive argon ions from the plasma strike the negatively charged insulator and positive charge would accumulate. The negative surface voltage would become less than that required to sustain the glow discharge, and the plasma would shut down. The time required for this to occur is only about 1–10 μsec .

The solution is to use a high-frequency alternating voltage instead of DC. 13.56 MHz is commonly used. RF voltages can be coupled capacitively through the insulating target to the plasma, so that conducting electrodes are not necessary, and the positive charge buildup is neutralized by electron bombardment over each cycle. The RF frequency is chosen to be high enough so that a continuous plasma discharge is maintained, that is, higher than the effective RC time constant of the system. The difference in mass and hence mobility between the electrons and ions allows for virtually continuous sputtering of the target throughout both half cycles of the RF voltage. The electrons have a high enough mobility to keep up with the changing electric field, but the heavier argon ions do not. During the first few complete cycles more electrons than ions are collected at each electrode, and this negative charge buildup on the electrodes is maintained during subsequent cycles. As a result, both electrodes maintain a self-biasing steady-state (DC) potential that is negative with respect to the plasma voltage, V_p . The value of V_p is set by the requirement that the DC current is zero. A positive V_p aids the transport of the slower positive ions and slows down the negative electrons. The spatial distribution of the voltage that develops is shown in Figure 9-27, with the solid curve for the symmetric RF case (that is, with equal area electrodes). This is similar to

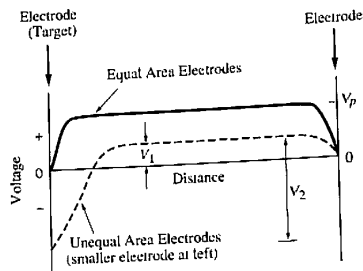


Figure 9-27 Steady-state voltage distribution in RF powered sputtering systems. The wafer sits on the right electrode.

the DC-powered case where electron depletion regions form and the plasma self-biases. However, V_p is generally larger and the net voltage drop at both electrodes would be equal in the symmetric RF powered case. The heavy ions, while they cannot travel across the sheath at the frequency of the RF signal, do respond to the average DC potential that develops across the sheath and strike the negatively biased target.

The induced negative biasing of the target due to the RF powering means that continuous sputtering of the target occurs throughout the RF cycle. But it also means that this occurs at both electrodes. The wafers will be sputtered at the same rate as the target since the voltage drops would be the same at both electrodes for a symmetric system. It would thus be very difficult to deposit any material this way. However, because the RF current must be continuous throughout the system, the voltage drops, V_1 and V_2 , across the dark space next to the two electrodes are related to the respective electrode areas, A_1 and A_2 , by [9.12]

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1} \right)^m \quad (9.27)$$

The smaller electrode requires a higher RF current density to maintain the same total current as the larger electrode. As a result, the fields must be higher at the smaller electrode. Simple theory gives $m = 4$, but experimentally the exponent is found to be between 1 and 2, perhaps due to voltage saturation and sheath collision issues, but the effect is the same. By making the area of the target electrode smaller than the area of the other electrode, the voltage drop at the target electrode will be much greater than at the other electrode. V_p in the asymmetric case would also be smaller. The resulting voltage distribution for unequal electrodes is also shown in Figure 9-27, and closely resembles that for the DC sputter deposition case (Figure 9-23 shown earlier). The volt-

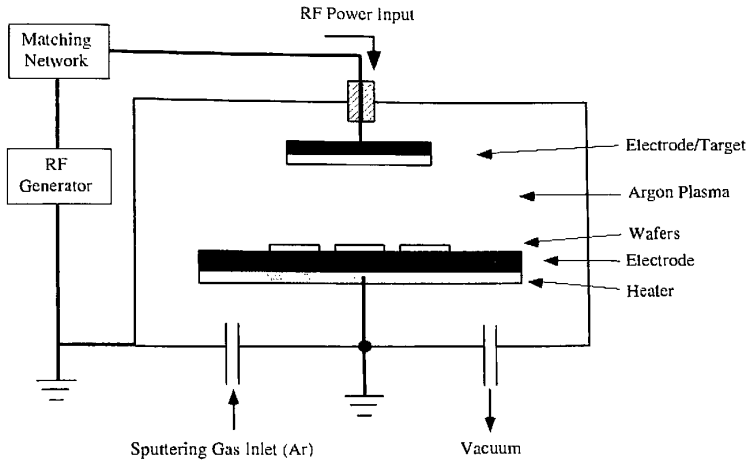


Figure 9-28 Schematic diagram of RF-powered sputter deposition system.

age drop from the plasma to the larger electrode is typically tens of volts while the voltage drop to the smaller electrode is several hundreds or even a thousand volts. Therefore almost all the sputtering will occur at the target electrode, as is desired.

Not only can the target electrode be made relatively small, but the wafer electrode can also be connected to the chamber walls (which would also be grounded), thus making the effective area of that electrode even larger. For typical RF-frequency power supplies (e.g., 13.56 MHz), an impedance matching network is also required. A typical setup for RF sputter deposition is shown in Figure 9-28.

In addition to allowing sputtering of nonconducting material, there are two other advantages of using RF powered systems. In RF plasmas, electrons gain energy directly from the oscillating RF fields because of low-loss elastic collisions with neutral atoms. Thus the plasma can in fact be sustained without the need for secondary electrons from the walls and electrodes. In addition, the efficiency of ionization is increased by these oscillating electrons, allowing for operation at somewhat lower pressures.

Bias Sputtering: Sputter Etching and Bias-Sputter Deposition Sometimes sputtering of the wafer is desirable. One application would be for precleaning the wafer before the actual deposition, or "sputter etching." Another application is in "bias sputtering," where deposition and sputtering of the wafer are done simultaneously. A negative bias relative to the plasma is applied to the wafer electrode, which is now electrically isolated from the chamber walls. Positive Ar ions from the plasma will now be accelerated to the wafers on the substrate carrier and sputter off atoms. The energy of the ions (and hence the sputtering yield) can be controlled by controlling the substrate bias. Usually an RF bias is used since the wafers often have insulating films on them, while the target is given either a DC or RF bias. The target bias is in the range of -700 to -2000 V, while the substrate bias is on the order of -50 to -300 V [9.6].

In sputter etching or cleaning, no deposition is allowed to occur on the wafer by using a shutter to block sputtered material from the target. During this step, a controlled thickness of surface material is sputtered off the wafer, removing any contaminants or native oxides. A film can then be sputter deposited immediately afterward without breaking the vacuum, inhibiting recontamination or reoxidation. Redeposition of contaminants due to sputtering of chamber walls and other parts of the wafer can occur during sputter etching and may be a concern. Charging damage can also be a problem.

In bias-sputter deposition, or sometimes just “bias sputtering,” sputter etching of the wafer and deposition on the wafer (by sputtering the target) are allowed to occur concurrently. Conditions are chosen so that more deposition occurs than etching. But by allowing some sputtering of the wafer surface to occur during the deposition, both the topography and the properties of the deposited film can be altered. The most common application is SiO_2 , called “bias-sputtered quartz” when deposited this way. By applying bias to the substrate, sharp edges of the SiO_2 due to the underlying structure are reduced during the sputter deposition. This allows for a more planarized film as well as for better filling between metal lines. The common reason given for these results is that the sputtering of the wafer is a function of the geometry of the surface features, with pointed and sloped features being sputtered, and removed, more easily than horizontal or vertical surfaces. Not only does this help planarize the film, but the overhang that can develop when depositing into a trench or hole is preferentially sputtered away, allowing for better filling of the hole, as illustrated in Figure 9–29. Another factor may be that the resputtered atoms allow for some redeposition on sidewalls. We will demonstrate these ideas with simulations of high-density plasma CVD and ionized PVD depositions which make use of these same phenomena later in this chapter. Another technique for filling spaces based on these same ideas is to do the deposition, usually by PECVD, and sputter etching sequentially, as opposed to simultaneously. This can be done in a cluster tool with separate chambers for each process. The wafer is physically moved back and forth between the two.

Because of the energetic ions bombarding the film surface during bias-sputter deposition, the properties of the films being deposited can also be altered by changing the substrate bias. These properties include resistivity, residual stress, density, etch rate, and dielectric constant. SiO_2 that is deposited by regular sputtering (no substrate bias) is very porous and unsuitable for use in microelectronics, while SiO_2 that is deposited using bias sputtering is much denser and has been used for intermetal dielectric applications.

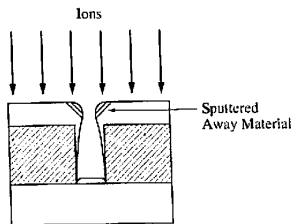


Figure 9–29 Illustration of angle-dependent sputtering which removes nonplanar features in bias-sputter deposition. Here the angled surfaces of the overhang are preferentially sputtered by the directed ions, allowing for better filling of the hole.

However, this method for depositing SiO_2 also has shortcomings. First, because of the combined deposition and etching of the wafer, the net deposition rate is low. Also, the SiO_2 will deposit on the chamber walls and other metal surfaces in the chamber and can flake off and cause contamination and particulate problems. For these reasons bias sputtering is not commonly used today. A method which combines bias sputtering with CVD deposition methods, the High-Density Plasma (HDP) CVD method described earlier, has recently been developed and does not suffer from these problems.

Magnetron Sputter Deposition In both conventional DC and RF sputtering, the efficiency of ionization from energetic collisions between the electrons and gas atoms is rather low. Most electrons lose their energy in non-ionizing collisions or are collected by the anode, and only a small percentage of them take part in ionization processes with the Ar atoms. In RF plasmas electrons gain energy directly from the oscillating RF fields, increasing the ionization efficiency somewhat. But the overall ionization efficiency is still fairly low. As a result, the deposition rates are usually rather low. Few Ar atoms are converted to positive ions and few positive ions strike the target.

In the technique known as magnetron sputtering, magnets are used to increase the percentage of electrons that take part in ionization events, and the ionization efficiency is increased significantly. A magnetic field is applied at right angles to the electric field, usually by placing large rectangular magnets behind the target. This traps the electrons near the target surface, and causes them to move in a spiral motion until they collide with an Ar atom. The ionization and sputtering efficiencies are increased significantly, and deposition rates of up to one μm per minute can be achieved—about 10 to 100 times faster than without the magnetron configuration. Unintentional wafer heating is also significantly reduced, since the dense plasma is confined to near the target and the ion loss to the wafers is less. In addition, a lower Ar pressure can be utilized since the ionization efficiency is so much larger. This can lead to better film quality since less argon will be incorporated in the film. Magnetron sputtering can be done in either DC or RF modes, but the former is more common. At these higher ion currents, cooling of the target can be a problem for RF systems which have insulating targets.

Collimated Sputter Deposition and Ionized Sputter Deposition A small range of arrival angles during deposition can cause nonuniform, or poor, step coverage over a step in topography. However, sometimes a small range of arrival angles is desirable. If material is required to be deposited into of a deep contact or via, as when filling it with Al or depositing Ti/TiN contact and barrier layers, a large arrival angle distribution can cause problems. As illustrated in Figure 9-30(a), a relatively large arrival angle distribution, when little surface diffusion or reemission occurs, can result in little deposition at the bottom of a hole due to shadowing effects. In addition, overhang formation can occur at the top corners of a deep hole, enhancing the shadowing effect. All this leads to poor coverage at the bottom of the hole or incomplete filling.

One way to improve this is by having a narrow range of arrival angles, with most of the depositing atoms arriving at the wafer perpendicular to the surface. Evaporation from a stationary small source would result in a narrow range of arrival angles compared to sputtering due to geometric arguments (Figure 9-26), as well as due to the fewer gas-phase collisions, but sputtering is preferred due to its ability to deposit uniform alloys and compounds and its in situ sputter etch capability. While standard sput-

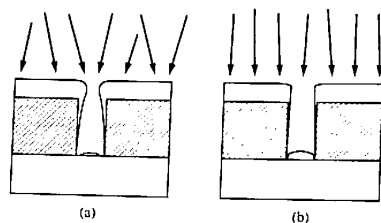


Figure 9-30 Effect of arrival angle distribution of depositing species on filling of trenches or holes in a PVD system. In (a) a relatively wide arrival angle distribution leads to poor bottom filling or coverage, while in (b) a narrower arrival angle distribution leads to better bottom filling. The higher the aspect ratio of the feature, the narrower the arrival angle distribution must be for adequate bottom coverage.

ter deposition equipment may adequately cover the bottoms of holes with low aspect ratios (less than 1 or so), its arrival angle distribution is too large for holes with high aspect ratios. Using a smaller sputtering target would result in a narrower arrival angle distribution at the wafer, with a single angle coming from a point source in the extreme case. But that angle would vary across the wafer. This would result in uneven asymmetric deposition, depositing more on one sidewall of the hole or the other for much of the wafer. (A small evaporation source would have the same problem.) It would be better to somehow filter or focus the incoming atoms in a sputtering system so that a narrow range of angles, all nearly perpendicular to the wafer surface, arrived at the wafer. Better filling of a hole, and more uniform directionality across a wafer, could thus be obtained as illustrated in Figure 9-30(b). Two methods to do this have been developed: collimated sputtering and ionized sputtering.

In collimated sputtering, a plate of circular or hexagonal holes, or cells, is placed between the target and the wafer. The height to width aspect ratio of these cells can range from 1:1 to 3:1. As the sputtered atoms travel through the collimator toward the wafer, only those with a nearly normal incidence trajectory will continue on to strike the wafer. Those at more oblique angles will strike the walls of the collimator, depositing there. The collimator thus acts as a physical filter to low-angle sputtered atoms. A problem with this method is that a significant percentage of the sputtered atoms are filtered out, up to 95 percent, and never make it to the wafers. Thus the deposition rate is significantly reduced and the collimator must be cleaned regularly. As a result, collimated sputtering is rarely used for completely filling contacts or vias where a lot of material needs to be deposited but is more suited for contact and barrier metal deposition where the amounts of material deposited are much less. It does do a fairly good job of covering the bottoms of contacts and vias with contact and barrier metals. Some sidewall coverage, important for via liners, is also obtained from the flux that is slightly off-normal, but the sidewall film is often less dense. In addition, uniformity across a wafer or from

wafer to wafer may be affected by shadowing by the collimator.

Another type of sputtering is called “long throw collimation,” or “long throw sputtering.” No collimation plate is actually used. Instead, a limited size target, about the size of the wafer in a single-wafer system, is used and the distance between the target and wafer is increased (up to about 30 cm compared to about 5 cm in more standard systems), decreasing the distribution of arrival angles at the wafers. In effect the sputtering chamber acts as one collimator cell, with the high-angle atoms not reaching the wafer but striking the chamber wall instead. However, as the spacing is increased, the deposition rate, the target utilization, and the uniformity all decrease; in addition, the chance of gas-phase collision and scattering increases. Low pressures are usually utilized in these systems to minimize gas-phase collisions. Because of the limited target size and target-to-wafer spacing, asymmetric depositions in trenches and vias can occur at the edges of wafers in these systems [9.13].

In ionized sputtering a narrow arrival angle distribution, perpendicular to the surface, is obtained as in collimated sputtering. However, most of the sputtered material reaches the wafers, giving it a significant advantage over collimated sputtering. In this technique, the sputtered atoms—mostly neutral as they leave the target—are ionized as they travel between the target and the wafers. An inductively coupled RF antenna is placed around the plasma region which ionizes the sputtered atoms through collisions with electrons in the plasma, as illustrated in Figure 9-31. Fifty to eighty-five percent of the atoms are ionized. The ionized atoms, Al^+ or Ti^+ , for example, will be accelerated and directed across the wafer sheath, striking the wafers in a nearly perpendicular direction. A DC or RF bias is usually applied to the wafer holder, similar to bias sputtering. This will further accelerate these ions to the wafer surface, resulting in a highly directional flux. In addition, the energy of the incoming ions (both Ar and the deposited element) can be tightly controlled, and the deposition behavior and film properties can

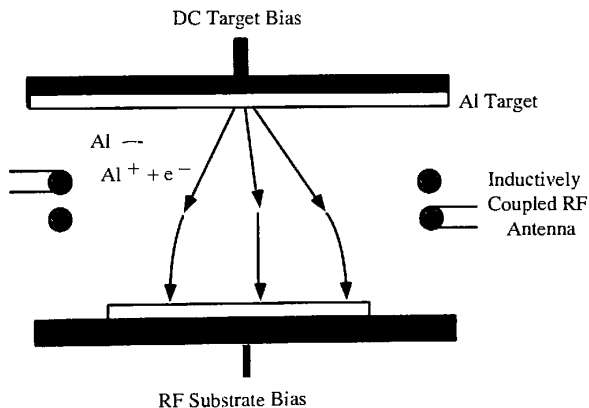


Figure 9-31 Schematic diagram of ionized sputter deposition system (or ionized PVD), showing atomic flux lines. (After [9.15].)

be altered, just as in bias sputtering. The highly directional ionized atoms and angle-dependent sputtering can result in very good coverage at the bottom of deep vias or contacts. However, the sidewall coverage may be poor. By increasing the ion energy, the resputtering and redeposition of the atoms can be increased, resulting in better sidewall coverage [9.14], often still thinner than the bottom, but at least continuous. Beveling of the top corners due to angle-dependant sputtering may also occur.

Ionized PVD (IPVD) is also referred to as "ionized metal plasma" (IMP) deposition. It is also sometimes called "high density plasma sputtering," analogous to high density plasma CVD, because of the higher density plasma produced which ionizes the metal atoms. With the high density plasma produced, much more efficient deposition occurs compared to bias-sputtering or collimated sputtering, and IPVD is beginning to be more commonly used in chip production. The main problems with ionized PVD have to do with uniformity and manufacturing issues. The ionizing coils are usually located inside the chamber, and not remotely as in HDP CVD and etching systems. (Otherwise metal deposited on the chamber windows would block the signal from a remote RF source.) Having the coils inside the chamber leads to deposition on the coils and other associated problems. Another type of IPVD system utilizes a hollow cathode magnetron source rather than RF coils.

High-Temperature or Hot-Sputter Deposition Another method used to fill spaces during deposition as well as to improve overall coverage is high-temperature, or hot-, sputtering. This will be discussed further in Chapter 11 since it is most often used for Al deposition. The basic idea is to heat the substrate to 450°C or higher during deposition, usually by a wafer backside gas heating system. Surface diffusion is significantly increased so that filling in spaces, smoothing edges, and planarization are accomplished, all driven by surface energy reduction. This can be considered as *in situ* reflow. Usually a thin "cold" deposition is done first with the substrate at room temperature, which serves as a nucleation-wetting layer. CVD deposition of Al may also be used for this. This is then followed by the hot-PVD deposition. High-temperature sputter deposition is still in the development stage. One of the main drawbacks is the relatively high temperature required. Excessive reaction between the hot Al and the liner can occur. The high temperature can also adversely affect other materials and layers already present on the wafer at this stage in the process.

Even more recently developed is high-pressure sputter deposition. In this method, called the aluminum extrusion fill, or the "ForceFill" [9.16] technique, the aluminum is first sputter deposited, which covers and seals the holes. Then the chamber is pressurized with argon to up to 700 atmospheres while the wafer is heated to about 400°C. This forces the aluminum to completely fill any hole and does so quite effectively. Applying the high argon pressure after the deposition minimizes any argon incorporation in the film. The high temperature and specialized equipment are drawbacks to this technique.

9.3 Manufacturing Methods

The different types of equipment used for the deposition of thin films in microelectronics were discussed in the previous section. In this section the techniques used today to deposit specific thin films will be briefly described. Table 9-1 provides a summary.

Table 9-1 Common deposition methods for thin films in integrated circuit fabrication

Thin Film	Equipment	Typical Reactions	Comments
Epitaxial silicon	APCVD, LPCVD	$\text{SiH}_4 \rightarrow \text{Si} + 2\text{H}_2$ $\text{SiCl}_4 + 2\text{H}_2 \rightarrow \text{Si} + 4\text{HCl}$ Also SiHCl_3 , SiH_2Cl_2	1000–1250°C. Reduce pressure for lower-temperature deposition.
Polysilicon	LPCVD	Same as epitaxial Si	575–650°C. Grain structure depends on deposition conditions and doping.
Si_3N_4	LPCVD, PECVD	$3\text{SiH}_4 + \text{NH}_3 \rightarrow \text{Si}_3\text{N}_4 + 12\text{H}_2$	650–800°C for oxidation mask. 200–400°C (PECVD) for passivation.
SiO_2	LPCVD, PECVD, HDPCVD, APCVD	$\text{SiH}_4 + \text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2$ $\text{Si}(\text{OC}_2\text{H}_5)_4 (+ \text{O}_3) \rightarrow \text{SiO}_2 + \text{byproducts}$	200–800°C. 200–500°C (LTO)—may require high T anneal. 25–400°C (TEOS-ozone, PECVD, HDPCVD).
Al	Magnetron sputter deposition		25–300°C (standard deposition). 440–550°C (hot Al for in situ reflow). CVD difficult for alloys (Al-Cu-Si).
Ti and Ti-W	Magnetron sputter deposition (standard, ionized, or collimated)		CVD difficult. Nitrogen can be added to Ti-W to stuff grain boundaries.
W	LPCVD	$2\text{WF}_6 + 3\text{SiH}_4 \rightarrow 2\text{W} + 3\text{SiF}_4 + 6\text{H}_2$ $\text{WF}_6 + 3\text{H}_2 \rightarrow \text{W} + 6\text{HF}$	250–500°C. Blanket deposition with two-step process using both reactions is common.
TiSi_2	Sputter and surface reaction Cosputtering or CVD	Ti(sputtered) + Si(exposed) $\rightarrow \text{TiSi}_2$	Sputter/reaction gives self-aligned silicide.
TiN	Reactive sputter deposition CVD	Ti + N_2 (in plasma) $\rightarrow \text{TiN}$ $6\text{TiCl}_4 + 8\text{NH}_3 \rightarrow 6\text{TiN} + 24\text{HCl} + \text{N}_2$	Organometallic source possible for MOCVD deposition. TiN can also be formed in TiSi_2 process.
Cu	Electroplating, electroless sputtering, CVD	$\text{Cu}^{2+} + 2\text{e}^- \rightarrow \text{Cu}$	Electroplating is most common method today.

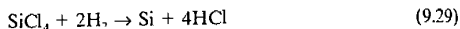
9.3.1 Epitaxial Silicon Deposition

“Epitaxy” is from the Greek for “arranges upon.” Epitaxial silicon is silicon that is deposited on top of single-crystal material (usually the silicon substrate) and “arranges upon” and takes the crystalline form of the substrate. If no underlying single-crystal silicon is present, or if the deposition conditions are not right, then either amorphous or polycrystalline silicon will result. An example of an epitaxial layer is shown in Figure 2-20 in the CMOS process in Chapter 2.

Epitaxial silicon is deposited by CVD by either decomposing silane via the reaction



or by the reaction of a silicon chloride compound with hydrogen. Silicon tetrachloride, SiCl_4 , is often used, and the overall reaction is



Two other chlorides, SiHCl_3 and SiH_2Cl_2 , are also sometimes used as the source species, with similar reactions. Eq. (9.29) is the overall, or net, reaction for the tetrachloride source. Actually many intermediate and competing reactions occur, and the process is quite complex. The silicon chloride species are all liquids at room temperature and are delivered to the chamber using hydrogen as a carrier gas. Using silicon chlorides rather than silane as sources of silicon is beneficial in that chlorine removes metal contaminants from the deposited silicon film.

Cold-walled CVD systems have traditionally been used for epitaxial silicon, either in the horizontal tube configuration shown earlier in Figure 9-4(a), or in a “pancake” configuration, with a round vertical wafer holder with source gases coming from above. Systems using quartz halogen radiant heating are also used. In each of these systems the wafers are placed side by side (end-to-end) to ensure uniform depositions.

For conventional epitaxial growth, relatively high deposition temperatures of 1000–1250°C are used, the lower temperature for silane and the higher for SiCl_4 . These high temperatures are used so that single-crystal layers are readily obtained. High temperatures are required to desorb any native oxide on the surface which would inhibit single-crystal growth. At temperatures over 1000°C at 1 atmosphere, thin silicon dioxide films that may form during the deposition become unstable and vaporize. Also at the higher temperatures the surface reactions for deposition are enhanced, especially the surface migration to kink sites and rearrangement of the atoms, allowing for single-crystal growth. It is important that the atoms on the surface have time to attach at the proper sites before too many more atoms arrive. If the deposition rate is too fast relative to the surface processes, polycrystalline material results. Because of the high temperatures and fast surface reactions, epitaxial growth in these systems usually occurs in the mass transfer limited regime. Therefore thickness uniformity requirements preclude the wafers from being stacked close together. Recently lower temperature systems have been developed, as will be described later in this section.

Conventional Si epitaxial deposition is done either at atmospheric pressure, or at somewhat reduced pressures of 80–200 torr, but still above normal LPCVD pressures

which are in the 0.5–2 torr range. Reducing the pressure gives better control over switching on and off dopant source flows and reduces autodoping so that more abrupt doping profiles may be obtained. This works well for arsenic and phosphorus, but not as well for boron. Decreasing the pressure too much may cause the deposition rate to be quite low, by perhaps limiting the reactant species, especially for the chlorine-based sources. In earlier years when most epitaxial deposition was done using these conventional systems, the epitaxial layers were up to 20 μm thick and higher deposition rates were required.

A clean surface is critical to obtaining epitaxial growth. If contaminants or native oxide films are present, single-crystal growth will not occur. Wafers must be carefully cleaned prior to loading into the system. In addition, after the wafers are loaded into the deposition chamber, they may be subjected to an in situ HCl gas clean. This will etch the top surface of the silicon and remove any contaminants or native oxide.

Doping of epitaxial Si films can be done during the deposition by introducing appropriate gas sources of the dopants into the chamber along with the silicon source gases, as discussed earlier in the section on APCVD. Common dopant source gases are arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6). Unintentional doping due to diffusion and autodoping can be a problem, as discussed earlier. Reducing the pressure and increasing the temperature seem to reduce the problem for arsenic and phosphorus, while increasing the pressure and reducing the temperatures work better for boron. Using antimony instead of arsenic or phosphorus helps due to antimony's low vapor pressure compared to arsenic and low diffusivity compared to phosphorus.

Selective deposition of epitaxial silicon is sometimes desired. Selective deposition means that the silicon deposits on exposed regions of silicon, but not on other films, such as silicon dioxide. Under non-selective conditions, epitaxial silicon would be deposited over the exposed silicon, and amorphous or polycrystalline silicon would be deposited over the other regions. But by using the right conditions, one can actually achieve selective deposition. Conditions are chosen to reduce the nucleation and growth rate on the less-favored nonsilicon regions and to enhance the surface migration of the silicon atoms. Chlorides in the source gas tend to inhibit nucleation on the oxide by Cl etching of Si and by increasing the surface migration of Si atoms. Therefore using SiCl_4 , SiHCl_3 , or SiH_2Cl_2 increases the selectivity over using SiH_4 . HCl gas is often added to improve selectivity. Reducing the pressure also helps.

Recently the need for low-temperature epitaxial growth of Si has greatly increased. In advanced process technologies, minimizing the thermal processing that the wafer is exposed to is crucial in order to decrease dopant diffusion, control abrupt interfaces and junctions, and reduce damage to device structures. Reducing the growth temperature from 1000°C or higher to temperatures in the 500–900°C range is of great benefit in these regards. Recently techniques have been developed to provide good quality epitaxial silicon films at low temperatures. The key to this is to provide an ultraclean environment so that oxides and other contaminants that would inhibit single-crystal growth are not present on the wafer surface. This can be achieved by using high-vacuum compatible CVD systems, ultrapure gases, and utilizing low pressures. A good predeposition clean is critical, and the wafers can also be initially heated to high temperatures in vacuum or H_2 in the chamber prior to the de-

position to desorb any native oxide. The low deposition temperatures and pressures usually place the deposition in the surface reaction limited regime. The deposition rates in these systems may be lower than in conventional epitaxial systems, but the layers that are grown today are generally much thinner than in older technologies.

Another technique for low-temperature Si epitaxy is Molecular Beam Epitaxy (MBE). Used for many years to produce III-V epilayers, this method is now being used more frequently for Si and SiGe films. In this technique, the silicon and dopant atoms are evaporated under ultrahigh vacuum conditions onto the Si substrate. Low-temperature epitaxial deposition, in the 500–800°C range, is obtainable, with ultraabrupt interfaces and doping profiles.

9.3.2 Polycrystalline Silicon Deposition

Polycrystalline silicon, unlike epitaxial silicon, can be deposited on arbitrary substrates and does not need exposed silicon underneath. It is used for the gate electrode in CMOS technology and for local interconnects and resistors. Its good thermal stability, good interface to silicon dioxide, good conformality, and ease of deposition and processing have made it a mainstay of silicon microelectronics technology. Polycrystalline silicon, as its name implies, is made up of regions or “grains” of crystalline material, but with each grain oriented differently from the neighboring grain. Separating each grain from the next one is a region called, appropriately, the grain boundary. Polycrystalline structures, grains, and grain boundaries will be discussed in detail in Chapter 11.

Polycrystalline silicon, or polysilicon, can be sputter deposited; however, CVD deposition is the usual method today. This is mainly due to two reasons, one historical and one technological. As we described above, epitaxial silicon must be deposited at relatively high temperatures in order for single-crystal material to be obtained. The low temperatures of sputtering (room temperature up to a maximum of about 400°C) preclude its use in depositing epitaxial film. Once CVD methods were developed for epi, it was a simple matter to use this method for polysilicon, where the conditions are much less restrictive. The other reason that CVD is used for polysilicon rather than sputtering is due to step coverage issues. In actual devices and circuits, polysilicon must often be deposited across steep topography. (See Figure 2–25 for example.) Good step coverage is important, and as we have mentioned many times, CVD usually gives better step coverage than sputtering. The conformality of LPCVD polysilicon is extremely good, with a sticking coefficient of 0.001 or less.

Silane is most often used for polysilicon deposition, following the decomposition reaction in Eq. (9.28). Silane is used rather than the silicon chlorides because better coverage over amorphous materials such as SiO_2 is obtained. This may be related to enhanced surface mobility in the presence of the chlorides or to the chlorine etching of Si over an insulator, leading to better nucleation and growth on the silicon and worse on the oxide. What is better for selective epitaxy is what is worse for depositing polysilicon over oxide layers, and vice versa. In addition, moderately low temperatures are usually used for polysilicon deposition, and silane decomposition occurs at lower temperatures than with the chloride reactions. If a dilutant or carrier gas is used, it is commonly nitrogen or hydrogen. However, using hydrogen increases the reverse reaction of

Eq. (9.28). and can decrease the deposition rate at lower temperatures. This decrease in the surface reaction shifts the deposition rate versus $1/T$ curve to the left compared to when nitrogen or no carrier gas is used, as shown in Figure 9-8 [9.17].

In order to achieve good throughput with uniform films, LPCVD systems with stacked wafers are used for polysilicon. Such a system was shown schematically in Figure 9-4(b). The deposition pressure is generally in the 0.2 to 1 torr range. A dilutant or carrier gas may or may not be used with the silane. By using a low pressure at the relatively low deposition temperatures described below, one ensures that the deposition occurs in the surface reaction limited regime while maintaining adequate deposition rates. The use of low pressure has been seen to greatly improve the uniformity of polysilicon depositions in closely stacked wafer systems at these moderate temperatures. Using lower pressure also results in fewer gas-phase reactions and less gas usage.

The deposition temperature for polysilicon films is moderately low, usually in the 575 to 650°C range. One reason for these low temperatures is to minimize the thermal budget of the wafer, reducing dopant diffusion and material degradation. Also, fewer gas phase reactions occur at the lower temperatures, leading to smoother and better adhering films, and silane depletion is less at lower temperatures. Another reason is that small grains are usually desired for polysilicon films. This is because films with finer grains are easier to mask and etch to give smooth and uniform edges. Deposition of polysilicon at higher temperatures usually results in larger grained films. For temperatures less than 575°C or so, the deposition rate is too low.

The deposition rate depends on the temperature, the silane partial pressure, and the presence of hydrogen or doping gases. At 620°C with a partial pressure of silane (with nitrogen carrier gas) of 0.5 torr, the deposition rate is about 15 nm per minute. According to our simple model for deposition, the deposition rate should be a linear function of the silane partial pressure, as given in Eq. (9.10). But the deposition rate for polysilicon has been found to be a sublinear function of the silane partial pressure, saturating as the silane is increased. This may be due to the desorption of hydrogen from the surface, which could be the limiting step in the deposition mechanism.

The as-deposited grain structure depends on these same factors: the temperature, the silane partial pressure, and the presence of other gases or species [9.17, 9.18]. Usually the polysilicon film is columnar in structure, meaning that most grains span the entire height of the film, with the grains quite tall and narrow in some cases. The average size at the top of the grains is often a little larger than at the bottom. Figure 9-32 shows several TEM cross-section images of deposited polysilicon films with the grain structure evident [9.19]. Polycrystalline films often exhibit a certain crystallographic texture. Crystallographic texture means that while the in-plane orientation of the columnar grains is more or less random, there is a preferred orientation of the grains in the vertical direction or out-of-plane orientation. For example in $\langle 110 \rangle$ textured films, which are common for typical polysilicon deposition conditions, most of the grains would have their $\langle 110 \rangle$ orientation pointing upward, or close to it. (And a more textured film means that a higher percent of the grains are at, or close to, a single orientation.) These orientations are a function of the deposition conditions, particularly deposition temperature and partial pressure of silane. If the deposition temperature is below 570–600°C, depending on the silane partial pressure, the as-deposited film will not be polycrystalline, but amorphous. Annealing the film at 600–1000°C will convert the struc-

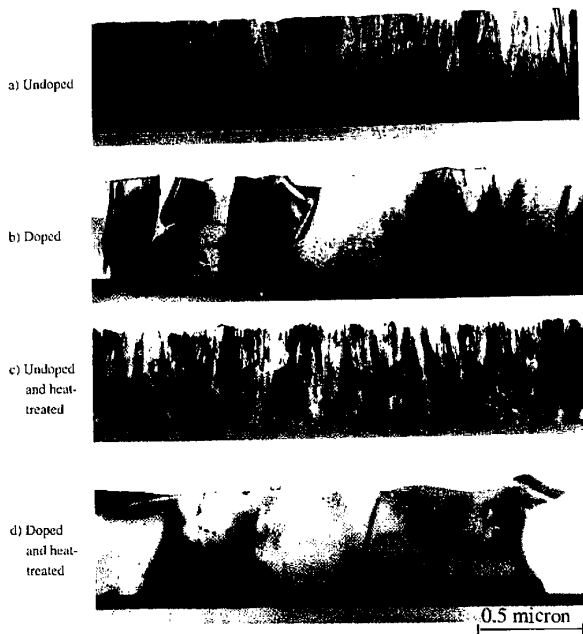


Figure 9-32 TEM cross sections of CVD polycrystalline films deposited at 625°C: (a) as-deposited, undoped film, showing the thin grains in a columnar structure; (b) as-deposited phosphorus-doped film, showing much larger grain size; (c) annealed (1000°C), undoped film, showing little grain growth as compared to (a); (d) annealed (1000°C), phosphorus-doped film, showing evidence of grain growth as compared to (b). Reprinted with permission of the Electrochemical Society [9,19].

ture to polycrystalline material.

The average diameter of the columnar grains, the grain size, also depends on the deposition conditions and can range from 0.01 to 0.4 μm , with a typical size of about 0.05–0.1 μm for undoped polysilicon deposited at 625°C. Higher deposition temperatures generally result in larger grain sizes, as do thicker film depositions. High doping with phosphorus or arsenic (greater than about $3 \times 10^{20} \text{ cm}^{-3}$ for P and greater than about $1 \times 10^{19} \text{ cm}^{-3}$ for As) can greatly increase the grain size of as-deposited films, as shown in Figure 9-32. During any subsequent high-temperature annealing, the average grain size will increase only slightly for as-deposited undoped polycrystalline silicon when annealed at 1000°C (more grain growth occurs at higher temperatures above 1200°C). For

as-deposited amorphous silicon, the crystalline grains will nucleate and grow much more, even exceeding the size of the grains of annealed polycrystalline films. They also end up being somewhat less columnar, with some grains nucleating and growing above and below each other. The final size in either case will depend on deposition and anneal temperatures, anneal time, dopants, impurities, and film thickness. For example, phosphorus and arsenic doping will significantly increase the grain growth, while boron does not. This may be related to the increased vacancy concentrations with n-doping [9.20]. Grain growth and the dependence on film thickness will be discussed in more detail in Chapter 11.

Polysilicon can be doped either n type or p type. This can be done either during the deposition or afterward by implantation or diffusion. Doping during deposition is done by adding dopant source gases, such as those used for epitaxial growth, into the chamber along with the silane. However, the addition of arsine (AsH_3) or phosphine (PH_3) greatly decreases the deposition rates, by competing for sites perhaps, as well as creates thickness variations in the films. For these reasons, n-type doping of polysilicon is usually done after the deposition by implantation or diffusion. Diborane (B_2H_6) on the other hand, increases the deposition rate slightly and therefore can be used to dope the film during the deposition. As and P dopant atoms in polysilicon tend to segregate to the grain boundary regions. Here they are electrically inactive, so that only the dopants that remain in the grain interiors contribute to the doping of the film. B does not appear to segregate to grain boundaries. In addition, the grain boundaries can trap carriers regardless of the dopant, also reducing the net doping. The carrier mobility is also decreased by the grain boundaries. As a result, polysilicon must be doped at higher than normal concentrations in order to achieve desired doping and resistivity levels. Figure 9-33 extends the resistivity data we discussed in Chapter 1 (Figure 1-18) to the polysilicon case and shows the resistivity of annealed films as a function of phosphorus doping concentration. One can see that the resistivity for polycrystalline material is higher than single-crystal silicon for the same total doping level. This difference gets smaller at higher doping concentrations, presumably because the finite number of traps at the grain boundaries get saturated above some critical concentration (about $1 \times 10^{17} \text{ cm}^{-3}$ here). One sees that for samples deposited at 580°C (indicated by the LP-580 label, for LPCVD deposition at 580°C), the critical doping concentration is less. This is consistent with a larger grain size for annealed amorphous films, as discussed above. Larger grains means less grain boundary area and thus fewer segregated dopant atoms and fewer carrier traps.

9.3.3 Silicon Nitride Deposition

Silicon nitride, Si_3N_4 , films are used primarily for two purposes. One is for a mask against oxidation. In earlier chapters we saw how Local Oxidation of Silicon (LOCOS) occurs using silicon nitride as a mask against oxidation (Figure 2-5). The diffusion of oxygen through the film is very slow, preventing the underlying silicon from being oxidized. The other main use of silicon nitride is as a final passivation layer on chips. Silicon nitride is a very good barrier against water and sodium diffusion (unlike silicon dioxide) and it can be deposited conformally, making it an excellent final passivation

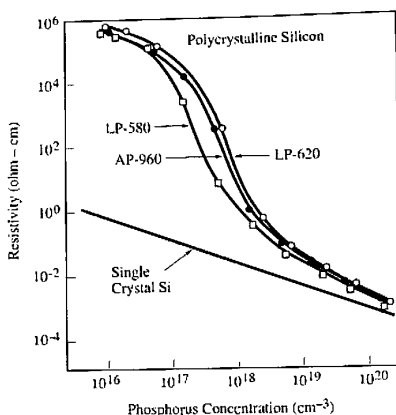
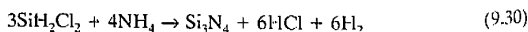


Figure 9-33 Resistivity of annealed phosphorus-implanted silicon films as a function of phosphorus concentration. LP refers to LPCVD deposition, AP refers to APCVD deposition, and the numbers in the labels refer to the deposition temperature. (After [9.21].)

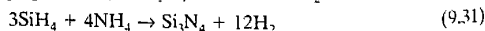
layer for silicon microelectronics. It is generally not used in direct contact with silicon because of its relatively poor interface properties, especially with regard to fixed or interface trapped charges, and its high stress.

The two uses of silicon nitride require different deposition techniques as dictated by the allowable deposition temperatures. The masking nitride film is deposited relatively early in the fabrication process, and the temperature can be high. However, the passivating nitride is deposited after the metallization is in place, and therefore the deposition and processing temperature is limited to 450°C.

Masking silicon nitride films for LOCOS processes are deposited by CVD in the 650–800°C range. Low-pressure CVD is frequently used for good uniformity and high wafer throughput. The source gasses are ammonia and either dichlorosilane or silane, following the reactions



or



To ensure the proper stoichiometry of the film, an excess of ammonia is used often by as much as 20-to-1 over the chloride or silane. Nitrogen can also be used in place of the ammonia. Significant amounts of hydrogen, up to 8 atomic percent, are incorporated into these films, bonded to both the silicon and nitrogen. This can affect the film properties, such as density and etch rate.

Silicon nitride films used for final passivation films must be deposited at tempera-

tures of 450°C or below, as mentioned above. Because of this, PECVD is the common technique employed here. Silane and ammonia are used, following Eq. (9.31), since silane works better at the lower temperatures than do the chlorides. Typical PECVD nitride deposition temperatures range from 200 to 400°C. The film that is deposited at these low temperatures using PECVD is often not stoichiometric Si_3N_4 and is often referred to as just SiN or Si_xN_y . In fact, even more hydrogen than in the high-temperature LPCVD films is incorporated, up to 25 atomic percent, and the film is often designated by $\text{Si}_x\text{N}_y\text{H}_z$. Oxygen can also be present in the film. The deposition rate and the film properties are very dependent on the deposition conditions, including the RF power and frequency, the gas flow, and the chamber pressure. The various film properties, such as composition, stress, dielectric constant, thermal expansion coefficient, and density, can be optimized, or at least traded off between each other, by varying the various deposition conditions.

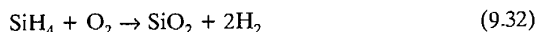
To reduce the Si/SiO_2 interface trapped charge density in the active devices, an anneal is done in a hydrogen ambient near the end of the fabrication process. Because a nitride film is a barrier to hydrogen diffusion, the hydrogen anneal must be done prior to the deposition of the passivation nitride film.

9.3.4 Silicon Dioxide Deposition

Deposited films of silicon dioxide are utilized throughout silicon microelectronic structures as described in the CMOS process flow in Chapter 2 (see, for example, Figures 2–29 and 2–38). They are used for dielectric layers to electrically separate conducting regions as well as for masks against implantation and diffusion. The requirements for the films will be discussed in more detail in Chapter 11, but obtaining films of good quality in terms of density, purity, and thermal stability, as well as good step coverage and filling properties are of major concern.

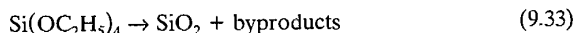
Most oxide films today are deposited at relatively low temperatures, from 200°C to 500°C using CVD techniques. Sputtered oxide films have worse coverage and more particulates (mostly due to flaking off of the metal chamber walls) than CVD oxides. Atmospheric systems have been used for many years, but today LPCVD and PECVD systems have become more commonly used.

Silane and oxygen are common source gases for CVD silicon dioxide deposition, following the net reaction



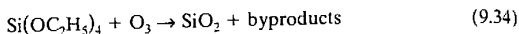
Oxidants other than O_2 are sometimes used, such as N_2O , NO , and CO_2 . When non-plasma, silane-based deposition of SiO_2 is done at temperatures less than 500°C, it is commonly referred to as Low-Temperature Oxide, or “LTO,” deposition.

Other sources of silicon are also used. One such source is tetraethoxysilane— $\text{Si}(\text{OC}_2\text{H}_5)_4$ —or “TEOS.” It can form silicon dioxide by decomposing via the reaction



(O_2 is often added as an extra oxidizer to reduce the carbon content in the film.) This

deposition occurs at higher temperatures, in the 650–800°C range, but because of the low sticking coefficient of TEOS compared to silane, more conformal films are produced. The deposition temperature can be lowered to below 400°C by either utilizing plasma enhanced deposition, or by including ozone, O_3 , in the chamber:



Because at low pressures (<150 torr) the deposition using TEOS and O_3 is limited by O_3 and gives worse film properties, a so-called Sub-Atmospheric CVD (SACVD) system has been developed for doped premetal films and undoped intermetal films [9.22, 9.23]. The deposition pressure in this system is at higher pressures (200–400 torr) and the deposition temperature is 300–550°C. Oxide films with good gap fill and reflow ability can be obtained with this system.

The major drawbacks to using TEOS are that it is a more difficult source to use (it is an organometallic liquid), the relatively high levels of carbon in the deposited film (from the TEOS itself), and the more porous films that are sometimes formed. But TEOS is still commonly used due to the conformality of the films it produces.

Silicon dioxide films deposited using LPCVD or APCVD at low temperatures (less than 500°C) can be nonstoichiometric and are often porous. They can also exhibit inadequate step coverage or filling. The films that are deposited before aluminum is on the wafer are usually annealed at high temperatures after their deposition. This densifies the films, as well as allowing them to flow, or “reflow,” in order to achieve better filling and step coverage, as discussed in Chapter 11. This is done at temperatures of 1000–1100°C for undoped films, and as low as 800°C for doped (with B and/or P) films.

For films that are deposited after aluminum is already on the wafer, other measures must be taken. No high temperature densification anneal can be done, so the films must be of adequate quality and have good coverage “as-deposited.” TEOS with ozone and plasma enhanced CVD TEOS are commonly used. While PECVD TEOS gives dense films and fairly good coverage and filling at low temperatures, the films usually contain a lot of hydrogen or nitrogen, up to 10 atomic percent, which can affect their properties. TEOS/ O_3 films may be more conformal and have better filling ability, but are also much more porous and can absorb moisture.

Another method of silicon dioxide deposition, bias sputtering, has been developed to improve filling and planarizing. But because of its low deposition rate and a high density of particulates, this method is not commonly used today. A relatively new technique called High-Density Plasma (HDP) CVD, which combines PECVD with bias sputtering, is being used for low-temperature deposition (150°C down to room temperature). Figure 9–34 shows examples of SiO_2 deposited by HDPCVD, displaying excellent filling and good planarizing properties. In this method, both silane- and TEOS-based chemistries can be used, giving good quality films in high aspect ratio structures.

Deposited silicon dioxide films are often doped with phosphorus or boron, or both. Phosphorus is used to reduce stress and enhance Na gettering, while the boron allows more reflow at lower temperatures, as will be discussed in Chapter 11. Doping can be done by using hydrides for the silane case and organometallic compounds for the TEOS case in APCVD, LPCVD and PECVD configurations.

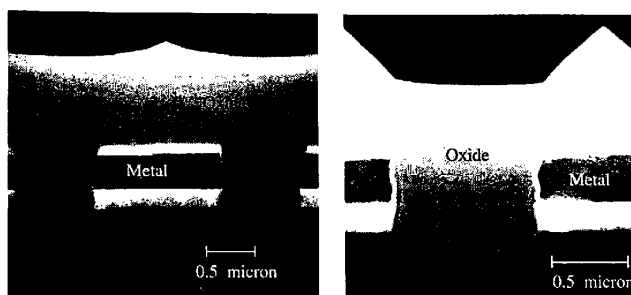


Figure 9-34 SEM images of High-Density Plasma (HDP) CVD deposited SiO_2 , showing good planarizing and excellent filling properties. Photos courtesy of VLSI Technology, Inc.

9.3.5 Al Deposition

Aluminum is used as the main interconnect material in silicon microelectronics. As explained in Chapter 11, pure Al is usually not used, but an alloy of Al with 0.5–1 wt. percent Cu (to reduce hillocking, voiding, and electromigration) and sometimes with about 1 wt. percent Si (to prevent spiking and other interactions). The technique almost always used is DC magnetron sputter deposition. CVD techniques are not currently used for Al (but are being investigated) because of the problems with CVD of metal alloys, and Al is almost always used as an alloy. These problems include unwanted reactions between the precursors for the various metals, carbon contamination, rough films, low deposition rate, and inefficient use of the Al precursor which can result in a very costly deposition. Sputter deposition does not have these problems. Magnetron sputtering is used primarily due to its high deposition rate of up to one μm per minute for Al.

The argon pressure and the presence of impurities can affect the deposition and properties of the film. While a higher pressure of argon will increase the number of collisions of the Al atoms in the gas phase, decreasing the mean free path and improving the step coverage, a higher pressure of argon also means that more Ar will be incorporated in the Al film, up to 1 atomic percent. This can have deleterious effects on the film properties, including higher electrical resistivity, more stress, and smaller grain size (the argon acting as nucleation sites for the grains). Therefore lower pressures are usually desired. The presence of other gas species also affects the deposition and film properties. Oxygen, nitrogen, hydrogen, and water vapor may all be present in small amounts in the deposition chamber or on the wafer surface. These can all be incorporated into the film, having the same effects on the film properties: higher electrical resistivity, more stress, and smaller grain size. It is important to minimize the amount of these impurities

in the deposition chamber. This is done by pumping down the chamber to very low pressures before the argon is introduced (to less than 1×10^{-10} torr in some cases), by using ultrapure argon, and by heating the wafers and wafer holders before the deposition occurs to desorb any moisture.

It is common to heat the wafers to 150–300°C during Al sputter deposition, mainly to improve step coverage by increasing the surface mobility. Heating the substrates to even higher temperatures, up to 450–550°C, is called high-temperature or hot sputter deposition, or just “hot Al”. The surface diffusion (and possibly bulk viscous flow) increases significantly, allowing for filling of deep contact and via holes.

There has been quite an interest recently in CVD Al. This is because of the technique’s inherently better step coverage and excellent filling ability. There has been some limited success at depositing an Al-Cu alloy by using different Al and Cu organometallic precursors [9.24]. However, a more common approach involves depositing a film of pure Al by CVD, then doping it with Cu afterward. Al itself can be deposited by CVD using metal-organic sources, such as tri-isobutyl-Al (TIBA), or dimethylaluminumhydride (DMAH). The presence of Al-H bonds in the latter precursor results in less carbon incorporation in the Al film. To alloy or dope the film with Cu, one method is to sputter a layer of Cu, or Al-Cu alloy, on top of or below the CVD Al layer. Then a 250–400°C anneal is done in which the Cu diffuses into the pure Al layer, producing a uniform layer of Al-Cu. However, controlling this process can be difficult.

9.3.6 Ti and Ti-W Deposition

Titanium is often used as an underlayer for contacts, vias, and interconnects because of its good adhesion to other materials, its ability to reduce native oxides, and its good electrical contacting properties (as discussed in Chapter 11). It is usually deposited by sputtering, either using standard magnetron sputtering, or by using collimated or ionized sputtering for good coverage in contact or via bottoms. CVD of Ti is difficult because the precursors are so thermodynamically unstable. Impurities incorporated into the film from the precursors can also be a problem.

Ti-W alloys have been used as barrier metals in contacts and as underlayers and anti-reflective layers in interconnects, as also discussed in Chapter 11. However, TiN is more commonly used today for these purposes due to its better thin film and barrier properties and will be discussed separately later. Ti-W films are deposited using magnetron sputtering, usually from a single target. The most common target is 10 wt. percent Ti (30 atomic percent) and 90 wt. percent W (70 atomic percent). Nitrogen is commonly added to the chamber and is incorporated into the polycrystalline film, plugging or “stuffing” the grain boundaries. This reduces atomic diffusion along these high diffusion paths and improves the film’s barrier performance. Film properties, such as resistivity, stress, composition, and barrier performance, depend mainly on four deposition conditions: sputtering pressure, power, temperature, and Ar/N₂ ratio. [9.25] The films can be rather brittle, leading to cracking on the wafer and to flaking off of deposition chamber walls onto the wafers.

9.3.7 W Deposition

Tungsten is commonly used as a contact or via metal, called “tungsten plugs,” or in some cases is used as a first-level metal interconnect. (See, for example, Figure 2–41 in Chapter 2.) It is usually deposited by CVD, due to its very good filling ability and conformal coverage, in a hot-wall, low-pressure system. Temperatures range from 250–500°C, and total pressures are in the 0.1-to 2-torr range.

Tungsten can be deposited either selectively or as a blanket deposition. In selective deposition, W is only deposited where Si is exposed, such as in contact holes, and not on the surrounding SiO₂. In this process, the exposed Si is used to reduce the W source gas, WF₆, depositing W only in those regions.

Much more commonly used is blanket tungsten deposition. Here tungsten is deposited everywhere, both in the contact or via holes, and over the dielectric. (CMP shown in Figure 2–42 or an etchback process described in Chapter 11 can then be used to remove all the tungsten except in the contact or via holes.) Tungsten can be deposited by either the hydrogen reduction reaction



or the silane reduction reaction



A two-step process utilizing both these reactions can also be used, as will be described in Chapter 11.

Tungsten films deposited by CVD have significant amounts of fluorine incorporated in them as a result of the WF₆ source, often 1 atomic percent or more. This increases their resistivity over bulk values by 1.5 to 2 times. The grain size is usually less than 0.2 μm, yet the films are rather rough. This is due to random crystal faceting. This makes the films harder to etch evenly and can also lead to problems in the dielectric overlays. Overall, good filling of contacts and vias with decent film properties can be achieved, and the use of tungsten plugs in current microelectronics is widespread.

9.3.8 TiSi₂ and WSi₂ Deposition

Titanium silicide, TiSi₂, and tungsten silicide, WSi₂, are common silicides used today in Si microelectronics. Other common silicides include CoSi₂, TaSi₂, MoSi₂, NiSi, and PtSi. We will focus on TiSi₂ and WSi₂ here, but many of the same deposition methods and ideas can apply to the other silicides as well. As we saw in the CMOS example in Chapter 2 and as we will discuss in much more detail in Chapter 11, silicides are used in CMOS technology extensively to reduce sheet resistance of polysilicon lines and n⁺ regions. (See Figure 2–36.) They are often part of contact structures, either because they are already on top of the gates, sources, and drains or because they are deposited in contact holes. Silicides can also be used as local interconnects themselves.

The formation of silicides can be done in two general ways: by direct deposition of the silicide; or by deposition of the metal on top of Si followed by the reaction between the metal and Si to form the silicide. The direct deposition method can be accomplished in several ways: (1) sputtering from a composite target, (2) cosputtering from two targets of the metal and Si, (3) co-evaporation of the metal and Si, and (4) CVD.

The reaction method is more common today than the direct deposition method. (See Figures 2-35 to 2-37.) Here the metal, Ti for example, is deposited by sputtering on the exposed gate and/or source drain regions, all of which are silicon. The wafer is annealed and the silicide forming reaction occurs wherever the silicon and metal are in contact, that is, on the gate and source/drain regions. The unreacted metal is then selectively etched away. This method produces self-aligned silicide structures on the gate and/or source drains. This is often called the "salicide" technique, for "self-aligned silicide," and will be discussed more in Chapter 11. Another reaction method deposits the metal and then amorphous Si, both usually by sputtering, then patterns and selectively etches the amorphous Si. The wafer is then annealed so that the metal and amorphous Si react to form the silicide, and the unreacted metal is selectively etched off. This nonself-aligned method produces local interconnects independent of either gate or source/drain regions. For either type of reaction method, it is imperative to have the equipment (both sputtering chamber and annealing chamber) and the wafer surfaces free of oxygen, water, and native oxides. Their presence can result in films that are high resistivity, have uneven thickness and rough surfaces, and have poor adhesion properties.

WSi_2 has historically been used in a "polycide" structure. Here the silicide is blanket deposited over a blanket polysilicon layer, and both polysilicon and silicide are masked and etched at the same time to produce a low-resistivity gate structure. Both sputtered and CVD deposition of the WSi_2 have been used. Sputtering from a composite (WSi_2) target has been used, but CVD has the advantage of better step coverage, lower stress, and better adhesion. (WF_6 and SiH_4 are used in the CVD reaction, but a high $\text{SiH}_4:\text{WF}_6$ ratio is used so that WSi_2 is deposited instead of W.) However the trend today is to use the self-aligned silicides (namely TiSi_2 or perhaps CoSi_2) instead and form the silicide on the gate and source/drain regions all at once.

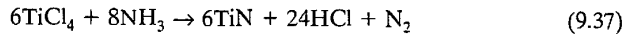
9.3.9 TiN Deposition

TiN films are used as barrier layers in contacts, and as underlayers and antireflective layers in interconnects, as discussed in Chapter 11. They have mostly replaced Ti-W in these applications due to their better barrier and film qualities. TiN films are often deposited or formed on top of Ti films, which have better contact and adhesion properties to the films underneath.

TiN is commonly sputtered using the reactive sputtering technique. As discussed earlier, reactive sputtering involves sputtering a metal in the presence of a reactive gas, such as nitrogen or oxygen, to form the metal oxide or nitride. In this case, nitrogen gas is mixed with the argon gas while Ti is being sputtered. The plasma breaks up the N_2 molecules into atomic nitrogen, which then react with the Ti to form the compound both on the target and on the wafer surface. The properties vary significantly with the

deposition conditions, which include the nitrogen flow or partial pressure, the total pressure, the sputtering power, the substrate temperature, and the substrate bias, if one is applied. While excess nitrogen, or the presence of oxygen, increase the film's resistivity, they also enhance the barrier properties by saturating the grain boundaries. The deposited film is typically fine grained, columnar, and with $\langle 111 \rangle$ orientation, although $\langle 100 \rangle$ orientation can be achieved under certain deposition conditions [9.25]. The microstructure and orientation also depend on the film underneath. For example, the crystallographic orientation or texture often follows that of the Ti film that is commonly underneath TiN films. TiN is a very hard metal, and the films can be highly stressed and brittle, like Ti-W. TiN is often sputter deposited when the TiN is part of the Al interconnect stack. Step coverage and filling are not usually issues here. Generally it is better to deposit the whole stack by one technique without leaving the deposition chamber to avoid surface oxidation and contamination. Ionized PVD of TiN can be used when TiN is utilized as a liner/barrier in vias.

To improve the step coverage and filling ability of TiN films, CVD methods for deposition of these films have gained a lot of interest recently [9.26]. Titanium tetrachloride, TiCl_4 , can be reacted with ammonia, NH_3 , to form the films at 400–700°C via the reaction



In general, the higher the deposition temperature, the better the film properties. As the temperature increases, the resistivity of the films decreases, the film density increases, and the chlorine content in the films decreases. Chlorine incorporation can be significant, 0.5–5%, which can lead to corrosion problems in Al interconnects. CVD processes using Metal-Organic (MOCVD) precursors instead of ammonia have also been developed. These include tetrakis-dimethylamino titanium (TDMAT) and tetrakis-diethylamino titanium (TDEAT). TiN films can be deposited with these sources at lower temperatures and with no Cl contamination. However, the films typically have carbon and oxygen incorporation, lower density, particulates, and not as good step coverage or filling ability. Methods to improve these properties, including postdeposition annealing treatments, or plasma treatments, are being developed. PECVD methods using ammonia are also being developed.

Another method to deposit TiN films is by first sputtering Ti on the wafers and then annealing them in a nitrogen or ammonia ambient above 600°C [9.25]. This was the process used in the CMOS technology described in Chapter 2. The Ti film reacts with the nitrogen to form the TiN film. The reaction is limited by the diffusion of the nitrogen through the forming TiN film, and the formation rate exhibits a diffusion-limited square root of time dependence, as expected. The reaction is very sensitive to the presence of oxygen, which can actually stop the reaction. The anneal is often done in an RTA system in which the heating time and exposure to oxygen are minimized. This nitridation method is often done in conjunction with silicide formation (Figure 2–36), as discussed in more detail in Chapter 11. Here the anneal done to react the deposited Ti with the underlying Si is performed in a nitrogen ambient. While TiSi_2 forms on the bottom of the Ti, TiN forms on top.

9.3.10 Cu Deposition

As we will discuss in Chapter 11, manufacturing processes that use copper instead of Al as a global interconnect metal have recently been developed. Cu has a lower electrical resistivity than Al and has better electromigration resistance. However, there are several processing and process integration issues concerning the use of Cu. One problem with Cu is that it is difficult to plasma etch due to the low volatility of the etch byproducts. Therefore a process that does not require etching, such as selective deposition, would be attractive. However, by using the damascene process, which was described in the CMOS technology in Chapter 2, the deposition and patterning of Cu can be accomplished by doing a blanket deposition in an etched-out dielectric layer, followed by Chemical-Mechanical Polishing (CMP). In this manner, neither etching or selective deposition is required. The damascene process will be described further in Chapter 11.

Copper can be deposited by PVD methods, including both sputter deposition and evaporation. However, the step coverage for these is quite poor and filling small vias is difficult. A high-temperature reflow step can be done, but that raises the thermal budget of the process.

CVD processes for Cu deposition are being developed to give better step coverage and filling properties. Like Al CVD processes, metal-organic sources are commonly used. One such source is $2\text{Cu}(\text{hfac})\text{L}$, where (hfac) is hexafluoroacetylacetonate and L is a neutral ligand weakly bonded to the Cu. One ligand is vinyltrimethylsilane. Both blanket and selective depositions can be done depending on the ligands and deposition conditions. However, it is difficult to simultaneously achieve low resistivity, good step coverage and via filling, and a high deposition rate with these methods [9.27].

An alternative deposition method that has been developed for Cu is electrolytic plating. This can be done with the use of external electrodes and applied current. Applying an external voltage is generally referred to as electroplating. Deposition without an applied field is called electroless deposition. Both processes are used commonly in printed circuit board manufacturing, and work is being done to apply these processes to on-chip interconnect fabrication. While the electroless method has received a lot of attention in recent years, mostly due to its selective deposition capabilities, the electroplating method has become the current method of choice for Cu deposition in integrated circuits.

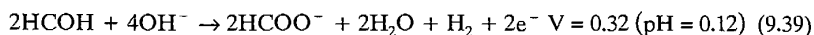
In electroplating, the wafers are mounted on a cathode and immersed into a plating solution that contains Cu ions. An inert anode, made of platinum, for example, is also immersed into the solution. A voltage is applied between the two electrodes and the current drives the Cu ions toward the wafer, forming metallic Cu on the surface, according to the electrochemical reaction



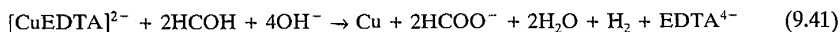
A thin seed layer, commonly Cu as well, is deposited first over the wafer so that electrical contact can be made to the surface of the wafer and so that Cu is electroplated over the entire wafer surface. This can be done by sputtering or CVD. Excellent filling can be achieved with electroplating of Cu. Filling of gaps is enhanced by using "level-

ing" agents or additives in the solution. These chemicals inhibit the deposition. They preferentially adsorb on the top surfaces of the topography rather than on the bottom of trenches or vias due to diffusion limitations. The top surfaces and corners then receive less deposition than the bottoms of trenches or vias, resulting in excellent filling ability. Low resistivity and high deposition rate can also be achieved with electroplating, and this method is becoming the most common method to deposit Cu for interconnects.

The electroless deposition technique involves the formation of thin metallic films from electrolytic solution without the use of external electrodes and without applying an external current. Simultaneous oxidation-reduction reactions between two half-reactions occur to deposit the film from solution. By this method, both blanket and selective Cu films of good quality, low resistivity, and good filling properties can be obtained. In a typical process, the solution contains formaldehyde (HCHO) as a reducing agent, and CuEDTA as the oxidizing agent and source of Cu. The two half-reactions with their redox potentials are



The net potential is positive so that the combined reaction is thermodynamically favorable:



Therefore by placing the wafer in a very basic solution containing Cu-EDTA and formaldehyde, copper will be deposited from the solution. Complexing and surfactant agents are also commonly added to the solution.

In order for the copper to be deposited on the wafer surface, thin seeding layers are required in the electroless process, as in the electroplating process. Here the seed layer is needed to catalyze the electroless process. These are often Pd, Au, or Cu layers, deposited by CVD, sputtering, or evaporation. For selective electroless deposition, the seed layers should only be in the locations where the Cu deposition is desired. This can be done by selectively depositing the seed layer material itself, doing very directional PVD deposition with subsequent etchback or CMP, or using lift-off techniques [9.28]. For nonselective Cu deposition, the seed layer is deposited everywhere followed by the Cu electroless deposition. A problem with the electroless method has to do with contamination. The seed layer materials, Pd, Au, or Cu, can all act as deep-level traps in the underlying Si. Of more concern are the alkali ions needed for the highly basic solution. The hydroxide ion, OH^- , is commonly produced from sodium or potassium hydroxide. Sodium and potassium ions can drift in SiO_2 and alter the electrical properties of the active devices as we saw in Chapter 4. Tetra-methyl-ammonium-hydroxide ($\text{N}(\text{C}_2\text{H}_5)_4\text{OH}$), or TMAH, is often used instead of sodium or potassium hydroxide. However, the complex solution and reactions involved with the electroless technique compared to the electroplating method and the slower deposition rate, along with the trend toward blanket deposition of Cu in conjunction with the damascene process, have made the electroless process less attractive than the electroplating method for Cu deposition in IC technology.

9.4 Measurement Methods

The properties of the deposited films can be measured either during the deposition (in situ) or after the deposition is complete. The latter case is much more common. Properties such as thickness, composition, impurity concentration, step coverage, and stress, which often depend on the underlying layers and topography, are usually measured on the films after the deposition is complete, outside the deposition chamber. The measurement of these properties is discussed in the measurement methods section of Chapter 11 for films used in back-end technology. These films include silicon oxide, silicon nitride, silicide, and metal films, but the techniques are also used for other dielectric and interconnect materials, such as polysilicon, as well. In this section we will briefly describe some in situ measurement techniques.

The measurement of properties during the deposition is not very common today during IC fabrication and is mostly limited to film thickness and deposition rate measurements or to research applications. But as the technology progresses and more control is needed over film properties, there is much more interest in having automatic feedback during deposition. This requires in situ measurement of the film properties so that the deposition conditions can be adjusted as the deposition proceeds. The two most common methods of in situ measurement techniques are ellipsometry and the quartz-crystal microbalance.

Ellipsometry was discussed in Chapter 6 and can measure the thicknesses, and hence the deposition rates, of thin films on a reflective substrate. This technique is based on the change of polarization that occurs when a polarized light beam is reflected by a film. When data for multiple wavelengths are acquired, the technique is known as spectroscopic ellipsometry. This provides better measurement when multiple layers are present on the wafer. Ellipsometers are well suited for in situ deposition measurements. This is because the ellipsometer can be placed outside a chamber, using a window to allow the measurement. Even if the window becomes coated with the deposition material or byproducts, the measurement can still be done since the technique relies on relative intensities of light, and not on an absolute intensity which can change with time. The technique is only useful for transparent thin films.

The other common in situ measurement technique, QCM (Quartz-Crystal Microbalance), is based on the fact that the oscillating frequency of a crystal depends on its mass. A quartz crystal, sandwiched between two electrodes, is mounted in the deposition chamber in the same area as the wafers. As the film material is deposited on the wafers and the crystal, the mass of the crystal increases. This changes the oscillating frequency of the crystal in a well-understood manner, which is monitored through an external circuit, and the change in frequency is determined. This is then converted to a change in mass, which is then converted to a change in thickness. This technique is so sensitive that it is often used to calibrate other, less common, techniques, such as atomic emission and atomic absorption methods. This method can be applied to the deposition of any thin film.

9.5 Models and Simulation

While one important goal in thin film deposition is to obtain films with the desirable physical and chemical properties, an equally important goal is to achieve proper coverage and filling of the film over the underlying topography. This becomes even more critical as lateral dimensions decrease and aspect ratios get larger.

In this section we will focus on the modeling of thin film deposition as it relates to film coverage and filling over nonplanar topography. In the first part of the section we will utilize SPEEDIE [9.29], a 2D simulator developed at Stanford University. SPEEDIE uses physically based models for deposition and etching. Many other simulation tools have been developed for deposition and etching in integrated circuit fabrication, such as EVOLVE at Arizona State University, SAMPLE at the University of California, Berkeley, SIMBAD at the University of Alberta, and SHADE at IBM. Many of these are also physically based and contain many of the same basic mechanisms and models as in SPEEDIE. We will use SPEEDIE's simulation capability to predict the thickness and shape, in two dimensions, of deposited thin films as a function of various deposition conditions and configurations. These simulations will be used to illustrate many of the concepts discussed earlier in this chapter. Later in this section we will describe simulations using some commercially available simulation tools.

9.5.1 Models for Deposition Simulations

Figure 9-35 shows a generalized deposition picture of both CVD and PVD techniques. PVD systems use sputtering or heating of a source, whereas CVD systems use gas flow introduction of chemical precursors to provide the reactant species for deposition. The surface processes may also be different. Nevertheless, the steps are in general terms the same and can be modeled with the same components. Only the values used for the specific terms of the model will vary depending on the deposition technique and conditions. This is in fact a good way to analyze and compare the different deposition techniques. By starting with the same basic components and equations, and modeling the different techniques by varying the different parameters, one can more clearly see the similarities and differences between them.

In Figure 9-35 we focus on the wafer surface and the gas phase adjacent to the wafer surface. In the figure we see the direct flux of species, both neutrals and ions, arriving from the gas phase. The deposition species, Al in this case, can be either neutral or ionic. The ions, both Al^+ and Ar^+ in the figure, can also sputter atoms which have been deposited on the surface. In addition the ions, by their surface bombardment, can enhance the deposition rate of the neutral species. We also see in the figure species being desorbed, or "emitted." We see atoms being resputtered by ions. Both emitted and resputtered species can then be redeposited elsewhere, as shown at the bottom of the trench, and are called indirect fluxes. We also see another type of indirect flux: atoms moving due to surface diffusion. The indirect fluxes are all surface, or near-surface, processes.

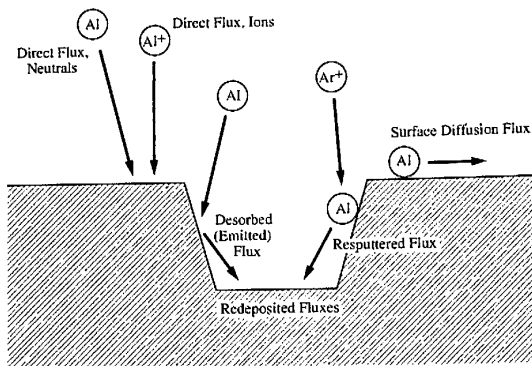


Figure 9-35 Generalized surface region processes in both CVD and PVD deposition systems used in simulation models.

A gas-phase boundary layer is not considered in this formulation since the gas pressure for most deposition systems is low enough so that diffusion through a boundary layer is not important. Because the mean free path in these systems is large compared to the feature sizes of the topography, free molecular flow (i.e., collisionless, nondiffusive flow) is assumed as the transport mechanism near the surface. (Diffusional flow through a boundary layer near the surface is important in APCVD systems, and SPEEDIE has a separate model for that. With many gas-phase collisions occurring within the topographical features in that case, good coverage and filling can result for processes that are reaction limited, such as with TEOS/O₃ SiO₂ [9.5].)

9.5.1.1 Models in Physically Based Simulators Such as SPEEDIE

In physically based topography models such as SPEEDIE all the phenomena depicted in Figure 9-35 should be taken into account. In SPEEDIE the net flux at each position on the surface is calculated based on all the fluxes arriving at and leaving that spot, either from the gas phase—the direct flux—or from other points on the surface—the indirect fluxes. The surface is then moved in the perpendicular direction according to the net flux, normal to the surface, at each point as the film grows. The net flux at each point i , F_{net}^i , can be expressed as the sum of the fluxes that are positive (due to atoms arriving at each point), minus each negative flux (due to atoms leaving each point):

$$F_{\text{net}}^i = F_{\text{direct(neutrals)}}^i + F_{\text{direct(ions)}}^i + F_{\text{reddep}}^i + F_{\text{diff.in}}^i - F_{\text{emitted}}^i - F_{\text{sputtered}}^i - F_{\text{diff.out}}^i \quad (9.42)$$

The fluxes are all in units of atoms (or molecules) $\text{cm}^{-2} \text{sec}^{-1}$.

The *direct fluxes*, $F_{\text{direct(neutrals)}}$ and $F_{\text{direct(ions)}}$, are the fluxes of the different species at each position (or small area) i on that surface that arrive without any interaction with the wafer surface topography. These arrive directly from the gas phase above the wafer and include both neutral and ionic species. In calculating the direct fluxes, one has to know the distribution of arrival angles of the incoming species relative to the orientation of the surface at that point. Then the distribution can be integrated over all angles to obtain the total direct flux, normal to the surface, at each point.

There are two ways of modeling the direct fluxes. One is to start with the angular distribution of fluxes as the species depart from the target or gas source, or the emitted angle distribution at each source point. Then by taking into account the gas-phase collisions and the geometric configuration of the chamber (distance from target to source, etc.), as well as the surface topography, one can determine the distribution of angles of the precursor arriving at each point on the wafer surface. This is similar in a qualitative way to Figure 9–30 which illustrates deposition from different size sources. But including the gas-phase reactions and the various equipment geometries can get quite complicated. Techniques such as Monte Carlo methods are often used for this and are available in various deposition simulators, but even so, fitting parameters are usually needed.

A simpler way to account for the incoming flux is to just assume some distribution function of fluxes as a function of angle just above the surface of the wafer. This is illustrated in Figure 9–36. If the fluxes were coming in equally at all angles, that is isotropic arrival fluxes, then one would assume a constant flux for all angles in the distribution at each position in that plane. However, as we discussed earlier, we define the arrival distribution in terms of the normal component of the incoming flux relative to a unit area on a surface. In this case the surface is a horizontal plane just above the wafer surface. By simple geometry, the normal component would be proportional to the cosine of the vertical angle. We would therefore express this

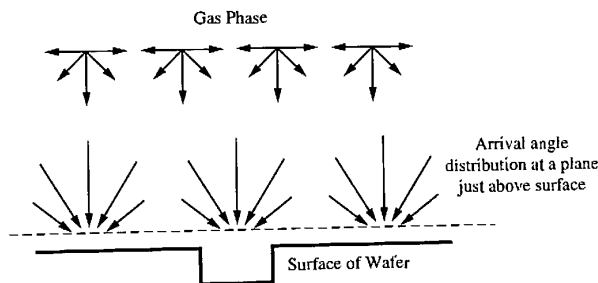


Figure 9–36 Simplified way of considering incoming flux distribution. A distribution of fluxes arriving from the gas phase, with a $\cos^2\theta$ distribution, is assumed in a plane just above the wafer surface. From that plane to the surface features below, no gas-phase collisions occur, just line-of-sight molecular flow.

isotropic arrival as a $\cos\theta$ distribution, where θ is relative to the vertical direction. To account for nonisotropic distributions as well, this can then be generalized to the form.

$$F_{\text{direct}}(\theta) = F^{\circ} \cos^n \theta \quad (9.43)$$

Here F_{direct} is the normal component of the flux coming in at angle θ , F° is the flux at $\theta = 0^{\circ}$, and θ is the angle from the vertical direction, as was indicated in Figure 9-25. By adjusting n from 1 to large numbers, one can obtain isotropic to very anisotropic distributions of arrival angles, as shown in Figure 9-37. With $n = 1$, the distribution is isotropic, with fluxes coming from above at all directions equally. Higher pressure, more collisions in the gas phase, and shorter mean free path would lead to a more isotropic arrival distribution. A larger source and shorter source-to-wafer distance also give more isotropic arrival. For $n > 1$, a tighter distribution of arrival angles would occur, as shown in the graph in Figure 9-37. This means that the fluxes arriving are more directed in the vertical direction. Lower pressure, fewer collisions in the gas phase, and a longer mean free path would lead to a more anisotropic arrival distribution, as would a smaller source and a longer source-to-wafer distance. For ionic species in biased systems (bias sputtering, ionized sputter deposition, HDPCVD), the n values would be quite high since the ions are accelerated and directed toward the surface by the imposed electric field. In practice, the value of n for the various species is first approximated according to the type of species (ionic or neutral, for example) and type of deposition system, as we will discuss later. The exact value of n for the various species is then often determined by fitting simulations to experimental profiles in calibration structures.

One must also specify the maximum, unobstructed flux of each species to normalize the flux distribution. This is the integrated flux of Eq. (9.43) over all directions above the surface and corresponds to deposition on an unobstructed flat surface.

Knowing the magnitude and distribution of fluxes coming from the gas phase, one must then take into account the local surface topography. This includes the relative orientation of the surface at each position i on the wafer surface, as well as any shadowing

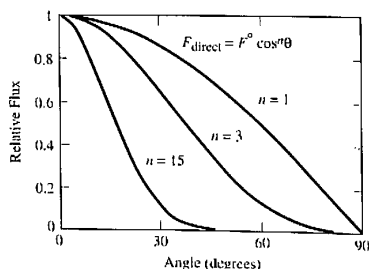


Figure 9-37 Distribution of arrival angles as function of angle to vertical direction, θ , for three values of n . The relative flux is $(F_{\text{direct}}/F^{\circ})$, where F_{direct} is the normal component of flux arriving at point i at angle θ . The $n = 1$ curve corresponds to isotropic arrival, with higher values of n representing more vertically directed arrival.

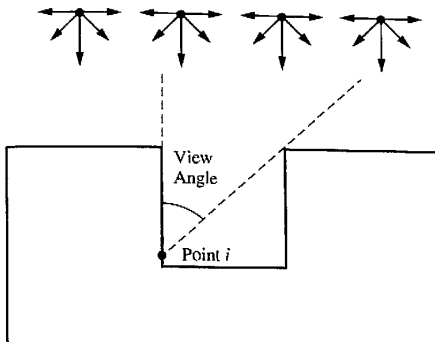


Figure 9-38 Illustration of viewing angle concept. A point, i , inside a trench would only see fluxes arriving within a limited range of angles due to geometrical considerations and shadowing.

effects from the topographical features. For example, the position may be on the side-wall of a step or trench and receive little flux from directly above with more coming in at an angle. Or the position may be at the bottom or side of a trench and see a limited amount of direct flux coming in since the fluxes outside the line of sight will be blocked or shadowed. Figure 9-38 illustrates this. One must take into account the line of sight, the “view angle,” from the reference plane in the gas phase just above the surface to each position on the surface. (In fact, each small area on the reference plane can be assumed to be an emission source to the topography below, equivalent to a downward pointing Knudsen cell, for these calculations.) Assuming collisionless, line-of-sight transport in this region, all this geometry and shadowing can be taken into account and the total direct flux can be calculated at each position on the wafer surface for each species.

The *indirect fluxes* include the processes that occur on the surface, or near the surface of the wafer during deposition. Surface diffusion, emission, sputtering, and redeposition are all indirect fluxes.

The *surface diffusion fluxes*, $F_{\text{diff.in}}$ and $F_{\text{diff.out}}$, are driven by differences in the curvature in the surface. Here we are not concerned with the short-range diffusion of atoms to attachment sites as part of the surface reaction process, but rather the longer-range diffusion in order to reduce surface energy. The net diffusion flux ($F_{\text{diff.net}} = F_{\text{diff.in}} - F_{\text{diff.out}}$) at each point is calculated by

$$F_{\text{diff.net}} = \frac{D_s}{kT} \gamma_s \Omega v \frac{\partial^2 K}{\partial s^2} \quad (9.44)$$

D_s is the surface diffusivity of the material, γ_s is the per area surface energy, Ω is the atomic volume of the atom or molecule, v is the number of atoms per unit area, or the surface density, K is the curvature, and s is the unit length along the surface the atom or molecule travels. In Chapter 11 in the section on reflow modeling we derive the expression for the surface flux and the normal velocity of a surface due to surface diffusion. Here we are interested in the flux relative to the surface normal direction, which is the velocity in that direction times the density. To get this, we multiply the expression

for the normal velocity due to surface diffusion of Chapter 11, Eq. (11.37), by the density, which is the same as dividing by Ω , and we end up with Eq. (9.44).

Because of curvature-induced surface diffusion during deposition, filling in of corners is enhanced and smoother and conformal depositions can result, as we will see in the simulations. Because atoms can diffuse along the surfaces to areas in less direct view of the source, better step coverage can be obtained. Surface diffusion also helps with filling vias and trenches and can lead to more planarized depositions as well. Fast surface diffusion is often used to explain, and model, conformal coverage in particular deposition cases. However, as we will see, emission and redeposition can model the same effects and may be more physically correct in some cases.

The flux due to emission of species from the surface, F_{emitted} , occurs because not all deposition species or precursors stay, or stick, where they land. A finite number, usually assumed to be a constant fraction of the incident species, will be emitted or desorbed from the surface without properly reacting or depositing there. The emitted flux at each point i is thus

$$F_{\text{emitted}}^i = (1 - S_c) F_{\text{incident}}^i \quad (9.45)$$

where F_{incident}^i includes both direct and indirect deposition fluxes incident at i and S_c is the sticking coefficient.

The sticking coefficient here is defined as the ratio of the flux of species that properly react and stay on the surface to the flux of incident deposition species:

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \quad (9.46)$$

This is really an effective sticking coefficient or probability, since factors such as site availability, reaction competition and completion, and surface recombination are implicitly included in it. It is also sometimes referred to as the reactive sticking coefficient, and is often empirically determined from deposition profiles. The sticking coefficient depends on the type of deposition system, the chemistry, and material being deposited. It can also depend on substrate temperature. The temperature dependence is often difficult to predict because of all the processes that may be involved. A lower value is seen at higher temperatures in some cases, such as in the case of LPCVD SiO_2 [9.30]. In other cases, such as with CVD W deposition, a lower value is sometimes seen at lower deposition temperatures. In more advanced models S_c may not be a constant, as is assumed here.

The fraction that does not stay at a particular position and is emitted is therefore equal to $(1 - S_c)$. The F_{emitted} flux is thus subtracted in the total flux equation [Eq. (9.42)]. It is often assumed that the emitted species are emitted in a range of directions away from the surface in a perfectly diffuse angular distribution, with a cosine emission distribution, independent of the incident angle [9.31, 9.32]. Thus the emission process is modeled as a small planar surface source at each position with $n = 1$ in the $\cos^m \theta$ -emitted angle distribution, and the distribution is centered around the surface normal. This implicitly assumes that there is appreciable residence time on the surface so that the

species lose their memory of their incident angle. It is also generally assumed that only the neutral species can be emitted from the surface. For ionic species the sticking coefficient is thus taken to be one, except when ion reflection effects are considered.

Those emitted species can then be deposited on other surfaces of the structure, leading to the *redeposited (emitted) flux*, $F_{\text{redep(emitted)}}$, term. So for each point i , there is a flux coming from each of the other points, termed k . It is again assumed that the number of gas-phase collisions near the surface area is negligible, so that the emitted species travel in straight lines to other points on the surface. Knowing the emitted flux distribution from each point k [Eq. (9.45) for k and a cosine distribution] and the geometrical relationship to point i , the flux emitted from each point k to point i can be determined. This is similar to what we did earlier in Section 9.2.2.1 when we calculated the flux and deposition rate onto a surface from a planar evaporation source. Of course, only those points that are in line of sight to point i are included. The redeposited flux at point i due to emission from point k is thus:

$$F_{\text{redep(emitted)}}^{ik} = g^{ik} * F_{\text{emitted}}^k = g^{ik} * (1 - S_c) * F_{\text{incident}}^k \quad (9.47)$$

Here g^{ik} is the geometrical factor that accounts for the distance and orientation between points i and k , similar to Eq. (9.23) where we derived the expression for deposition on a surface from a planar source. Then all the emitted fluxes from each of the points k to i are summed up and contribute to $F_{\text{redep(emitted)}}$.

We mentioned earlier that fast surface diffusion of deposited species is often used to explain good step coverage. However, the emission and redeposition process can result in the same effect. If the sticking coefficient is less than one, species landing on surfaces can be emitted and deposited elsewhere, including vertical surfaces (recall Figure 9-21). The diffuse emission behavior means that surfaces in all orientations can get material deposited on them. The lower the sticking coefficient, the more conformal step coverage becomes. Whether this redistribution of material and better coverage occur by the emission/redeposition process or the surface diffusion process is sometimes hard to distinguish although this has been done in some studies [9.32, 9.33, 9.34]. Using a clever overhang test structure, these experiments concluded that the emission-redeposition process, and not surface diffusion, was dominant for LPCVD and PECVD deposition of SiO_2 . Thus a lower sticking coefficient, and not faster surface diffusion, is the appropriate model, at least in the specific cases studied. For other materials or techniques, surface diffusion may be important. For example, for sputter deposited Al, the sticking coefficient is found to be one [9.33], which implies no emission/redeposition, but surface diffusion at high temperatures is significant. In our simulations later we will show how both the sticking coefficient and surface diffusivity affect the degree of coverage.

The *sputtered flux*, $F_{\text{sputtered}}$, term is from the atoms that are sputtered off the wafer surface. This is a momentum-transfer process in which incoming particles knock and displace surface atoms from the surface of the wafer. This is due mainly to accelerated ions, both argon and ionized deposition atoms, coming directly from the gas phase. These ions can achieve energies of several hundred eV. The flux term at each point i due to sputtering (a negative term since atoms are being taken away from the surface) is therefore

$$F_{\text{sputtered}}^i = Y * (F_{\text{argon}}^i + F_{\text{direct(ions)}}^i) = Y * F_{\text{ions}}^i \quad (9.48)$$

Y is the sputtering yield and is defined as the ratio of atoms that are sputtered or knocked off the surface to the number of incident ions or atoms. Y is a function of the target material (the wafer surface material, in this case) and the mass of the incoming ion. However, we assume here for simplicity that it is the same for argon and other ions. It is also a function of ion energy and the incident angle, θ , at point i , with a normal incidence being 0° . Ions coming in at low angles are reflected from the surface and do not transfer significant energy to the surface atoms, while ions coming perpendicular to the surface knock surface atoms further into the film and not off. The maximum sputtering yield occurs usually between 40 and 70° .

A typical sputtering yield distribution can be seen in Figure 9–39. The specific characteristics of the curve, such as the maximum yield and angle for which it occurs, could vary for different materials and equipment conditions. Also shown is the normalized etch rate due to sputtering, which equals the product of the sputter yield and $\cos\theta$. The latter term is due to the fact that it is the component of the flux normal to the surface that determines the etch or deposition rate on a surface, as we have described. The sputter etch rate also has a maximum at some angle between 40 and 70° . This angular dependence on sputtering yield and sputter etch rate is often utilized to achieve more planar surfaces during deposition. The sloped surfaces (with incident angles less than 90°) sputter off faster than horizontal or vertical surfaces (with incident angles near 0 or 90°), leading to more planar surfaces.

The emitted angle distribution of the sputtered atoms has been modeled with different types of distributions [9.35]. The sputtered atoms often exhibit an ideal, diffuse angular distribution, being sputtered away at many angles centered around the surface normal, like desorbed or emitted species discussed above. This is illustrated in Figure 9–40(a). The distribution may in fact be more sideways, described by an “under-cosine” distribution (where n is less than 1 in a $\cos^n\theta$ distribution), shown in Figure 9–40(b). In some cases though they can exhibit a specular distribution, being sputtered at just a few angles near an angle opposite that of the incident angle. In this case the angular distribution of the sputtered atoms is described by a $\cos^n\theta$ distribution, with a very large

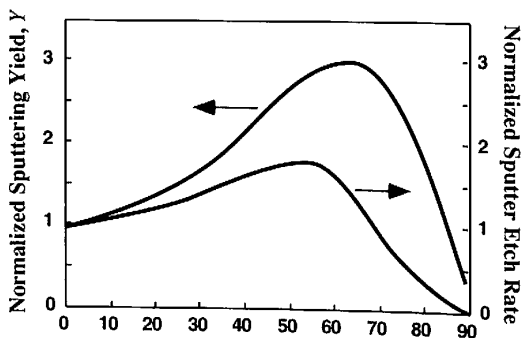


Figure 9–39 Typical example of sputtering yield and sputter etch rate as functions of incident angle, with respect to surface normal.

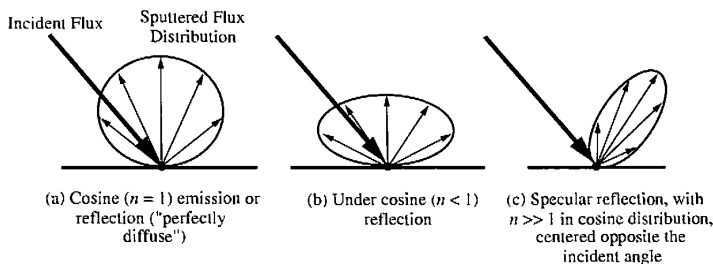


Figure 9-40 Different emitted angle distributions for sputtered atoms.

value of n , and the distribution of sputtered atoms is centered opposite the incident angle. This is illustrated in Figure 9-40(c). Generally distributions like Figure 9-40(a) or (c) are used in simulations. The first type is common when the incident ion energy is high and there are collisions and interactions deeper in the sputtered layer, while the latter type is more common when the incident ion energy is lower and single surface collisions dominate. Also, amorphous materials like SiO_2 tend to sputter off with the first type of distribution.

As was the case with desorbed or emitted species, sputtered atoms can be redeposited elsewhere on the surface. The method to calculate this *redeposited (sputtered) flux* and the equations used are analogous to the redeposition due to emission. At each point i the redeposited flux coming from each other point k is determined:

$$F_{\text{redep(sput)}}^{ik} = g^{ik} * F_{\text{sputtered}}^k = g^{ik} * Y * F_{\text{ions}}^k \quad (9.49)$$

Again, g^{ik} is a geometrical factor that accounts for the distance and orientation between points i and k , as well as for the angular distribution of sputtered atoms from k . Only those points, k , that are in line of sight of i would contribute.

In addition to sputtering off atoms from the surface, ions striking the surface can also enhance the deposition rate. Ion bombardment can transfer energy to the species on the surface and enhance the different surface processes and reactions. One way to model this is by adding an ion-induced deposition rate enhancement term to the total flux expression (rather than trying to modify the other terms). This term is of the form of

$$F_{\text{ion-induced}}^i = K_i * F_{\text{ions}}^i \quad (9.50)$$

where K_i is the ion-induced deposition enhancement factor and is usually a fitting factor. Remember that F_{ions} includes all ions, both nonprecursor ions, such as Ar^+ , and any ionized precursor species (Al^+ or Ti^+ , for example). Non-ionized precursor species are still needed for this ion-enhanced process, but they are not a limiting factor in this process so only the ion flux is included in this term. The competition between ion sputtering and ion-enhanced deposition can be very important in obtaining the final topography.

9.5.1.2 Models for Different Types of Deposition Systems

To simulate deposition, one should include all the terms in Eq. (9.42) and calculate the total net flux at each point i on the surface. The deposition rate, or the movement of the surface, at each point is then calculated by dividing the total flux by the film density. For ease of use and computation, one could simplify the net flux equation for the various types of deposition methods (LPCVD, standard PVD, ionized PVD, etc.), including only the important terms for each technique. This is in fact what is done in SPEEDIE. Relative rate constants (K 's) are often included in front of some of the flux terms which makes it easier to adjust the parameters when multiple processes are occurring. By examining the expressions that SPEEDIE uses for each type of deposition, one can gain a better understanding of the different deposition techniques.

First we consider *LPCVD deposition*. Figure 9-41 shows an LPCVD system and the relevant fluxes needed to model this type of deposition. In this case we have the direct flux of neutral species, as well as the desorbed or emitted flux and redeposition of these neutral species. There is no flux of ionic species in this system, and no sputtering or redeposition of sputtered species. In addition, long-range surface diffusion is usually not important, as discussed earlier.

Eq. (9.42) becomes

$$F_{\text{net}}^i = F_{\text{direct(neutrals)}}^i + F_{\text{redepos(emitted)}}^i - F_{\text{emitted}}^i \quad (9.51)$$

Substituting in Eqs. (9.45) and (9.47) for the emitted and redeposited fluxes leads to

$$\begin{aligned} F_{\text{net}}^i &= F_{\text{direct(neutrals)}}^i + F_{\text{redepos(emitted)}}^i - (1 - S_c)(F_{\text{direct(neutrals)}}^i + F_{\text{redepos(emitted)}}^i) \\ &= S_c * (F_{\text{direct(neutrals)}}^i + F_{\text{redepos(emitted)}}^i) \\ &= S_c * (F_{\text{direct(neutrals)}}^i + g^{ik} * (1 - S_c) * F^k) \end{aligned} \quad (9.52)$$

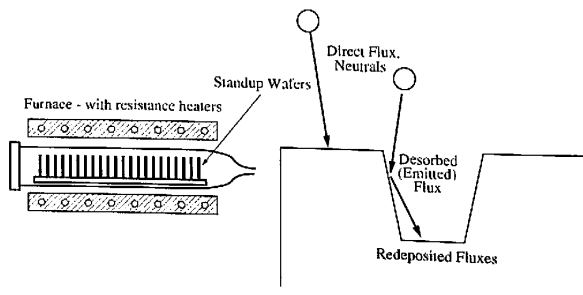


Figure 9-41 LPCVD system and relevant fluxes needed to model this type of deposition.

This must be solved for all points i simultaneously and consistently—emitted flux from one point affects the redeposition flux at another, which in turn affects the flux at the first point, and so on. In the deposition rate equation, the direct neutral and redeposited flux terms ($F_{\text{direct(neutrals)}}^i + F_{\text{redeposited}}^i$) are combined into one term, and called the local deposition flux, F_d . The net deposition rate at each point is just the net flux, F_{net}^i , divided by the film density, N . So for LPCVD, the deposition rate equation is given as

$$\text{rate} = \frac{S_c F_d}{N} \quad (9.53)$$

The other important parameter would be the value of n in the $\cos^n \theta$ distribution for the incoming (direct) species.

Example Calculate the deposition rate using the expression for LPCVD deposition for SiO_2 on a flat surface with no topography. S_c is equal to 0.3, the maximum, unobstructed flux is equal to 3×10^{15} molecules $\text{cm}^{-2} \text{sec}^{-1}$, and the density of the deposited oxide film is 2.27 gm cm^{-3} .

Answer We are interested in this case in the deposition rate on a flat surface. Therefore there is no shadowing, no redeposited flux, and each surface position is in a horizontal orientation. Thus F_d is just equal everywhere to the maximum, unobstructed flux. The deposition rate for LPCVD in SPEEDIE is given by Eq. (9.53). The density of SiO_2 in terms of molecules cm^{-3} is

$$\begin{aligned} \text{density}(\text{SiO}_2) = N &= 2.27 \text{ gm cm}^{-3} * \frac{\text{Avogadro's number}}{\# \text{ of gms per mole for SiO}_2} \\ &= 2.27 \text{ gm cm}^{-3} * \frac{6.02 \times 10^{23} \text{ molecules/mole}}{60 \text{ gms/mole}} \\ &= 2.3 \times 10^{22} \text{ molecules cm}^{-3} \end{aligned}$$

Therefore the deposition rate on the flat, unobstructed surface is

$$\begin{aligned} \text{rate} &= \frac{S_c F_d}{N} = \frac{0.3 * 3 \times 10^{15} \text{ molecules cm}^{-2} \text{sec}^{-1}}{2.3 \times 10^{22} \text{ molecules cm}^{-3}} \\ &= 3.9 \times 10^{-8} \text{ cm sec}^{-1} = 0.023 \text{ } \mu\text{m min}^{-1} \end{aligned}$$

For deposition on nonflat or obstructed surface locations, on sidewalls of trenches, for example, the local deposition rate can be less than this. The same expression is used but the local flux, F_d , can be less at these locations. Figure 9-46 to be discussed later illustrates LPCVD simulation on more complex topography.

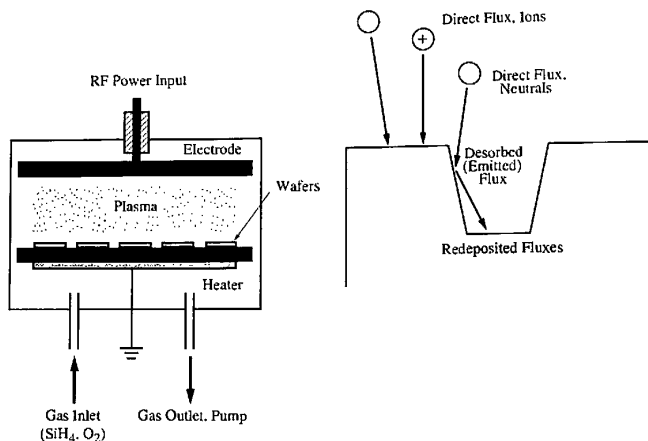


Figure 9-42 PECVD system and relevant fluxes needed to model this type of deposition.

For *PECVD deposition* there is an ion flux onto the wafer which can accelerate the deposition process. Figure 9-42 shows a PECVD system and the relevant fluxes needed to model this type of deposition. As in LPCVD there is a direct flux of neutral species and the emission and redeposition of those species. Also shown is the direct flux of ions, accelerated across the plasma sheath toward the wafer, which can enhance the deposition. The ion energy in this type of system is assumed low enough that sputtering is not significant and direct deposition of ions is also not appreciable. Again surface diffusion is not considered important.

Thus the ion-induced enhancement term, Eq. (9.50), is included in the rate equation, which becomes

$$\text{rate} = \frac{(S_c K_d F_d) + (K_i F_i)}{N} \quad (9.54)$$

K_d and K_i are relative rate constants for the non-ion-enhanced component and the ion-enhanced component, respectively. F_i is the net local ion flux, or F_{ion} as given above. N is the film density. Surface diffusion is not explicitly modeled here; however a low value of S_c can have a similar effect, as discussed earlier, and has been used in SPEEDIE in modeling PECVD oxide deposition [9.4, 9.5, 9.34, 9.36]. Both the non-ion-enhanced and the ion-enhanced terms are important in PECVD depositions [9.5, 9.36].

For *standard PVD*, the same fluxes and terms are used as for LPCVD. Figure 9-43 shows a PVD system, a standard DC sputtering system in this case, and the relevant fluxes needed to model PVD deposition. The rate equation is thus

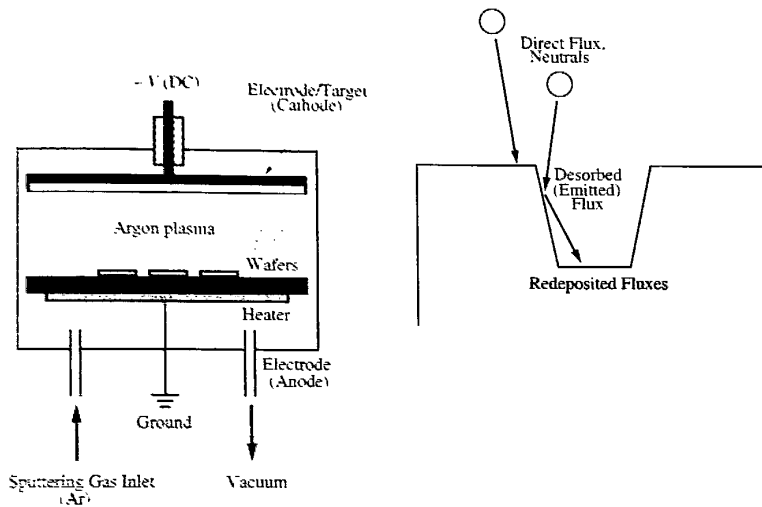


Figure 9-43 PVD (sputter deposition) system and relevant fluxes needed to model this type of deposition. The specific system illustrated here is a DC sputtering system.

$$\text{rate} = \frac{S_c F_d}{N} \quad (9.55)$$

However, different values for the sticking coefficient and the arrival angle distribution would be used for PVD versus CVD methods, as discussed below. There is no significant ion bombardment onto the wafer surface for standard sputter deposition systems, either DC or RF based, as described earlier. This model would also be applicable to collimated and long throw sputter deposition systems, but a narrower arrival angle distribution would be used than for standard sputtering systems. In addition, evaporation can be modeled by this model.

For *high-temperature PVD*, the relevant fluxes are the same as those illustrated in Figure 9-43, except that we now add the surface diffusion flux term, Eq. (9.44), to the rate equation for standard PVD, giving

$$\text{rate} = \frac{S_c F_d + \frac{D_s}{kT} \gamma_s \Omega v \frac{\partial^2 K}{\partial s^2}}{N} \quad (9.56)$$

For *ionized PVD*, there are several fluxes in addition to those in the standard PVD model. Figure 9-44 shows an ionized sputtering system and the relevant fluxes. In addition to the direct flux of neutral species, there is a significant flux of ionized species—both ionized Ar atoms and ionized precursor species (Al^+ or Ti^+ , for example). These

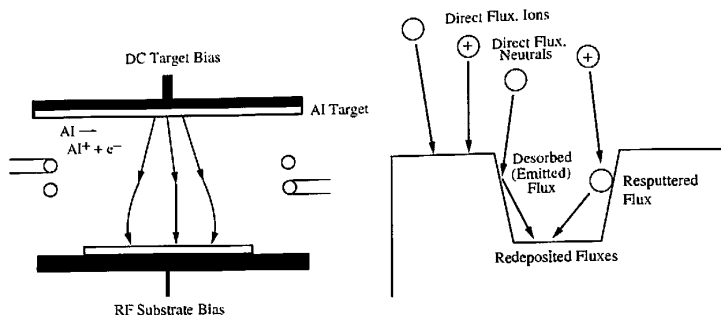


Figure 9-44 Ionized PVD system and relevant fluxes needed to model this type of deposition.

ionized species can be deposited as well as sputter surface atoms on the wafer. Therefore, for ionized PVD, several terms are added to the rate equation for PVD deposition: the term due to direct flux of precursor ions, $F_{\text{direct(ions)}}$; the ion-induced enhancement term [Eq. (9.50)]; and the sputtering and redeposition terms [Eqs. (9.48) and (9.49)]. The direct deposition of precursor ions and the ion-induced enhancement terms are combined into one term, equal simply to F_i , the total ion flux. This is done in SPEEDIE partly because SPEEDIE does not currently distinguish between Ar ions and ionized precursor species.

This leads to the expression given for the rate to be

$$\text{rate} = \frac{(S_c F_d) + F_i - (K_{sp} Y F_i) + (K_{rd} F_{rd})}{N} \quad (9.57)$$

Here F_{rd} is the flux given by Eq. (9.49) for redeposition of sputtered material. K_{sp} and K_{rd} are relative rate constants, used mostly as fitting parameters. A term for reflected Al ions may also be important for coverage issues and may be added in future models (A model for bias sputtering would be similar to this, but the ions would be mostly Ar and not the deposited species.)

Similarly, for *high-density plasma (HDP) CVD* the sputtering and redeposition terms are added to the PECVD case, without any non-ion direct term. Figure 9-45 shows a HDP CVD system and the relevant fluxes used to model this deposition. This gives the following expression for the rate:

$$\text{rate} = \frac{(K_i F_i) - (K_{sp} Y F_i) + (K_{rd} F_{rd})}{N} \quad (9.58)$$

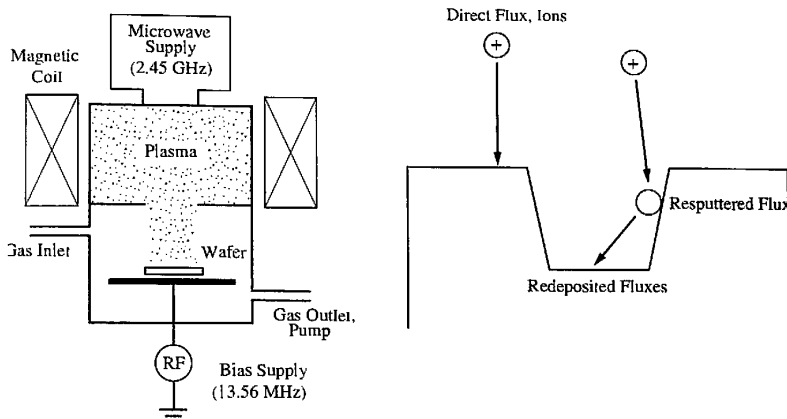


Figure 9-45 HDPCVD system (with a microwave ECR source) and relevant fluxes needed to model this type of deposition.

similar to that for ionized PVD. The ion flux would be quite high in these systems, about $1 \times 10^{17} \text{ cm}^{-2}$. The non-ion (neutral) direct term is not included here because its flux is often too low at these low pressures to be important on its own, that is, when not enhanced by the ions. (However, more recent SPEEDIE models include other terms such as the neutral flux as secondary effects.) There could also be direct deposition of ionized precursor species; this is included in the ion-enhanced term for simplicity.

9.5.1.3 Comparing CVD and PVD and Typical Parameter Values

Before presenting some simulation results, we first want to make some general comments regarding some of the values used for CVD and PVD models. This will be done in terms of the standard models for low-pressure systems, that is, most CVD and PVD systems used today. The discussion will center around the parameter values summarized in Table 9-2.

In CVD deposition systems, the arrival angle distribution for neutral species is basically isotropic, while for PVD systems the arriving flux of neutrals is often more anisotropic and directed. This is due to the different types of sources and the different pressures.

In CVD systems the source gas fills the whole chamber and arrives at the wafers from many directions. This is partly due to injection systems and cages that may be used, as well as gas-phase reactions that may occur. More importantly, the precursor species in the gas phase in CVD systems undergo significantly more collisions than in PVD depositions before the species strike the wafers. This is due to the much higher gas pressure in CVD systems, even for low-pressure CVD systems, compared to PVD systems

Table 9-2 Typical values of modeling parameters for various PVD and CVD depositions

	n (exponent in cosine arrival angle distribution)	S_c (sticking coefficient)
Sputter deposition		
standard	~1.0-4	~1
ionized or collimated	8-80	~1
Evaporation	3-80	~1
LPCVD silicon dioxide		
silane	1	0.2-0.4
TEOS	1	0.05-0.1
LPCVD tungsten	1	0.01 or less
LPCVD polysilicon	1	0.001 or less

(HDPCVD systems with its very low gas pressure being the notable exception). More collisions result in a shorter mean free path and a wider arrival angle distribution. As a result, the arrival angle distribution can be assumed to be isotropic in most CVD systems. This means a value of 1 in the $\cos^{\theta}$ arrival angle distribution given by Eq. (9.43) for the direct flux of deposition species, as shown in Table 9-2. (Higher values for n would be used for the ionic species in plasma CVD systems, where the ions are accelerated and directed toward the wafer surface.)

For PVD depositions, the depositing material comes off a remotely located and finite area source or target, and due to geometrical considerations often arrives at the wafers in a more directed manner. In addition, fewer collisions occur during the transport, as mentioned above, because of the much lower pressures used—usually in the 1- to 10-mtorr range for sputtering and even lower for evaporation, as compared to 0.25 to 2.0 torr for LPCVD. As a result of all this, the arrival angle distribution for PVD depositions can be less isotropic and the atoms arriving at the wafer surface can therefore be more directed. For standard sputter PVD the value of n in the $\cos^{\theta}$ arrival distribution will often be greater than 1, representing a more directed distribution arriving from above. The exact value will depend on the system geometry (e.g., target size, target-to-wafer distance, rotation of substrate), the operating conditions (e.g., pressure) and the depositing material. Typical values range from very close to 1 (virtually isotropic) to about 4 (fairly directed), as shown in Table 9-2, with higher pressures, larger target sizes, and smaller target-to-wafer distances giving the lower n values.

For ionized species (most of the depositing species in ionized PVD systems and some of them in a plasma CVD system) as well as neutral species in collimated PVD systems, the species are very directed. The n value can be from about 8 to 80 or even higher, depending on the system and the plasma conditions.

The sticking coefficient is also generally different for CVD and PVD systems. For CVD systems, the sticking coefficients used in modeling are almost always less than one, as shown in Table 9-2. Typical values would be 0.05 to 0.1 for TEOS-based CVD of SiO_2 , and somewhat higher—up to about 0.4—for silane-based CVD. Similar values would apply to the neutral component in PECVD systems. The value for CVD W (via

hydrogen reduction) and polysilicon is even lower, 0.01 or less. Surface reactions are important for CVD depositions. The species can desorb before the reactions required for deposition can take place. Another precursor required for the reaction may not be present at that spot at that time. In addition, competing reactions and site occupation by other species and byproducts can occur. These all contribute to the effective sticking coefficient for CVD depositions being low. The species not reacted and deposited are desorbed, or emitted, and are usually assumed to have a diffuse emitted angle distribution, as opposed to a specular emission.

For conventional PVD, where physical processes dominate, the atoms generally attach and then stay attached, without desorbing. No surface reactions occur other than simple condensation. The sticking coefficient is therefore usually assumed to equal 1, or close to it, for PVD depositions. (One exception to this may be PVD of Ta, which seems to have a somewhat lower sticking coefficient.) For the ionized species in either PVD or CVD systems, the sticking coefficient is often assumed to equal one.

In PVD and CVD systems where large fluxes of ions are present, such as in ionized PVD and high-density plasma CVD, sputtering of deposited material can occur. This, like desorption, leads to redeposition of species elsewhere. The angular distribution for sputtered atoms is usually assumed to be perfectly diffuse, with n equal to 1, but is sometimes assumed to be specular. For example, sputtered Al is often modeled with specular redeposition with n equal to 2 or more for a cosine distribution that is directed away from the incident angle. HDPCVD deposition is dominated by the highly directional ion flux and actually acts more like PVD deposition rather than other CVD systems in regards to the sticking coefficient (which is effectively equal to one since no neutral species are considered) and arrival angle distribution. It is its angle-dependent sputtering and redeposition action as well as its very directed ion deposition, at the very low pressures, that give HDPCVD its good filling and planarization abilities.

We did not discuss models for evaporation deposition. However, by using a very directed angular arrival distribution (n equal to 15 or higher in the $\cos^2\theta$ distribution) for a stationary substrate or a less directed distribution for a rotated substrate, and using a sticking coefficient equal to one, one could simulate the evaporation deposition processes using the equations given above for conventional PVD.

Based on these generalizations we can now understand why CVD depositions usually give more conformal coverage than PVD depositions in these low pressure systems. An isotropic arrival distribution may lead to better step coverage. More importantly, a lower sticking coefficient, allowing more desorption and redeposition occurrences, leads to better step coverage, and a lower sticking coefficient is usually more important than the arrival angle in obtaining good coverage. Whatever the arrival angle distribution, a low sticking coefficient can lead to good coverage (even in a deep hole) since the deposition species, wherever they land, will bounce around and deposit on all sides of a structure. Thus CVD generally gives much more conformal coverage than PVD.

Having said that CVD techniques generally deposit films with better step coverage, it should be reiterated that a low sticking coefficient and good step coverage are not always required. This is the case in applications like covering the bottoms of deep, narrow contact or via holes with materials like Ti or filling narrow spaces with SiO_2 or metal. While a low sticking coefficient can greatly help in filling or lining holes, such as with

CVD W or TiN, there is another approach. The idea behind collimated and ionized PVD techniques, and even in HDPCVD with its strong ion component, is to deposit material from the gas phase in a very vertical direction. In this case a narrow arrival angle distribution is desired, not a wide distribution associated with good step coverage, and a low sticking coefficient is not required. Step coverage per se may not be great, but the bottoms of vias will be covered. In addition, good filling of holes or gaps can be achieved by the use of directed ions and sputtering action—filling up the holes from the bottom and sputtering away any overhang that may develop. In addition, the resputtering may help with sidewall covering. Directed ions and sputtering are the keys to ionized PVD and HDPCVD. Of course, CVD techniques with very low sticking coefficients are also good at covering the bottoms of deep vias and filling holes and gaps. But viable CVD techniques have not been developed for some materials, and these other techniques have been shown to do a good job at achieving bottom coverage and gap filling.

9.5.2 Simulations of Deposition Using a Physically Based Simulator, SPEEDIE

We now describe a series of deposition simulations using the SPEEDIE simulator and the equations and ideas presented above. These simulations will be used to demonstrate and further explain some of the concepts covered earlier in this chapter. The main focus will be on issues of film coverage and filling the underlying topography.

The first example is a simple simulation of LPCVD deposition of SiO_2 following Eq. (9.53). We will use parameter values in some cases that may not be realistic for LPCVD deposition but are more appropriate for PVD deposition. We do this to illustrate some of the important concepts of the models.

The initial topography is a wide trench—3 μm wide and 1 μm deep—in underlying Si (Figure 9-46). 0.5- μm -thick SiO_2 layer is deposited over this topography. A $\cos^2\theta$ distribution for the arrival angle distribution [Eq. (9.43)] is used, with $n = 1$ in the first simulation. This corresponds to a uniform or isotropic distribution of arrival angles, that is, molecules coming from all directions equally. This would be the case when many collisions occur in the gas phase so that the mean free path is very short and the direction of molecules becomes randomized. As we mentioned above, an n value of 1 is used for most CVD depositions. The other important parameter is the sticking coefficient, S_s , and a value of 1 is used here. This means that there is no reemission or redeposition of depositing species once they strike the surface. This is unrealistic for most CVD processes, as discussed earlier, but we will use it here for illustrative purposes. Later we will use lower values for S_s .

Figure 9-46(a) shows the SPEEDIE simulation of this simple LPCVD simulation. The different contour lines of the deposited SiO_2 represent different time intervals, which makes the deposition process easier to follow. It is apparent that the thickness is reduced on the sidewalls and corners by about 50% or more, even for an isotropic arrival angle distribution. This is because the species are arriving from one side only to each wall (a 90° maximum view angle on the side versus 180° on top). Also the sidewall deposition is sloped inward, which could cause problems for the deposition of subsequent layers.

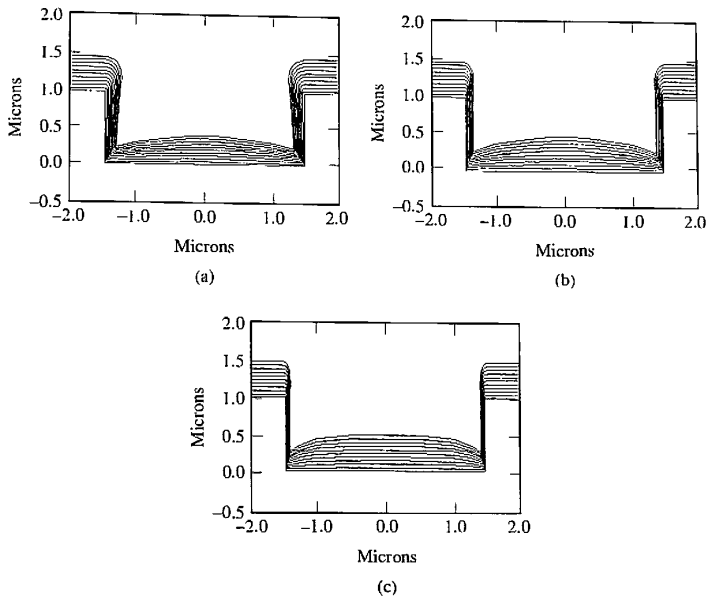


Figure 9-46 SPEEDIE simulations for LPCVD deposition of SiO_2 with $S_r = 1$ (which is more typical of PVD than LPCVD) and varying values of n , the arrival angle distribution factor: (a) $n = 1$; (b) $n = 3$; (c) $n = 10$. Worse step coverage results as n increases (the arrival angle distribution narrows).

The next simulation, shown in Figure 9-46(b) is the same as the previous one, except that the angular distribution of arrival atoms is changed. The value of n used in the $\cos^n\theta$ distribution is 3 here, compared to 1 in the previous simulation. This means that the arrival angle distribution is narrower, with more material coming in closer to the vertical direction. This might be the case when the gas pressure is lower and fewer gas-phase collisions occur, resulting in a longer mean free path. There is even less material deposited on the side walls and corners, as expected. Figure 9-46(c) shows a similar simulation, except with $n = 10$. (Again, this would be unrealistic for CVD processes, but is still illustrative since the same basic model is used for PVD where n can vary greatly—from close to 1 to very large values.) Here the sidewalls receive much less deposition. If the material deposited were a metal conductor, one can see that a narrow arrival distribution could cause problems with conduction of the interconnect line when it travels over steep topography. From these first three simulations we demonstrate the importance of the arrival angle distribution.

In the next example a narrower trench is utilized in the underlying layer (Figure 9-47). This corresponds to a higher aspect ratio (about 1.25 in this example compared to about 0.33 in the previous one). The same model (LPCVD) and parameters are used as

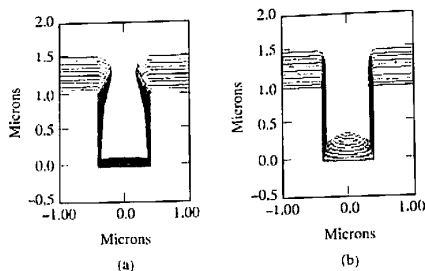


Figure 9-47 SPEEDIE simulations for LPCVD deposition of SiO₂ with $S_s = 1$ (again, more typical of PVD than CVD) and varying values of n , the arrival angle distribution factor, but for a narrower trench: (a) $n = 1$; (b) $n = 10$. Less deposition on sidewalls again results as n increases (the arrival angle distribution narrows), but better bottom filling is achieved due to more vertically directed deposition and less overhanging and shadowing at top corners.

in the previous simulations. Figure 9-47(a) shows the simulation results for the case of $n = 1$, a wide arrival distribution. Here the sidewalls and bottom of the trench have much less SiO₂ deposited on them as compared to Figure 9-46(a), which also had $n = 1$. Clearly geometry and shadowing effects came into play here. Those molecules arriving at large angles to the vertical cannot reach the inside of the trench. Instead they are blocked by the top corners of the trench. Another way to look at it is in terms of the view angle: For locations near the top of the trench the view angle is much larger than for points at the bottom, and more deposition occurs at the top corner and less at the bottom. This becomes even worse as the corners build up due to further deposition, leading to large overhangs. Eventually the top part will seal off, leaving a void in the trench. This angular-dependent blocking effect is called shadowing, for obvious reasons.

In Figure 9-47(b) the value of n is changed to 10. With the more directed incoming molecules, there is less deposition on the top part of the sidewalls. However, there is much more deposition on the bottom of the trench compared to when $n = 1$ [Figure 9-47(a)]. This is because there are more molecules coming vertically into the trench which are not blocked by the top corners or overhangs. Eventually the whole trench could be filled without the development of overhangs or voids.

We see from these examples where the sticking coefficient is unity, and therefore no reemission and redeposition occur, that geometry and line-of-sight considerations are very important. This would be the case for PVD depositions, where the sticking coefficients are usually assumed to be close to one.

We now change the value of the sticking coefficient, S_s , to see its effect. In the previous simulations it was set to 1. Next we set it to 0.1. This means that only one-tenth of the deposition species stay at each arrival point. The rest are reemitted and can be deposited elsewhere. In fact, each redeposited species also has a sticking coefficient

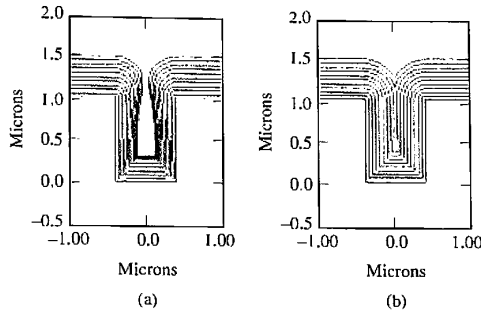


Figure 9-48 SPEEDIE simulations for LPCVD deposition of SiO_2 in a narrow trench with the same isotropic arrival angle distribution ($n = 1$) but with different values of the sticking coefficient: (a) $S_c = 0.1$, and (b) $S_c = 0.01$. Lowering the sticking coefficient from 1 to 0.1 results in much more conformal coverage. (a) has model parameters typical of CVD SiO_2 depositions, while (b) is typical of CVD W depositions.

of 0.1, so the effect is compounded. The net result is that the species bounce around a lot within the topographical features, giving much improved coverage.

Figure 9-48(a) shows the results for $S_c = 0.1$ and $n = 1$. In this case, these would be realistic parameters for LPCVD SiO_2 . Comparing this with Figure 9-47(a), we see that much more uniform deposition occurs, with much more material deposited on both the sidewalls and the bottom. In fact better overall coverage occurs here than in any of the previous simulations (where S_c was set to one in all of them and n was varied), indicating that the sticking coefficient has a larger impact on the profile than the arrival angle distribution. Lowering the sticking coefficient improves both the step or sidewall coverage and the filling ability, and a low sticking coefficient can compensate for orientation and shadowing effects to a very large degree. Figure 9-48(b) shows the simulation results for an even lower value of $S_c = 0.01$, such as would be typical of W or polysilicon depositions, again with $n = 1$. Even more uniform coverage and better filling result.

Figure 9-49 shows the results of some SPEEDIE simulations in which the angle of the sidewall is varied. For all three simulations shown, a sticking coefficient of 0.2 and an n value of 1 are used. One can see that better filling is obtained as the slope is changed from completely vertical (90°) to slightly sloped. Contact and via holes are often sloped slightly to enhance filling. Sloping a step also helps sidewall step coverage.

While the above series of simulations utilized the LPCVD model in SPEEDIE, the same results and trends would occur with the standard PVD model since the same equations are used. Only the values used for the sticking coefficient and the arrival angle distribution would be different. As discussed earlier, CVD processes have sticking coefficients in the 0.001 to 0.3 range and an n value of 1. PVD processes generally have sticking coefficients close to 1, while n is often greater than 1. The above simulations help illustrate the difference between CVD and PVD by varying those parameter values.

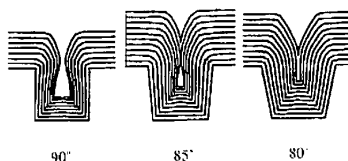


Figure 9-49 Results of SPEEDIE LPCVD simulations in which the sidewall angle is changed. Decreasing the angle from 90 to 80° greatly improves the trench filling. (After [9,37])

We next compare the methods of LPCVD and HDPCVD. In the latter case there is a strong ion component that not only leads to good filling of holes (since the effective n value of the species would be high) but also causes angle-dependent sputtering of the surface. This causes sloped surfaces to etch away faster than flat surfaces, which can lead to more planar surfaces when depositing material over nonplanar underlying topography. We shall now demonstrate this with SPEEDIE simulations. We first show a simulation using the LPCVD model. The initial topography is a small, flat rectangle on a flat surface, such as would be the case for a polysilicon or metal interconnect line on a flat underlayer (Figure 9-50). We now deposit LPCVD SiO_2 on top. A cosine arrival distribution with $n = 1$ is used, and a sticking coefficient of 0.1 is specified, typical for LPCVD SiO_2 using TEOS. The simulated results are shown in Figure 9-50(a). One sees the very conformal coverage due to the isotropic arrival and, more importantly, the low sticking coefficient [compare this to Figure 9-46(a), which has isotropic arrival but $S_s = 1$]. Conformal coverage is good for obtaining a uniform thickness everywhere, but bad for planarizing the structure.

Next we simulate the same situation, but with the HDP model and typical parameters for that method. A very large and very directed ion flux with a high n value in the arrival angle distribution is used, typical for HDP systems. Sputtering and redeposition of material occur, with a maximum sputtering yield at 62° relative to the surface. This causes sloped surfaces, such as at corners, to be sputtered faster than vertical or horizontal surfaces. The simulation results are shown in Figure 9-50(b). One sees the characteristic pyramid shape that starts to form due to these effects, which then gets reduced to a small bump on the surface. The final topography is more planar than the final topography in the standard LPCVD case. For wider features in the topography, wider and taller pyramidal shapes would form and a thicker HDP oxide layer would have to be deposited for planar topography to be achieved. For this reason, other planarizing techniques, such as CMP, are often used in conjunction with HDP depositions to completely planarize the topography over all the features.

The ability of HDP deposition to fill spaces and to avoid the formation of voids, usually the main reasons for using HDPCVD, are demonstrated in the next two simulations. First in Figure 9-50(c) we show LPCVD deposition in a trench, similar to that described earlier in Figure 9-48(a). The sticking coefficient is still 0.1 and the angular distribution factor, n , is still 1, but the maximum deposition flux is increased a little more to try to fill the space. One sees that a void is formed in the narrow trench, even with a fairly small sticking coefficient. While low sticking coefficients of CVD processes help with filling, the shadowing effects are still important and voids can still form for high-aspect

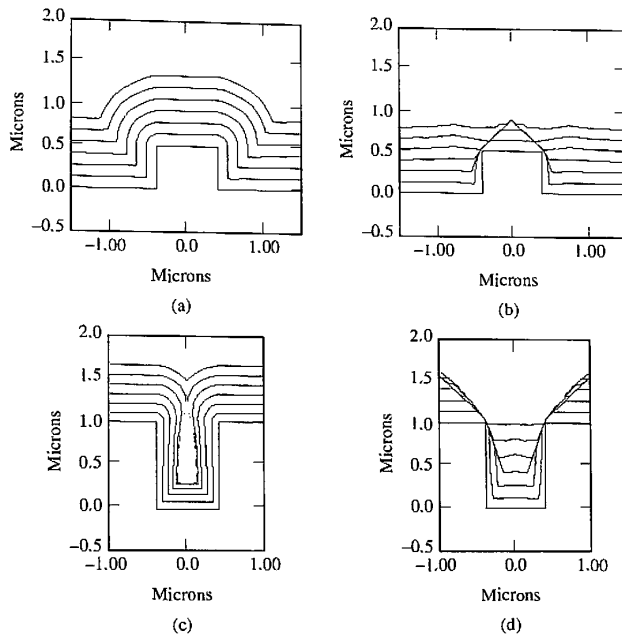


Figure 9-50 SPEEDIE simulations comparing LPCVD and HDPCVD depositions. (a) LPCVD deposition of SiO_2 over rectangular line. (b) HDPCVD deposition, with directed ionic flux and angle-dependent sputtering, over rectangular line showing much more planar topography. (c) LPCVD deposition in trench, showing void formation. (d) HDPCVD deposition in trench, showing much better filling.

ratio structures. The use of standard PECVD oxide deposition can help with the filling of gaps and spaces. With its directional ion component [Eq. (9.54)] which can be approximately equal in magnitude to the purely chemical component for oxide deposition [9.34], a little better filling of the gap occurs as compared to LPCVD deposition. But a void still results for this aspect ratio gap (though slightly higher up in the gap) and gets worse as the aspect ratio increases (see Figure 9-3). Using TEOS/O_3 films or the dep/etch/dep method can also improve the filling.

In Figure 9-50(d) we show how HDP deposition completely fills the gap, doing the same simulation with the trench structure but with the HDP model. No voids are formed. This is due to a combination of the highly directed ion flux component of deposition and the angle-dependent sputtering action. The deposition is very directed, and any overhang that does develop on the top corners is sputtered away, preventing any shadowing due to an overhang. Redeposition of sputtered material also helps with cov-

erage and preventing voids. This is obviously a very attractive technique for filling narrow spaces with dielectric material. Simulations show that void-free depositions can result even for aspect ratios of 3:1 and higher [9.4]. Actual SEM images of HDP deposition of SiO_2 were shown earlier in Figure 9-34 illustrating the filling and planarizing abilities of HDP deposition.

We next describe some PVD simulations. The standard PVD deposition SPEEDIE model uses Eq. (9.55). This is identical to that used for LPCVD, but usually different values for the sticking coefficient and the arrival angle distribution are used, as discussed earlier. In the first simulation we deposit Al onto a very narrow SiO_2 via structure with an aspect ratio of about 3, indicative of more advanced technology. We use parameter values more typical of PVD (sputter deposition) here. For S_c , we use 1. For the arrival angle distribution, we assume a $\cos^n\theta$ distribution with $n = 4$ (at the high end for standard sputter depositions), giving a more directed incoming flux than LPCVD systems. In this case, we specify that the hole is a via (a cylinder) rather than a trench, as was the case in the LPCVD simulations shown earlier. SPEEDIE calculates slightly different normalized direct fluxes because of this, and a via is generally harder to fill than a trench. The simulation results are shown in Figure 9-51(a). Very little Al is deposited inside the narrow via. Even though the flux is somewhat more directed than in LPCVD systems, the shadowing effect and the absence of any reemission (or resputtering) and redepositing of material allow very little material to be deposited in and on the bottom of such high-aspect ratio vias.

We next simulate ionized PVD, or IPVD, of Al. We assume that 80% of the Al atoms being deposited are ionized. For the un-ionized atoms, we assume a $\cos^n\theta$ distribution with $n = 4$. But for the ionized Al, we use a $\cos^n\theta$ distribution with n equal to 80. Eq. (9.57) is used for the local deposition rate, similar to the HDPCVD case [Eq. (9.58)]. We use typical parameter values for PVD, such as a sticking coefficient of 1. The same sputtering yield dependence is used as was used in the HDPCVD simulation, that is, with a

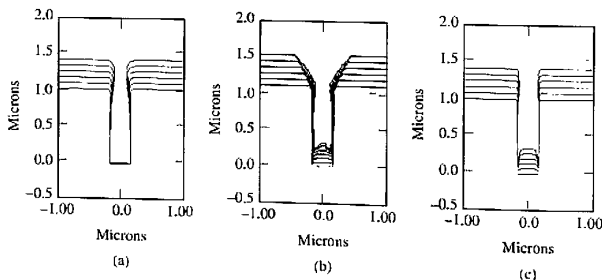


Figure 9-51 SPEEDIE simulations for sputtering (PVD) and ionized sputtering (IPVD) deposition of Al. (a) conventional PVD deposition with $S_c = 1$ and $n = 4$. (b) IPVD deposition with 80% of Al atoms being ionized (with near vertical arrival angles) and angle-dependent sputtering yield, showing much better filling of narrow via. (c) IPVD with no resputtering.

maximum sputtering yield at 62° . The simulation is shown in Figure 9-51(b). Two differences from the standard PVD simulation are evident. First, the bottom of the via gets filled much more than the standard PVD case. This is due to the very narrow (and vertical) ion angle distribution and is the reason why ionized PVD is an attractive alternative to standard PVD for some applications, such as for Ti or TiN deposition in today's deep contacts and vias. The second difference is the sloped sides at the top corners. This is due to the resputtering action, which not only bevels the corners but also redeposits material elsewhere. Resputtering from the bottom also occurs.

To see the effect of the resputtering and redeposition, another simulation is done. Everything is the same as in the first IPVD case except that the sputtering is turned off. The results are shown in Figure 9-51(c). With the very directed Al ion flux, a good amount of Al gets deposited down into the hole. But without the sputtering, the top corners are no longer sloped. The shadowing from the top corners and the lack of redeposition result in virtually no coverage on the sides of the via. Coverage on the side of a via or trench is often important, for example when depositing a TiN liner in a via, and the sputtering and redeposition action in IPVD is seen to help with this, especially at higher ion energies [9.14, 9.35].

In the final SPEEDIE simulations, High-Temperature PVD (or HTPVD) of Al in a trench is simulated. Eq. (9.56) is used in the HTPVD model. This includes the flux term that is due to surface diffusion driven by surface energy reduction, causing reflow during the deposition. For comparison we first show a standard PVD deposition into a trench, equivalent to setting the surface diffusivity in the HTPVD model to zero, in Figure 9-52(a). Next we do a HTPVD deposition with the deposition temperature specified to be 400°C , which determines the surface diffusion coefficient. The surface diffusivity and energy are taken from [9.38]. The resulting SPEEDIE simulated profile is shown in Figure 9-52(b).

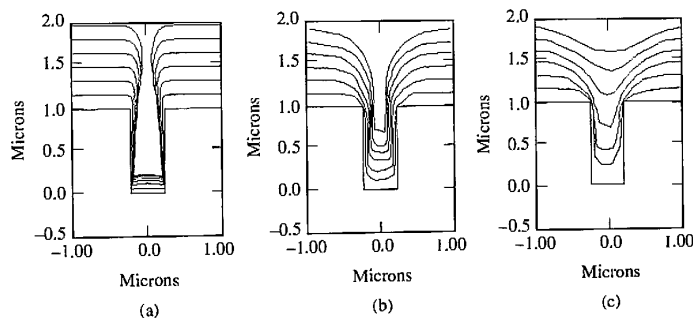


Figure 9-52 SPEEDIE simulations for High-Temperature PVD (HTPVD). (a) Standard PVD deposition with $S_c = 1$, $n = 4$, for comparison, showing poor filling. (b) HTPVD with surface diffusion at a deposition temperature of 400°C . Better filling is seen with the HTPVD. (c) HTPVD again but at a higher deposition temperature of 550°C , showing even more reflow during deposition than in (b), and even better filling and a smoother topography.

One can see that with the surface diffusion, much better filling of the hole and better step or sidewall coverage occur. Next we raise the deposition temperature even more to 550°C, resulting in a higher surface diffusivity, and the results are shown in Figure 9-52(c). The hole is completely filled, and much better planarization of the metal layer results. One can see from these simulations the advantages of high-temperature PVD deposition. More simulations involving surface diffusion driven reflow will be presented in Chapter 11. In that case, reflow of the metal after deposition will be simulated.

9.5.3 Other Deposition Simulations

We now describe some simulations done with a commercially available simulation tool. Here we use Silvaco's ATHENA simulator (using the ELITE topography tool) to simulate deposition of CVD SiO₂ and sputtered Al. While most of Silvaco's front-end simulation models for diffusion, oxidation, and so forth, are directly based on SUPREM IV models, the deposition and etch models used here and in the next chapter from ATHENA (many based on models originally in the SAMPLE simulator [9.39]) are more empirical in nature than SPEEDIE's models, and more user friendly. For example, instead of the user specifying the various fluxes, the user specifies the deposition rate and various parameters to obtain the shape and type of deposition profile.

First the ATHENA CVD deposition model is illustrated. Here the deposition rate, time, and step coverage are the input parameters. In the first simulation where we deposit oxide into a trench, the step coverage parameter is designated to be 1. This means that the same thickness is deposited everywhere. The resulting profile is shown in Figure 9-53(a). Next the step coverage is set to 0.2 instead of 1. The deposition rate, R , for the CVD model is given by:

$$R = \text{dep.rate}((1 - \text{step.cov})\cos\theta + \text{step.cov}) \quad (9.59)$$

where θ is the angle between the surface segment and the horizontal. For a step.cov term less than 1, this results in the full deposition rate on horizontal surfaces (where $\theta = 0$),

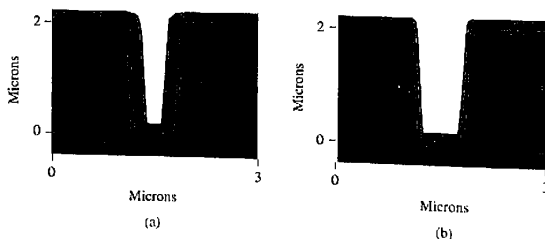


Figure 9-53 Silvaco's ATHENA [9.40] topography simulation of SiO₂ CVD deposition in a via for different values of the step coverage parameter, step.cov , selected by the user: (a) step coverage parameter = 1; (b) step coverage parameter = 0.2.

but less deposition on nonhorizontal surfaces. The result for the step coverage parameter equaling 0.2 is shown in Figure 9-53(b), exhibiting nonconformal coverage. Similar trends in step coverage using SPEEDIE were obtained by varying the value of the sticking coefficient in Figures 9-47(b) and 9-48(b).

The next example simulates the deposition of sputtered or evaporated Al using one of the Silvaco PVD models, the hemispheric deposition model. Here the incoming material flux arrives within a range of directions, coming from an extended, but limited, source such as in a planar sputtering system. The two angles designated by the user are the limits within which the flux is uniformly distributed. In this first case, the flux arrives at angles from 10 to -10 degrees from the normal to the substrate. The growth rate at each point is calculated after [9.41] by

$$R = C(\cos\omega_1 - \cos\omega_2)i + C(\sin\omega_1 - \sin\omega_2)j \quad (9.60)$$

where C is the deposition rate of an unshadowed surface normal to the flux direction, ω_1 and ω_2 are the upper and lower bounds of angles, and i and j are the unit vectors in the x and y directions, respectively. Figure 9-54(a) shows the simulation for such a deposition. One can see the directed deposition with most of the film being deposited on the horizontal surfaces. This is different from the ATHENA CVD model with a low step coverage value shown in the previous example in that now shadowing, overhangs, and cusping are observed on the corners. Figure 9-54(b) shows the result when the upper and lower angles are 45 and -45 degrees. More uniform deposition is observed, with more film deposited on the sidewalls, and less on the bottom of the trench. This model is similar to the SPEEDIE PVD model in which the distribution of arrival angles is controlled by the n term in the $\cos^n\theta$ distribution. However, unlike the SPEEDIE model, the sticking coefficient cannot be specified. The upper and lower bounds of angles used will depend on machine specifications, such as source and wafer geometry and characteristics, and operating pressure.

With this model one could also simulate asymmetric depositions such as might occur at the edge of a wafer in an evaporation system with a small source or in a sputtering

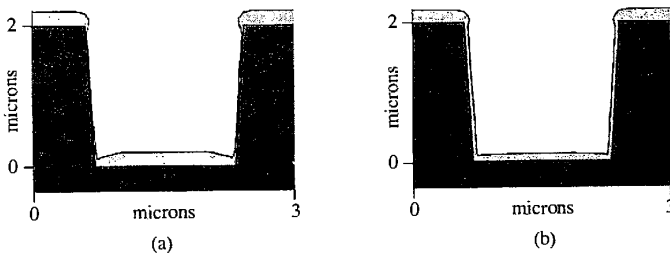


Figure 9-54 Silvaco's ATHENA [9.40] topography simulation of Al PVD deposition in a wide trench using the hemisphere model for different ranges of arrival angles: (a) the arrival angles are from -10 to $+10^\circ$; (b) the arrival angles are from -45 to $+45^\circ$. The wider angle range results in better step coverage, similar to a low n value in SPEEDIE's arrival angle distribution function.

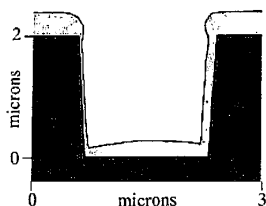


Figure 9-55 Silvaco's ATHENA [9.40] topography simulation of Al PVD deposition in a wide trench using the hemisphere model for asymmetric arrival angles: -25° to $+45^{\circ}$. Asymmetric deposition is seen, as might be the case for the right side edge of a wafer in a PVD system with a small source or target.

system with a relatively small target, such as in a long-throw collimation system. In that case the absolute value of the two angles in the hemispheric deposition model could be different. Figure 9-55 shows an ATHENA simulation in which the upper angle is 45° and the lower angle is -25° . Asymmetric deposition is clearly seen in the profile.

The advantages of these models are their ease of use, as well as the ease of calibrating them. For a certain piece of deposition equipment, a given material to be deposited, and specific deposition conditions, one matches the profiles of some test structures to the simulations by varying the model parameters. One can then build up a library of deposition equipment, materials, and normal deposition conditions to use in simulations. Model parameter libraries of various commonly used pieces of deposition and etch equipment with model parameters for various films and normal operating conditions are often included with these simulators [9.40]. One can easily use these models to optimize the deposition process without worrying about all the details of the process.

The advantage of the SPEEDIE-type models is that they are more physically correct and the parameters are more closely linked to actual processes and mechanisms. One inputs the actual flux information in terms of flux rate, type, angular distribution. In addition one specifies physical properties such as sticking coefficients, surface diffusivities, sputtering and reemission coefficients, and the like. The models then calculate the combined effects of the various processes and phenomena in a 2D/3D structure. Obviously, these types of models are more apt to be valid for conditions outside those for which they are calibrated and tested. It should be noted that Silvaco and other commercial simulation tool vendors such as Avant! [9.42] also now offer other deposition and etch models which are less empirical and more physically based, such as those using Monte Carlo calculations. Avant!, for example, in their commercially available TAURUS-TOPOGRAPHY deposition and etch simulator, offers both the simpler, more empirical models as well as models and algorithms similar to those in SPEEDIE involving actual direct and indirect flux calculations. Silvaco has developed such models as well.

Models have also been developed for deposition processes that take into account the specific chemical reactions in CVD systems and the mechanisms for how ion bombardment enhances the deposition rate in plasma systems (see [9.43] to see this how all this has recently been done for HDPCVD of SiO_2 , which considers 167 gas-phase reactions and 96 surface reactions!). These, or parts of these models, are sometimes incorporated into physically based simulators to make them even more complete.

9.6 Limits and Future Trends in Technologies and Models

A major problem that process developers face as the technology advances is the ever-increasing aspect ratios of the features. Depositing into smaller, deeper holes and spaces will only get more difficult. The NTRS roadmap [9.44] projects that the aspect ratios of metal lines, the spaces between the lines, and the via and contacts holes will all continue to increase over the next generations. (This is primarily due to the metal thicknesses not shrinking as fast as the lateral dimensions.) This means that the spaces and holes will get smaller and relatively deeper. We saw in this chapter how good coverage and filling are more difficult as the aspect ratio increases. High-density plasma deposition systems have proven successful in filling spaces between metal lines with oxide, and continued use and development of these types of systems are likely. New processes such as dual damascene described in Chapter 11, could place less importance on filling spaces with dielectrics in the future. Most metal films are still being deposited by sputtering techniques. The trend has been to try to use CVD because better filling and conformal step coverage are often achieved by CVD. However there are still problems with CVD deposition of metals which must be solved. These include contamination (impurities, such as carbon, from the precursors mostly) and difficulty with alloys (Al with Cu, for example). A PECVD process for Ti using TiCl_4 for filling small holes has recently been announced by one of the equipment manufacturers. New sputter techniques such as ionized, collimated, and hot-sputter depositions have been developed for metal deposition in high aspect ratio features. Future processes may use electroplating and spin-on methods for metals and dielectrics respectively. A potentially new type of deposition involves layer by layer adsorption and reaction of reactant species on the surface, utilizing sequentially pulsed precursors and self-limiting reactions for example. Commonly called Atomic Layer Deposition, or ALD, this can result in extremely conformal coverage of very thin layers.

A trend in deposition equipment is the use of cluster tools. Here various deposition (and etch) steps are performed sequentially with the wafers passed automatically from one process chamber to another, without breaking vacuum. As a result multiple deposition and etch steps can be done without exposing the wafers to the atmosphere between steps. This minimizes exposure of the wafers to ambient particulate contamination and surface oxidation. It also minimizes wafer handling as well as increases throughput. Both PVD and CVD chambers can be connected sequentially, as well as plasma etch chambers. Most of these tools are single-wafer, as opposed to batch, systems. More control over each wafer is possible, but the deposition rates must be high since each wafer must be processed individually. Depositing films on one wafer at a time does give a little more flexibility in choosing deposition conditions. For example, one can operate in the mass transfer regime in CVD deposition since gas-phase mass transfer is easier to control in a single wafer system.

In regards to modeling and simulation, we have been mostly concerned in this chapter with topology. That is, we discussed models and simulations that predict the shape of the deposited films over nonplanar underlying topography—over steps or in vias for example. These have been useful in explaining some of the basic concepts of deposition and the different types of equipment. However, there are other important properties of

the deposited films, such as grain structure, composition, density, mechanical properties, and stress. It is important to model these physical and compositional properties as a function of deposition technique and conditions, as these will affect the film behavior and performance during subsequent processing and operation.

There have been deposition models and simulators developed for physical and compositional properties. An example is the SIMBAD/GROFILMS [9.45, 9.46] group that simulate as-deposited film structure. In SIMBAD, two-dimensional discs rather than atoms or molecules are used in simulating deposition, with each disc representing about 1000 atoms. In these Monte Carlo simulations, regions of dense, closely packed discs form and evolve into a columnar grain structure—the grains separated by the less dense grain boundaries. The exact structure depends on the amount of surface diffusion and flux shadowing, consistent with Thornton's zone model for microstructure evolution [9.47]. (For a good discussion of zone models for microstructure, see [9.48].) GROFILMS extends this modeling to include other processes that can occur during deposition such as wetting, reflow, grain boundary grooving, and crystal faceting. Many of these processes will be discussed in Chapter 11. Deposition simulators that include modeling the physical structure and composition of thin films will be increasingly important in the future.

Making the models more physically correct by including the chemistry and physics of specific deposition systems is important. The use of atomic-level models and simulators could also greatly help in developing more physically correct models. Simulations using Monte Carlo or Molecular Dynamics (MD) calculations, based on atomic-level, first principles models, have already revealed new insights into the deposition processes, as they have in oxidation, diffusion, and implantation modeling. The goal is to make these simulations more efficient and scalable to realistic dimensions and number of atoms. In addition they need to be linked to higher-level, continuum-type simulation tools to make their results more easily accessible.

9.7 Summary of Key Ideas

In this chapter we have examined how thin films are deposited as part of the fabrication of integrated circuits. Important issues in thin film deposition include the physical and chemical properties of the film, and coverage of nonplanar underlying topography, including step coverage and the filling of holes or spaces. Coverage and filling issues are especially important with the ever-decreasing lateral feature sizes, as well as with the increasing use of multilayer film structures.

The two main deposition techniques are Chemical Vapor Deposition (CVD) and Physical Vapor Deposition (PVD). CVD depositions are mostly done at elevated temperatures and are chemical reaction based. In the simple model for CVD presented, the deposition process is seen to be limited by the surface reactions or by mass transfer through a gas-phase boundary layer depending on the temperature, pressure, and system chemistry. At low pressures, mass transfer through the boundary layer is not a limiting step, and the surface reactions become rate limiting, allowing for fewer constraints on reactor geometry. However, shadowing by surface topographical features must still be considered to model coverage and fill profiles. In PVD, the processes are physically,

rather than chemically, based. Arrival angle distribution of the source material at the wafer surface is important. There are fewer gas-phase collisions between the source and the wafer and virtually no reactions on the surface, so the species can arrive more vertical to the wafer from the source and stick where they hit without desorbing and redepositing. Shadowing by topographical features can be very important in PVD depositions. Many different types of both CVD and PVD systems have evolved and are still being used for different applications and different types of films. Other deposition techniques include spin-on techniques for dielectrics and electrolytic deposition of metals.

Most simulation of thin film deposition is based on topography modeling, that is, predicting the shape of the deposited film over nonplanar topography. SPEEDIE is one such topography simulator for both CVD and PVD. Both direct and indirect fluxes of neutral and ionic species are used to model the deposition process in low pressure systems. The direct fluxes are modeled by the arrival angle distribution. Indirect fluxes are due to surface or near-surface processes, such as surface diffusion or reemission and resputtering of surface species with subsequent deposition elsewhere. The sticking coefficient is used to model the reemission process. CVD processes, used mostly for dielectrics, in general have a low sticking coefficient and isotropic arrival angle distribution, leading to good step coverage. A low sticking coefficient also leads to better filling of holes. PVD processes, used more for metals, have a high sticking coefficient. A narrower arrival angle distribution, more characteristic of PVD processes leads to worse coverage over a step but can result in better filling of the bottoms of holes. Concurrent sputtering and redeposition of material along with the directed deposition of ionized species can lead to good gap or hole filling of relatively high aspect ratio features in either PVD or CVD systems, such as in ionized sputtering or HDPCVD. Simulations show the importance of different components and different parameters, especially in terms of arrival angle, sticking coefficient, angle-dependent sputtering, and coverage and filling issues.

9.8 References

- [9.1]. F. C. Eversteyn, "Chemical-Reaction Engineering in the Semiconductor Industry," *Philips Res. Rep.*, vol. 29, p. 45, 1974.
- [9.2]. A. S. Grove, *Physics and Technology for Semiconductor Devices*, John Wiley & Sons, 1967, p. 12.
- [9.3]. T. J. Rogers, "Advanced Integrated Circuit Technology for Micropower Integrated Circuits," Ph.D. thesis, Stanford University, 1975.
- [9.4]. J. Li, "Topography Simulation of Intermetal Dielectric Deposition and Interconnection Metal Deposition Processes," Ph.D. thesis, Stanford University, 1996.
- [9.5]. J. Li, J. P. McVittie, J. Ferziger, K. C. Saraswat, and J. Dong, "Optimization of Intermetal Dielectric Deposition Module Using Simulation," *J. Vac. Sci. Tech. B*, vol. 13, p. 1867, 1995.
- [9.6]. M. Ohring, *The Materials Science of Thin Films*, Academic Press, 1992, p. 90.
- [9.7]. *Handbook of Thin Film Technology*, L. I. Maissel and R. Glang, eds., McGraw-Hill, 1970.
- [9.8]. M. Knudsen, "Die Verdampfung von Kristalloberflächen," *Ann. Physik*, vol. 52, p. 105, 1917.
- [9.9]. R. Siegel and J. R. Howell, *Thermal Radiation Heat Transfer*, 3d ed., Hemisphere Pub. Corp., 1992.
- [9.10]. S. A. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996, p. 284.

- [9.11]. B. N. Chapman. *Glow Discharge Processes*, John Wiley & Sons, 1980.
- [9.12]. M. A. Lieberman and A. J. Lichtenberg. *Principles of Plasma Discharges and Materials Processing*, John Wiley and Sons, 1994.
- [9.13]. A. A. Mayo, S. Hamaguchi, J. H. Joo, and S. M. Rossnagel, "Across-Wafer Nonuniformity of Long Throw Sputter Deposition," *J. Vac. Sci. Tech. B.*, vol. 15, p. 1788, 1997.
- [9.14]. S. Hamaguchi and S. M. Rossnagel, "Liner Conformality Ionized Magnetron Sputter Metal Deposition Process," *J. Vac. Sci. Tech. B.*, vol. 14, p. 2603, 1996.
- [9.15]. S. M. Rossnagel, "Ionized Magnetron Sputtering for Lining and Filling Trenches and Vias," *Semiconductor International*, February 1996, p. 99.
- [9.16]. G. A. Dixit, M. F. Chisholm, M. K. Jain, T. Weaver, L. M. Ting, S. Poarch, K. Mizobuchi, R. H. Havemann, C. D. Dobson, A. I. Jeffryes, P. J. Holverson, P. Rich, D. C. Butler, and J. Hems, "A Novel High Pressure Low Temperature Aluminum Plug Technology for Sub-0.5 Micron Contact/Via Geometries," *IEDM Technical Digest*, 1994, p. 105.
- [9.17]. T. Kamins, *Polycrystalline Silicon for Integrated Circuit Applications*, 2d ed., Kluwer Academic Publishers, 1998.
- [9.18]. P. Joubert, "The Effect of Low Pressure on the Structure of LPCVD Polycrystalline Si Films," *J. Electrochem. Soc.*, vol. 134, p. 2541, 1987.
- [9.19]. R. Falckenberg, E. Doering, and H. Oppolzer, "Surface Roughness and Grain Growth of Thin P-Doped Polycrystalline Si-Films," *Proc. of Fall 1979 Electrochem. Soc. Meeting* (Los Angeles, October 1979), abstract 570, p. 1492.
- [9.20]. H. J. Kim and C. V. Thompson, "Kinetic Modeling of Grain Growth in Polycrystalline Silicon Films Doped with Phosphorus or Boron," *J. Electrochem. Soc.*, vol. 135, p. 2312, 1988.
- [9.21]. M. M. Manduarh, K. C. Saraswat, and T. I. Kamins, "Phosphorus Doping of Low Pressure Chemically Vapor-Deposited Silicon Films," *J. Electrochem. Soc.*, vol. 126, p. 1019, 1979.
- [9.22]. S. Robles, K. Russell, M. Galiano, V. Siva, V. Kithcart, and B. C. Nguyen, "Gap Fill and Film Reflow Capability of Subatmospheric Chemical Vapor Deposited Borophosphosilicate Glass," *J. Electrochem. Soc.*, vol. 143, p. 1414, 1996.
- [9.23]. L. Q. Xia, E. Yieh, P. Gee, F. Campana, and B. C. Nguyen, "Process Characteristics for Sub-atmospheric Chemical Vapor Deposited Borophosphosilicate Glass and Effect of Carrier Gas," *J. Electrochem. Soc.*, vol. 144, p. 3208, 1997.
- [9.24]. E. Kondoh, Y. Kawano, N. Takeyasu, T. Katagiri, H. Yamamoto, and T. Ohta, "Interconnection Formation by Simultaneous Cu Doping in CVD Al (Al-Cu CVD)," *IEDM Technical Digest*, p. 277, 1993.
- [9.25]. D. Pramanik and V. Jain, "Barrier Metals for ULSI: Deposition and Manufacturing," *Solid State Technology*, January 1993, p. 73.
- [9.26]. R. Liu, "Metallization," in *ULSI Technology*, ed. C. Y. Chang and S. M. Sze, McGraw-Hill, NY, 1996, p. 388.
- [9.27]. J. Li, T. E. Seidel, and J. W. Mayer, "Copper-based Metallization in ULSI Structures," *MRS Bulletin*, vol. 19, no. 8, p. 15, 1994.
- [9.28]. Y. Shacham-Diamand, V. Dubin, and M. Angyal, "Electroless Copper Deposition for ULSI," *Thin Solid Films*, vol. 262, p. 93, 1995.
- [9.29]. SPEEDIE 3.0 Manual, Stanford University, 1995.
- [9.30]. K. Watanabe and H. Komiyama, "Micro/Macrocavity Method Applied to the Study of the Step Coverage Formation Mechanism of SiO₂ Films by LPCVD," *J. Electrochem. Soc.*, vol. 137, p. 1222, 1990.

- [9.31]. H. C. Wulu, K. C. Saraswat, and J. P. McVittie, "Simulation of Mass Transport for Deposition in Via Holes and Trenches," *J. Electrochem. Soc.*, vol. 138, p. 1831, 1991.
- [9.32]. J. C. Rey, L. Y. Cheng, J. P. McVittie, and K. C. Saraswat, "Monte Carlo Low Pressure Deposition Profile Simulations," *J. Vac. Sci. Tech. A*, vol. 9, p. 1083, 1991.
- [9.33]. L. Y. Cheng, J. P. McVittie, and K. C. Saraswat, "New Test Structure to Identify Step Coverage Mechanisms in CVD of Silicon Dioxide," *Appl. Phys. Lett.*, vol. 58, p. 2147, 1991.
- [9.34]. C. Y. Chang, "Experiments and Simulation of Plasma Deposition and Sputter Etching Processes," Ph.D. thesis, Stanford University, 1995.
- [9.35]. David Bang, "Modeling and Simulation of Physical Vapor Deposition Systems for VLSI Metallization," Ph.D. thesis, Stanford University, 1997.
- [9.36]. C. Y. Chang, J. P. McVittie, J. Li, K. C. Saraswat, S. E. Lassig, and J. Dong, "Profile Simulation of Plasma Enhanced and ECR Oxide Deposition and Sputtering," *IEDM Technical Digest*, p. 853, 1993.
- [9.37]. J. C. Rey, L. Y. Cheng, J. P. McVittie, and K. C. Saraswat, "Numerical Simulation of CVD Trench Filling Using a Surface Reaction Coefficient Model," *Proc. of Seventh Int'l. IEEE Multilevel Inter. Conf.*, Santa Clara, CA, June 1990, p. 425.
- [9.38]. T. S. Cale, M. K. Jain, D. S. Taylor, R. L. Duffin, and C. J. Tracy, "Model for Surface Diffusion of Aluminum-(1.5%) Copper during Sputter Deposition," *J. Vac. Sci. Tech. B*, vol. B11, p. 311, 1993.
- [9.39]. W. G. Oldham, An. R. Neureuther, C. Sung, J. L. Reynolds, and S. N. Nandgaonkar, "A General Simulator for VLSI Lithography and Etching Processes: Part II—Application to Deposition and Etching," *IEEE Trans. Elec. Dev.*, vol. ED-27, p. 1455, 1980.
- [9.40]. ATHENA is a process simulator from Silvaco Inc.
- [9.41]. I. A. Blech, "Evaporated Film Profiles Over Steps in Substrates," *Thin Solid Films*, vol. 6, p. 113, 1970.
- [9.42]. TAURUS-TOPOGRAPHY is a topography simulator from Avant! Inc.
- [9.43]. E. Meeks, R. S. Larson, P. Ho, C. Apblett, S. M. Han, E. Edelberg, and E. S. Aydil, "Modeling of SiO₂ Deposition in High Density Plasma Reactors and Comparisons of Model Predictions with Experimental Measurements," *J. Vac. Sci. Tech. A*, vol. 16, p. 544, 1998.
- [9.44]. "National Technology Roadmap for Semiconductors," SIA, 1997.
- [9.45]. S. K. Dew, T. Smy, and M. J. Brett, "Simulation of Elevated Temperature Al Metalization Using SIMBAD," *IEEE Trans. Elec. Dev.*, vol. 39, p. 1599, 1992.
- [9.46]. L. J. Frederick, S. K. Dew, M. Brett, and T. Smy, "Thin Film Microstructure Modelling Through Line-Segment Simulation," *Thin Solid Films*, vol. 266, p. 83, 1995.
- [9.47]. J. A. Thorton, "Influence of Apparatus Geometry and Deposition Conditions on the Structure and Topology of Thick Sputtered Films," *J. Vac. Sci. Tech.*, vol. 11, p. 666, 1974.
- [9.48]. M. Ohring, *The Materials Science of Thin Films*, Academic Press, 1992, p. 223 ff.

9.9 Problems

- 9.1. What would you expect would happen to the threshold voltage of a MOS transistor if the gate oxide were deposited using $\text{SiH}_4 + 2\text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2\text{O}$ rather than thermally grown?

- 9.2. What are the two commonly observed rate limiting steps in silicon epitaxial growth? Under what conditions do they normally dominate the overall deposition rate?
- 9.3. In an epitaxial deposition, under mass transfer limited conditions, is it more important to control the reactor temperature or the source gas composition in the gas stream to obtain reproducible results? Why?
- 9.4. In a reactor used for epitaxial growth, the wafers are normally placed flat on the susceptor, and epi grows on the top side only. If the same reactor were used to oxidize wafers, by introducing O_2 rather than SiH_4 (or another Si gas source), SiO_2 would grow on both sides of the wafer. Explain why SiO_2 grows on both sides and epi grows only on the top side.
- 9.5. For CVD deposition of a film, it is found that the mass transfer coefficient $h_G = 10 \text{ cm sec}^{-1}$ and the surface reaction rate coefficient $k_s = 1 \times 10^7 \exp(-1.9 \text{ eV}/kT) \text{ cm sec}^{-1}$. For a deposition at 900°C , which CVD system would you recommend using: (a) a cold-walled, graphite susceptor type; or (b) a hot-walled, stacked wafer type? Explain your answer.
- 9.6. Calculate the deposition rate for an LPCVD system with the same parameter values given in the example following Eq. (9.10), but at a reduced total pressure so that h_G is increased by 100 times. Assume that the partial pressure of the incorporating species, P_G , remains the same, and C_T decreases by the same factor as the total pressure.
- 9.7a. Plot the deposition rate (on a log scale) versus $1/T$ (Kelvin), for $600\text{--}1200^\circ\text{C}$, for a CVD system with the following parameter values:

$$h_G = 0.5 \text{ cm sec}^{-1}$$

$$k_s = 4 \times 10^6 \exp(-1.45 \text{ eV}/kT) \text{ cm sec}^{-1}$$

$$\text{Partial pressure of incorporating species} = 1 \text{ torr}$$

$$\text{Total pressure} = 1 \text{ atm}$$

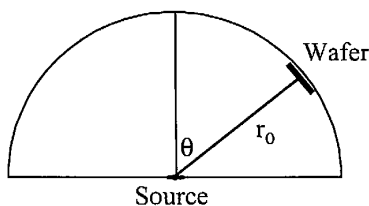
$$C_T/N = 1/10,000$$

Identify the reaction and mass transfer limited regimes.

- b. Redo the problem when the total pressure is decreased to 1 torr, so that h_G increases by 100 times. Assume that the partial pressure of the incorporating species remains the same, and C_T decreases by the same factor as the total pressure.
- 9.8. Show that placing the source in an evaporation system at the inside surface of a sphere facing the center, with the wafers also at the inside surface of the sphere facing the center, leads to uniform deposition wafer to wafer, assuming a small planar surface source.
- 9.9. Calculate the deposition rate for a small planar surface evaporation source in which $\theta_i = 30^\circ$, $\theta_k = 45^\circ$, the evaporation rate is $1 \times 10^{-3} \text{ gm sec}^{-1}$, the distance

from the source to the wafer is 5 cm, and the density of the material being deposited equal 5 gm cm^{-3} .

- 9.10. The new deposition engineer installed the company's new evaporation system. Hoping to get uniform depositions on all the wafers mounted on the inside of the spherical wafer holder, he installed the evaporation source crucible at the center of the sphere. If the evaporation source behaves like an ideal small-area planar source, what will be the deposition rate as a function of θ , as defined below? (Let the evaporation rate equal $3 \times 10^{-3} \text{ gm sec}^{-1}$, radius r_0 equal 10 cm, and the density of the material being deposited equal 10 gm cm^{-3} .) Sketch a plot of d versus θ for θ from -90° to $+90^\circ$, and specify the deposition rate at a point directly facing the planar source (at $\theta = 0^\circ$) and at $\theta = 90^\circ$. Will uniform depositions result?
- 9.11. Calculate the mean free path of a particle in the gas phase of a deposition system



and estimate the number of collisions it experiences in traveling from the source to the substrate in each of the cases below. Assume that in each case the molecular collisional diameter is 0.4 nm, the source-to-substrate distance is 5 cm, and that the number of collisions is approximately equal to the source-to-substrate distance divided by the mean free path.

- An evaporation system in which the pressure is 10^{-5} torr and the temperature is 25°C .
 - A sputter deposition system in which the pressure is 3 mtorr and the temperature is 25°C .
 - An LPCVD system in which the pressure is 1 torr and the temperature is 600°C .
 - An APCVD system in which the pressure is 1 atm and the temperature is 600°C .
- 9.12. a. Using schematic diagrams, show qualitatively how changing the gas pressure in a standard PVD system (using an appropriate S_c value for this type of system) can affect the deposition profile when depositing a material on a narrow trench structure.
- b. In a similar fashion, show how changing the sticking coefficient affects the deposition profile when depositing a material on topography with a step.

- 9.13. How does the ability to fill the bottom of a narrow trench using sputter deposition change as the target is moved further away from the wafer? Neglect any gas phase collision effects.
- 9.14. Explain how asymmetric depositions can occur on a wafer in a sputter deposition system. Asymmetric deposition means that thicker deposition occurs on one side of a feature (a step or side of a trench, for example) than the other. Also suggest a way to decrease the asymmetry. Use schematic diagrams in your explanations.
- 9.15. What value of n in the arrival angle distribution is desired for good step coverage over a step in topography? For good bottom filling of a via? Explain with schematic diagrams. What value of S_c , the sticking coefficient, do you want in each case? Explain with diagrams.
- 9.16. We have seen how conformal coverage in low pressure systems results from a low sticking coefficient, which causes multiple reemission and redeposition processes within the feature topography. Describe how in atmospheric CVD systems, gas phase collisions by the reactants within the topographical features can lead to good coverage and filling.