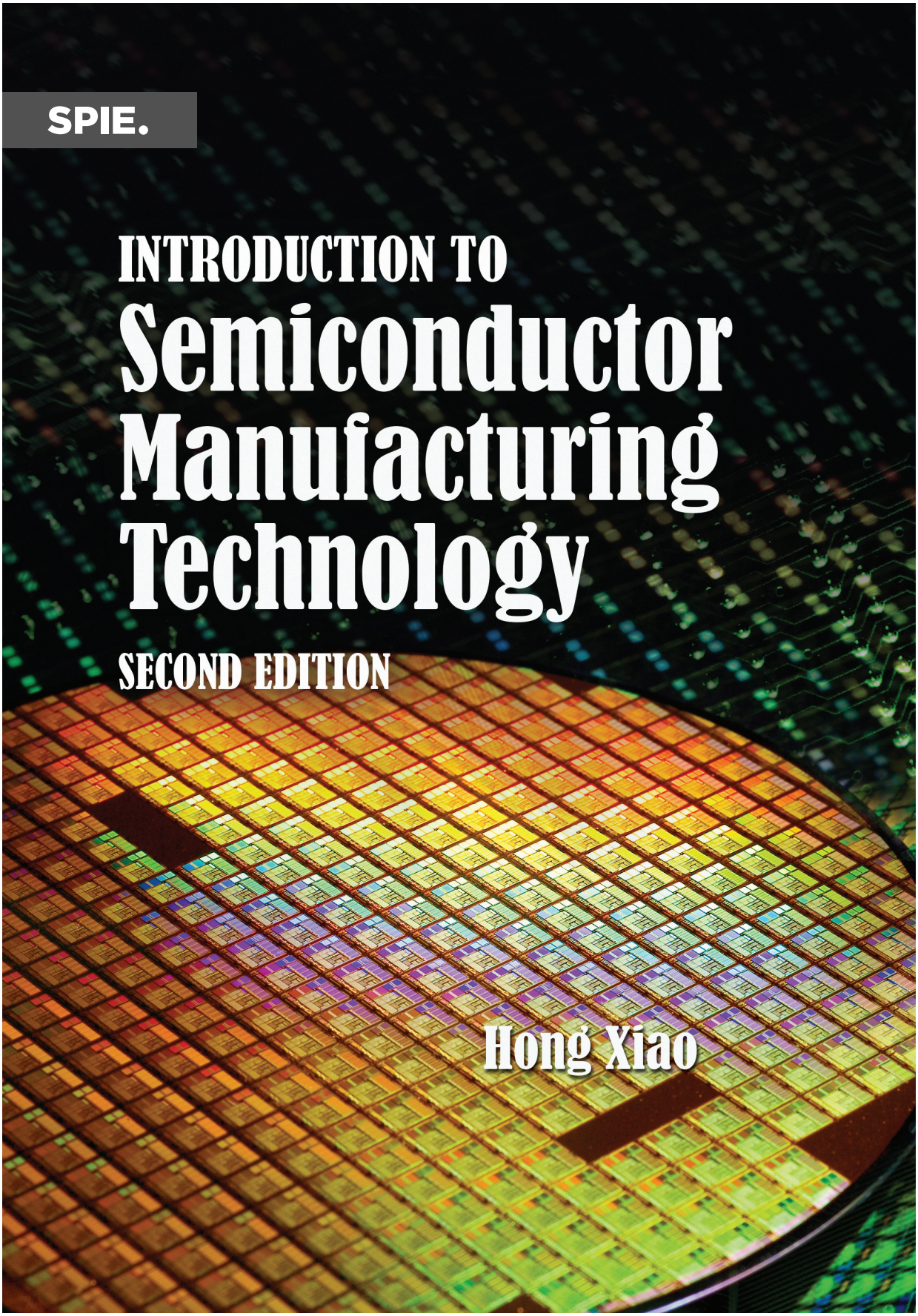


TSMC EXHIBIT 1020 (Part 1 of 3)

SPIE Terms of Use: This SPIE eBook is DRM-free for your convenience. You may install this eBook on any device you own, but not post it publicly or transmit it to others. SPIE eBooks are for personal use only. For details, see the SPIE Terms of Use. To order a print version, visit SPIE.



INTRODUCTION TO
Semiconductor Manufacturing Technology
SECOND EDITION

INTRODUCTION TO Semiconductor Manufacturing Technology

SECOND EDITION

Hong Xiao

SPIE
PRESS

Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Xiao, Hong.
Introduction to semiconductor technology / Hong Xiao. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-8194-9092-6
1. Semiconductors–Design and construction. 2. Semiconductor industry. I. Title.
TK7871.85.X53 2012
621.3815'2–dc23

2012013231

Published by

SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: +1 360.676.3290
Fax: +1 360.647.1445
Email: Books@spie.org
Web: http://spie.org

Copyright © 2012 Society of Photo-Optical Instrumentation Engineers (SPIE)

All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means without written permission of the publisher.

The content of this book reflects the work and thought of the author(s). Every effort has been made to publish reliable and accurate information herein, but the publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Printed in the United States of America.
First printing



This textbook is dedicated to my wife, Liu Huang, and my sons, Jarry Xiao and Colin Xiao.

Contents

Preface to the First Edition ..... **xxi**

Preface to the Second Edition ..... **xxiii**

List of Acroynms ..... **xxv**

Chapter 1 Introduction **1**

1.1 Brief History of Integrated Circuits 2
1.1.1 First transistor 2
1.1.2 First integrated circuit 3
1.1.3 Moore's law 5
1.1.4 Feature and wafer size 5
1.1.5 Definition of the integrated circuit technology node 10
1.1.6 Moore's law or the law of more 11
1.2 Brief Overview of Integrated Circuits 12
1.2.1 Manufacturing materials 12
1.2.2 Processing equipment 12
1.2.3 Metrology tools 13
1.2.4 Wafer manufacturing 13
1.2.5 Circuit design 14
1.2.6 Mask formation 16
1.2.7 Wafer processing 19
1.3 Summary 19
1.4 Bibliography 20
1.5 Review Questions 21

Chapter 2 Introduction to Integrated Circuit Fabrication **23**

2.1 Introduction 23
2.2 Yield 25
2.2.1 Definition of yield 25
2.2.2 Yield and profit margin 26
2.2.3 Defects and yield 27
2.3 Cleanroom Basics 28
2.3.1 Definition of a cleanroom 29

2.3.2 Contamination control and yield ... 30
2.3.3 Basic cleanroom structure ... 32
2.3.4 Basic cleanroom gowning procedures ... 33
2.3.5 Basic cleanroom protocols ... 35
2.4 Basic Structure of an Integrated Circuit Fabrication Facility ... 36
2.4.1 Wafer processing area ... 37
2.4.1.1 Wet bay ... 37
2.4.1.2 Diffusion bay ... 38
2.4.1.3 Photo bay ... 40
2.4.1.4 Etch bay ... 40
2.4.1.5 Implant bay ... 42
2.4.1.6 Thin-film bay ... 42
2.4.1.7 Chemical mechanical polishing bay ... 44
2.4.2 Equipment area ... 45
2.4.3 Facility area ... 45
2.5 Testing and Packaging ... 46
2.5.1 Die testing ... 47
2.5.2 Chip packaging ... 47
2.5.3 Final test ... 52
2.5.4 3D packaging ... 53
2.6 Future Trends ... 53
2.7 Summary ... 55
2.8 Bibliography ... 56
2.9 Review Questions ... 56

Chapter 3 Semiconductor Basics ... 59

3.1 What Is a Semiconductor? ... 59
3.1.1 Bandgap ... 59
3.1.2 Crystal structure ... 61
3.1.3 Doping semiconductor ... 61
3.1.4 Dopant concentration and conductivity ... 63
3.1.5 Summary of semiconductors ... 64
3.2 Basic Devices ... 64
3.2.1 Resistor ... 64
3.2.2 Capacitor ... 66
3.2.3 Diode ... 69
3.2.4 Bipolar transistor ... 70
3.2.5 Metal-oxide-semiconductor field effect transistor ... 72
3.3 Integrated Circuit Chips ... 74
3.3.1 Memory ... 75
3.3.1.1 Dynamic random access memory ... 75
3.3.1.2 Static random access memory ... 75
3.3.1.3 Erasable programmable read-only memory, electric erasable programmable read-only memory, and flash ... 76

3.3.2 Microprocessor ... 78
3.3.3 Application-specific integrated circuits ... 79
3.4 Basis Integrated Circuit Processes ... 79
3.4.1 Conventional bipolar transistor process ... 80
3.4.2 p-Channel metal-oxide-semiconductor process (1960s technology) ... 81
3.4.3 n-Channel metal-oxide-semiconductor process (1970s technology) ... 82
3.5 Complementary Metal-Oxide Semiconductor ... 82
3.5.1 Complementary metal-oxide-semiconductor circuit ... 85
3.5.2 Complementary metal-oxide-semiconductor circuit process (1980s technology) ... 87
3.5.3 Complementary metal-oxide-semiconductor process (1990s technology) ... 89
3.6 Technology Trends after 2000 ... 89
3.7 Summary ... 90
3.8 Bibliography ... 92
3.9 Review Questions ... 93

Chapter 4 Wafer Manufacturing, Epitaxy, and Substrate Engineering ... 95

4.1 Introduction ... 95
4.2 Why Silicon? ... 96
4.3 Crystal Structures and Defects ... 97
4.3.1 Crystal orientation ... 97
4.3.2 Crystal defects ... 99
4.4 Sand to Wafer ... 100
4.4.1 Crude silicon ... 100
4.4.2 Silicon purification ... 101
4.4.3 Crystal pulling ... 102
4.4.3.1 Czochralski method ... 102
4.4.3.2 Floating zone method ... 104
4.4.4 Wafering ... 105
4.4.5 Wafer finishing ... 107
4.5 Epitaxial Silicon Deposition ... 108
4.5.1 Gas phase epitaxy ... 110
4.5.2 Epitaxial growth process ... 111
4.5.3 Epitaxy hardware ... 112
4.5.4 Epitaxy process ... 114
4.5.5 Future trends of epitaxy ... 115
4.5.6 Selective epitaxy ... 116
4.6 Substrate Engineering ... 116
4.6.1 Silicon-on-insulator wafer ... 116
4.6.2 Hybrid orientation technology ... 118
4.6.3 Strained silicon wafer ... 119

4.6.4 Strained silicon-on-insulator wafer ... 119
4.6.5 Strained silicon in integrated circuit manufacturing ... 120
4.7 Summary ... 120
4.8 Bibliography ... 122
4.9 Review Questions ... 122

Chapter 5 Thermal Processes ... 125

5.1 Introduction ... 125
5.2 Thermal Process Hardware ... 126
5.2.1 Introduction ... 126
5.2.2 Control system ... 127
5.2.3 Gas delivery system ... 128
5.2.4 Loading system ... 128
5.2.5 Exhaust system ... 129
5.2.6 Processing tube ... 129
5.3 Oxidation ... 129
5.3.1 Applications ... 131
5.3.2 Preoxidation cleaning ... 135
5.3.3 Oxidation rate ... 137
5.3.4 Dry oxidation ... 139
5.3.5 Wet oxidation ... 142
5.3.6 High-pressure oxidation ... 145
5.3.7 Oxide measurement ... 146
5.3.8 Recent oxidation trends ... 149
5.4 Diffusion ... 150
5.4.1 Deposition and drive-in ... 152
5.4.2 Doping measurement ... 154
5.5 Annealing ... 156
5.5.1 Postimplantation annealing ... 156
5.5.2 Alloy annealing ... 157
5.5.3 Reflow ... 158
5.6 High-Temperature Chemical Vapor Deposition ... 159
5.6.1 Epitaxial silicon deposition ... 159
5.6.2 Selective epitaxial growth processes ... 160
5.6.3 Polycrystalline silicon deposition ... 160
5.6.4 Silicon nitride deposition ... 164
5.7 Rapid Thermal Processing ... 167
5.7.1 Rapid thermal annealing ... 168
5.7.2 Rapid thermal oxidation ... 171
5.7.3 Rapid thermal chemical vapor deposition ... 173
5.8 Recent Developments ... 174
5.9 Summary ... 176
5.10 Bibliography ... 177
5.11 Review Questions ... 177

Chapter 6 Photolithography 179

6.1 Introduction 179
6.2 Photoresist 181
6.3 Photolithography Process 184
6.3.1 Wafer cleaning 185
6.3.2 Preparation 187
6.3.3 Photoresist coating 188
6.3.4 Soft bake 193
6.3.5 Alignment and exposure 194
6.3.5.1 Contact and proximity printers 194
6.3.5.2 Projection printer 195
6.3.5.3 Stepper/scanner 196
6.3.5.4 Exposure light sources 199
6.3.5.5 Exposure control 199
6.3.6 Postexposure bake 201
6.3.7 Development 203
6.3.8 Hard bake 205
6.3.9 Metrology and defect inspection 206
6.3.10 Track-stepper integration system 211
6.4 Lithographic Technology Trends 212
6.4.1 Resolution and depth of focus 212
6.4.2 Mercury lamps and excimer lasers 215
6.4.3 Resolution enhancement techniques 216
6.4.3.1 Phase shift mask 217
6.4.3.2 Optical proximity correction 218
6.4.3.3 Off-axis illumination 219
6.4.4 Immersion lithography 221
6.4.5 Double, triple, and multiple patterning 222
6.4.6 Extreme-ultraviolet lithography 226
6.4.7 Nanoimprint lithography 228
6.4.8 X-ray lithography 229
6.4.9 Electron beam lithography 230
6.4.10 Ion beam lithography 231
6.5 Safety 232
6.6 Summary 233
6.7 Bibliography 234
6.8 Review questions 235

Chapter 7 Plasma Basics 237

7.1 Introduction 237
7.2 Definition of Plasma 237
7.2.1 Components of plasma 238
7.2.2 Generation of plasma 238

7.3 Collisions in Plasma ... 239
7.3.1 Ionization ... 239
7.3.2 Excitation-relaxation ... 240
7.3.3 Dissociation ... 241
7.3.4 Other collisions ... 243
7.4 Plasma Parameters ... 243
7.4.1 Mean free path ... 244
7.4.2 Thermal velocity ... 245
7.4.3 Magnetic field ... 247
7.4.4 Boltzmann distribution ... 248
7.5 Ion Bombardment ... 248
7.6 Direct-Current Bias ... 250
7.7 Advantage of Plasma Processes ... 253
7.7.1 Chemical vapor deposition ... 254
7.7.1.1 Plasma-enhanced chemical vapor deposition ... 254
7.7.1.2 Stress control ... 255
7.7.1.3 Chamber cleaning ... 255
7.7.1.4 Gap fill ... 256
7.7.2 Plasma etch ... 256
7.7.2.1 Etch profile control ... 256
7.7.2.2 Etch rate and selectivity ... 257
7.7.2.3 Endpoint ... 257
7.7.2.4 Chemical use ... 257
7.7.3 Sputtering deposition ... 257
7.8 Plasma-Enhanced Chemical Vapor Deposition and Plasma Etch Chambers ... 257
7.8.1 Processing differences ... 257
7.8.2 Chemical vapor deposition chamber design ... 258
7.8.3 Etch chamber design ... 258
7.9 Remote Plasma Processes ... 260
7.9.1 Photoresist strip ... 260
7.9.2 Remote plasma etch ... 260
7.9.3 Remote plasma cleaning ... 261
7.9.4 Remote plasma chemical vapor deposition ... 262
7.10 High-Density Plasma ... 262
7.10.1 Inductively coupled plasma ... 263
7.10.2 Electron cyclotron resonance ... 264
7.11 Summary ... 264
7.12 Bibliography ... 265
7.13 Review Questions ... 265

Chapter 8 Ion Implantation ... 267

8.1 Introduction ... 267
8.1.1 Brief history ... 267

8.1.2 Advantages of implantation 270
8.1.3 Applications 271
8.2 Ion Implantation Basics 272
8.2.1 Stopping mechanisms 272
8.2.2 Ion range 275
8.2.3 Channeling effect 278
8.2.4 Damage and annealing 281
8.3 Ion Implantation Hardware 284
8.3.1 Gas system 284
8.3.2 Electrical system 284
8.3.3 Vacuum system 285
8.3.4 Control system 285
8.3.5 Beam line system 285
8.3.5.1 Ion source 285
8.3.5.2 Extraction system 286
8.3.5.3 Mass analyzer 288
8.3.5.4 Postacceleration 289
8.3.5.5 Charge neutralization system 290
8.3.5.6 Wafer handler 292
8.3.5.7 Beam stop 294
8.4 Ion Implantation Process 294
8.4.1 Device applications 294
8.4.2 Other ion implantation applications 299
8.4.3 Processing issues 302
8.4.3.1 Wafer charging 302
8.4.3.2 Particle contamination 303
8.4.3.3 Elemental contamination 304
8.4.4 Process evaluation 305
8.4.4.1 Secondary ion mass spectroscopy 305
8.4.4.2 Four-point probe 305
8.4.4.3 Thermal wave 306
8.4.4.4 Optical-electrical sheet resistance measurement 308
8.5 Safety 308
8.5.1 Chemical hazards 308
8.5.2 Electrical hazards 309
8.5.3 Radiation hazards 310
8.5.4 Mechanical hazards 310
8.6 Recent Developments and Applications 310
8.7 Summary 312
8.8 Bibliography 312
8.9 Review Questions 313

Chapter 9 Etch 315

9.1 Introduction 315

9.2 Etch Basics ... 317
9.2.1 Etch rate ... 317
9.2.2 Uniformity ... 318
9.2.3 Selectivity ... 320
9.2.4 Profile ... 320
9.2.5 Etch bias ... 321
9.2.6 Loading effects ... 321
9.2.6.1 Macroloading ... 321
9.2.6.2 Microloading ... 321
9.2.7 Overetch ... 321
9.2.8 Residue ... 324
9.3 Wet Etch Process ... 325
9.3.1 Introduction ... 325
9.3.2 Oxide wet etch ... 325
9.3.3 Silicon etch ... 327
9.3.4 Nitride etch ... 328
9.3.5 Metal etch ... 329
9.4 Plasma (Dry) Etch ... 330
9.4.1 Introduction ... 330
9.4.2 Plasma review ... 330
9.4.3 Chemical, physical, and reactive ion etches ... 331
9.4.4 Etch mechanisms ... 333
9.4.5 Plasma etch chamber ... 335
9.4.6 Endpoint ... 340
9.5 Plasma Etch Processes ... 341
9.5.1 Dielectric etch ... 341
9.5.2 Single-crystal silicon etch ... 349
9.5.3 Polysilicon etch ... 353
9.5.4 Metal etch ... 357
9.5.5 Photoresist strip ... 358
9.5.6 Dry chemical etch processes ... 359
9.5.7 Blanket dry etch processes ... 360
9.5.8 Plasma etch safety ... 360
9.6 Process Trends ... 361
9.7 Recent Developments ... 362
9.8 Summary ... 365
9.9 Bibliography ... 365
9.10 Review Questions ... 366

Chapter 10 Chemical Vapor Deposition and Dielectric Thin Films ... 369

10.1 Introduction ... 369
10.2 Chemical Vapor Deposition ... 371
10.2.1 Chemical vapor deposition process description ... 372

10.2.2 Chemical vapor deposition reactor types ... 374
10.2.2.1 Atmospheric pressure chemical vapor deposition 374
10.2.2.2 Low-pressure chemical vapor deposition ... 375
10.2.2.3 Plasma-enhanced chemical vapor deposition ... 377
10.2.3 Chemical vapor deposition basics ... 377
10.2.3.1 Step coverage ... 377
10.2.3.2 Gap fill ... 380
10.2.4 Surface adsorption ... 382
10.2.4.1 Chemisorption ... 382
10.2.4.2 Physisorption ... 383
10.2.5 Chemical vapor deposition precursors and their adsorption 383
10.2.5.1 Sticking coefficient ... 386
10.2.6 Chemical vapor deposition kinetics ... 387
10.2.6.1 Chemical reaction rate ... 387
10.2.6.2 Surface-reaction-limited regime ... 388
10.2.6.3 Mass-transport-limited regime ... 389
10.2.6.4 Chemical vapor deposition reactor deposition regime ... 389
10.3 Applications of Dielectric Thin Films ... 390
10.3.1 Shallow trench isolation ... 391
10.3.2 Sidewall spacer ... 391
10.3.3 Interlayer dielectric-0 ... 392
10.3.4 Interlayer dielectric-1 and up ... 396
10.3.5 Passivation dielectrics ... 398
10.4 Dielectric thin-film characteristics ... 398
10.4.1 Refractive index ... 398
10.4.1.1 Ellipsometry ... 399
10.4.1.2 Prism coupler ... 401
10.4.2 Thickness ... 402
10.4.2.1 Color chart ... 402
10.4.2.2 Spectroreflectometry ... 404
10.4.2.3 Deposition rate ... 405
10.4.2.4 Wet etch rate ... 406
10.4.2.5 Shrinkage ... 406
10.4.2.6 Uniformity ... 407
10.4.3 Stress ... 408
10.5 Dielectric Chemical Vapor Deposition Processes ... 411
10.5.1 Thermal silane chemical vapor definition process ... 411
10.5.2 Thermal tetraethoxysilane chemical vapor deposition process ... 412
10.5.3 Plasma-enhanced chemical vapor deposition silane process ... 412
10.5.3.1 Passivation ... 413
10.5.3.2 Interlayer dielectric-0 barrier layer ... 415
10.5.3.3 Dielectric antireflective coating ... 415

10.5.4 Plasma-enhanced chemical vapor deposition tetraethoxysilane process ... 416
10.5.5 Dielectric etchback process ... 418
10.5.6 Ozone-tetraethoxysilane process ... 421
10.5.6.1 Ozonator ... 422
10.5.6.2 Ozone-tetraethoxysilane undoped silicate glass process ... 423
10.5.6.3 Ozone-tetraethoxysilane phosphosilicate glass and borophosphosilicate glass processes ... 424
10.6 Spin-On Glass ... 424
10.7 High-Density Plasma Chemical Vapor Deposition ... 426
10.8 Dielectric Chemical Vapor Deposition Chamber Cleaning ... 428
10.8.1 Radio-frequency plasma cleaning ... 428
10.8.2 Remote plasma cleaning ... 431
10.9 Process Trends and Troubleshooting ... 433
10.9.1 Silane plasma-enhanced chemical vapor deposition processing trends ... 433
10.9.2 Plasma-enhanced tetraethylorthosilicate trends ... 434
10.9.3 Ozone-tetraethoxysilane trends ... 435
10.9.4 Troubleshooting guide ... 437
10.10 Recent Developments ... 439
10.10.1 Low-κ dielectrics ... 441
10.10.2 Air gap ... 441
10.10.3 Atomic layer deposition ... 443
10.10.4 High-κ dielectrics ... 444
10.11 Summary ... 446
10.12 Bibliography ... 448
10.13 Review Questions ... 449

Chapter 11 Metallization ... 451

11.1 Introduction ... 451
11.2 Conducting Thin Films ... 454
11.2.1 Polysilicon ... 454
11.2.2 Silicides ... 454
11.2.3 Aluminum ... 455
11.2.4 Titanium ... 458
11.2.5 Titanium nitride ... 459
11.2.6 Tungsten ... 461
11.2.7 Copper ... 462
11.2.8 Tantalum ... 464
11.2.9 Cobalt ... 464
11.2.10 Nickel ... 465

11.3 Metal Thin-Film Characteristics 465
11.3.1 Thickness and deposition rate 467
11.3.2 Uniformity 469
11.3.3 Stress 470
11.3.4 Reflectivity 472
11.3.5 Sheet resistance 472
11.4 Metal Chemical Vapor Deposition 475
11.4.1 Introduction 475
11.4.2 Tungsten 477
11.4.3 Tungsten silicide 480
11.4.4 Titanium 481
11.4.5 Titanium nitride 481
11.4.6 Aluminum 483
11.5 Physical Vapor Deposition 483
11.5.1 Introduction 483
11.5.2 Evaporation processes 485
11.5.2.1 Thermal evaporation 485
11.5.2.2 Electron beam evaporation 486
11.5.3 Sputtering 487
11.5.4 Basic metallization processes 490
11.5.4.1 Degas 492
11.5.4.2 Preclean 492
11.5.4.3 Titanium physical vapor deposition 493
11.5.4.4 Titanium nitride physical vapor deposition 494
11.5.4.5 Aluminum–copper physical vapor deposition 495
11.6 Copper Metallization 497
11.6.1 Precleaning 497
11.6.2 Barrier layer 499
11.6.3 Copper seed layer 499
11.6.4 Copper electrochemical plating 500
11.6.5 Copper chemical vapor deposition 502
11.7 Safety 503
11.8 Summary 503
11.9 Bibliography 504
11.10 Review Questions 505

Chapter 12 Chemical Mechanical Polishing 507

12.1 Introduction 507
12.1.1 Overview 508
12.1.2 Definition of planarization 511
12.1.3 Other planarization methods 511
12.1.4 Necessity of chemical mechanical polishing 514
12.1.5 Advantages of chemical mechanical polishing 515
12.1.6 Applications of chemical mechanical polishing 516

12.2 Chemical Mechanical Polishing Hardware 519
12.2.1 Introduction 519
12.2.2 Polishing pad 519
12.2.3 Polishing head 521
12.2.4 Pad conditioner 521
12.3 Chemical Mechanical Polishing Slurries 522
12.3.1 Oxide slurry 523
12.3.2 Metal polishing slurry 525
12.3.3 Tungsten slurry 526
12.3.4 Aluminum and copper slurries 527
12.4 Chemical Mechanical Polishing Basics 528
12.4.1 Removal rate 528
12.4.2 Uniformity 529
12.4.3 Selectivity 529
12.4.4 Defects 530
12.5 Chemical Mechanical Polishing Processes 534
12.5.1 Oxide chemical mechanical polishing 534
12.5.2 Tungsten chemical mechanical polishing 535
12.5.3 Copper chemical mechanical polishing 537
12.5.4 Chemical mechanical polishing endpoint detection 538
12.5.5 Post-chemical-mechanical-polishing wafer cleaning 540
12.5.6 Process issues 542
12.6 Recent Developments 543
12.7 Summary 545
12.8 Bibliography 546
12.9 Review Questions 547

Chapter 13 Process Integration 549

13.1 Introduction 549
13.2 Wafer Preparation 550
13.3 Isolations 551
13.3.1 Blanket field oxide 551
13.3.2 Local oxidation of silicon 552
13.3.3 Shallow trench isolation 554
13.3.4 Self-aligned shallow trench isolation 557
13.4 Well Formation 557
13.4.1 Single well 557
13.4.2 Self-aligned twin wells 557
13.4.3 Twin wells 559
13.5 Transistor Formation 559
13.5.1 Metal gate process 559
13.5.2 Self-aligned source/drain process 560
13.5.3 Lightly doped drain 561
13.5.4 Threshold adjustment 563
13.5.5 Anti-punch-through 564

13.6 Metal-Oxide-Semiconductor Field-Effect Transistors with High-κ and Metal Gates ... 564
13.6.1 Gate-first process ... 565
13.6.2 Gate-last process ... 567
13.6.3 Hybrid high-κ metal gates ... 570
13.7 Interconnections ... 570
13.7.1 Local interconnections ... 570
13.7.2 Early interconnections ... 572
13.7.3 Aluminum alloy multilevel interconnections ... 573
13.7.4 Copper interconnections ... 574
13.7.5 Copper and low-κ dielectrics ... 576
13.8 Passivation ... 577
13.9 Summary ... 578
13.10 Bibliography ... 580
13.11 Review Questions ... 580

Chapter 14 Integrated Circuit Processing Technologies ... 581

14.1 Introduction ... 581
14.2 Complementary Metal-Oxide-Semiconductor Process Flow of the Early 1980s ... 581
14.3 Complementary Metal-Oxide-Semiconductor Process Flow with 1990s Technology ... 582
14.3.1 Comments ... 583
14.4 Complementary Metal-Oxide-Semiconductor Process Flow with Technology after 2000 ... 598
14.5 Complementary Metal-Oxide-Semiconductor Process Flow with Technology after 2010 ... 616
14.6 Memory Chip Manufacturing Processes ... 625
14.6.1 Dynamic random access memory process flow ... 627
14.6.2 Stacked dynamic random access memory process flow ... 629
14.6.3 NAND flash process ... 636
14.7 Summary ... 646
14.8 Bibliography ... 647
14.9 Review Questions ... 648

Chapter 15 Future Trends and Summary ... 649

15.1 Bibliography ... 655

Index ... 657

Preface to the First Edition

The semiconductor industry is developing rapidly with new technology introduced almost on a daily basis. The device feature size is shrinking continuously and the number of transistors on an integrated circuit (IC) chip is increasing rapidly, as predicted by Moore's law. Compared with only a decade ago, IC fabrication processing technology has become more complicated.

This book thoroughly describes the complicated IC chip manufacturing processes in a semiconductor fab, using minimum mathematics, chemistry, and physics. It covers the advanced technologies while keeping the contents simple and easy to understand for readers without science and engineering degrees. It focuses on the newest IC fabrication technologies and describes the older technologies to provide better understanding of the historical development. The processes chosen for the book are very close to those used in real fabs, especially process troubleshooting and process and hardware relations.

This book is intended for technical and college students who need an in-depth understanding of the technology as they prepare to find a job in the field. It is also intended as a reference book for engineering students and to provide a more realistic picture of the semiconductor industry. Industry operators, technicians, engineers, and personnel in sales, marketing, administration, and management can also benefit. This book can help them to learn more about their jobs, improve their troubleshooting and problem-solving skills, and raise their career development potential.

Chapter 1 briefly reviews the history of the semiconductor industry and describes semiconductor manufacturing processes. Chapter 2 introduces basic semiconductor fabrication including yield, cleanroom, semiconductor fab, and IC chip test and packaging. Chapter 3 gives a brief review of fabrication semiconductor devices, IC chips, and early technologies in semiconductor processing. Crystal structure, single-crystal silicon wafer manufacturing, and epitaxial silicon growth are described in Chapter 4. Chapter 5 lists and discusses thermal processes, including oxidation, diffusion, annealing, alloying, and reflow processes. Rapid thermal processes (RTPs) and conventional furnace thermal processes are discussed. Chapter 6 details the photolithography process. Fundamentals of plasmas used in semiconductor processing are covered in Chapter 7, which introduces plasma applications, DC bias, and plasma-process relations. Chapter 8 discusses the ion implantation process. Chapter 9 gives a detailed description of etch processes including wet and dry etches; chemical, reactive ion etch (RIE), and

physical etches; and patterned and blanket etch processes. Basic chemical vapor deposition (CVD) and dielectric thin-film deposition processes, including dielectric CVD processes, process trends, troubleshooting, and future trends are discussed in detail in Chapter 10. Chapter 11 covers metallization, metal CVD, and physical vapor deposition (PVD) processes. It also describes the copper metallization process. Planarization processes including chemical mechanical polishing (CMP) are discussed in Chapter 12. Chapter 13 discusses process integration. Chapter 14 diagrams CMOS process flows including an advanced CMOS process flow with copper and low-κ interconnection. Chapter 15 predicts the future development of the semiconductor industry.

Many people helped me to write this book. I especially appreciate the useful information provided by my current and former colleagues: Lou Frenzel, Thomas E. Thompson, Ole Krogh, Tony Shi, Alberto Quiñonez, Lance Kinney, Scott Bolton, and Steve Reedy. Many of my students helped me by proofreading and improving the book. I especially express my thanks to Wayne Parent, Jeffrey Carroll, Boyd Woods, and Ronald Tabery.

I would also like to thank the following reviewers for their valuable suggestions: Professor Dave Hata, Portland Community College; Professor Fred Lavender, Albuquerque Technical Institute; Professor Gene Stouder, Southwest Texas State University; Professor Bassam Matar, Glendale Community College; Professor Carlo Sapijaszko, DeVry Institute of Technology; Professor George Shaiffer, Pikes Peak Community College; Professor Val Shires, Gwinnett Technical Institute; and Professor Devinder Sud, DeVry Institute.

Hong Xiao
2001

Preface to the Second Edition

When the first edition was published in 2001, the leading-edge IC technology node was about 130 nm. Shortly after the publication of the first edition, I attended an international IC technology conference where 90 nm was the leading-edge technology node of IC manufacturing. Former Honda CEO Hiroyuki Yoshino gave a keynote speech in which he talked about ASIMO, the robot Honda introduced in 2000. At that time, ASIMO could understand some simple words and follow few verbal instructions to walk slowly and speak simple words. Mr. Yoshino envisioned that in the future, humanoids like ASIMO would be able to run, walk forward and backward, and navigate stairs. Not only would they be able to understand speech, but they could also understand the speaker's mood. They would be able to identify individuals and human emotions through facial-image recognition software. However, 90-nm technology was not sufficient to achieve these technologies, and Mr. Yoshino believed that 22-nm-technology IC chips would be needed.

A decade later, cutting-edge IC technology has reached 22 nm. An all-new ASIMO was introduced in 2011 that can run, dance, and use sign language. Whereas the older version of ASIMO is an "automatic machine" that needs an operator, the new ASIMO is an "autonomous machine," which means it can make its own decisions and actions based on environmental sensors. It has the intelligence to walk among a group of people without collision by adjusting its movement through the observation and prediction of others. It has the capability to recognize the voices of multiple people, and its image sensor is capable of facial recognition. Although the new ASIMO has made a huge improvement, it is still far from having the envisioned emotion and mood recognition capabilities; to achieve that, perhaps sub-10-nm-technology IC chips are needed.

There have been many changes in the semiconductor industry and its manufacturing technologies over the last decade. Although Intel still keeps the bulk of its IC manufacturing technology in North American fabs, the center of IC manufacturing has shifted to East Asia, in nations such as Taiwan, North and South Korea, and China. The IC manufacturing fabs in Europe, Japan, and North America are declining at an alarming rate.

The biggest challenge for IC-technology node scaling has always been patterning technology. As suggested by my predictions in the first edition, unforeseen innovations such as immersion lithography and multiple patterning

have extended the application of optical lithography and further delayed the development of next-generation lithography.

Many people helped me write this second edition. The book would not have been possible without the encouragement and support from my wife, Liu (Lucy) Huang. My elder son, Jarry Xiao, proofread several chapters. My current and former colleagues provided many useful suggestions; I would like to express my deepest appreciation to Paul MacDonald, Pierre Lefebvre, and Alan Liang.

Hong Xiao
October 2012

List of Acronyms

3MS	trimethylsilane or $(CH_3)_3SiH$
4MS	tetramethylsilane or $(CH_3)_4Si$
α-Si	amorphous silicon
AA	active area
AAPSM	alternating aperture PSM
ADC	automatic diameter control
ADI	after-development inspection
AFM	atomic force microscope
ALD	atomic layer deposition
ALU	arithmetic-logic unit
AMU	atomic mass unit
APCVD	atmospheric pressure CVD
AR	adsorption rate
ARC	antireflective coating
ASIC	application-specific integrated circuits
AttPSM	attenuated PSM
BARC	bottom antireflective coating
BEoL	back-end of the line
BL	bitline
BLC	bitline contact
BOE	buffered oxide etch
BOX	buried oxide
BPSG	borophosphosilicate glass
BSG	borosilicate glass
BST	$Ba_{0.6}Sr_{0.4}TiO_3$
BTBAS	bis-(tertiary-butylamino)silane] or $SiH_2[NH(C_4H_9)]_2$
bWL	buried wordline
C2C	cell-to-cell (inspection)
CB	contact to bit line
CBH	carborane or $C_2B_{10}H_{12}$
CD	critical dimension
CHO	cyclohexene oxide or $C_6H_{10}O$
CISC	complete instruction set computers
CMOS	complementary metal-oxide semiconductor
CMP	chemical mechanical polishing/planarization
CNT	carbon nanotube
CPU	central processing unit
CR	chemical reaction rate

CS	contact to source line
CVD	chemical vapor deposition
CZ	Czochralski (method)
D2D	die-to-die (inspection)
D2DB	die-to-database (inspection)
DEMS	diethoxymethylsilane or $C_5H_{14}Si$
DFM	design for manufacturing
DHF	diluted HF
DI	deionized
DMAH	dimethylaluminum hydride or $Al(CH_3)_2H$
DOF	depth of focus
DPP	discharge-produced plasma
DPT	double-patterning technology
DR	deposition rate
DRAM	dynamic random access memory
DSC	dichlorosilane (or dichloride silane) or SiH_2Cl_2
DSP	digital signal processing
DUV	deep ultraviolet
EBDW	electron beam direct write
EBI	electron beam inspection
EBR	edge-bead removal
ECD	electrochemical deposition
ECMP	electrochemical mechanical polishing
ECP	electrochemical plating
ECR	electron cyclotron resonance
EDA	electronic design automation
EEPROM	electric erasable programmable read-only memory
EGS	electronic-grade silicon
EMR	electromigration resistance
EOT	effective oxide thickness
EPD	electroplating deposition
EPROM	erasable programmable read-only memory
ESL	etch-stop layer
EUV	extreme ultraviolet
F/C	fluorine-to-carbon (ratio)
FEM	focus exposure matrix
FEoL	front-end of line
FET	field effect transistor
FOUP	front-opening unified pod
FPD	flat-panel display
FPGA	field-programmable gate array
FSG	fluorinated silicate glass
FTES	fluorotriethyloxysilane or $FSi(OC_2H_5)_3$

FTIR	Fourier transform infrared
FZ	floating zone
GPS	global positioning system
HDD	hard disc drive
HDP	high-density plasma
HDP-CVD	high-density plasma CVD
HEPA	high-efficiency particulate air (filter)
HKMG	high-κ and metal gate
HMDS	hexamethyldisilazane or $(CH_3)_3SiNHSi(CH_3)_3$
HOT	hybrid orientation technology
HP	half-pitch
IC	integrated chip
ICP	inductively coupled plasma
IDLH	immediate danger to life and health
IDM	integrated device manufacturer
ILD	interlayer dielectric
IMD	intermetal dielectric
IR	infrared
KGD	known good die
LAT	large angle tilt
LCD	liquid crystal display
LDD	lightly doped drain
LED	light-emitting diode
LELE or LE^2	litho-etch-litho-etch
LER	line edge roughness
LFLE	litho-freeze-litho-etch
LOCOS	local oxidation of silicon
LPC	landing pad contact
LPC	large-particle count
LPCVD	low-pressure chemical vapor deposition
LPP	landing pad poly
LPP	laser-produced plasma
LTE	low-temperature epitaxy
LWR	line width roughness
MBE	molecular beam epitaxy
MEMS	microelectromechanical systems
MEoL	mid-end of line
MERIE	magnetically enhanced RIE
MFC	mass flow controler
MFP	mean free path
MGS	metallurgical-grade silicon
MOCVD	metal organic CVD

MOS	metal-oxide semiconductor
MOSFET	metal-oxide-semiconductor field effect transistor
MRAM	magnetoresistive random access memory
MuGFET	multiple-gate field effect transistor
MW	microwave
NA	numerical aperture
NGL	next-generation lithography
NIL	nanoimprint lithography
nMOS	n-type MOS
NU	nonuniformity
NVM	nonvolatile memory
OMS	optical measurement system
OPC	optical proximity correction
OSG	organosilicate glass
OSHA	Occupational Safety and Health Administration
PBL	poly-buffered LOCOS
PC	personal computer
PCB	printed circuit board
PD	passivation dielectric
PEB	postexposure bake
PECVD	plasma-enhanced CVD
PEL	permissible exposure limit
PFC	perfluorocarbon
PIII	plasma immersion ion implantation
PLAD	plasma doping
PM	preventive maintenance
PMD	premetal dielectric
PMMA	polymethylmethacrylate
pMOS	p-type MOS
POCL	phosphorus oxychloride or $POCl_3$
PR	photoresist
PSG	phosphorous silicate (or phosphosilicate) glass
PSM	phase shift mask
PVA	poly-vinyl-alcohol
PVD	physical vapor deposition
RC	resistance $\times$ capacitance
RCA	Radio Corporation of America
RG	recessed gate
RIE	reactive ion etch
RISC	reduced instruction set computer
RP	remote plasma
RPCVD	remote plasma CVD

RTA	rapid thermal annealing
RTCVD	rapid thermal CVD
RTO	rapid thermal oxidation
RTP	rapid thermal process
SAC	self-aligned contact
SA-CVD	subatomic CVD
SAPD	spacer (self-) aligned double patterning
SA-STI	self-aligned shallow trench isolation
SC	standard cleaning
SCE	short channel effect
S/D	source/drain
SDE	S/D extension
SEG	selective epitaxial growth
SEM	scanning electron microscope
SIMOX	separation by implantation of oxygen
SIMS	secondary ion mass spectroscopy
SMIF	standard mechanical interface
SMO	source-mask optimization
SMT	stress memory technique
SN	storage node
SNC	storage node contact
SOC	system on chip
SOD	spin-on dielectric
SOG	spin-on glass
SOI	silicon-on-insulator
SPC	statistic process control
SRAM	static random access memory
SSD	solid-state disk/drive
SSOI	strained silicon-on-insulator
STI	shallow trench isolation
STM	scanning tunneling microscope
TCE	trichloethylene
TCP	transformer-coupled plasma
TCS	trichloride silane (or trichlorosilane) or $SiHCl_3$
TDEAT	tetrakis-diethylamidotitanium-titanium or $Ti[N(C_2H_5)_2]_4$
TDMAT	tetrakis-dimethylamido-titanium or $Ti[N(CH_3)_2]_4$
TEM	transmission electron microscope
TEOS	tetraethoxysilane (or tetraethylorthosilicate) or $Si(OC_2H_5)_4$
TMA	trimethylaluminum or $Al(CH)_3$
TMAH	tetramethyl ammonium hydroxide or $(CH_3)_4NOH$

TMB	trimethylborate or $PO(OCH_3)_3$
TMP	trimethylphosphite or $P(OCH_3)_3$
TSV	through-silica via
TW	thermal wave
UHV	ultrahigh vacuum
UHV-CVD	ultrahigh-vacuum CVD
ULK	ultralow-κ
ULSI	ultralarge-scale integration
USG	undoped silicate glass
USJ	ultrashallow junction
UV	ultraviolet
VLSI	very large-scale integration
WCVD	tungsten CVD
WERR	wet etch rate ratio
WID	within die
WIW	within wafer
WL	wordline
WTW	wafer to wafer
XRR	x-ray reflectometry

INTRODUCTION TO
Semiconductor Manufacturing Technology
SECOND EDITION

Chapter 1
Introduction

Integrated circuit (IC) chip technology has been changing our lives dramatically for more than 50 years. Since their introduction in the 1960s, IC chips have developed in complexity and usefulness to the point where hundreds, if not thousands, can be found in the average households of developed countries. IC chips instigated a technical revolution considered to be more significant than any other period of invention in human history. IC chips are the backbone of the computer industry and have spurred related technologies such as software, mobile electronics, and the Internet. Every product of the information age is an offspring of IC technology.

IC chips can be found in automobiles, televisions, Blu-ray and DVD players, digital cameras, smartphones, appliances, and game consoles. The list of applications and seemingly miraculous behind-the-scene functions of IC chips could easily fill an entire book.

The future of IC technology is full of possibilities, such as humanized robotics. It is not hard to imagine in the near future a robot that assists the elderly or disabled, with the ability to react to voice command, do chores, hold a conversation, and respond to facial expressions.

When dealing with 3D graphics and artificial intelligence, the thirst for computational power and memory space will never end and will further drive demands for more-powerful microprocessors and memory chips with larger storage size.

Objectives

After finishing this chapter, the reader will be able to:

- name the three scientists who invented the first transistor
- identify the two people who shared the patent for the invention of integrated circuits
- explain the difference between a discrete device and an IC device
- describe Moore's law
- explain the effects of feature size and wafer size on IC chip manufacturing
- define the term semiconductor technology node.

1.1 Brief History of Integrated Circuits

1.1.1 First transistor

The semiconductor era started in December 1947. Two AT&T Bell Laboratories scientists, John Bardeen and Walter Brattain, demonstrated a solid-state electrical device made from germanium, a semiconductor material. They observed that when electrical signals were applied to contacts on a crystal of germanium, the output power was larger than the input. These results were published in 1948. This first transistor, a point contact type, is shown in Fig. 1.1. The word transistor comes from the combination of two words: transfer and resistor.

Bardeen and Brattain's supervisor William Shockley, not wanting to be left out of such an important invention, was determined to make his imprint on the discovery. Shockley worked very hard that December to find out how the bipolar transistor functioned. He solved the problem and published his theory in 1949. He also predicted another kind of transistor, the junction bipolar transistor, which would prove easier to mass produce. In 1956 William Shockley, John Bardeen, and Walter Brattain shared the Nobel Prize in physics for the invention of the transistor. Figure 1.2 shows these three coinventors.

Driven by military and civilian demands for electronic devices, the semiconductor industry developed rapidly in the 1950s. Germanium-based transistors quickly

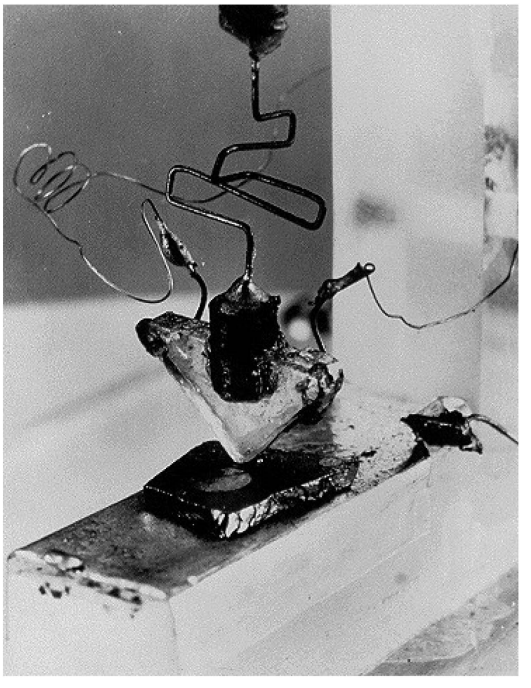


Figure 1.1 The first transistor made in AT&T Bell Laboratories (AT&T Archives, reprinted with permission from AT&T).

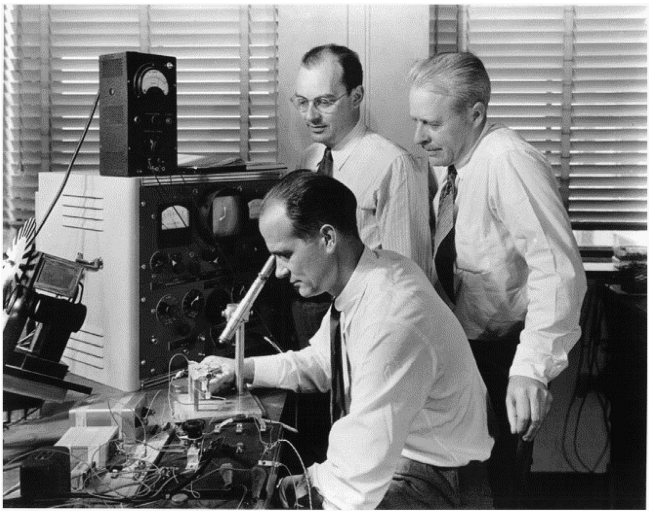


Figure 1.2 Three coinventors: William Shockley (front), John Bardeen, (back left), and Walter Brattain (right).

replaced vacuum tubes in most electronics equipment because of their smaller size, lower power consumption, lower operating temperature, and shorter response time. Transistor manufacturing was significantly accelerated by the introduction of technologies to manufacture pure, single-crystal semiconductor materials. The first single-crystal germanium was made in 1950, and the first single-crystal silicon followed in 1952. Throughout the 1950s, the semiconductor industry provided discrete devices that the electronics industry used to make radios, computers, and many other civilian and military products. A *discrete device* is an electronic device, such as a resistor, capacitor, diode, or transistor, that contains only one device per piece. They are still widely used in today's electronic products. One can easily find discrete devices on almost every printed circuit board (PCB) in any electronic system.

1.1.2 First integrated circuit

In 1957, Jack Kilby, one of the attendees at a workshop commemorating the tenth anniversary of the invention of the transistor, noticed that most discrete devices such as resistors, capacitors, diodes, and transistors could be made from a piece of semiconductor material like silicon. Therefore, he thought it would be possible to make them on the same piece of semiconductor substrate and connect them to form a circuit. This would make a much smaller circuit and reduce the cost of the electronic circuit. Kilby joined Texas Instruments (TI) in 1958 to pursue his new idea. As a newly hired employee, he did not have vacation time, thus, when most of his coworkers took the mandatory summer vacation, he worked to consolidate his idea of an IC in the deserted research-and-development laboratory. When his coworkers came back from summer vacation, he presented his idea and started

to put it into action. Because no silicon substrate was readily available, he used what material he could find: a strip of germanium with one transistor already built onto it. He added a capacitor and used a germanium bar to form three resistors. By connecting the transistor, capacitor, and three resistors from the 1.5-in. long germanium bar to fine metal wires, Kilby made the first IC device, as shown in Fig. 1.3. At TI, IC devices have been called bars instead of chips or dies for a long time because of the shape of the first IC device made by Kilby.

About the same time, Robert Noyce of Fairchild Semiconductor was working on the same idea: making more for less. Different from Kilby's IC chip (or bar), where real metal wires were used to connect different components, Noyce's chip employed aluminum patterns etched from an evaporated aluminum thin film coated onto the wafer surface to form the metal interconnections between the different devices. By using silicon instead of germanium, Noyce applied planar technology, developed by his colleague Jean Horni, to make junction transistors, which took advantage of the silicon and its natural oxide, silicon dioxide. Highly stable silicon dioxide can be easily grown on a silicon wafer surface in a high-temperature oxidation furnace and can be used as the electric isolation and diffusion mask.

One of the first silicon IC chips Noyce designed in 1960 (shown in Fig. 1.4) was made from a 2/5-in. (~10-mm) silicon wafer. Noyce's chip had the basic processing techniques of modern IC chips and served as a model for all future ICs.

In 1961, Fairchild Semiconductor made the first commercially available IC, which consisted of only four transistors, and sold them for $150 each. Ironically, it was much more expensive than purchasing four transistors and connecting them into the same circuit on a circuit board. NASA was the main customer for the newly available IC chips because rocket scientists and engineers were willing to pay higher prices to reduce even a gram of weight from a space rocket.

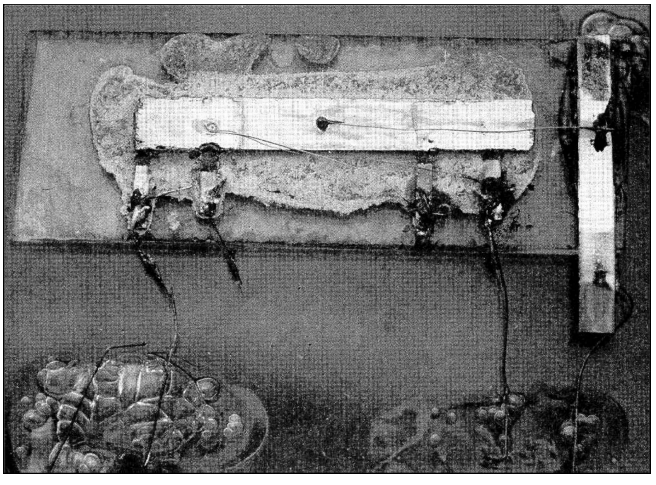


Figure 1.3 The first integrated circuit made by Jack Kilby (Texas Instruments).

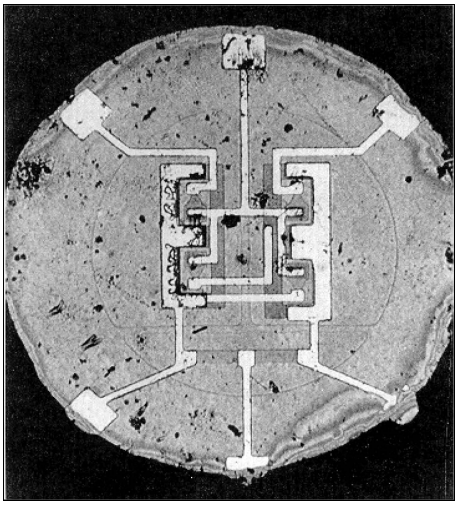


Figure 1.4 The first IC made on a silicon wafer by Fairchild Semiconductor (Fairchild Semiconductor International).

After several years of legal battle, TI and Fairchild Semiconductor settled their case by agreeing to cross-license their technologies. Kilby and Noyce shared the title of coinventor of the IC. In 2000, Kilby (seen in Fig. 1.5) was awarded the Nobel Prize in physics for the invention of the IC. Noyce (seen in Fig. 1.6) left Fairchild Semiconductor and cofounded Intel Corporation with Andrew Grove and Gordon Moore in 1968. He later served as chief executive officer of SEMATECH, an international consortium of semiconductor manufacturers, in Austin, Texas.

1.1.3 Moore's law

In the 1960s, the IC industry developed very rapidly. In 1964, Gordon Moore, one of the cofounders of Intel Corporation, noticed that the number of components on a computer chip doubled every 12 months while the price stayed the same. He predicted that trend would hold in the future. His vision has become well known in the semiconductor industry as Moore's law. Amazingly, Moore's law has been proven accurate for more than 40 years, with only a slight adjustment made in 1975 when he changed 12 months to 18 months. Figure 1.7 shows Moore's prediction in 1965, and Fig. 1.8 is Moore's law for microprocessors. The scales of integration level for IC chips used in the semiconductor industry are listed in Table 1.1.

1.1.4 Feature and wafer size

Before 2000, the semiconductor industry feature size was usually measured in microns, which is equal to one-millionth (10^{-6}) of a meter and is also noted as μm. As a reference, the diameter of a human hair is about 50 to 100 μm. After 2000, semiconductor technology advanced to nanometer (nm) technology nodes.

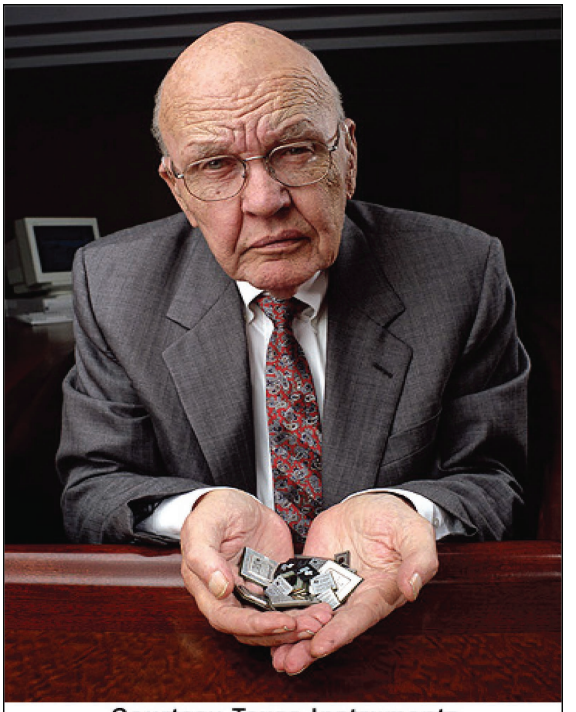


Figure 1.5 Jack St. Clair Kilby (November 8, 1923–June 20, 2005). (See http://media.aol.hk/drupal/files/images/200811/25/kilby_jack2.jpg.)

Table 1.1 Scales for different integration levels for IC chips.

Integration level	Abbreviation	Number of devices on a chip
Small-scale integration	SSI	2 to 50
Medium-scale integration	MSI	50 to 5,000
Large-scale integration	LSI	5,000 to 100,000
Very large-scale integration	VLSI	100,000 to 10,000,000
Ultra-large-scale integration	ULSI	> 10,000,000

A nanometer is one-billionth (10^{-9}) of a meter. In less than 50 years, the minimum feature size of IC chips has shrunk dramatically, from about 50 μm in the 1960s to 32 nm in 2010. By reducing the minimum feature size, a smaller device can be made that allows more chips per wafer, or a more powerful chip can be made with the same die size. Both ways help IC fabrication facilities (fabs) gain a greater profit in the manufacturing of IC chips, profit being the most important driving force of IC technology development.

For example, when the technology node shrank from 28 to 20 nm, the size of a chip shrank by a factor of $(20/28)^2 \sim 0.51$. This means that the number of chips

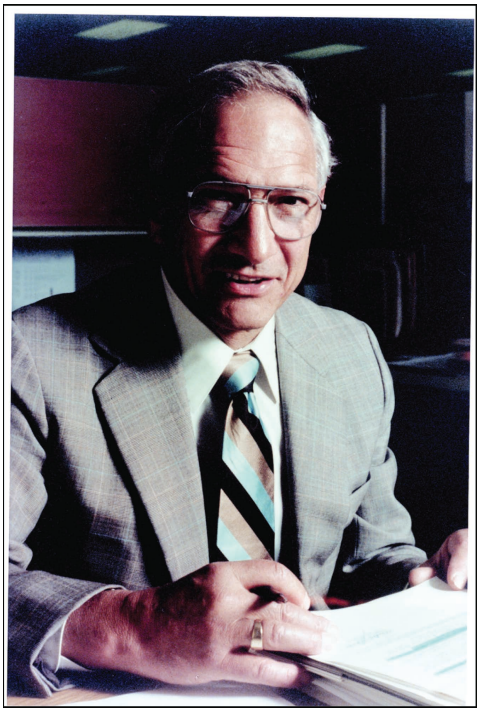


Figure 1.6 Robert Norton Noyce (December 12, 1927–June 3, 1990).
(See http://download.intel.com/museum/research/arc_collect/history_docs/pix/noyce1.jpg.)

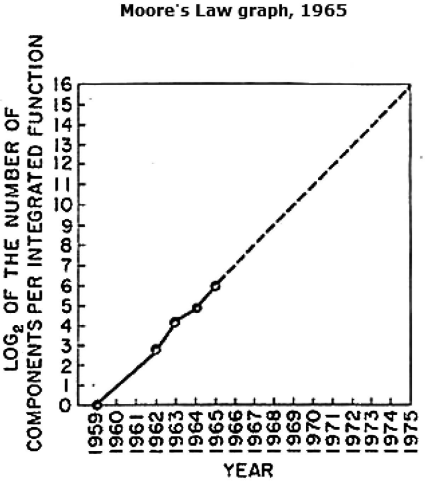


Figure 1.7 Moore's observation and prediction in 1965
(source: http://www.intel.com/technology/mooreslaw/).

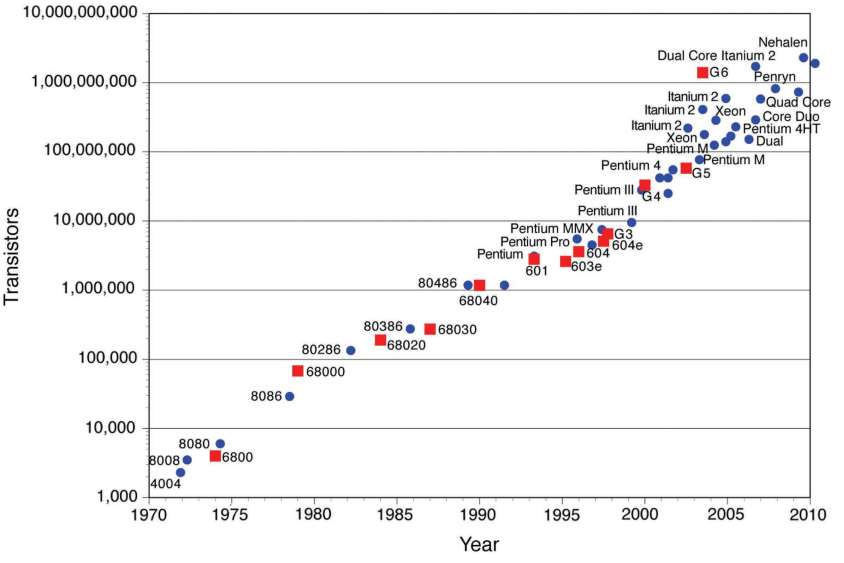


Figure 1.8 Moore's law of microprocessors
(source: http://www.is.umk.pl/~duch/Wyklady/komput/w03/Moores_Law.jpg).

almost doubles if both the chip and wafer are squares. Since a silicon wafer is round, the edge effect can cut the increase to about 50%. Similarly, by further shrinking the feature size to 14 nm, the number of chips almost quadruples compared with the 28-nm technology, as shown in Fig. 1.9.

Figure 1.10 shows the smallest known operational metal-oxide-semiconductor (MOS) transistor with an effective gate length of 0.004 μm, or 4 nm, made by NEC and published in the proceedings of the 2003 International Electron Devices Meeting of IEEE).

Question: Is there a limit for minimum feature size?

Answer: Yes. The minimum feature size of a microelectronic device made on a single-crystalline silicon substrate cannot be smaller than the silicon lattice spacing, which is 5.43 Å. Notice that 1 Å = 0.1 nm = 1×10^{-10} m.

Question: What is the achievable minimum feature size of an IC chip?

Answer: History has disproved many predictions on the achievable minimum feature size of an IC device, and the prediction given here may also fall into that category. One silicon atom is not enough to form the feature of an electronic device. Therefore, if one needs ten silicon lattices to form a minimum feature, the achievable minimum feature size of an IC device might be about 50 Å, or 5 nm.

There will be many technological challenges to overcome before minimum feature size can reach its final physical limit. Most notable is the patterning process, which is the fundamental IC manufacturing step used to transfer the

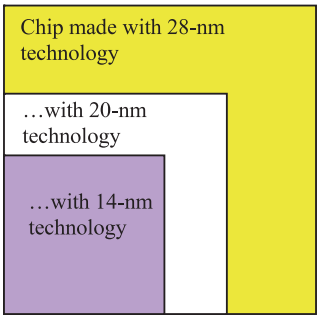


Figure 1.9 Relative chip size with different technology nodes.

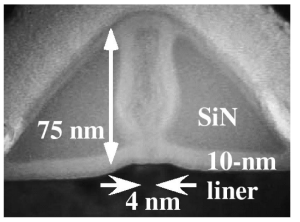


Figure 1.10 The world's smallest known metal-oxide-semiconductor transistor (H. Wakabayashi, et al., *IEEE Proc. IEDM*, 2003).

designed pattern to a wafer surface and form IC devices. Most likely, optical photolithography currently being used will have to switch to alternative lithography technologies—such as extreme ultraviolet (EUV) lithography, nano-imprint lithography (NIL), or electron beam direct write (EBDW) lithography—before minimum feature size can reach its final physical limit. This is discussed in more detail in Chapter 7.

While minimum feature size is shrinking, wafer size is also increasing. Starting from 10 mm (2/5 in.) in the 1960s, wafer size has increased to 300 mm (12 in.). By increasing wafer size, one can put more chips on a wafer. From 200 to 300 mm, the wafer area increases by a factor of $(3/2)^2 = 2.25$, which means the number of chips per wafer can be doubled on a 300-mm wafer. Figure 1.11 illustrates relative wafer sizes of 150, 200, 300, and 450 mm.

Question: What is the maximum wafer size?

Answer: No one knows for sure yet. Because flat-panel display (FPD) manufacturing equipment has already started to handle a tenth-generation glass substrate of size 2850 × 3050 mm, mechanical handling of 1000-mm-diameter (1 m) silicon wafers should not be a big problem. However, the size of a wafer is limited by many factors, such as crystal pulling, wafer slicing technologies, process equipment development, and most

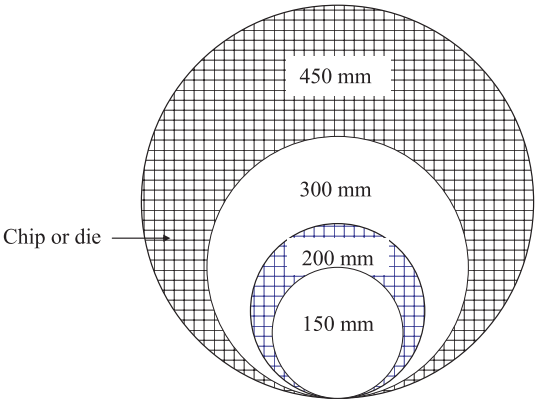


Figure 1.11 Relative wafer sizes.

importantly, the demands of IC fabrication. Because huge amounts of initial capital investment is needed for research and development to increase wafer size, not every IC manufacturer is enthusiastic about doing it, especially doing it first. Currently, the largest wafer used in IC production is the 300-mm wafer, and this is becoming the mainstream for advanced IC fabs. The largest wafer on display is 450 mm in diameter, which was proposed for use in IC fabrication starting in 2012. However, the majority of semiconductor manufacturers and most semiconductor equipment manufacturers are not excited about investing the large amount of capital needed for 450-mm wafer processing, and the targeted year for implementing 450-mm wafers in IC production almost certainly will be delayed. If 450-mm wafers eventually do go into IC production, they will most likely be the largest wafer size.

1.1.5 Definition of the integrated circuit technology node

Technology nodes such as 45, 40, 32, 28, 22, 20 nm, etc., are not defined as the minimum critical dimension (CD) of the devices. The definition of technology node is related to the pitch of the gate pattern, as shown in Fig. 1.12. Although one can easily reduce pattern CD with techniques such as photoresist trimming [shown in Fig. 1.12(b)], this technique will not reduce the pitch of the pattern. To reduce pattern pitch, the pattern technologies (which include photolithography and etch processes) need to be upgraded.

Different IC devices have different relationships of technology node to pattern pitch. For example, in NAND flash devices (which do not have contact between gates), the technology node is defined as a half pitch of the gate pattern. Therefore, a 20-nm NAND flash device has gate patterns of 40-nm pitch (gate CD plus gap) or a 20-nm half pitch. For a logic IC device with contact between gates (called

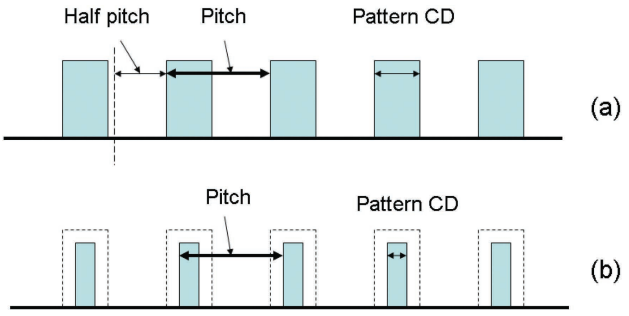


Figure 1.12 Relationship of pattern CD and pattern pitch. (a) Initial photoresist pattern and (b) photoresist pattern after trimming. Although pattern CD is reduced by the trimming process, pattern pitch remains the same.

a contact gate), the technology node usually is one quarter of the gate pitch. For example, a 20-nm logic device usually has an 80-nm gate pitch.

1.1.6 Moore's law or the law of more

Since the invention of the IC, its manufacturing technology has developed rapidly and matched Moore's law very well, as shown in Fig. 1.8. However, the actual driving force of semiconductor technology advancement is not Moore's law, but the law of more (profit). By shrinking the minimum feature size, one can place more chips onto a wafer or place more devices in a chip. In the "good old days," reducing the size of the device improved device speed, reduced power consumption, and enhanced overall device performance. Therefore, scaling down the minimum feature size would reduce manufacturing cost, improve profit margin, and strengthen competitive position. When the cost of research and development was justified by the benefit of feature size reduction, IC manufacturers had strong incentive to heavily invest in new technologies and push device scaling. For the past 50 years, Moore's law overlapped with the law of more, giving the illusion that IC technology was driven by Moore's law. However, when the IC technology node reached the nanometer range, simply shrinking the minimum feature size no longer improved device performance due to the leakage issue, unless very expensive high-κ gate dielectric and metal gates were used in the MOS device. In the nanometer technology era, research and development costs increased almost exponentially with the scaling of the IC technology node. At the 45-nm technology node, some semiconductor companies could not afford the estimated $1.5 billion in research and development costs alone. As IC technology nodes progress to 32/28 nm, 22/20 nm, 14 nm, and beyond, fewer and fewer IC manufacturers will be able to afford research and development costs all by themselves. In the foreseeable future, Moore's law will become history, and IC technology advancement will be determined by the law of more. The semiconductor industry will become a mature industry, like the automobile industry, where technology continues to improve at a more moderate rate.

1.2 Brief Overview of Integrated Circuits

IC chip manufacturing is a very complex system. It involves the manufacture of materials and wafers, circuit design, cleanroom technology, processing equipment, metrology tools, wafer processing, die testing, chip packaging, and final testing.

1.2.1 Manufacturing materials

Semiconductor manufacturing requires a great amount of raw materials for wafer processing. To ensure IC production yield, gaseous, liquid, and solid materials used for wafer processes such as chemical vapor deposition (CVD), etch, physical vapor deposition (PVD), and chemical mechanical polishing (CMP) need to be of ultrahigh purity and have extremely low particle density.

Many semiconductor process materials are poisonous, flammable, explosive, and/or corrosive, and a few are strong oxidizers. Many of the chemicals require that employees receive special training to ensure proper handling.

1.2.2 Processing equipment

Semiconductor processing requires highly specialized tools for a variety of processes. There are epitaxial silicon deposition reactors, CVD and etch equipment, ion implanters, furnaces and rapid thermal process systems, metal deposition reactors, chemical mechanical polishers, and photolithography tools. These process tools are very sophisticated, complicated, and expensive. They require that technicians receive a great deal of special training before being properly operated, maintained, and diagnosed. Because these tools are so expensive and the square footage of a cleanroom is also very costly, semiconductor fabs attempt to keep them running 24 hours a day, seven days a week, nonstop until scheduled preventive maintenance or a breakdown occurs. It is very important to reduce tool downtime to boost productivity and increase throughput to improve profit margin. Well-trained and experienced engineers and technicians play a crucial role in this aspect.

Before the 1970s, most IC manufacturers made their own process equipment. Now semiconductor equipment companies provide the majority of manufacturing equipment for IC manufacturers. They provide not only the sophisticated equipment, but also the production-ready processes. Batch systems, which process many wafers simultaneously, are still widely used, although single-wafer, multichamber cluster process tools are now receiving more attention and application. Cluster tools with multiple process integration capabilities can help improve process throughput and process yield. Another trend is to stack process chambers or process stations in a vertical direction to reduce the footprint of a tool and save cleanroom space. Cleanroom space is very expensive, especially the high-grade cleanrooms of advanced IC fabs. Integration of metrology tools on production equipment for in-situ measurement and real-time process control is another trend of process equipment development.

1.2.3 Metrology tools

Every semiconductor manufacturing step requires specialized tools to measure, monitor, maintain, and control the process. There are tools to measure thin-film characteristics, such as thickness, uniformity, stress, refractivity, reflectivity, and sheet resistance. There are tools to measure device characteristics, such as current-voltage (I–V) curve, capacitance-voltage (C–V) curve, and breakdown voltage. Optical and electron microscopes are widely used to check patterns, profiles, and alignments. Infrared and x-ray radiation are used in some metrology tools to measure and analyze chemical components and their concentrations.

It is very important to keep metrology tools in proper working condition, otherwise false readings can be obtained from time to time, inducing unwanted process troubleshooting and unnecessary tool downtimes, both of which process throughput. Therefore, training employees to understand the basic principles of metrology tools will keep the system well calibrated and prevent unnecessary process interruptions.

The development of semiconductor processes poses the greatest challenge to the expansion of metrology tools. Faster, more accurate measurements, ultrathin-film (<10 Å) measurements, nondestructive pattern and profile measurements, and real-time, in-situ measurements are required for better process monitoring and control.

To enhance and maintain manufacturing yield, defect inspection and monitoring techniques also need to be developed rapidly. Optical inspection systems use photons to detect and capture physical defects on both blank and patterned wafers. Electron beam inspection systems use electrons to capture both tiny physical defects and some electrical defects, such as open or short circuits. To control defect density and maintain production yield, one must have the capability to capture a defect that is half the size of the technology node. For example, in the 14-nm technology node, the capability to capture a 7-nm defect becomes necessary.

1.2.4 Wafer manufacturing

Wafer manufacturing begins with common quartz sand. First, carbon reacts with sand at a high temperature to generate crude silicon or metallurgical-grade silicon (MGS) with purity of at least 98%. Then the MGS is ground into a powder to react with hydrogen chloride to form liquid silicon hydrochloride, such as $SiHCl_3$ (TCS), which can be purified to a level of up to 99.9999999%. Then TCS reacts with hydrogen at high temperatures to deposit high-purity polycrystalline silicon, called electronic-grade silicon (EGS). Placed in a rotating quartz crucible, EGS is melted at 1415 °C, and a rotating seed crystal is slowly pushed in and pulled out to attain an ultrapure single-crystal silicon ingot. Single-crystal silicon wafers are then obtained by sawing the ingot into slices. Because of this process, single-crystal silicon wafers are always round in shape. The wafers are then lapped, cleaned, etched, polished, labeled, and shipped to IC chip manufacturers. Many wafer manufacturers even deposit a thin layer of single-crystal silicon, called epitaxial silicon, on the single-crystal silicon wafer surface for IC chip manufacturers. The

processes of wafer fabrication and epitaxy silicon deposition are covered in more detail in Chapter 4.

1.2.5 Circuit design

When Kilby made the first IC with five components, he drew the design circuit by hand, as shown in Fig. 1.13. At the 22-nm technology node, a test chip with 364 million bits (Mb) of static random access memory (SRAM) with more than 2.9 billion transistors has been demonstrated. 64-Gb NAND flash memory chips, which have more than 64 billion components, have also been fabricated with 19-nm technology. It is impossible to design these chips without the help of powerful computer design tools. Even with computer design tools, a complicated IC such as a high-end microprocessor chip takes dozens, if not hundreds, of engineers and designers several months for design, test, and layout.

Main design considerations are performance, die size (cost of chip manufacture), design time (cost of IC designing and scheduling), and testability (cost of testing and scheduling). IC design is always the tradeoff among those factors to achieve adequate performance and profitable results. Figure 1.14(a) illustrates the circuit of a complementary metal-oxide-semiconductor (CMOS) inverter. Figure 1.14(b) is a textbook layout of a CMOS inverter. The advantage of this layout is that it can give the device cross section of both n-type MOS (nMOS) and p-type MOS (pMOS) on the same surface, as shown in Fig. 1.14(c).

In a normal IC design, a CMOS inverter layout is more compact, as shown in Fig. 1.15. It basically rotates the pMOS in Fig. 1.14(b) 180 deg and places it above the nMOS. By doing that, the shared gate of the nMOS and pMOS is shortened and straightened. Compared with the U-shaped gate shown in Fig. 1.14(b), the benefit of this layout is obvious. Of course, this layout will not present both nMOS and pMOS in one cross section. To achieve that, the textbook-style layout is needed.

IC design includes architectural, logic, and transistor level design. Architectural design defines the application operation system and divides modules for the system. Logic design puts logic units such as adders, gates, inverters, and registers into each module. It also runs subroutines in each module. In transistor level design,

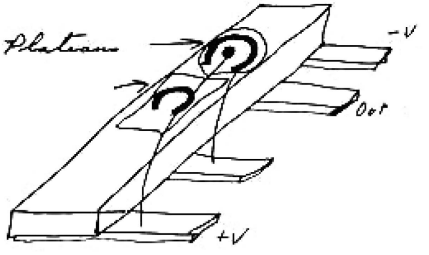


Figure 1.13 The first IC design Kilby sketched in his notebook on September 12, 1958 (Texas Instruments).

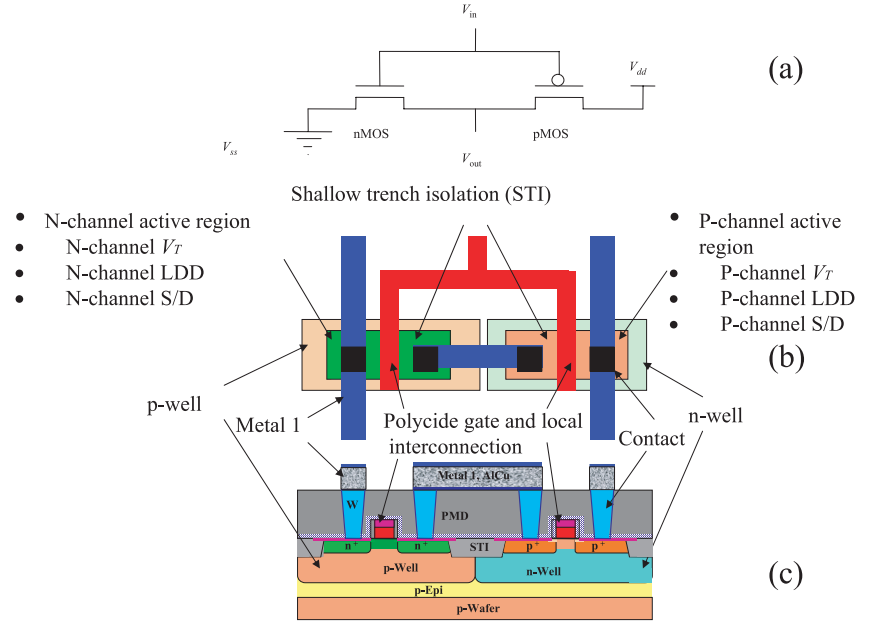


Figure 1.14 (a) The circuit of a CMOS inverter, (b) an example of a textbook-style design layout of a CMOS inverter, and (c) the cross section of the textbook layout.

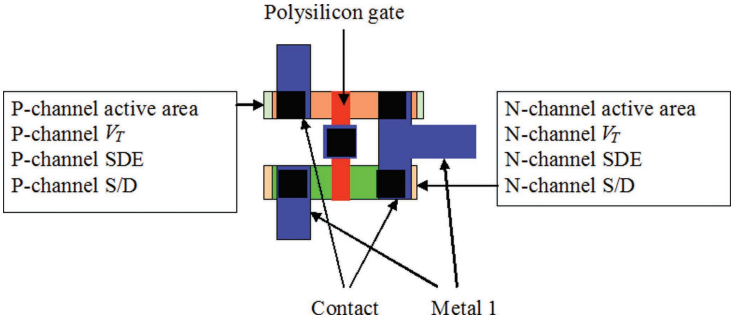


Figure 1.15 CMOS inverter in a real-life IC layout.

individual transistors are laid out in each logic unit. Binary instructions (0 and 1) are given to the logic units to test the circuit design.

After test procedures eliminate design faults or bugs to make certain that there are no design rule violations, and have verified that the design meets yield rule requirements, the layouts of the design can be taped out. In a mask shop, the designed patterns are precisely printed on a piece of chromium glass to make the masks or reticles. Masks or reticles are used in the photolithography process to transfer designed patterns to the photoresist, which is temporarily coated on the semiconductor wafers by means of a photochemical reaction during the exposure process.

Before the 1980s, most semiconductor companies designed, fabricated, and tested IC chips by themselves. These traditional semiconductor companies are called integrated device manufacturers (IDMs). During the 1990s, two kinds of semiconductor companies rapidly developed in the IC industry. One is called a foundry company, which owns wafer processing fabs and does not design its own chips. It receives orders from other companies, makes the masks/reticles or acquires the masks/reticles from mask shops, processes the wafers, and manufactures chips for its customers. Another type is called a fabless semiconductor company, which designs chips for its customers based on customer specifications. These fabless semiconductor companies, or IC design houses, usually place wafer orders with foundry companies to manufacture chips based on their design. Some design houses have their own test center to test the chips after the foundry fabricates and delivers them. Some other fabless companies rely completely on their vendors to do all of the fabrication and testing.

IC design has a direct impact on IC processing. For example, if someone designs metal interconnections on a chip, with one area densely packed with metal lines and another area with very little or no metal lines, this variation could cause what is called a loading effect in the etch process and a dishing effect in the CMP process. Production engineers in the design for manufacturing (DFM) group of a fab usually work closely with both design and process teams to solve these types of problems.

When IC technology arrived in the nanometer era, because printed patterns on the wafer were significantly smaller than the wavelength used to expose these patterns, an optical proximity correction (OPC) had to be used, and design-related process issues became more significant. Design tool companies, which provide electronic design automation (EDA) software for designers to plan IC chips, are now working more closely with foundries to make sure that their products help designers create IC chips with high manufacturability and high production yields in the first silicon run.

1.2.6 Mask formation

After the IC design is finished, the layout image generated by EDA software is printed on a piece of quartz glass coated with a layer of chromium. A computer-controlled laser beam projects the layout image onto the photoresist-coated chrome glass surface. Photons change the chemistry of the exposed photoresist via

photochemical reaction, and the photoresist is later dissolved in a base developer solution. A patterned etching process removes the chromium at the location where the photoresist has been dissolved. Therefore, the image of the IC layout is transferred to the chromium on the quartz glass.

To keep the surface of the mask clean, a thin sheet of plastic called a pellicle is mounted a short distance away from the chrome glass surface. The pellicle provides some protection for the tiny chrome patterns. More importantly, particles that fall onto the pellicle of the mask are out of focus during the pattern exposure process and do not print on the wafer surface (causing repeating defects). Figure 1.16(a) illustrates the basic structure of a conventional or binary photomask, and Fig. 1.16(b) illustrates the basic structure of an attenuation phase shift mask. Figure 1.17 shows the relationship of the IC layout and photomasks for a CMOS inverter. It can be seen that to make a functional CMOS inverter, one needs at least ten masks.

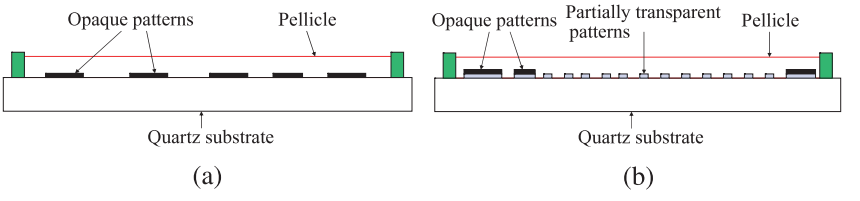


Figure 1.16 (a) Binary mask and (b) an attenuated phase shift mask.

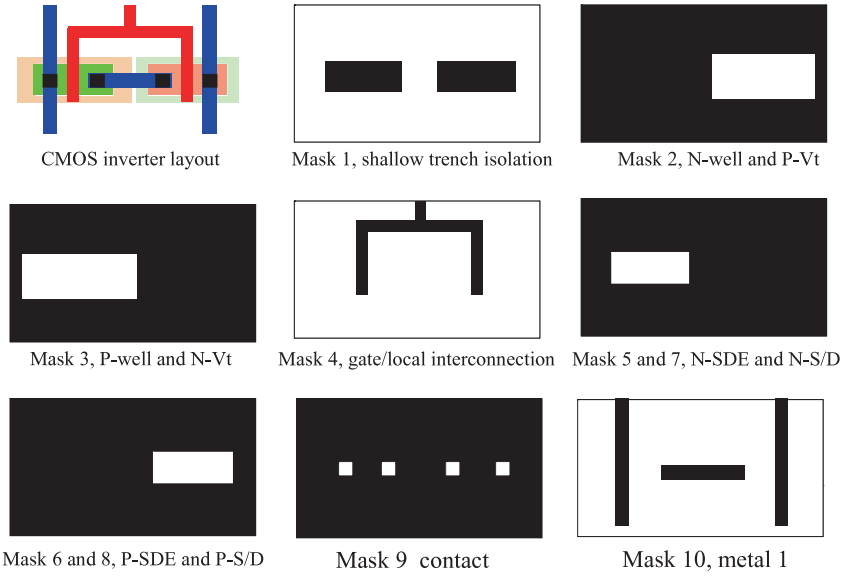


Figure 1.17 Layout and binary masks of a CMOS inverter.

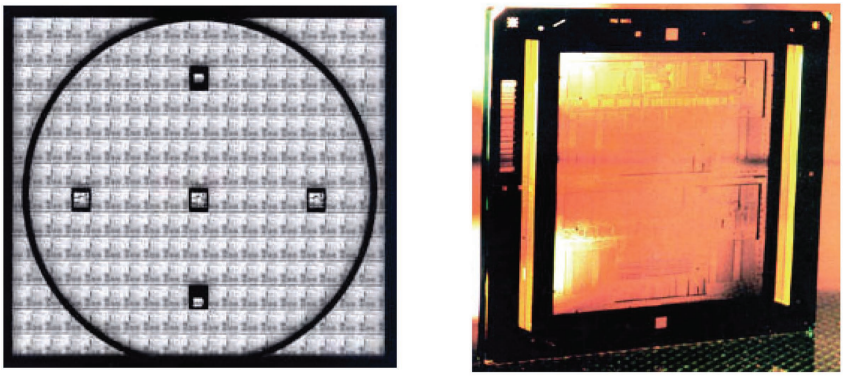
A computer-controlled electron beam (e-beam) is also used to expose photoresist in order to achieve pattern transfer. Since an energetic e-beam has a shorter wavelength than UV light, it has higher resolution and generates a sharper image in the photoresist coated on the chrome glass. As feature size shrinks, more and more masks require an e-beam writer to pattern.

Normally, when the chrome glass has an image covering the entire wafer, it is called a mask. A mask normally transfers an image to the wafer surface at a 1:1 ratio. Exposure systems such as projection, proximity, and contact printers use masks. The highest resolution for a mask is about 1.5 μm.

When the image on the chrome glass covers only part of a wafer, it is called a reticle. A reticle has a larger image and feature size than the image it projects onto the wafer surface, usually with a 4:1 (4×) reduction ratio. Exposure systems that utilize reticles need to expose the reticle multiple times to cover the entire wafer. The process is called step-and-repeat or step-and-scan, and the alignment/exposure systems are called steppers. Some people in fabs also call the system a scanner because of the way it exposes the reticle. Advanced semiconductor fabs usually use steppers for exposure during the photolithography process to achieve the resolution required for patterning. Figure 1.18 illustrates both a mask and reticle. In an IC fab, the reticle is usually referred to as a mask; therefore, the subject of Fig. 1.18(b) can be called both a mask and a reticle, while the subject of Fig. 1.18(a) can only be called a mask.

Partially transparent patterns are formed to allow light to pass through them, with phase shifting that causes destructive interference with the light passing through neighboring clear patterns. This improves photolithography resolution for patterning the tiny features of nanometer technology nodes, as described in Chapter 6.

At least five masks are required to make the simplest MOS transistor. It can take more than 30 masks/reticles to make an advanced IC chip.



(a) (b)

Figure 1.18 (a) A mask and (b) reticle. (SGS Thompson).

1.2.7 Wafer processing

First, IC designers design the circuits with the help of EDA tools. Then the mask shop uses the layout files provided by the designers to print the designed features on the photoresist coated onto the chrome glass with a laser or e-beam writer, and then etch the chromium to make the masks. The masks are then shipped to the photo bay of the fab. Wafer manufactures produce and provide single-crystal silicon wafers with different crystalline orientations, different dopant types, and different dopant concentrations, with or without epitaxial layers, based on requests from the IC fabs. Material manufactures make a variety of ultrapure materials, which are needed in IC manufacturing.

Once wafers are inside a fab, they usually are laser scribed, cleaned, and thermally grown on a thin layer of silicon dioxide. In the front-end of the line (FEoL) process, where transistors are made, wafers go through photolithography many times, and most of them are followed by different ion implantations to form well junctions, source/drain (S/D) extension junctions, polysilicon gate doping, and S/D junctions. Only two FEoL photolithography processes are followed by patterned etch processes—one forms shallow trench isolation and another forms the gate electrodes.

In the back-end of the line (BEoL) process, all photolithography processes are followed by etch processes. In copper metallization, this dual damascene process is repeated multiple times, depending on how many metal layers there are: dielectric CVD, photolithography, dielectric etch, photoresist strip and clean, photolithography, dielectric etch, photoresist strip and clean, deposition of metal layers, metal anneal, and CMP. After all of the metal layers are formed, CVD silicon oxide and silicon nitride are deposited as the passivation layer, and the last photolithography process defines the pads for wire bonding or bump formation. After that, the chips are tested, die sawed, sorted, packaged, burned-in, and shipped to customers.

It takes an IC fab a few weeks and several hundreds of processing steps to create the microelectronic devices of ICs on silicon wafers. Wafer processing includes wet clean, oxidation, photolithography, ion implantation, rapid thermal process (RTP), etch, photoresist stripping, CVD, PVD, CMP, etc. Figure 1.19 shows the flow chart of IC chip fabrication in an advanced semiconductor fab. These processes in the wafer fabs are covered in detail in later chapters with a minimum of math, chemistry, and physics.

1.3 Summary

- Three scientists who invented the first transistor are William Shockley, John Bardeen, and Walter Brattain.
- Jack Kilby and Robert Noyce are the coinventors of the IC.
- A discrete device is an individual electronic device such as resistor, capacitor, diode, or transistor. An IC device is a functional circuit built on a single piece of substrate; IC devices exist in many electronics products.

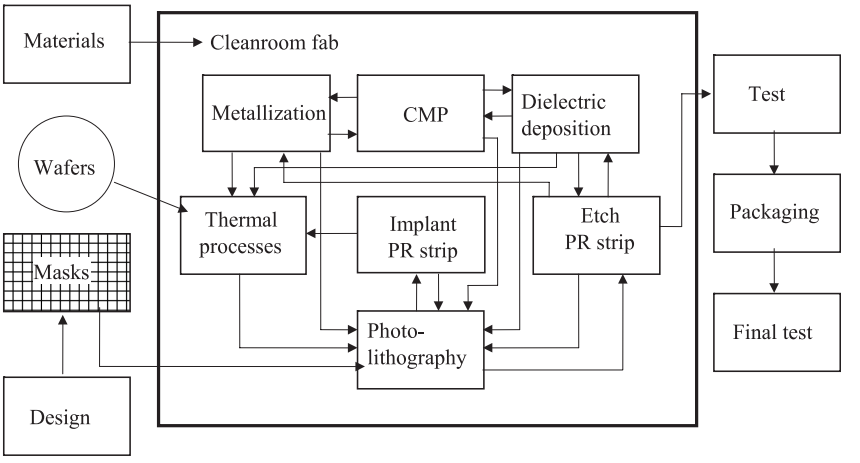


Figure 1.19 Flow chart of a wafer process.

- Moore's law predicts that the number of components on a chip doubles every 12 to 18 months, while the price stays the same.
- By reducing feature size, die size is reduced, which allows more dies on a wafer. By increasing wafer size, more chips can be made on a wafer. Both allow IC chip manufacturers to make a greater profit when research and development costs are manageable.

1.4 Bibliography

J. Bardeen and W. H. Brattain, "The Transistor, A Semiconductor Triode," *Physics Rev.* **74**, 230–231 (1948).

C. Y. Chang and S. M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

J. S. Kilby, "Miniaturized electronic circuit," US Patent No. 3,138,743, filed Feb. 6, 1959, granted June 23, 1964.

G. E. Moore, "Cramming more components onto integrated circuits," *Electron. Mag.* **38**, 114 (1965).

R. N. Noyce, "Semiconductor device-and-lead structure," U.S. Patent No. 2,981,877, filed Jul. 30, 1959, granted Apr. 25, 1961.

M. Riordan and L. Hoddeson, *Crystal Fire*, W. W. Norton, New York (1997).

W. Shockley, "The theory of p-n junctions in semiconductor and p-n junction transistors," *Bell Cyst. Tech. J.* **28**, 435–489 (1949).

H. Wakabayashi, S. Yamagami, N. Ikezawa, A. Ogura, M. Narihiro, K. Arai, Y. Ochiai, K. Takeuchi, T. Yamamoto, and T. Mogami, "Sub-10-nm planar-bulk-CMOS devices using lateral junction control," *Proc. IEEE IEDM Tech. Digest* **20**(7), 1–3 (1993).

N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, second edition, Addison-Wesley, Reading, MA (1993).

1.5 Review Questions

1. When and where was the first transistor made?
2. What is the difference between discrete devices and the IC?
3. Who were the coinventors of the IC?
4. What are the main differences between TI's first IC bar and Fairchild's first IC chip? Which one is closer to a modern IC chip?
5. How many components did the first IC have? Can you identify them in Fig. 1.13?
6. What is the difference between a mask and a reticle? Which one is used with a stepper for higher-resolution photolithography?

Chapter 2
Introduction to Integrated Circuit Fabrication

This chapter introduces the basics of semiconductor fabrication, including the basics of cleanroom, contamination control, yield, IC fab layout, testing, and packaging processes.

Objectives

After finishing this chapter, the reader will be able to:

- define yield
- explain the importance of yield
- describe the basic structure of a minienvironment cleanroom
- explain the importance of cleanroom protocols
- list four basic operations of IC processing
- name at least six processing areas of an IC fab
- list the commonly used facility systems in IC fabrications
- explain the purposes of chip packaging
- compare ceramic and plastic packaging
- describe standard wire bonding and flip-chip bump bonding processes
- list temperature requirements for packaging processes
- describe the purpose of an induced failure test.

2.1 Introduction

IC chip fabrication is a very complicated and time-consuming process. It begins with an IC design, formed with the help of powerful electronic design automation (EDA) software. The layouts of the design are then verified and taped out. In a mask shop, a photoresist (PR) layer is coated onto a chrome glass plate, and the designed pattern is printed by an electron beam writer or a laser writer. After PR development, a chromium etch process is used to transfer the PR pattern to the chrome glass to form the mask or reticle. After cleaning and inspection, the masks are ready to be shipped to IC fabs. A single IC chip requires 20 to 30 masks, depending on the type of device and the technology node.

Initially, wafer providers obtain rough silicon from quartz sand, then purify the silicon, pull it into a single-crystal ingot, and slice the ingot into wafers. After edge rounding, wet etching, surface polishing, and sometimes epitaxial silicon growth, the silicon wafers are ready to be shipped to the semiconductor fabs. A typical 300-mm IC fab can process more than 10,000 wafers per month. For a "mega lab," production can be more than 50,000 300-mm wafers per month.

Many chemical materials with ultrahigh purity are needed as consumables for IC manufacturing. Gases such as oxygen, nitrogen, and hydrogen; liquids such as deionized water, sulfuric acid, nitric acid, and hydrofluoric acid; and solids such as phosphorus, boron, aluminum, and copper are heavily used in IC fabs.

After wafers are shipped to the IC fab, they are cleaned and go through many processing steps, such as thermal processes, photolithography, etch, ion implantation, dielectric thin-film deposition, chemical mechanical polishing (CMP), metallization, etc. After IC fabrication, the processed wafers are shipped to a testing and packaging house, where they are tested, packaged, and given a final test. A generic IC fabrication process flow is illustrated in Fig. 2.1.

Wafers are always kept in a cleanroom, where the size and amount of free particles are carefully controlled and maintained at very low levels. Even in a cleanroom, wafers are usually stored in specially designed containers to minimize possible contamination. For 200-mm or smaller wafers, many fabs use open cassettes with carrier boxes that can hold a cassette with wafers in the slots in a vertical direction. Some advanced 200-mm fabs use containers called standard mechanical interface (SMIF) boxes. Containers for 300-mm wafers are called front-opening unified pods or FOUPs.

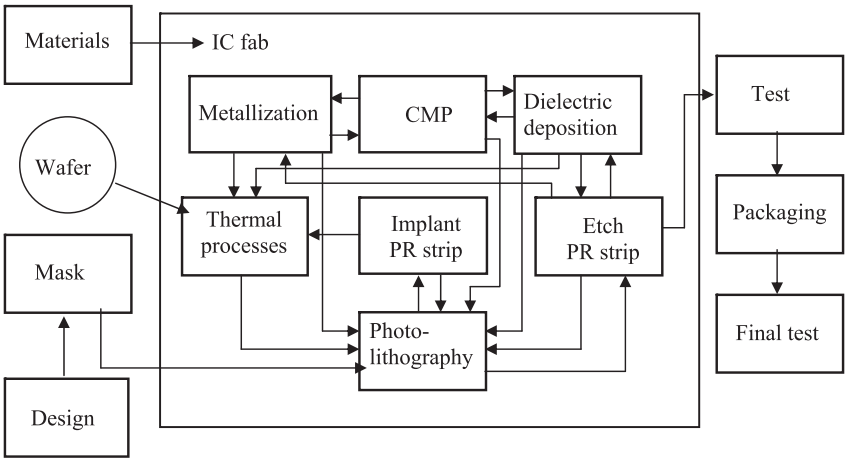


Figure 2.1 IC manufacturing process flow.

2.2 Yield

Yield is one of most important factors of IC fabrication. It can determine whether a fab is making a profit or losing money. Yield relates to many factors, including environment, materials, equipment, processes, and people working in the fab. Improving yield is so critical that semiconductor fabs always hire yield engineers (yield-enhancement engineers) to work in their IC fabs.

2.2.1 Definition of yield

There are three different yields in IC chip fabrication.

Wafer yield: the ratio between the number of good wafers after process completion and total number of wafers used for IC chip fabrication:

$$Y_W = \frac{wafers_{good}}{wafers_{total}}.$$

Die yield: the ratio between the number of good dies per good wafer after process completion and total number of dies on the wafer:

$$Y_D = \frac{dies_{good}}{dies_{total}}.$$

Packaging yield: the ratio between the number of good chips after packaging completion and total number of chips packaged:

$$Y_C = \frac{chips_{good}}{chips_{total}}.$$

Wafer yield mainly relates to processing and wafer handling. Careless human wafer handling, robot malfunction, and robot miscalibration can break the brittle silicon wafers. Incorrect processing, such as misaligned photolithography followed by etch or ion implantation, wrong dopant concentration, nonuniformity of film thickness in thin-film deposition, large amounts of particles on a wafer, etc., can also cause scrapping of the wafers. Die yield relates to many factors such as particle contamination, process maintenance, total processing steps, design-related processing window, etc. Chip yield relates to the wire bonding quality and specification difference between the die test and the chip final test:

$$Y_T = Y_W \times Y_D \times Y_C. \tag{2.1}$$

Overall yield Y_T of the fab is the product of the three, which is a very important factor in semiconductor fabrication because it can determine whether a fab is making or losing money.

2.2.2 Yield and profit margin

The definition of fab yield is the ratio of the number of good chips to the total chips at the start of the process. A simple example can help explain why yield is so important to semiconductor fabs.

The cost of a 300-mm wafer varies from time to time, mainly determined by the condition of supply and demand. Depending on circuit requirements, a wafer requires several-hundred processing steps before it can be sent to a packaging house to be tested and packaged. Each processing step adds some cost, typically about $1 USD per wafer. If we assume a wafer costs $200, and it takes 500 steps to complete the processing, the total cost per processed wafer is $200 (wafer cost) + $500 (processing cost) = $700. We can assume 100% wafer yield, which means no wafer scrap during the 500 processing steps. We can also assume the cost of test and package for each good die is $10, and no dies failed during final testing after package completion (100% packaging yield). In the case of 500 chips per wafer, and each chip selling for $30, 35 good dies per wafer or a 7% die yield is needed to achieve breakeven. This can be expressed as:

$$\$700 \text{ (wafer and processing costs)} + 35 \text{ (good dies)} \times \$10 \text{ (test and package cost)} = \$1050 = 35 \text{ (good dies)} \times \$30 \text{ (revenue/die)}.$$

If the die yield increases to 50% while wafer yield and package yield both stay at 100%, the total cost per wafer will be $700 (wafer and processing cost) + 250 × $10 (test and packaging cost) = $3200, while the revenue per wafer will be 250 × $30 = $7500. The profit margin per wafer is $7500 − $3200 = $4300. This means that if a wafer fab can process 10,000 wafers a month with 100% wafer yield, 50% die yield, and 100% packaging yield, it can make $43 million in profit a month.

While $43 million a month is indeed a very impressive number, we must remember the approximate $3 billion it costs to build a 300-mm wafer fab, and the thousands of employees necessary to keep it running 24 hours a day, 7 days a week (24/7) nonstop. A total yield of 50% with 10,000 wafers per month might not be high enough to generate sufficient cash flow to pay all the bills! Therefore, improving yield and increasing throughput is vital for an IC fab.

Question: For the same IC fab, if the die yield is 90%, wafer and packaging yield are 100%, and capacity is 20,000 wafers per month, what is the total profit margin per month? (The hardest math in this book.)

Answer: $20,000 \times [500 \times 90\% \times \$30$ (revenue per wafer) − ($700 + 500 × 90% × $10) (cost per wafer)] = $166 million/month.

We can see that the higher the yield and capacity, the larger the profit margin. Of course, the math becomes much more complicated when the wafer yield and packaging yield are no longer 100%, as they are in the real world of IC manufacturing.

As mentioned earlier, advanced IC chip fabrication involves about 500 processing steps. To achieve a reasonably high overall yield, every step has to have very high yield, close to 100%.

Question: If the die yield for every processing step is 99%, and there are 500 processing steps for IC fabrication, what is the overall die yield?
Answer: The overall die yield is 99% × 99% for 500 times, which equals $0.99^{500} = 0.0066 = 0.66\%$.

For advanced IC chips, the yield of most processing steps needs to be 100%, or very close to a perfect value to ensure overall yield. Yield improvement is a nonstop process in semiconductor fabs. At the beginning of a new processing technology development, the launch of a brand new product, or the introduction of a new set of tools, overall yield is normally not very high. After a few wafer production cycles, systematic factors that affect the yield are mined out and eliminated, and the yield is continuously improved and stabilized when it is limited by random factors such as a few airborne particles. Excursion of yield can happen from time to time due to some unforeseen issues such as a slight mismatch of processing equipment, human error, etc. The cycle starts all over when new processes, new products, or new tools are introduced. Figure 2.2 illustrates a typical yield ramp curve of IC technology.

2.2.3 Defects and yield

Equation (2.2) shows the relationship of overall yield and the killer defect density, chip size, and number of processing steps:

$$Y \propto \frac{1}{(1 + DA)^n}, \tag{2.2}$$

where Y is the overall yield, D is the killer defect density, A is the chip area, and n is the number of processing steps. From Eq. (2.2), we can see that to achieve 100% yield, the killer defect density must be zero for every processing step. At the same defect density and chip size, the more processing steps there are, the lower the yield. This equation also indicates that at the same defect density level, the larger the chip size is, the lower the yield will be, as is illustrated in Fig. 2.3.

Equation (2.2) assumes that the defect density of each processing step is the same, making it an obviously oversimplified model. However, it provides a simple

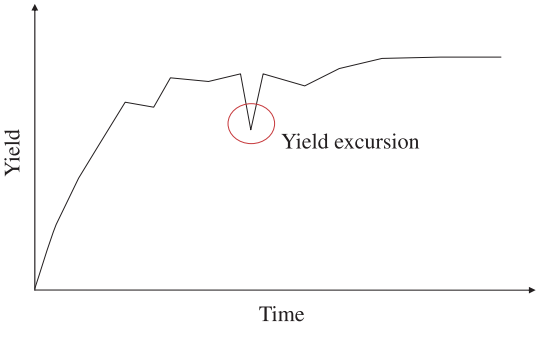


Figure 2.2 Typical yield ramp curve of IC technology.

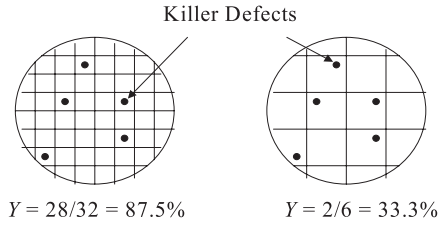


Figure 2.3 Relationship between die size and die yield.

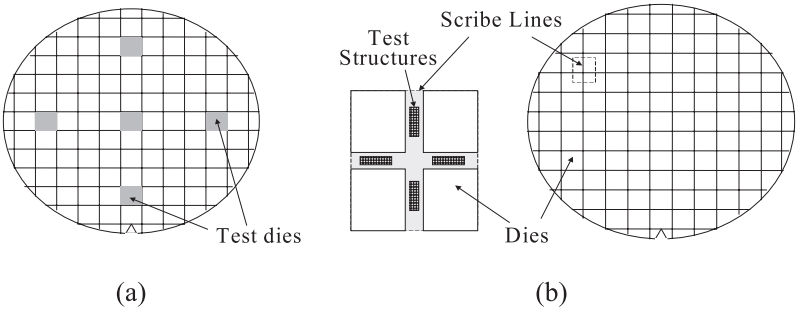


Figure 2.4 Wafer with (a) test dies and (b) test structures in scribe lines.

and straightforward explanation of the relationship between defect and yield. (For more detail on yield modeling, see Chang and Sze in the Bibliography at the end of this chapter.)

Some production wafers are designed with several test dies, in which transistors and test circuits are built during wafer processing, as shown in Fig. 2.4(a). Due to the improvement of technology and shrinking of feature size, test structures—including devices and circuits—are built on the scribe line between the dies to save space on the silicon wafer for making more chips, as shown in Fig. 2.4(b). Sampling tests are performed on these test structures throughout wafer processing to ensure the yield of IC fabrication. If the results of the electrical tests confirm that most transistors and circuits in the test structures do not function as designed, transistors in the real devices will not likely meet design requirements. Wafer processing is stopped, and usually the whole batch of affected wafers is scrapped, which contributes to the loss of wafer yield. A physical failure analysis is performed quickly to pinpoint the failure mechanism and the root cause.

2.3 Cleanroom Basics

The requirement of a cleanroom is one reason IC fabs are so expensive. Since tiny particles can cause defects on a microelectronic device and circuit (affecting chip yield), IC chips must be manufactured inside a cleanroom to achieve acceptable

yield. When feature size shrinks, so does the size of killer particles. A smaller feature size requires a higher grade of cleanliness in the cleanroom.

Since particles have a significant impact on yield, considerable efforts have been made to improve cleanroom conditions, which in turn reduces particulate counts and improves IC production yield. Developments of cleanroom furniture, gowns, and gowning procedures have all contributed to this improvement. Tight cleanroom protocol helps keep contaminants out and prevent yield degradation due to particulate and other contamination.

2.3.1 Definition of a cleanroom

A cleanroom is a manmade minienvironment that has much lower particle counts than the normal environment. The first cleanrooms were built for hospital surgery rooms to control airborne bacteria contamination and reduce postsurgery infection. Soon after the semiconductor industry started, people realized the importance of contamination control, and cleanroom technology was adapted in transistor and IC fabrication. The first silicon IC chip made by Noyce of Fairchild Camera (see Fig. 1.4) shows much particle contamination.

The standard definition of cleanroom class is a strange combination of metric and English units: class 10 is defined as less than 10 particles with diameters larger than 0.5 μm per cubic foot. Class 1 is defined as less than 1 particle with diameters larger than 0.5 μm per cubic foot. A fab making IC chips with a minimum feature size of 0.25 μm needs a class-1 cleanroom to achieve an acceptable yield. For comparison, there are more than 500,000 particles with diameters larger than 0.5 μm per cubic foot inside an ordinary "clean" house. Figure 2.5 illustrates the number of particles in a cubic foot of air of different cleanroom classes.

The cleanest class of cleanroom, M-1, only uses metric units. According to the definition of U.S. Federal Standard 209E, the M-1 class cleanroom allows less than 10 particles larger than 0.5 μm in a cubic meter, or less than 0.28 particles larger than 0.5 μm per cubic foot. Table 2.1 gives the definition of cleanroom classes.

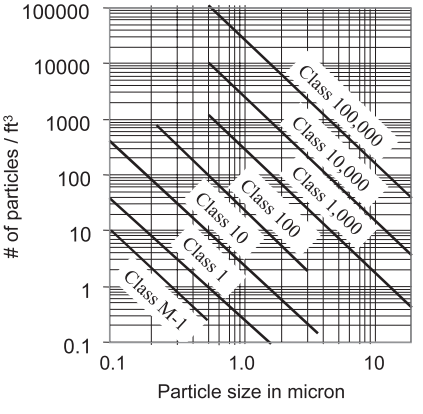


Figure 2.5 Number of airborne particles in the cleanroom.

Table 2.1 Definition of airborne particulate cleanliness class per U.S. Federal Standard 209E.

Class	Particles/ft^3				
	0.1 μm	0.2 μm	0.3 μm	0.5 μm	5 μm
M-1	9.8	2.12	0.865	0.28	
1	35	7.5	3	1	
10	350	75	30	10	
100		750	300	100	
1,000				1,000	7
10,000				10,000	70

Table 2.2 Particle count and profit margin.

Wafer starts/month	10,000	10,000
Fab yield	0.85	0.85
Wafer diameter (mm)	100	100
Edge exclusion (mm)	4	4
Particle count	**20**	**19**
Monthly fixed costs $M	$0.53	$0.53
Various costs/wafer	$76.11	$76.11
Total cost/wafer	$129.00	$129.00
Mask steps	7	7
Defect density (cm^2)	0.30	0.29
Die Size (cm^2)	0.5	.05
Random yield	0.37	0.39
System yield	0.70	0.70
Die yield	0.26	0.27
Die/wafer	113	113
Good die	30	31
Cost/die	$4.35	$4.15
Package cost	$1.00	$1.00
Burn in/test	$0.50	$0.50
Test yield	0.90	0.90
Circuit cost	$6.50	$6.28
Circuit price	$12.00	$12.00
Price/cost ratio	1.85	1.91
Sales/wafer	$320.32	$335.37
Annual sales ($M)	$32.67	$34.21
Annual manufacturing cost ($M)	$17.70	$17.91
Annual gross margin ($M)	$14.98	$16.30
Annual gross margin difference ($M)		**$1,321,943**

2.3.2 Contamination control and yield

Particles on a wafer can cause defects, which can reduce yield and affect the profitability of an IC fab. Table 2.2 gives a good example of the effects of particles. It shows that just one more particle count on each wafer could have cost a 4-in. wafer fab more than $1.3 million a year in the early 1980s. Although these data are old, they still demonstrate the effects of particle contamination on the die yield and the profit margin of an IC fab.

Particles can cause many different defects for different processes. For instance, imaging of particles on the clear area of a mask or reticle can cause pinholes for

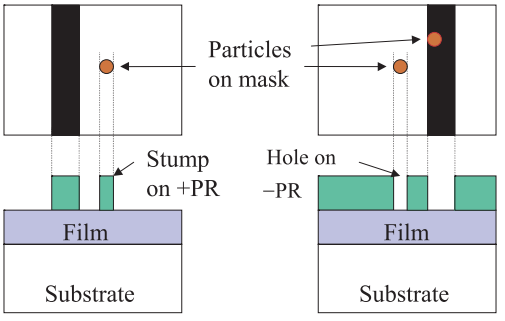


Figure 2.6 Effects of particle contamination on a photomask.

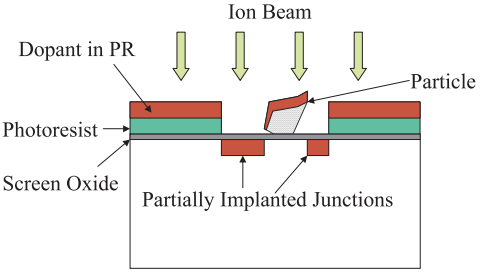


Figure 2.7 Effects of particle contamination during the ion implantation process.

negative photoresist, or stumps for positive photoresist, during photolithography processing. During etch processing, these pinholes and stumps are transferred to the wafer surface and cause defects. Since masks/reticles are used for every wafer for many times during IC fabrication, particle contamination of masks can cause serious problems with chip yield. Masks or reticles must be kept in the cleanest environment to prevent particle contamination. The effects of a photomask particle contaminant on the photolithography process are shown in Fig. 2.6. Note that particles on the dark part of the mask do not affect the photolithography process.

Particle contamination can cause other problems in different processes; for example, it can cause a broken metal line, or a short between neighboring metal lines. Particles can block implanting ions and create an incomplete junction during ion implantation, possibly affecting device performance, as shown in Fig. 2.7.

A particle with half the size of the technology node could be a killer particle. For example, an 11-nm particle can be a killer for a 22-nm IC chip. If a particle falls onto a critical area, it could be a killer even if it is smaller than half the size of the technology node. As feature size reduces, so does the size of the killer particle. Particles of different sizes behave quite differently; therefore, cleanrooms of different classes need different design concepts and require different protocols.

For example, large particles (>1 μm in diameter) can be blown away with high-pressure airflow, while smaller particles cannot. Therefore, high-pressure air or nitrogen blowers are widely used in older, 100-mm (4-in.) fabs, and they are seldom used in the advanced 300-mm fabs. An air blower can remove large particles from a wafer surface; however, it adds more small particles to it. These small particles cause no problem when the feature size is larger than several microns, but they can cause many problems when the feature size shrinks to submicron.

2.3.3 Basic cleanroom structure

The basic structure of a cleanroom for a 300-mm wafer fab is illustrated in Fig. 2.8. A cleanroom is usually located on a raised floor with grid panels, allowing airstream to flow vertically from the ceiling to the facility area underneath the cleanroom. Airflow returns to the cleanroom through a high-efficiency particulate air (HEPA) filter, which removes most particles carried by the airflow. To cut costs, only equipment interfaces are designated as the highest class of cleanroom (where wafers are loaded into processing or metrology equipment), while fabs are built with lower-class cleanroom designs, and most facilities are installed in a subfab beneath the cleanroom floor. Wafers are carried inside a sealed FOUP and are only exposed to airflow in the processing or metrology equipment.

For a less advanced IC fab (illustrated in Fig. 2.9), wafers are carried in an open cassette inside a carrier box. Because wafers will be exposed to airflow before they are loaded into the processing or metrology equipment in an open cassette, the wafer processing areas are designed as a high-class cleanroom. To save costs, equipment areas are made with lower-class cleanroom designs, and facilities are installed in a subfab beneath the cleanroom. For the 0.13-μm technology node, a class-1 or higher cleanroom is required for processing areas. Normally, a class-1000 cleanroom is good enough for the equipment area and costs significantly less to build than a class-1 cleanroom. However, with the shrinking IC feature size,

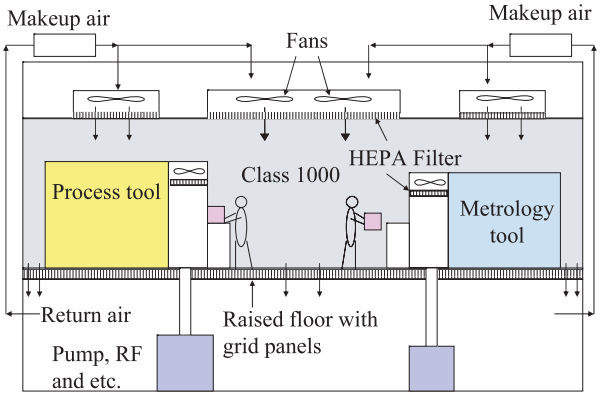


Figure 2.8 Basic structure of an advanced cleanroom.

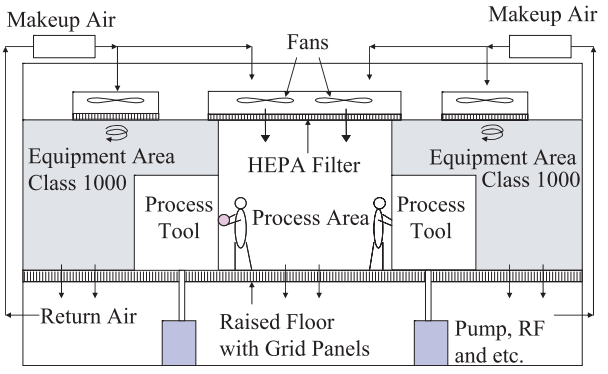


Figure 2.9 A less advanced fab.

it becomes less cost effective to further improve the class level of the processing areas in this kind of cleanroom; thus, the minienvironment cleanroom, illustrated in Fig. 2.8, has become the mainstream for advanced 300-mm fabs.

To achieve better than class-100 cleanliness, it is very important to keep a linear airflow, usually called laminar flow, and avoid air turbulence. Air turbulence can drive out particles sticking to the surfaces of walls, ceilings, tables, and tools, causing them to become airborne, and helps these particles stay airborne for a longer time, increasing the number of particles in the air. With a laminar flow, airborne particles are carried away quickly by the airflow. Class 100 is the dividing line between laminar flow and turbulent flow, a choice that is a very important factor of cleanroom cost. A lower class than class 100 can be achieved with turbulent flow, which costs much less than achieving and maintaining laminar flow.

A cleanroom is always kept at a higher pressure than nonclean areas; this allows continual outward airflow, preventing the entry of airborne particles from the outside when a door is opened to let people, tools, or other materials enter the cleanroom. The same principle is also applied to the different class areas inside the cleanroom. The higher-class areas have a higher pressure than the lower-class areas. Variation of temperature, airflow rate, and humidity can also disturb particles from a surface and cause them to become airborne; therefore, temperature, airflow rate, and humidity are strictly controlled in the cleanroom.

2.3.4 Basic cleanroom gowning procedures

It is very important to maintain rigid cleanroom protocols to minimize the yield loss due to contamination. Human beings emit a great deal of particles, and people are also the major source of sodium, which can cause mobile ion contamination. Therefore, people working inside the cleanroom must wear specially designed garments. Because people are the main contamination sources, some companies have even limited the number of people working in the cleanroom to control contamination. Improvement of cleanroom garments and strict cleanroom gowning procedures have sharply reduced contamination from people working inside the

cleanroom. Proper gowning and degowning procedures are important parts of cleanroom protocol.

Although different companies, even different fabs within the same company, have different gowning procedures, the basic idea is the same: to prevent particles and other contaminants from being carried into the cleanroom by people.

Some fabs require people to put on linen gloves before they even enter the gowning room. Linen gloves are usually made of composite fibers that do not tend to attract particles. They can prevent sodium and particles on the hands from contaminating cleanroom garments, and they are also more comfortable when putting on an outer layer of latex gloves.

The bottoms of shoes carry the largest number of particles. Some fabs use sticky pads or shoe brushes to remove dirt from shoes before putting on shoe covers or booties. It is important to keep bare shoes at the entrance side of the bench and covered shoes on the other side of the bench to prevent large numbers of particles from being carried into the gowning room. Figure 2.10 shows the gowning area of a cleanroom.

Hair covers are required before entering a growing room. Human hairs become positively charged via friction, which attracts negatively charged particles. Certain pressure, humidity, and temperature discharge can neutralize these particles so that they are no longer attracted by hairs. As a consequence, discharged particles from hair could easily become airborne and cause contamination. Hair covers block the emission of these particles.

Generally the first cleanroom garment to put on is the head hood with mask, which further covers the hair and face to prevent particles and other contaminants associated with breathing, coughing, and sneezing. Badges, two-way pagers, cell phones, and radios must be removed before putting on the cleanroom suit, so that they will be on the outside of the suit (otherwise they will be sealed inside the suit and cannot be used in the cleanroom). It is forbidden to unzip a cleanroom suit inside the cleanroom, since that activity can release many particles originally sealed

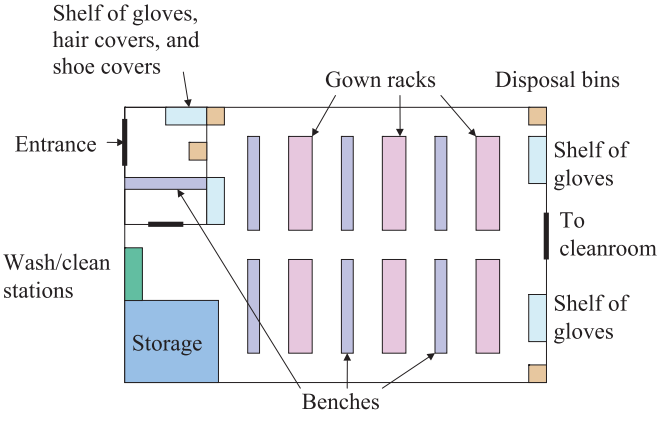


Figure 2.10 Cleanroom gowning area.

in by the suit. After putting on the suit and attaching it to the hood, a pair of boots is put on and attached over the calf of the legs to completely seal particles on the shoes. Safety glasses are required, since IC fabs always use corrosive chemicals in wafer processing, and some processing tools have moving parts, which pose hazards for eye injury. Before entering a cleanroom, a pair of latex gloves is put on top of the linen gloves to prevent particle and mobile-ion contamination. They can also provide minimum protection for the hands from corrosive chemicals. For some processes, such as the CVD chamber and implanter beam line wet clean, two layers of latex gloves are required for better protection from corrosive materials, such as hydrofluoric (HF) acid. For some wet chemical processes, full acid gear is mandatory for protection. Before entering a cleanroom, a person should look in the mirror to make sure hair and street clothes are not exposed.

Before entering a cleanroom, some fabs require people to go through an air shower, in which high-pressure airflow blows away particles on the surface of the garments. To achieve the best result, a person should raise both arms and rotate the body slowly during the air shower. Now the person is ready to enter the cleanroom.

The degowning procedure is almost the reverse of the gowning procedure. First the boots are taken off, then the suit and the head hood are removed. They are usually hung on the garment rack at the next entry. Cleanroom suits are normally washed every week at a specialty laundry, and some fabs are now using disposable cleanroom garments that can be thrown away after being used. Once outside the gowning room, hair covers and gloves are taken off and thrown into trash cans. Shoe covers are the last items to be removed after exiting the cleanroom. Some fabs recycle linen gloves for washing and reusing, while other fabs do not require linen gloves at all.

2.3.5 Basic cleanroom protocols

While cleanroom protocols are different from company to company, the basic concepts are the same: keeping particles where they are and not letting them become airborne, and preventing other contaminants from reaching the wafers.

Once inside a cleanroom, people need to walk steadily. Running or jumping is prohibited because it can disturb particles sticking to the surfaces of the floor, walls, and ceiling. There are very few chairs inside a cleanroom, since particles on the surface of a chair can easily become airborne when a person sits down or stands up. For the same reason, sitting on a table or leaning against a wall is not allowed in a cleanroom, since these actions can dislodge particles from the surfaces.

Ordinary paper carries many small fiber fragments, which can cause particle contamination. Therefore, only specially made cleanroom paper is allowed in a cleanroom. For a class-1 or better cleanroom, even cleanroom paper is not allowed; all all data must be recorded electronically, by using hand-held or notebook computers.

There are other types of contamination, such as mobile ions. For example, a trace amount of sodium contamination can cause the malfunction of MOS transistors and affect the reliability of the integrated circuits. Sodium contamination must

be tightly controlled. Therefore, total isolation between people who process the wafer and the wafer itself is required. The human body is a sodium carrier (sodium chloride, NaCl); thus, if one's glove touches the skin of the face, the latex gloves should be replaced immediately with a fresh pair, or a clean pair can be placed over the dirty pair. Similarly, workers should replace or cover latex gloves as soon as possible after covering their mouth during sneezing or coughing, because high-pressure airflow from the mouth can force small amount of saliva through the facemask. Ideally, one should attempt to leave the processing area and go to equipment areas before sneezing or coughing. It goes without saying that eating is strictly forbidden in the cleanroom.

People who work in cleanrooms are not allowed to use cosmetics, perfume, or cologne, because these materials can emit particles and cause contamination. Technicians cannot wear contact lenses, since trace amounts of chlorine in the fab can react with the lenses and cause eye injury. Smoking is forbidden in the cleanroom and adjacent buildings, and smokers working in cleanrooms are strongly encouraged to quit smoking. Even after they have finished smoking, they can emit particles they have inhaled.

Question: In a 200 mm fab, after finishing the scheduled preventive maintenance (PM) of the processing equipment, a technician knocks on the wall of the processing area to warm another technician in the equipment area that the tool is going to be turned on. Is this act against cleanroom protocol?

Answer: Yes. Knocking on the wall can disturb particles adhering to the wall and ceiling surface, and cause them to become airborne. A radio is a better tool for communication between coworkers in the fab. Some processing tools have secondary terminals, so that the system can be turned on in the equipment area after finishing PM while the terminal on the processing side is locked.

2.4 Basic Structure of an Integrated Circuit Fabrication Facility

Semiconductor manufacturing has many different processes, which can be characterized by the following four basic operations: adding, removing, patterning, and heating. Doping, layer growth, and deposition are adding processes; etch, clean, and polish are removing processes; photolithography is the patterning process; and annealing, alloying, and reflow are the heating processes. Figure 2.11 illustrates the IC processing flow module.

An IC fab usually consists of offices, facilities, storage areas, equipment, and wafer and processing areas. Some fabs have their own chip test and packaging facilities. Wafer processing, equipment, chip testing areas, and packaging are located in a cleanroom, while other areas such as offices, facilities, and process-materials storage are located elsewhere on site. Wafers are always in the wafer processing area, which maintains the highest class of all cleanrooms. A large number of processing equipment is installed in a lower class cleanroom, sometimes

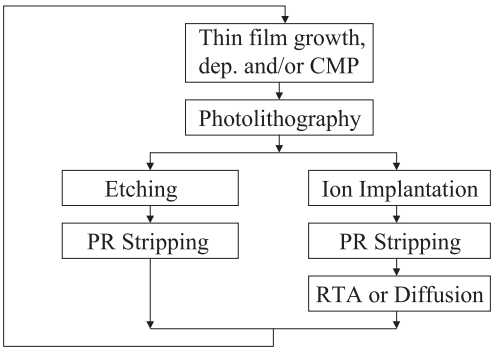


Figure 2.11 IC process flow.

called a gray area, while still interfaced with the high-class cleanroom so that wafers can be transferred into it for processing. Chip testing and packaging are normally performed in an even lower-class cleanroom, since these processes deal with much larger feature sizes that are much less sensitive to particles compared to wafer processing.

2.4.1 Wafer processing area

The cleanroom used for wafer processing always has the highest class in an IC fab, and a wafer does not leave this area unless all of the processing steps are finished. After a wafer is sealed with the final passivation layer, the last masking and nitride/oxide etching processes open the bonding pads (or bump connections) and the photoresist is stripped. The wafers then are sent out to the testing and packaging area to finish the chip fabrication.

Wafer processing areas are usually separated into several processing areas, or bays, as shown in Fig. 2.12. They are: the wet bay, diffusion bay, photo bay, etch bay, diffusion bay, thin-film bay, and chemical mechanical polishing (CMP) bay. In a traditional fab, process engineers, process technicians, and production operators mainly work in the processing areas, while equipment engineers and maintenance technicians mainly work in the equipment areas. Application engineers from the process-equipment manufacturers also work in the processing area during tool start-ups, as they can help to troubleshoot processes when they do not perform as expected.

2.4.1.1 Wet bay

The wet bay is where wet processes take place. Photoresist stripping, wet etching, and wet chemical clean processes are the most common processes in the wet bay. Corrosive chemicals and strong oxidizers such as hydrofluoric acid (HF), hydrochloric acid (HCl), sulfuric acid (H_2SO_4), nitric acid (HNO_3), phosphoric acid (H_3PO_4), and hydrogen peroxide (H_2O_2) are commonly used in this area. Huge amounts of high-purity deionized (DI) water are used for wafer rinsing after the wet processes. Most acids used in the wet bay are corrosive. Nitric acid and

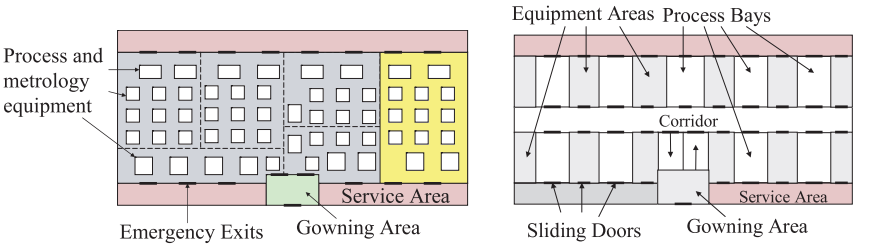


Figure 2.12 Floor plan of a semiconductor fab with (a) a minienvironment and (b) traditional layout.

hydrogen peroxide are strong oxidizers. In an IC fab, it should never be assumed that a clear drop of liquid is water. To be safe, a clear droplet should be treated as HF, because HF has no smell, looks like water, and feels like water. Wet bays always have shower stations and eye washers nearby so that people can use them immediately in case of accidental contact, such as from chemical spilling.

Wet processes are removal processes that usually take three steps—process, rinse, and dry—as illustrated in Fig. 2.13. Wet process tools usually are batch process tools, which can process whole cassettes of 25 wafers at the same time. Robots take the cassette with wafers from the loading station and dip it into a processing solution. After the required process time, robots pick the cassette up and put it into a rinse sink, where DI water washes away the chemicals from the wafer surface. Then the cassette is picked up and sent into a dryer, where wafers are spun dry with the cassette. After that, the cassettes are picked up and sent back to the loading station for wafer unloading. Some wet stations can process several cassettes of wafers at one time. Depending on the chemical used in the process, sometimes wafers need to be picked up from the cassettes and put on a quartz boat for the wet process and rinse.

To reduce chemical usage, single-wafer wet processing tools have also been developed. By processing wafers one at a time, only a small amount of chemical needs to be sprayed onto the wafer surface; thus, the amount of chemical used in the wet process can be significantly reduced in comparison to the batch system, in which the whole cassette is submerged in the wet chemical.

Some fabs do not have a separate wet bay; wet chemical stations are simply placed alongside the oxidation and LPCVD tools, since these processes require a wet clean beforehand.

2.4.1.2 Diffusion bay

The diffusion bay is where thermal processes are performed. These processes can be adding processes, such as oxidation, low-pressure chemical vapor deposition (LPCVD), and diffusion doping, or heating processes, such as postimplantation annealing, dopant drive-in, alloy annealing, and dielectric reflow. High-temperature furnaces are installed for oxidation, LPCVD, diffusion/drive-in

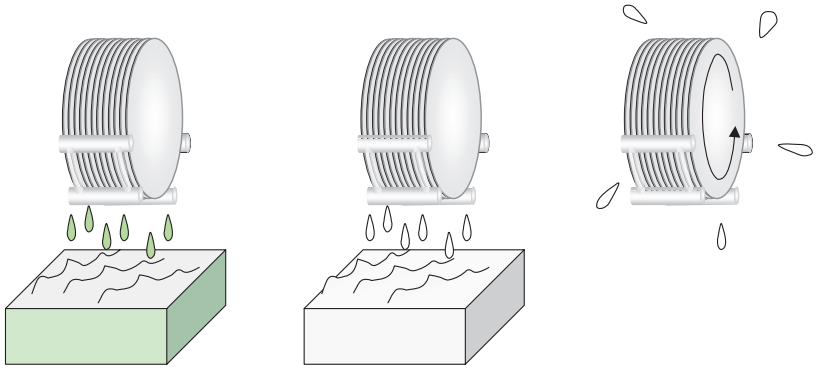


Figure 2.13 Example of wet processes.

processes, and annealing processes. Some fabs also have epitaxy reactors in their diffusion bays. Oxidation and diffusion doping in high-temperature furnaces were the most frequently used processes of IC fabrication before the introduction of ion implantation in the mid-1970s. Although very little diffusion doping processes are performed in advanced IC processing, the name diffusion bay is still used for historical reasons.

Furnaces are batch process tools with the ability to process more than 100 wafers at a time. Figure 2.14(a) gives the schematics of a vertical furnace, and Fig. 2.14(b) shows a horizontal furnace. Due to a smaller footprint and better contamination control, only vertical furnaces are used in advanced 200- and 300-mm IC fabs. Horizontal furnaces are still used in some older fabs with smaller wafer sizes. Most 300-mm fabs also have cluster tools with single-wafer chambers for deposition of polysilicon and silicon nitride films, and silicide annealing processes in their diffusion bays.

Gases commonly used in diffusion bays are oxygen (O_2), nitrogen (N_2), anhydrous hydrogen chloride (HCl), hydrogen (H_2), silane (SiH_4), dichlorosilane (DSC, SiH_2Cl_2), trichlorosilane (TSC, $SiHCl_3$), phosphine (PH_3), diborane (B_2H_6), and ammonia (NH_3). Nitrogen is a safe gas. Oxygen is an oxidizer, which can cause fire or explosions if mixed with other flammable, explosive materials at certain conditions. HCl is corrosive; hydrogen is flammable and explosive; silane is pyrophoric (self-igniting), explosive, and toxic; both DSC and TSC are flammable; and ammonia is corrosive. Both phosphine and diborane are very poisonous, flammable, and explosive. Nitrogen is used in almost every processing tool in an IC fab as the purge gas. O_2 and anhydrous HCl are used for the dry oxidation process, while H_2 and O_2 are used for the wet oxidation process. Silane, DCS, or TCS can be used as a silicon precursor for polysilicon deposition, and they can also be used with NH_3 for silicon nitride deposition. Phosphine and diborane are used as dopant gases for polysilicon deposition.

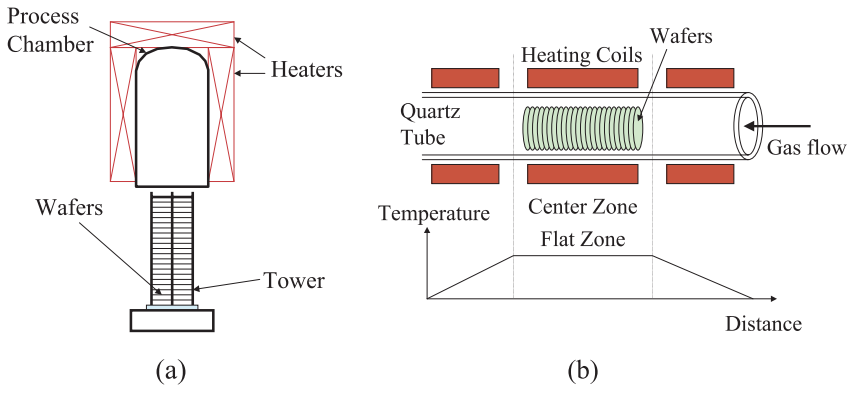


Figure 2.14 Schematics of (a) vertical and (b) horizontal furnaces.

2.4.1.3 Photo bay

The photolithography process is one of the most important processes in IC fabrication. It transfers the designed pattern from a mask or reticle to the photoresist coated onto the wafer surface. In the photo bay are located the integrated track-stepper systems, which perform primer and photoresist coating, baking, alignment and exposure, and photoresist development processing steps. A stepper is the most commonly used tool to pattern the photoresist coated onto the wafer surface by exposing the photoresist to ultraviolet (UV) light or deep UV (DUV) light, thereby inducing photochemical reactions. It is the most expensive tool in advanced IC fabs. For example, an advanced 193-nm immersion scanner could cost more than $40 million. There are also many metrology tools in photo bays, such as ellipsometry and reflectometry systems, for measuring the thickness and thickness uniformity of transparent films such as photoresists. There are also optical microscopes for visual inspection, and scatterometry systems for CD measurements. Scanning electron microscopes (SEMs) are also widely used for CD measurements, and optical overlay systems are used to measure the alignment between different mask layers. Figure 2.15 shows the schematic of an integrated track-stepper system, which is also called a photocell in an IC fab.

In an advanced IC fab, the track systems do not look like the schematic in Fig. 2.15. Many fabs prefer to use a stacked track system, which makes a much smaller footprint by stacking the hot plates and chill plates instead of placing them on the same plain. Some systems even stack the spin coaters and development stations to further reduce the footprint and save precious cleanroom space.

2.4.1.4 Etch bay

After the photoresist is patterned and passes inspection, the wafers are sent either to the implant bay or to the etch bay. In an etch bay, patterns are etched based on the pattern defined by the photoresist, permanently transferring the designed pattern to the layer on the wafer surface. Etching is a removal process in which materials on the wafer surface are selectively removed chemically, physically, or

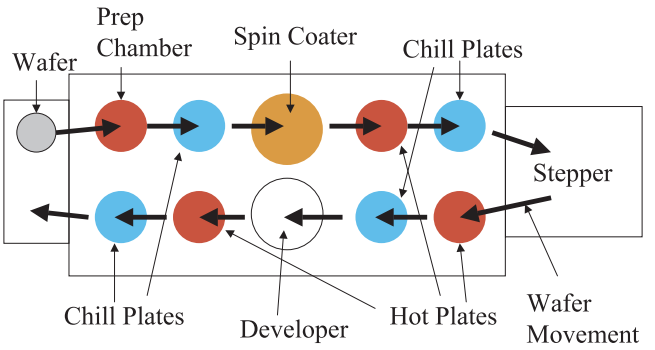


Figure 2.15 Schematic of a track-stepper integrated system in a photo bay.

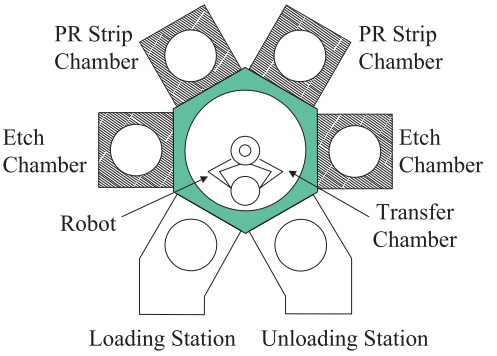


Figure 2.16 Schematic of a cluster tool with etch and photoresist strip chambers.

most commonly by a combination of both, to create patterns on the wafer surface. In the etch bay, only plasma etchers are used for the patterned etch, since wet etching cannot be used to etch features smaller than 3 μm. A plasma etcher usually consists of a vacuum chamber, rf system, wafer transfer mechanism, and gas delivery system. Many etch systems have in-situ photoresist strip chambers, which allow photoresists to be stripped within the system before exposure of the wafer to the atmosphere. Figure 2.16 illustrates a cluster tool with both etch and photoresist strip chambers.

There are four different kinds of etch processes commonly used in semiconductor fabrication. The dielectric process etches silicon oxide and nitride film to form contact via a bonding pad or silicon etch hard mask. The silicon etch process etches single-crystal silicon to form shallow trench isolation (STI). It also has been used to form deep trenches for capacitors in deep trench dynamic random access memory (DRAM). The poly process etches polysilicon or silicide and polysilicon stacked films to form gate and local interconnections. And the metal process etches metal stacks to form metal lines for global interconnections.

Because more fabs have started using copper metallization, less metal etch is required.

Each etch process has different requirements and chamber designs and uses different chemical gases. Carbon fluorides such as CF_4, C_2F_6, C_3F_8, and CHF_3 are commonly used along with argon for dielectric etch processes. Hydrogen bromide (HBr) is widely used to etch single-crystal silicon, and chlorine (Cl_2) is commonly used for both polysilicon and metal etch processes. None of these chemicals are user friendly. Although carbon fluoride gases are very stable, they are considered global warming gases; hydrogen bromide is corrosive; and chlorine is an oxidizer and poisonous.

2.4.1.5 Implant bay

Besides the etch bay, the implant bay is the only other place wafers are sent after the photolithography process. Ion implantation and rapid thermal annealing (RTA) systems can be found in this bay. Ion implantation is an adding process in which dopant is added to the semiconductor substrate to change its conductivity. RTA is a heating process that repairs lattice damage of the substrate at high temperatures while neither removing nor adding materials to the wafer surface. The ion implanter is usually the largest and heaviest processing tool in a semiconductor fab. It also involves many safety hazards, such as high voltage (to 100 kV), a strong magnetic field (which can affect pacemakers), and generation of strong x-ray radiation. It also uses poisonous, flammable, and explosive gases such as arsine (AsH_3) and phosphine (PH_3), poisonous solid materials such as boron (B), phosphorus (P), and antimony (Sb), and the corrosive gas boron trifluoride (BF_3).

In the implant bay, there are several different kinds of implanters used for different processes. For example, well formation in the CMOS process needs a high-energy, low-current implanter. To form the source and drain of the metal-oxide-semiconductor field effect transistor (MOSFET) requires a low-energy, high-current implanter. Other applications require a mid-current, mid-energy implanter. One of the properties of the ion implanter is that it can run different processes with different chemicals in the same implanter, without too much concern regarding cross-contamination due to high-purity ion beam selection by the magnetic analyzer. Metrology tools commonly used in the implant bay are the four-point probe, thermal wave system, and optical measurement system (OMS).

2.4.1.6 Thin-film bay

The main processes performed in the thin-film bay are dielectric and metal thin-film deposition for interconnection applications. Some fabs separate these two processes; therefore they have dielectric and metal bays instead of a single thin-film bay. Obviously, thin-film depositions are additive processes. CVD processes are always used for dielectric thin-film deposition. Plasma-enhanced CVD (PECVD) is commonly used due to the lower temperature requirement for dielectric layers in multilevel interconnection applications. O_3-TEOS [ozone tetra ethyl oxy silane $Si(OC_2H_5)_4$]-based CVD processes are also widely used to deposit silicate glasses

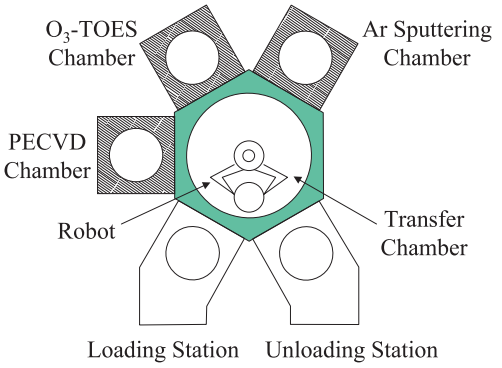


Figure 2.17 Schematic of cluster tool with dielectric CVD and etchback chambers.

due to its excellent gap-fill abilities. An argon (Ar) sputtering etch process is also performed in the thin-film bay to help gap-fill in dielectric deposition. Figure 2.17 illustrates a cluster tool with PECVD, O_3-TEOS, and Ar sputtering chambers.

Before 2000, metals used for connecting the millions of transistors in most IC chips were aluminum-copper alloy (Al·Cu), tungsten (W), titanium (Ti), and titanium nitride (TiN). Still underway is a transition from traditional interconnections to copper interconnections, in which copper (Cu) with a tantalum (Ta) or tantalum nitride (TaN) barrier layer is used in IC chips for interconnection. Copper replaced the aluminum-copper alloy in logic IC and microprocessors after the 0.18-μm technology node because it helps improve the speed and reliability of IC chips due to its lower resistivity and higher electromigration resistance. Copper is replacing aluminum-copper alloys in DRAM and flash memory for metal interconnects.

In metal bays, PVD tools (or mainly sputtering deposition tools) are used to deposit AlCu, Ti, and TiN, and CVD tools are widely used for W and TiN deposition. PVD tools are also used to deposit the Ta or TaN barrier layer and copper seed layer, while electrochemical plating deposition (EPD) tools are used to deposit bulk copper. PVD processes usually operate in a vacuum chamber, with a very high base vacuum to get rid of moisture in the chamber and minimize metal oxidation. Figure 2.18 illustrates a cluster tool with a PVD chamber for a Ta or TaN barrier layer and Cu seed layer deposition.

The most commonly used metrology tools in the thin-film bay are the reflectospectrometer, ellipsometer, stress gauge, laser scattering particle mapping tool for dielectric thin-film measurement, and four-point probe and profilometer for metal thin-film monitoring.

TEOS is the most commonly used silicon precursor for silicon oxide deposition, while silane (SiH_4) is commonly used for silicon nitride PECVD processes. Oxygen, ozone, and nitrous oxide (N_2O) are commonly used oxygen source gases, and nitrogen and ammonia are commonly used nitrogen sources. Nitrogen trifluoride (NF_3), or one of the carbon fluoride gases such as CF_4, C_2F_6, and C_3F_8,

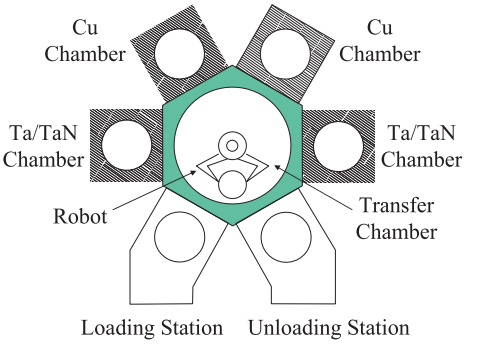


Figure 2.18 Schematic of a cluster tool with AlCu, Ti, and TiN PVD chambers.

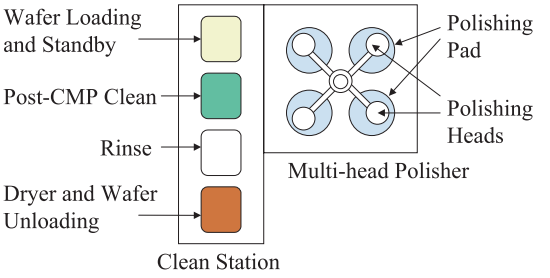


Figure 2.19 Schematic of a dry-in/dry-out multihead CMP system.

is widely used along with oxygen or nitrous oxide for dielectric CVD chamber cleaning. Silane is pyrophoric, explosive, and toxic; TEOS is combustible; O_3 and NF_3 are strong oxidizers; N_2O (laughing gas) can cause numbness; and carbon fluoride gases are considered global warming gases.

Metal PVD only uses argon and nitrogen, and neither of them is considered a safety hazard. Tungsten CVD widely uses tungsten hexafluoride (WF_6), silane (SiH_4), and hydrogen (H_2). WF_6 is corrosive, as it forms HF when it reacts with water (H_2O).

2.4.1.7 Chemical mechanical polishing bay

CMP is a removal process in which materials are moved from a wafer surface by the combination of mechanical grinding and wet chemical reaction. Commonly used CMP processes are silicate glass CMP, tungsten CMP, and most recently copper CMP. Timely post-CMP cleaning is a critical step to ensure the yield of CMP processing; therefore, some CMP tools are integrated with wet clean stations to form the so-called dry-in/dry-out CMP systems. Figure 2.19 illustrates a dry-in/dry-out multipolishing head CMP system.

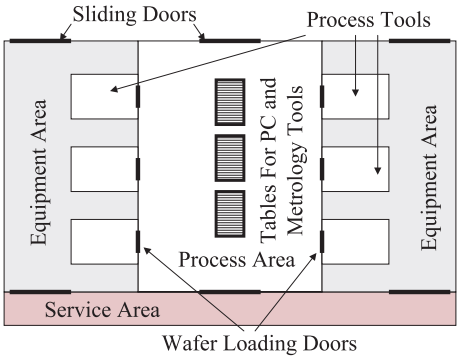


Figure 2.20 Process and equipment areas.

Fine particles, such as silica (SiO_2) or celica (CeO_2) in silicate glass CMP slurries and alumina in metal CMP slurries, are used as abrasives and play very important roles in CMP processes.

2.4.2 Equipment area

Many processing tools such as ion implanters, plasma etchers, CVD reactors, and PVD and CMP tools are located in gray areas of a fab, which are cleanrooms with higher airborne particle counts, or lower-class cleanrooms, typically at class 1000. By putting processing tools in a lower-class cleanroom instead of a high-class cleanroom such as the processing area, both fab building and maintenance costs can be significantly reduced.

Figure 2.20 illustrates the location of the processing and equipment areas in a semiconductor fab. Equipment engineers and technicians mainly work in the equipment area. Equipment support personnel from tool manufacturers also work in this area to help to install, start up, and maintain processing tools.

For the minienvironment fab illustrated in Fig. 2.12(a), there is no such separation of process and equipment areas. The entire cleanroom is laid out as one big ballroom.

2.4.3 Facility area

The facility area hosts facilities that support the processing tools and metrology instruments. It is not located in the cleanroom, and some facilities are not even in the same building as the cleanroom. Facilities needed for the wafer processes are gases, water, electric power, and subsystems that support the processing equipment. Facility engineers, technicians, and support personnel from facility providers mainly work in this area. They also need to work inside the cleanroom during installation and maintenance of the processing equipment.

Gases include process gases with ultra high purity, purge gas such as nitrogen with a lower purity than process gases, and dry air used for driving the pneumatic systems of the processing tools. A large amount of clean air is needed to constantly

feed into the cleanroom to make up for the loss of air while maintaining the positive pressure of the cleanroom. Large amounts of gases are consumed every day in IC fabrication, especially nitrogen, which is the most heavily used purge gas for purging process reactors and gas delivery lines. High-purity nitrogen is also used as a process gas in IC processing. Many fabs must have their own nitrogen plants to manufacture and purify the nitrogen needed from the atmosphere by freezing and vaporization of compressed air. Most process gases are held in high-pressure gas bottles, which are stored in specially designed gas cabinets located in small rooms built with thick concrete walls and doors that open only toward the outside. This design can prevent major damage to the fab in case of a gas explosion.

A large quantity of water is used in IC manufacturing. High-purity DI water is used for rinsing during wafer wet clean and wet etch processes. DI water is also used for the cooling of processing systems. IC fabs must have their own plants to generate the DI water needed in wafer processing. A large amount of city water is also used to cool the subsystems and other facility equipment, such as air conditioners. An IC fab also requires strong air-conditioning systems to maintain a constant temperature. Temperature variation can cause a drift of measurement results in metrology systems and induce yield loss due to process drift.

Electric power breaks and distribution systems are located in the facility area. Electric power consumption of an IC fab is enormous. For example, a diffusion furnace can consume about 28.8 kW (480 V by 60 A). Keep in mind that a furnace is always on unless it requires PM, and an IC fab usually has more than 50 of these furnaces.

Question: Assuming the average power of a furnace is 20 kW and a fab has 50 such furnaces, calculate the electricity bill for a year to keep these 50 furnaces operating if electric power costs 10 cents per kilowatt-hour.

Answer: $20 \times 50 \times 24 \times 365 \times 0.1 = \$876{,}000$.

Many subsystems of the processing equipment, such as vacuum pumps, rf power generators, gas and liquid delivery systems, and heat exchangers, are installed in the facility area, usually directly beneath the processing and equipment area, as illustrated in Fig. 2.8. Vacuum pumps are needed to evacuate air from the reactors and keep them under vacuum environments required by the processes. Radio frequency generators are needed to strike and maintain stable plasmas in many PECVD and plasma etch processing reactors. Gas and liquid delivery systems flow processing gases or liquids to the processing tools as required by the process. Heat exchangers provide chilled or heated water to the processing systems to maintain constant temperatures required by the different processes.

2.5 Testing and Packaging

After wafer processing is finished, wafers are sent out for testing and packaging. Some fabs do testing and packaging on site, and some fabs send the finished wafers out to a packaging plant. Die testing is a very labor-intensive process,

since it requires operators to carefully contact the fine pins of the test tool to tiny bonding pads by hand (about 100×100 μm^2) under an optical microscope for every die on the wafer. Some IC manufacturers build testing and packaging plants in developing countries to take advantage of lower labor costs. This could change due to rapid improvement of automation technology in IC testing equipment and the continuous shrinking of the percentage of labor costs in the total cost of IC manufacturing.

2.5.1 Die testing

It is very important to detect a faulty chip as early as possible in the IC manufacturing process. The earlier a problem is caught, the lower the costs for the manufacturer. First, a design team must verify the design and make sure there are no design defects before the layouts are sent out to the mask shop to make masks or reticles. Detecting a design fault costs a semiconductor fab nothing. Mask/reticles also need to be carefully inspected, and their condition is constantly monitored in IC fabs during wafer processing.

Throughout wafer processing, specially designed test devices and test circuits on the testing structures are randomly sampled and tested, to ensure yield for certain key processing steps, and to catch any process specification shift that affects the yield. Detecting a bad chip at the wafer level costs the semiconductor chip manufacturer about $0.10 per chip. The cost for detecting a malfunctioning chip at the packaging level is about $1. This cost increases rapidly from $1 to $10 per chip if a faulty chip is used at the electric board level. The cost further increases at the electronics system level ($10 to $100 per chip) and customer level (>$100 per chip). In the early 1990s, some malfunctioning microprocessors caused errors in certain division calculations running on their personal computers (PC), and it cost millions of dollars for the manufacturer to replace these faulty microprocessors.

After finishing all of the processes in a high-class cleanroom, wafers are transferred to a lower-class cleanroom (typically about a class 1000) for die testing, and even a lower-class cleanroom (typically about a class 10,000) for chip packaging. Specially designed test tools are used to test each die on the wafer where tiny pins are connected to the bonding pads or bumps on the chip, and the test programs are carried out to test and verify whether the IC chip meets the design requirements. Faulty die are inked with a dot, as shown in Fig. 2.21, and are not picked up for packaging after die separation. Advanced test-sorting tools store the information of die condition in an on-board computer and do not use ink to mark the bad dies because ink can cause contamination and must be constantly refilled. Mistakenly ink-marking good dies or missing the mark on bad dies due to loss of ink can cause unwanted costs for the IC manufacturer.

2.5.2 Chip packaging

There are four main reasons for chip packaging: to provide physical protection for the IC chip, to provide a barrier layer against chemical impurities and moisture,

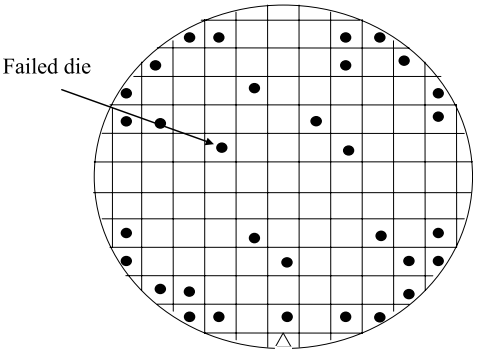


Figure 2.21 Tested wafer.

to allow the IC chip to connect to an electric circuit with a sturdy set of leads, and to dissipate heat generated during chip operation. The main concerns with chip packaging are process cost, process temperature, and packaging materials, including their thermal conductivity, coefficient of thermal expansion, and Young's elastic modulus.

After chip testing, wafers are normally coated with a protective layer on the surface and mechanically polished on the backside to reduce wafer thickness. The backside thinning process can help remove the backside coating from the wafer process and improve the contact of metal coating to the silicon substrate. It also reduces the thickness of the chip, generally from 600 to 775 μm to 250 to 350 μm, so that chips can easily fit into the shallow pocket on the lead frame. Some chips need to be thinned to about 100 μm to fit into a smart card. Other techniques, such as wet etch and plasma etch, are also developed for the wafer thinning process. A thin metal layer usually is coated onto the backside of the wafer after wafer thinning. Gold is commonly used for backside metallization, and both sputtering and evaporation processes can be used for gold metallization.

After removal of the protective layer from the wafer surface, the wafer is taped onto a solid frame from the backside with a sticky, flexible material such as Mylar. The tape helps prevent wafer movement and holds the wafer in one piece during die separation. Diamond-impregnated saws operating at very high speeds (up to 20,000 rpm) with constant coolant flow are used to separate the individual die from the wafer along the scribe lines. Other technologies such as scribing and mechanical breaking have been used; however, they are no longer in use because they can cause chipping and cracking.

In the die sorting process, good chips (without dots) are picked up, while bad ones (with dots) are left behind. Good dies are attached to a packaging socket, which is a good conductor for both heat and electricity. Metal or metal-coated ceramic are commonly used as die-attaching materials. In a thermal process, solder between the metal-coated chip backside and metal-coated substrate surface melts and attaches the chip to the substrate when it cools down. Temperature for the die-attaching process is limited by the fact that tiny aluminum alloy wires are

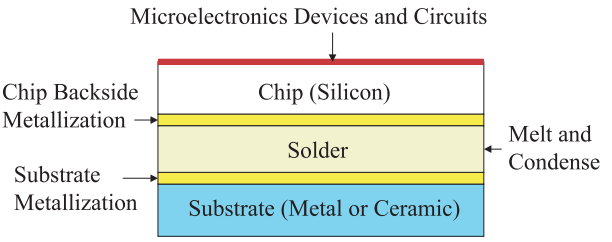


Figure 2.22 Chip bond structure.

used in many IC chips, and aluminum alloy cannot sustain a temperature higher than 550 °C due to the eutectic of aluminum and silicon. Rule of thumb is that no thermal process with a temperature higher than 450 °C is allowed after aluminum metallization. A gold-silicon eutectic from the heat is commonly used in the IC industry, because a mixture of gold and silicon creates a low melting temperature of the alloy. A mixture of 97.15% Au and 2.85% Si by weight melts at 363 °C. Figure 2.22 illustrates the basic structure of die attachment.

After the chip is attached to the lead frame, a soldering machine is used to make the connection between bonding pads on the chip and lead pins on the lead frame with thin metal wires. A traditional wire bonding process is illustrated in Fig. 2.23.

First, a molten ball is formed by a high-temperature heater at the tip of the metal wire in the soldering head. Then the soldering head presses the molten metal ball to the bonding pad surface to solder the metal wire with the bonding pad. The head then retreats and bends the wire to the pin lead, heats, and then presses the wire to solder it with the pin lead. The clamp closes as the soldering head moves away, breaking the metal wire due to excessive extension stress. After a short time, the tip of the broken metal wire forms a molten ball, again, due to surface tension, and the soldering head is ready for the next soldering step. Gold wire is commonly used in the wire bonding process. It is important that the wire bonding temperature is not higher than the chip bonding solder melting point. Figure 2.24 illustrates a chip with bonding pads, and the traditional packaging technology with wire bonding of the IC chip to the lead pins of the chip socket.

An alternative packaging technology has been developed and is becoming widely used in IC manufacturing. Different from the traditional wire bonding process, flip-chip technology forms metal bumps instead of bonding pads on the IC chip surface after passivation dielectric etch. After die separation, the chip is placed into a socket face down, and metal bumps on the chip surface and metal leads on the socket surface are precisely aligned and contacted. In a heated environment, the metal bumps and pin leads melt and join together after the chip is cooled, forming connections between the IC chip and the pins on the socket. This technology is called flip-chip packaging because in traditional wire bonding packaging, chips are placed face up in the lead frame, while in flip-chip packaging, the chip must be mounted face down in the sockets. One advantage of flip-chip packaging is that it

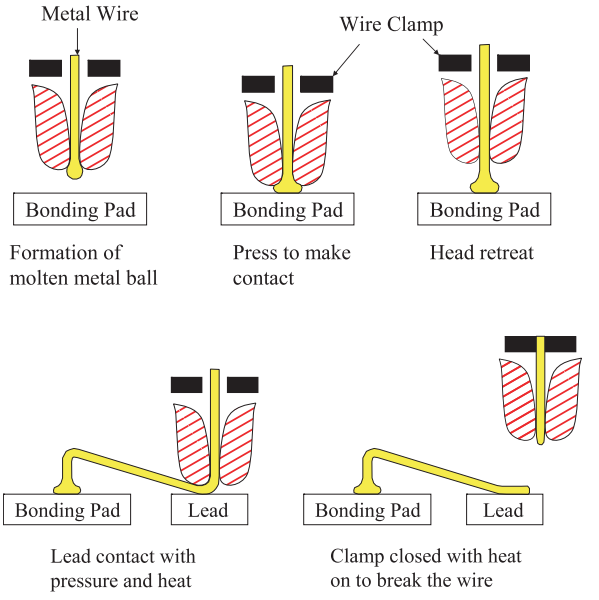


Figure 2.23 Wire bonding process steps.

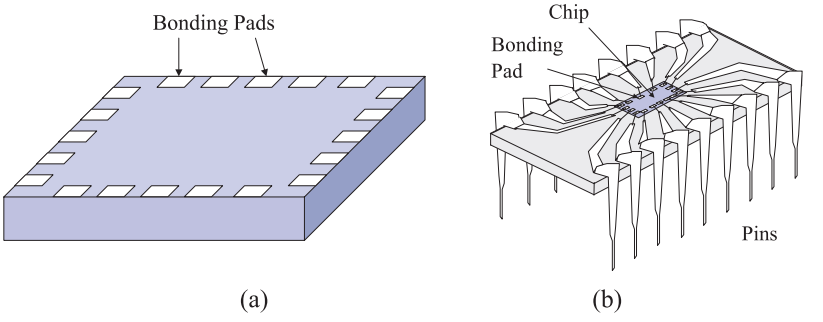


Figure 2.24 IC chip with (a) bonding pads and (b) wire bonding.

can significantly reduce package size. Figure 2.25 illustrates the chip with bumps, and Fig. 2.26 illustrates flip-chip packaging process steps.

After the IC chip is connected to the lead pins of the chip socket through wire bonding or bump connection, the chip and lead frame are ready to be sealed. A preseal inspection is conducted to eliminate chips broken during the mechanical handling and thermal processes of the die attaching and wire bonding (or bumps) connection. There are two kinds of chip sealing processes: ceramic and plastic. Ceramic is a much better barrier material for impurities such as mobile ions and moisture, and it also has much higher thermal stability, higher thermal conductivity,

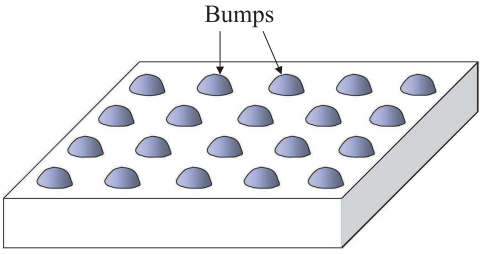


Figure 2.25 IC chip with metal bumps.

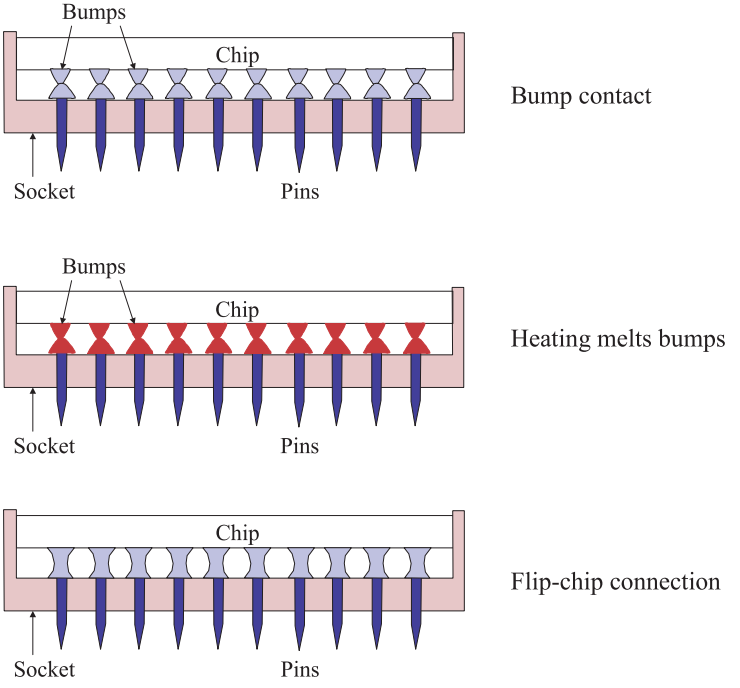


Figure 2.26 Flip-chip packaging technology.

and lower coefficients of thermal expansion. However, it is much more expensive than the plastic packaging, by about a factor of 5 per chip packaging, and it is also much heavier. Therefore, to reduce packaging costs, people like to use plastic packaging whenever it is applicable. For example, most memory chips and logic chips use plastic packaging. Some IC chips, especially microprocessors for computers, or CPUs, require ceramic packaging because of the large amount of heat generated during their operation. Ceramic packaging was previously the standard in the IC industry; now, more and more chips use plastic packaging.

Ceramic packaging usually uses metal solder to seal the cap and the lead frame. During the sealing process, the ceramic cap is heated and pressed to the lead frame.

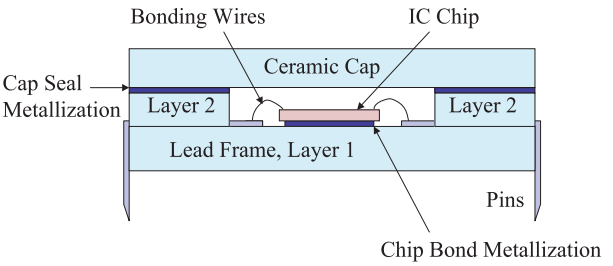


Figure 2.27 Cross-section of IC chip ceramic packaging.

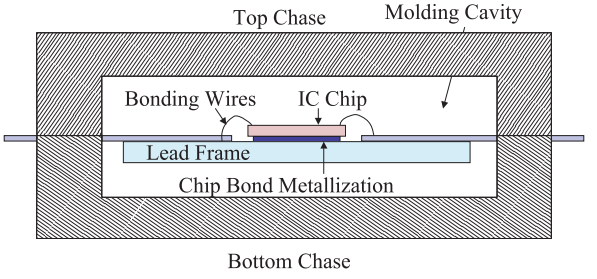


Figure 2.28 Cross section of the molding cavity for plastic packaging.

The heat melts the metal layers coated on surfaces and solders them together. Figure 2.27 illustrates the cross section of the ceramic packaging of the IC chip.

Plastic packaging uses a molding technique to seal the IC chip and lead frame with plastic. After wire bonding, the lead frame is placed between the top and bottom chases of the packaging tool. The heated chases close to form a cavity surrounding the chip and the lead frame. Molten plastic flows into the cavity and seals the chip after cooling and plastic condensation. Figure 2.28 shows the molding system for chip plastic packaging.

Plastic is not a very good barrier for moisture and other impurities such as mobile sodium ions. These materials can penetrate the plastic seal and affect IC performance, which can affect IC chip reliability. However, the improvement of packaging technology helps extend the lifetime of a plastic packaged IC chip to more than 5000 h in the failure-inducing test, a time period that is equivalent to more than ten years of life in normal conditions, enough for most electronic applications.

2.5.3 Final test

After packaging is finished, chips are sent to a final test, which places them in an unfavorable environment to force any unreliable chips to fail. High acceleration

in a spinner can mechanically cause any loosely bonded wires to break off from the bonding pad, and high-temperature thermal stress can accelerate the electrical failure of a chip. Chips then are put on racks to run their normal operation continuously for several days, in what is called a burn-in process. Some microprocessor chips that fail to operate at their designed frequency can work fine at lower frequencies and hence are often sold as lower-speed processor chips at lower prices. Chips that pass the burn-in process can be shipped to customers (mainly the electronic industry) for placement on electronic boards to build electronic systems.

2.5.4 3D packaging

In the future, 3D packaging technology, which allows IC manufacturers to stack multiple chips in a package, will be developed and implemented in production. Stacking two chips is almost the equivalent of reducing feature size by a technology node and doubling device density. By stacking four or eight chips, the effect is more significant. When the cost of scaling feature size skyrockets, 3D packaging will become more attractive and could be more cost effective.

3D packaging has already been used in CMOS image sensors and other devices. Through-silicon via (TSV) is one of these enabling technologies that could be used in the implementation of 3D packaging. There are several options of TSV 3D packaging processes in development. Figure 2.29 illustrates one of them, in which the TSV is formed in an IC fab after transistor formation and before the forming of the interconnection, as shown in Figs. 2.29(a) to 2.29(e). After wafers are sent to a packaging house, the backside of the wafer can be thinned to expose the buried TSV plugs used to form the backside bumps. The thickness of the wafer after thinning depends on the depth of the TSV holes. Figure 2.29(f) shows the wafer with backside bumps, which is ready for 3D packaging.

If the die yield of the wafers is high, fabs have the option to perform wafer-to-wafer stacking, which aligns wafers so that the backside bumps of the top wafer accurately connect with the front-side bond pad of the wafer underneath. A heating process melts the solder bumps to form an electrical connection, bonding the two wafers together, as shown in Fig. 2.29(g). In case die yield of the wafers is not very high, die to known-good-die (KGD) stacking is more cost effective. In this case, a good die is picked up and aligned with a KGD on the handling wafer so that the backside bump of the die can accurately contact with the front-side bond pad of the KGD. A heating process melts the solder bumps to form an electrical connection and bond the two dies.

2.6 Future Trends

In the near future, IC processing materials will have some significant changes. High-dielectric-constant (κ) dielectrics, or high-k dielectrics, will replace silicon dioxide and silicon oxynitride as the gate dielectrics in advanced logic IC chips. Low-κ dielectrics have replaced commonly used silicon oxide as insulators for

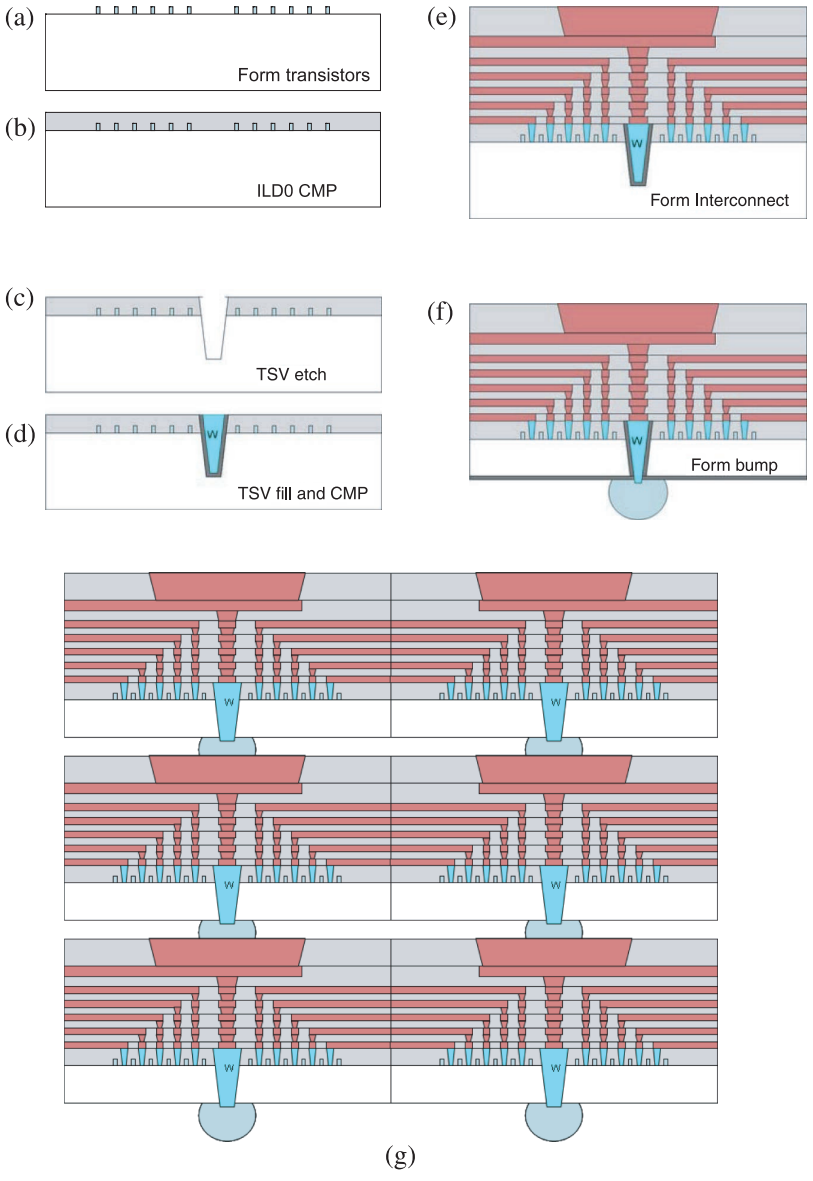


Figure 2.29 TSV 3D packaging processes.

the interconnection of IC chips since the 90-nm technology node, and they are being used to replace silicon oxide in memory devices. The standard κ-value to determine high-κ or low-κ is 3.9, which is the dielectric constant of silicon dioxide. Copper interconnects have become mainstream for the interconnection of logic devices since the 0.18-μm technology node and are gradually replacing aluminum alloy interconnects in memory devices in advanced nanometer technology nodes. Immersion lithography, which can improve patterning resolution, became widely used after the 45-nm technology node. A double-patterning technique that can further improve pattern density has been adopted in the 32-nm technology node and beyond. CMP processes will be more widely used in IC manufacturing, especially when memory devices start adopting copper metallization. The combination of copper and low-κ dielectrics can help increase the speed of IC chips. Facilities that support IC processes will change accordingly.

Faster, more precise, more reliable, and fully automatic IC test tools will be developed for more-complicated IC testing in the future. More IC manufacturers will adopt the flip-chip packaging technique for chip packaging processes. Multilayer lead frames for multichip packaging will be more commonly used, especially stacked multichip packaging. 3D packaging will see more development, and very likely 3D packaging using TSV technology will go to production in the near future.

2.7 Summary

- Overall yield is the ratio between the total number of good chips after the final test and total number of dies on all of the wafers used to manufacture this chip.
- Yield determines whether IC chip fabrication is losing money or making a profit.
- The cleanroom is a controlled environment in which particle concentration in air is kept very low by using filters and controlling airflow, pressure, temperature, and humidity.
- It is very important to strictly follow cleanroom protocols to reduce airborne particulates and other contaminants; otherwise, these contaminants can affect chip yield and fab profit.
- An IC fab normally has a photo bay, diffusion bay, implantation bay, etch bay, thin-film bay, CMP bay, and wet bay.
- The processing area consists of several processing bays, and is a high-grade cleanroom. For manufacturing IC chips with 0.25-μm feature size, the processing area requires a class-1 cleanroom. For equipment areas, class 1000 is good enough and costs much less to build and maintain. To achieve better than a class-100 cleanroom, laminar flow is required. Turbulent airflow can be used for a class-1000 cleanroom.
- Gas delivery systems, DI water systems, electric power distribution systems, pump and exhaust systems, and rf power systems are the most commonly used facility systems in IC fabrications.
- Chip packaging provides physical and chemical protections for IC chips, provides the fine wiring that connects the chip to the lead pins on a sturdy base, and helps to dissipate the heat generated during chip operation.

- Compared with plastic packaging, ceramic packaging provides better protection for chemical and moisture contamination and better physical protection. It also has much better thermal stability and higher thermal conductivity. The main advantage for plastic packaging is its significantly lower cost. Most IC chips now use plastic packaging.
- Standard wire bonding uses fine metal (gold) wire to connect chip bonding pads to lead pin bonding pads. The flip-chip uses bump bonding to connect chips to lead pins.
- For standard wire bonding packaging processes, the temperature for chip sealing should be low enough to not affect wire bonding and die attachment. Wire bonding temperature should not affect die attachment, and die attachment temperature is limited by aluminum's melting point.
- In final testing, a packaged chip is tested in some abnormal conditions such as higher temperatures, high acceleration, high humidity, etc., to intentionally cause unreliable IC chips to fail before they are sent to customers.
- 3D packaging has already been used in CMOS image sensors. Development of technologies such as TSV will help more IC products adopt 3D packaging in the near future.

2.8 Bibliography

R. Carranza, *Silicon Run II* (video tape), Ruth Carranza Productions, Mountain View, CA (1993).

C. Y. Chang and S. M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

S. M. Sze, *VLSI Technology*, *Second Edition*, McGraw-Hill, New York (1988).

S. Wolf and R. N. Tauber, *Silicon Processing for the VLSI Era, Vol. 1, Process Technology*, *Second Edition*, Lattice Press, Sunset Beach, CA (2000).

See http://www.set3.com/papers/, paper 209e (last accessed Sept. 2012).

2.9 Review Questions

1. How many particles larger than 0.5 μm in a cubic foot are allowed in a class-1000 cleanroom?
2. Sodium is a mobile ion, and trace amounts of sodium ions can damage microelectronic devices. What is the main source of sodium?
3. Explain the relationship between die yield and defect density, die size, and number of process steps by using Eq. (2.2).
4. What is the yield of the wafer shown in Fig. 2.21? (Total dies should exclude those on the wafer edge; these should be treated as incomplete dies.)
5. Why does a class-1 cleanroom need laminar flow?
6. List at least two pieces of processing equipment that can be found in a photo bay.

7. List at least two pieces of metrology equipment that can be found in photo bay.
8. Explain why a class-1000 cleanroom is usually used in a minienvironment IC fab instead of a higher class-1 cleanroom.
9. List at least two processing systems in a diffusion bay.
10. CVD and PVD are _____ processes.
    (a) adding (b) removing (c) patterning (d) heating
11. Photolithography is a _____ process.
    (a) adding (b) removing (c) patterning (d) heating
12. Anneal is a _____ process.
    (a) adding (b) removing (c) patterning (d) heating
13. CMP is a _____ process.
    (a) adding (b) removing (c) patterning (d) heating
14. Nitrogen can be used both as a purge gas and processing gas in IC fabrication. Does nitrogen for these two applications require the same purity?
15. Explain why people no longer use ink to mark the failed dies.
16. List the main purposes of chip packaging.
17. Describe the chip bonding and wire bonding processes.
18. What are the advantages and disadvantages of ceramic packaging compared with plastic packaging?
19. What is the purpose for the failure-inducing final test?
20. What is the benefit of using multichip packaging?
21. Describe flip-chip packaging.
22. Describe the TSV process.

Chapter 3
Semiconductor Basics

This chapter provides the basics of a semiconductor, semiconductor devices, and semiconductor processes.

Objectives

After finishing this chapter, the reader will be able to:

- identify at least two semiconductor materials from the periodic table of elements
- list n-type and p-type dopants
- describe a diode and MOS transistor
- list three kinds of chips made in the semiconductor industry
- list at least four basic processes required for chip manufacturing.

3.1 What Is a Semiconductor?

Semiconductors are materials with electrical conductivity between good conductors such as metal (including copper, aluminum, tungsten, etc.) and good insulators such as rubber, plastic, and dry wood. The most commonly used semiconductor materials are silicon (Si) and germanium (Ge), both of which are located in column IVA of the periodic table of elements, shown in Fig. 3.1. Some compounds such as gallium arsenate (GaAs), silicon carbide (SiC), and silicon germanium (SiGe) are also semiconductor materials. One of the most important properties of a semiconductor is that its conductivity can be controlled by intentionally adding certain impurities, a process called doping, and by applying an electric field.

3.1.1 Bandgap

The fundamental difference between a semiconductor and an insulator or a conductor is the bandgap. Atoms form all materials, and every atom has its own orbital structure [see Fig. 3.2(a)]. Electron orbits of the atom are called shells because electrons orbit around the nucleus in a 3D shell. Thus, the orbits in Fig. 3.2(a) can be seen as the cross section of these shells. The outermost shell is called a *valence shell*. Electrons in a valence shell cannot conduct an electric

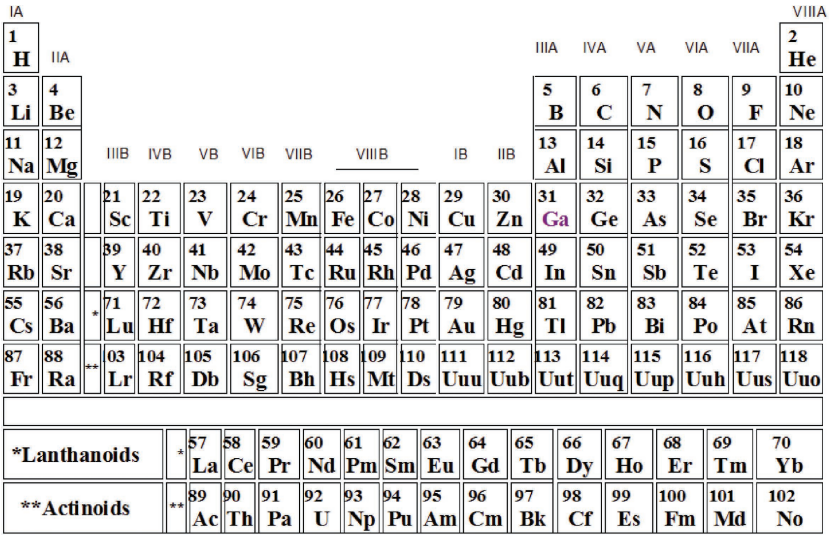


Figure 3.1 Periodic table of elements.

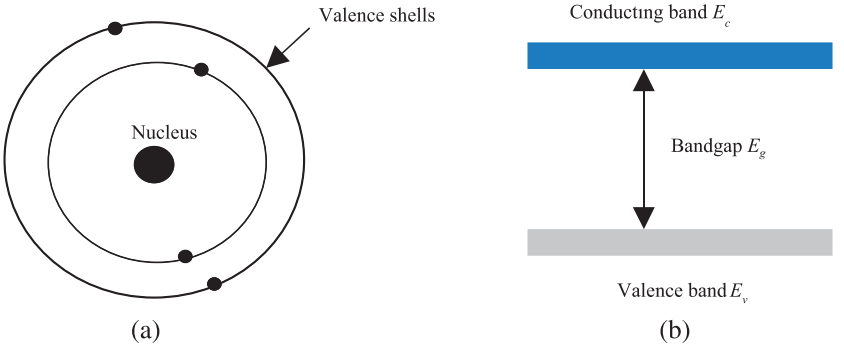


Figure 3.2 (a) Orbital structure of a single atom and (b) an energy band.

current. When an electron escapes the constriction of the nucleus and leaves the valence shell, it becomes a free electron and can conduct an electric current.

When many atoms bond together to make solid materials, their orbits overlap and form bands, as shown in Fig. 3.2(b). Electrons in conducting bands can move relatively freely inside solid materials and can conduct electric currents when an electric field is applied to the solid material. Electrons in the valence band are bound with the nuclei and cannot move freely; therefore, they cannot conduct electric currents. Since the valence band has lower electric potential, electrons always tend to stay in the valence band.

Resistivity is the ability of a material to resist an electric current. A good conductor has a very low resistivity, and a good insulator (or dielectric) has a very high resistivity. The most commonly used unit of resistivity is micro-ohm

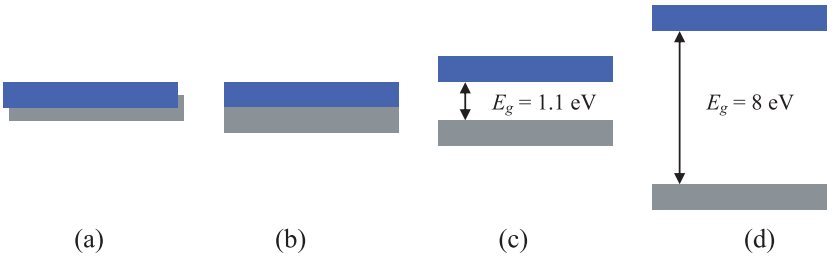


Figure 3.3 (a) Energy bands and bandgap of Al, (b) Na, (c) Si, and (d) SiO_2.

times centimeter ($\mu\Omega \cdot$ cm). At room temperature, the resistivity of aluminum is 2.7 $\mu\Omega \cdot$ cm, sodium is 4.7 $\mu\Omega \cdot$ cm, intrinsic silicon is about 10^{11} $\mu\Omega \cdot$ cm, and silicon dioxide is $>10^{18}$ $\mu\Omega \cdot$ cm.

For most metals, the conducting band and valence band overlap or have a very small bandgap, so small that electrons with thermal energy at room temperature (300 K $\approx$ 0.0259 eV) can jump across it. One electron volt (1 eV) is the energy gained by an electron when it passes through two points with one volt (1 V) of voltage difference. Therefore, a conducting band has many electrons, explaining why metals are always good electrical conductors. For dielectrics such as glass and plastic, the bandgap is so large that electrons cannot jump across it from the valance band, so the conducting band has very few electrons to conduct electric currents.

For semiconductors, the bandgap is somewhere between that of conductors and insulators; for instance, silicon is 1.1 eV, germanium is 0.67 eV, and gallium arsenate is 1.40 eV. While most electrons stay in the valance band, there are always some thermal electrons (from the tail of the Boltzmann distribution, discussed in Chapter 7) that can jump into the conducting band and conduct electric current. For intrinsic silicon at room temperature (300 K), there are about 1.5×10^{10} per cubic centimeter (cm^{-3}) electrons in the conducting band. This means at room temperature, only about one in ten trillion electrons is in the conducting band, while the majority of electrons remain in the valence band. Therefore, intrinsic semiconductors can conduct electric currents at room temperature better than dielectrics, but not as well as conductors.

3.1.2 Crystal structure

From their position in column IVA of the periodic table of elements, we can see that the most commonly used semiconductor materials—silicon and germanium—both have four electrons in the outermost shell. In a single-crystal structure, every atom is bonded with four atoms and shares a pair of electrons which each of them, (see Fig. 3.4). Since most IC chips are made from a silicon substrate, this text mainly covers the silicon process and uses silicon to explain doping.

3.1.3 Doping semiconductor

By intentionally adding dopants into pure single-crystal semiconductor materials such as silicon, conductivity can be improved. There are two kinds of dopants:

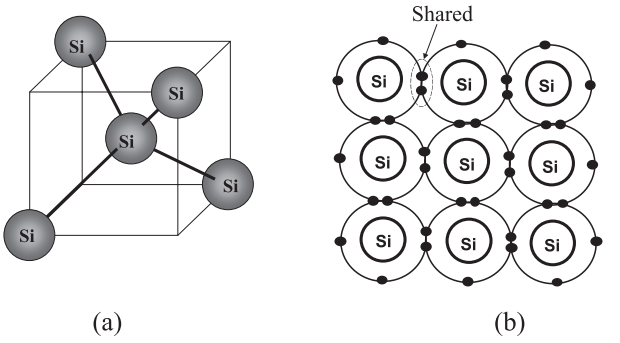


Figure 3.4 Actual tetrahedron silicon single-crystal structure and (b) its 2D simplification.

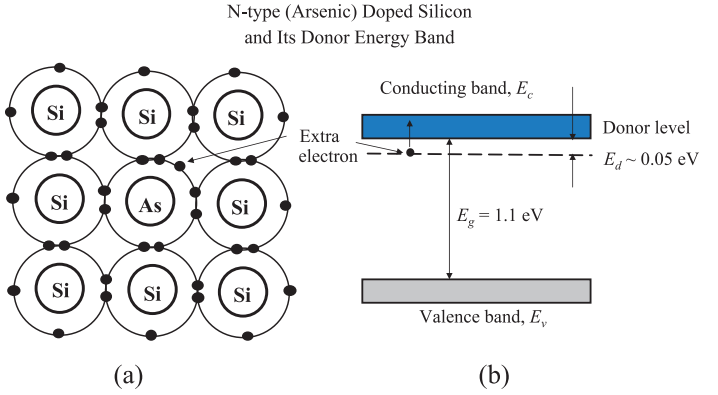


Figure 3.5 (a) n-type (arsenic) doped silicon and (b) its donor energy band.

p-type such as boron (B) from column IIIA of the periodic table, and n-type such as phosphorus (P), arsenic (As), and antimony (Sb) from column *VA*. Since P, As, and Sb provide an electron in semiconductor materials, they are called donors. Boron provides a hole, which allows other electrons to jump in and create a hole elsewhere. It is called an acceptor.

Phosphorus and arsenic have five electrons in their outermost shell. When P or As is doped into pure single-crystal Si or Ge, there is an extra electron in the outermost shell [see Fig. 3.5(a)]. This extra electron can easily jump into the conducting band and become a free electron to conduct an electric current. In this case, the majority carriers of electric currents are electrons, and the semiconductor is an n-type dopant (n stands for negative) because electrons have negative charges. P, As, and Sb are n-type dopants. The more n-type dopant atoms enter the semiconductor substrate, the more free electrons they will provide to conduct electric current, and therefore the better the conductivity of the semiconductor.

Figure 3.6(a) shows the empty spot (or hole) in the outermost shell left when boron is doped into pure single-crystal silicon or germanium. Figures 3.6(c),

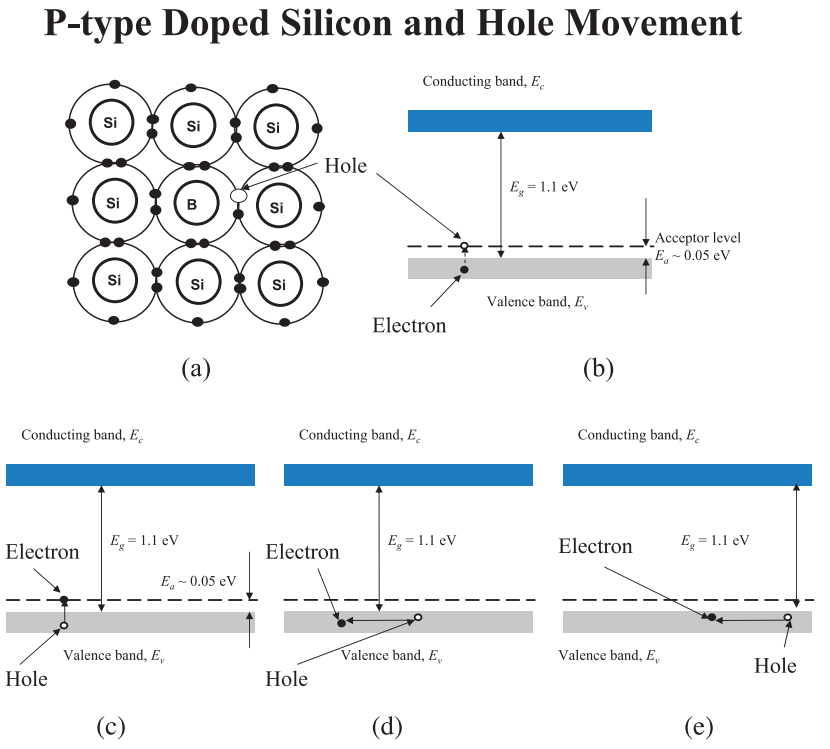


Figure 3.6 (a) p-type (boron) doped silicon, (b) its acceptor energy band, and (c), (d), and (e) illustrations of hole movement.

3.6(d), and 3.6(e) show that electrons in the valence band can easily jump to the acceptor energy band, creating holes in the valance band. In the electric field, other electrons in the valance band move and jump into these holes, while creating new holes where these electrons originated, allowing other electrons to jump in. The subsequent hole movement carries an electric current just like the movement of positive charges. A semiconductor with holes as majority carriers is a p-type dopant (p stands for positive). Boron is the main p-type dopant used in the IC industry.

3.1.4 Dopant concentration and conductivity

Figure 3.7 shows the relation between silicon resistivity and dopant concentration. From it we can see that the higher the dopant concentration, the lower the resistivity. This is because the more dopant atoms there are in silicon, the more carriers (electrons or holes, determined by type of dopant) they will provide to conduct electric current.

Figure 3.7 also shows that at the same dopant concentration, phosphorus-doped silicon (n-type) has lower resistivity than boron-doped silicon (p-type). This is because electrons move significantly faster in the conducting band than holes move in the valance band.

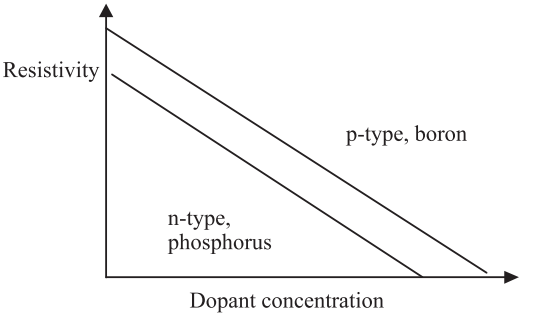


Figure 3.7 Dopant concentration and resistivity for silicon.

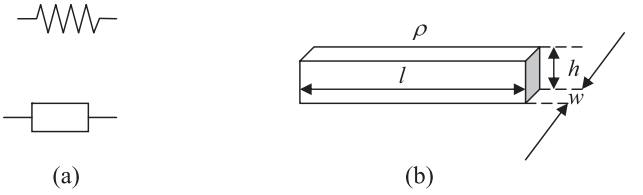


Figure 3.8 (a) Symbols and (b) structure of a resistor.

3.1.5 Summary of semiconductors

- Semiconductors have conductivity between conductors and insulators.
- Semiconductor conductivity can be controlled by dopant concentration: the higher the dopant concentration, the lower the semiconductor's resistivity.
- Holes form the majority carriers in p-type semiconductors. Boron is a p-type dopant.
- The majority carriers of n-type semiconductors are electrons. Phosphorus, arsenic, and antimony are p-type dopants.
- At the same dopant concentration and temperature, n-type semiconductors have lower resistivity than p-type semiconductors because their electrons have higher mobility than their holes.

3.2 Basic Devices

The basic devices in an IC chip are a resistor, capacitor, diode, transistor, and MOSFET.

3.2.1 Resistor

A resistor is the simplest electronic device. Its symbol in an electronic circuit and its structure are shown in Fig. 3.8(a).

Resistance of a resistor is:

$$R = \rho \frac{l}{wh}, \tag{3.1}$$

where R is the resistance, ρ is the resistivity, and w, h, and l are the conductor line width, height, and length, respectively [see Fig. 3.8(b)]. This equation is used again in Chapter 11.

In the past, patterned doped silicon was used to make resistors, with resistance determined by the length, line width, junction depth, and dopant concentration. Currently, doped polysilicon lines are used to make resistors on an IC chip, with resistance determined by the height, width length, and dopant concentration of the poly line.

Question: Many fabs use polysilicon to form gates and local interconnections. Resistivity of polysilicon is determined by dopant concentration, which is very high, about 10^{22} cm^{-3}, and $\rho \sim 200\ \mu\Omega \cdot$ cm. Assuming that the polysilicon gate and local interconnection line width, height, and length are 1, 1, and 100 μm, respectively (note: 1 μm = 10^{-6} m = 10^{-4} cm), what is the resistance?

Answer: $R = \rho \frac{l}{wh} = 200 \times \frac{100 \times 10^{-4}}{10^{-4} \times 10^{-4}} = 2 \times 10^8\ \mu\Omega = 200\ \Omega$.

Question: From the late 1980s to the late 1990s, the minimum feature size (gate width) shrank from 1 to 0.25 μm. What is the resistance of a polysilicon line with width, height, and length equal to 0.25, 0.25, and 25 μm, respectively?

Answer: $R = \rho \frac{l}{wh} = 200 \times \frac{25 \times 10^{-4}}{0.25 \times 10^{-4} \times 0.25 \times 10^{-4}} = 8 \times 10^8\ \mu\Omega = 800\ \Omega$.

It is evident that while device dimension shrinks, resistance increases! It is very important to keep resistance low, otherwise, it will slow down the device speed and increase power consumption and heat generation. Therefore, it is necessary to use a better conducting material such as silicide, which has a much lower resistivity, from 13 to 50 $\mu\Omega \cdot$ cm. In advanced CMOS IC chips, fabs normally use a polysilicon-silicide stack, also known as polycide, to form the gate and local interconnections.

Question: At one time, aluminum-copper alloy was the most commonly used material for metal interconnection in the IC industry. Assuming the resistance of aluminum-copper alloy $\rho \sim 3.2\ \mu\Omega \cdot$ cm, and metal line width, height, and length are 1, 1, and 100 μm, respectively, what is the resistance of the metal interconnection?

Answer: $R = \rho \frac{l}{wh} = 3.2 \times \frac{100 \times 10^{-4}}{10^{-4} \times 10^{-4}} = 3.2 \times 10^6\ \mu\Omega = 3.2\ \Omega$.

Question: Copper has been used for metal interconnections in semiconductor manufacturing since the late 1990s. Copper has a significantly lower resistivity than aluminum copper alloy, $\rho = 1.7\ \mu\Omega \cdot$ cm. What is the resistance of a copper line with the same geometry as the first question?

Answer: $R = 1.7\ \Omega$.

By using copper for metal interconnection, resistance can be reduced by almost half, which can be translated into higher device speed, lower power consumption, and less heat generation.

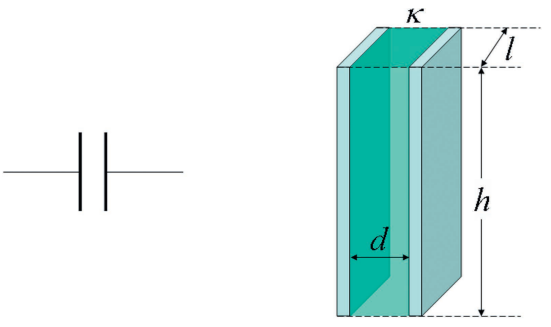


Figure 3.9 (a) Symbol and (b) structure of a capacitor.

3.2.2 Capacitor

The capacitor is one of the most important components in an IC chip, especially for a dynamic random access memory (DRAM) chip. Its symbol in an electronic circuit and its structure are shown in Fig. 3.9.

Capacitors are very important for DRAM chips that store electrical charges and maintain memory. When two conducting surfaces are separated by a dielectric, a capacitor is formed, and its capacitance can be expressed as

$$C = \kappa\varepsilon_0 \frac{hl}{d}, \tag{3.2}$$

where C is the capacitance of the capacitor, h and l are the conducting plate height and length, respectively, and d is the distance between the two parallel conducting plates. $\varepsilon_0 = 8.85 \times 1010^{-12}$ F/m is the absolute permittivity of a vacuum, and κ is the dielectric constant of the dielectric material between the two parallel plates. (In most academic disciplines, a dielectric constant is denoted as ε, but since this text is orientated toward industrial work, we use κ instead of ε.)

In an IC chip, the conductor part of the capacitor is mainly made with polysilicon. Dielectric materials vary from silicon dioxide and silicon nitride to more recent high-k dielectrics such as titanium dioxide (TiO_2), hafnium dioxide (HfO_2), and aluminum oxide (Al_2O_3). The purpose of using a high-κ dielectric is to shrink capacitor size while keeping the same capacitance. Capacitors can be made in planar, stack, and deep trench styles, as shown in Fig. 3.10. Stack and deep trench styles are favored by DRAM manufacturers.

Question: What is the capacitance for a capacitor with the same structure as Fig. 3.9(b) with $h = l = 10$ μm? (The dielectric between two conducting plates is silicon dioxide, with $\kappa = 3.9$, and $d = 1000$ Å.)

Answer:

$$C = \kappa\varepsilon_0 \frac{hl}{d} = 3.9 \times 8.85 \times 10^{-12} \times \frac{10 \times 10^{-6} \times 10 \times 10^{-6}}{1000 \times 10^{-10}}$$
$$= 3.45 \times 10^{-14}\text{F} = 34.5 \text{ fF}.$$

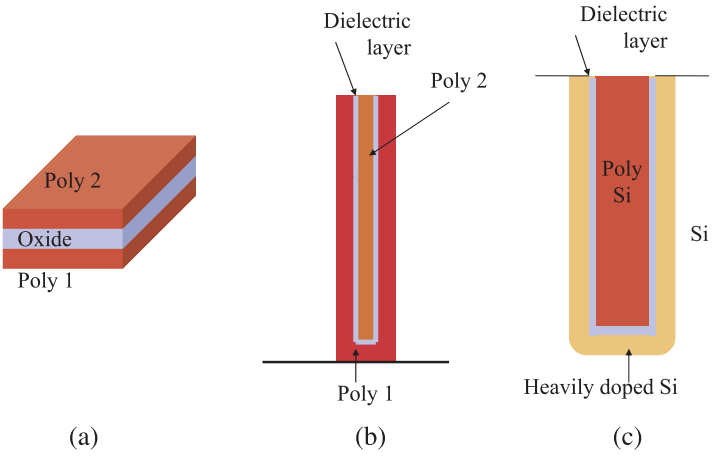


Figure 3.10 (a) Parallel plate, (b) stack, and (c) deep trench capacitors.

Question: By reducing the distance d between the two plates, capacitor dimensions h and l can be reduced while keeping the same capacitance. For $h = l = 1$ μm, what is d using the same capacitance as the first question?

Answer:

$$d = \kappa\varepsilon_0 \frac{hl}{C} = 3.9 \times 8.85 \times 10^{-12} \times \frac{10^{-6} \times 10^{-6}}{3.45 \times 10^{-14}} = 10^{-9} m = 10 \text{ Å}.$$

Unfortunately, a 10 Å silicon dioxide layer is too thin to work reliably, even at 1 V, because the dielectric strength of SiO_2 is about 10^7 V/cm, equivalent to 0.1 V/Å. Therefore, a high-κ dielectric such as HfO_2 ($\kappa \sim 25$) or ZrO_2 ($\kappa \sim 25$) is needed to reduce capacitor dimension while keeping a certain value of capacitance and preventing dielectric breakdown.

Question: What is the required dielectric constant κ for the capacitor in the previous question with $d = 100$ Å?

Answer: $$\kappa = \frac{Cd}{\varepsilon_0 hl} = \frac{3.45 \times 10^{-14} \times 100 \times 10^{-10}}{8.85 \times 10^{-12} \times 10^{-6} \times 10^{-6}} = 39.$$

On an IC chip, metal interconnection lines (shown in Fig. 3.11) form unwanted capacitors, called parasite capacitors. Currently, IC device speed is mainly limited by an RC (resistance multiplied by capacitance) time delay, which is the time electrons need to charge up the parasite capacitors and metal line resistance. Using low-κ dielectric materials and a better conduction metal (such as copper) can reduce RC time delay and increase device speed.

The time it takes for charges to reach the capacitor is about $t = Q/I$ for the first order of approximation. Here, charge $Q = CV$ is the amount of charge needed to

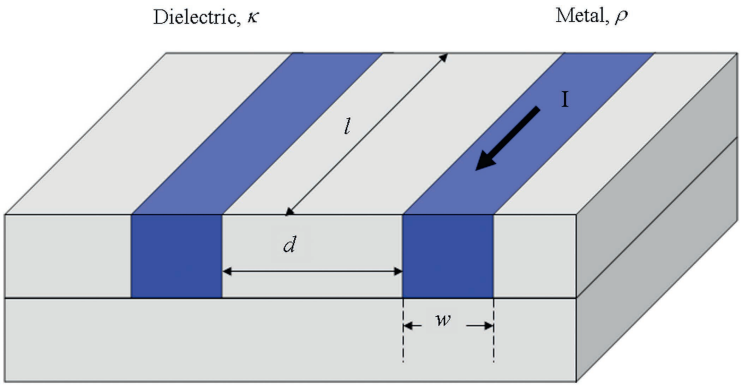


Figure 3.11 Metal interconnection.

fill the capacitor, V is the voltage drop on the metal line, and $I = V/R$ is the current in the metal line. Therefore, time delay $t = CV/(V/R) = \text{RC}$. An electrical signal with a frequency higher than 1/RC cannot pass through the metal interconnection lines. To achieve high device speed, it is necessary to reduce RC time.

Question: Most IC chips use aluminum-copper alloy metal interconnections. Assume resistivity $\rho = 3.2\ \mu\Omega \cdot \text{cm}$, and metal line geometry width w, height h, length l, and line spacing d are 1 μm, 1 μm, 1 cm, and 1 μm, respectively. CVD silicon oxide is between the metal line with dielectric constant $\kappa = 4.0$. What is the time delay RC?

Answer:

$$\text{RC} = \rho\frac{l}{wh}\kappa\varepsilon_0\frac{hl}{d} = \rho\kappa\varepsilon_0\frac{l^2}{wd} = 3.2\times10^{-8}\times4.0\times8.85\times10^{-12}\times\frac{0.01^2}{10^{-6}\times10^{-6}} = 1.133\times10^{-10}\ \text{sec}.$$

Therefore, an IC chip with this kind of interconnection cannot operate for frequencies higher than 1/RC = 8.83 GHz. To increase circuit speed, the RC time delay needs to be reduced by changing the interconnection line geometry and/or changing the conducting and dielectric materials. One way is to reduce l, which requires shrinking the device dimension. Another way is to increase w and d, which means more layers of metal interconnection. State-of-the-art technology in the late twentieth century employed more than nine layers of metal interconnection. Another option is to reduce both ρ and ε, by using a better conducting metal (such as copper) to replace the commonly used Al-Cu alloy, and employing low-κ dielectric materials to replace commonly used silicate glass.

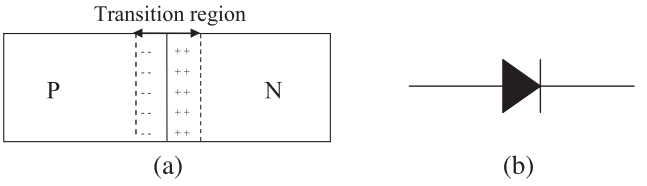


Figure 3.12 (a) p-n junction diode and (b) its symbol.

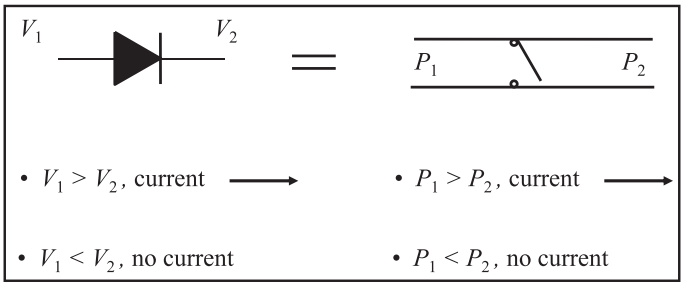


Figure 3.13 Mechanism of a diode and a one-way flow value.

3.2.3 Diode

Unlike the resistor and capacitor, a diode is a nonlinear device, so the response of current to voltage is no longer a linear relation. Figure 3.12 shows a p-n junction diode and its symbol.

A diode allows electric currents to pass through in only one direction when the voltage is forward biased. It does not allow currents to pass through if the applied potential is reverse biased. When applied voltage V_1 is higher than V_2, it is called forward bias, as the electric current goes through the diode (in the direction marked by the symbol in Fig. 3.13) with little resistance. When V_1 is lower than V_2, it is called reverse bias, as the resistance is very high and almost no electric current can pass through the diode. The mechanism is similar to a tire valve. When you pump air into the tire, and pump pressure P_1 is higher than tire pressure P_2; the pressure opens the valve and air flows into the tire. Without the pump, tire pressure P_2 is higher than atmospheric pressure P_1; therefore, the pressure difference closes the valve, and air stays inside the tire (see Fig. 3.13).

When p- and n-type semiconductors join together, they form a p-n junction diode, as shown in Fig. 3.14. Holes in the p-type region will diffuse to the n-type region, and electrons in the n-type region will diffuse to the p-type region. The charge separation causes electrostatic force, which eventually stops the minority carrier diffusion. The area dominated by minority carriers is called the transition region.

Voltage crossing the transition region can be expressed as

$$V_0 = \frac{\kappa T}{q} \ln \frac{N_a N_d}{n_i^2}, \tag{3.3}$$

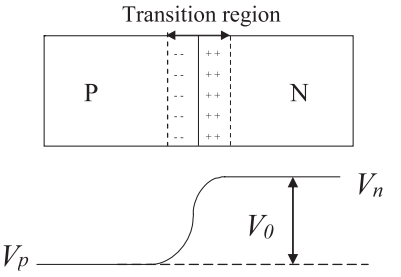


Figure 3.14 A p-n junction diode and its intrinsic voltage.

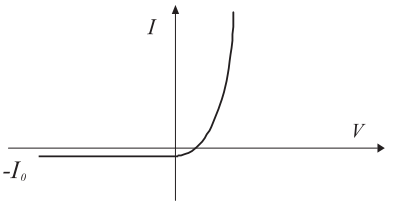


Figure 3.15 I-V curve of a diode.

where κ is the Boltzmann constant, T is the temperature, q is the charge, N_a is the acceptor (p-type dopant) concentration, N_d is the donor (n-type dopant) concentration, and n_i is the intrinsic carrier concentration.

For silicon at room temperature, $\kappa T/q = 0.0259$ V, $n_i = 1.5 \times 10^{10}$ cm^{-3}, and if $N_a = N_d = 10^{16}$ cm^{-3}, we can calculate that $V_0 \sim 0.7$ V. From Eq. (3.1), we can see that V_0 is not very sensitive to the dopant concentrations. Therefore, it always requires about 0.7 V of forward bias to flow current through a p-n junction. For the same reason, it requires about 0.7 V to turn on a bipolar transistor.

The I–V curve of a diode is illustrated in Fig. 3.15. When the diode is forward biased and V is larger than 0 (especially for $V > 0.7$ V), the current I going through the diode increases exponentially with V with very little resistance. When the diode is reverse biased, the current going through it is very low. It is almost a constant: I_0 when the voltage increases. Resistance in this case is very high. If the reverse bias voltage is too high, it could irreversibly break down the diode and cause it to no longer have its characteristic I–V curve.

3.2.4 Bipolar transistor

The symbols and bar structure of npn and pnp bipolar transistors are shown in Fig. 3.16.

The first transistor made in the Bell Laboratories was a point-contact bipolar transistor. Most bipolar transistors made in IC chips today are planar junction transistors, as shown in Fig. 3.17.

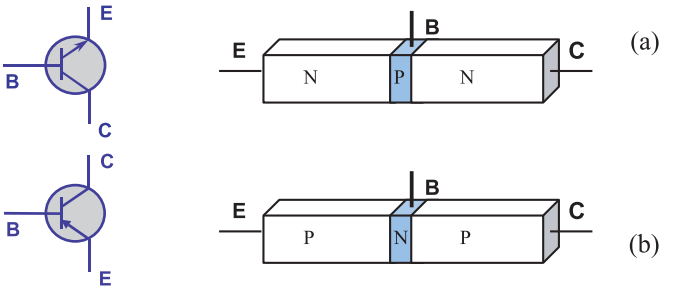


Figure 3.16 Symbols and bar structure of (a) npn and (b) and pnp bipolar transistors. E stands for emitter, B for base, and C for collector.

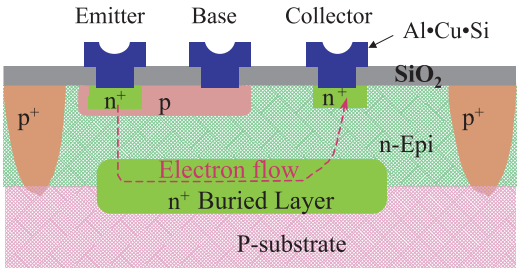


Figure 3.17 Cross section of a conventional npn planar bipolar transistor.

A bipolar transistor can be used as a switch because its emitter-to-collector current can be controlled by a base-to-emitter bias. For an npn transistor, when the base emitter is forward biased and $V_{BE} > 0.7$ V, electrons in the emitter can overcome the emitter-base n-p junction barrier potential, inject into the base, sweep across the thin base layer, and reach the collector.

When $V_{BE} = 0$ V, there is no electron injection from the emitter. Thus, there is no current between the emitter and collector, no matter how biased it is between the collector and emitter.

Most fabs use bipolar transistors for their ability to amplify an electric signal. Under the right conditions, a current between the emitter and collector equals β times the current flow into the base, or $I_{EC} = \beta I_B$, where β is the amplifying coefficient, normally between 30 to 100.

Figure 3.18 shows a cross section of an advanced sidewall base contact npn bipolar transistor. Bipolar transistors are mainly used for high-speed devices, analog circuits, and high-power devices.

From 1950 to 1980, bipolar transistors and bipolar-transistor-based IC chips were dominant in the semiconductor industry. After 1980, driven by the demands of logic circuits (especially demands for DRAM), MOS-transistor-based IC chips developed rapidly and surpassed bipolar-based IC chips. Today, CMOS-based IC chips dominate the semiconductor industry.

3.2.5 Metal-oxide-semiconductor field effect transistor

The symbols for n-channel MOS (nMOS) and p-channel MOS (pMOS) are shown in Figs. 3.19(a) and 3.19(b), respectively.

An nMOS structure is shown in Fig. 3.20. It has a conducting gate, silicon substrate, and a very thin layer of gate dielectric, usually silicon dioxide, sandwiched in between. For nMOS, the substrate is p-type, and the source and drain are heavily doped with an n-type dopant (which is why they are noted as n^+). Source and drain are symmetric; normally, the grounded side is called the source and the biased side is called the drain.

When no bias voltage is applied to the gate, no matter how the source/drain is biased, no current can pass through from source to drain, or vice versa. When the gate is biased positively, it generates positive charges at the metal-oxide surface. The gate dioxide is a very thin dielectric layer between the metal and semiconductor. Like a capacitor, the positive charge on the metal-oxide surface expels positive charges (holes, majority carriers) from the silicon-oxide surface and attracts negative charges (electrons, minority carriers) to that surface. When the gate voltage is higher than the threshold voltage, $V_G > V_T > 0$, the silicon-oxide surface accumulates enough electrons to form a channel and allow electrons from the source and drain to flow across it. This is why nMOS is also called an n-channel MOSFET or nMOSFET. This process is illustrated in Fig. 3.21. By controlling gate voltage, the electric field will affect the conductivity of the semiconductor device and switch the MOS transistor on and off. This is the reason it is called a field effect transistor (FET).

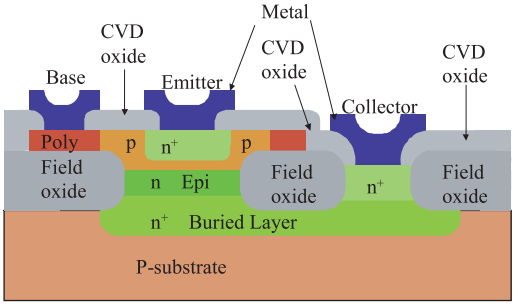


Figure 3.18 Sidewall base contact npn bipolar transistor.

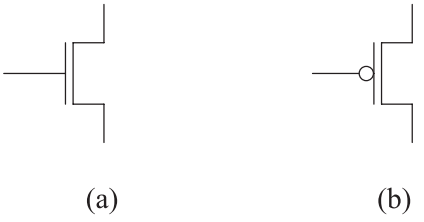


Figure 3.19 Symbols for (a) nMOS and (b) pMOS.

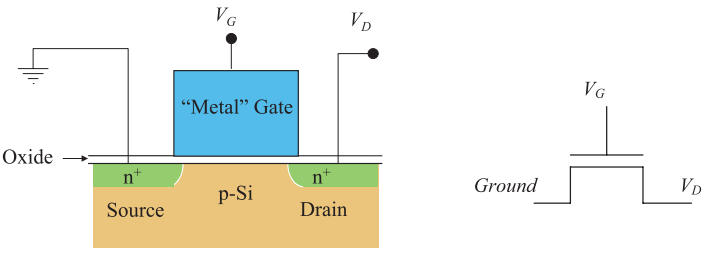


Figure 3.20 An nMOS transistor.

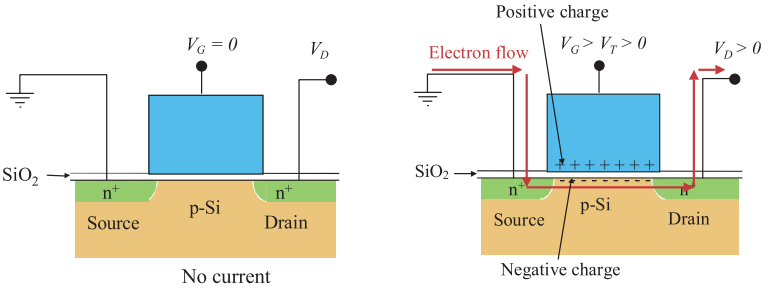


Figure 3.21 Switching process of the nMOS.

For a pMOS, the substrate is an n-type semiconductor, and the source and drain are heavily p-type doped. The pMOS uses a negative gate bias to generate negative charges (electrons) on the metal-oxide surface. This in turn drives away electrons (majority carriers of the substrate) and attracts holes (minority carriers) to the silicon-oxide surface to form a hole-channel underneath the gate. Holes from the source and drain then flow through the channel and conduct electric current between the source and drain (see Fig. 3.22). With positive bias, the pMOS transistor turns off.

By adjusting the dopant concentration of the substrate with ion implantation underneath the gate, the threshold voltage V_T can be adjusted. Therefore, the MOSFET can be set as normal-on and normal-off. In a CMOS circuit, both nMOS and pMOS are used; usually nMOS is set as normal-off, and pMOS is set as normal-on.

MOSFETs are mainly used for logic circuits such as microprocessors and memory chips. Figure 3.23 shows a tunneling electron microscope (TEM) image of a MOSFET manufactured with 32-nm technology by Intel Corporation.

The idea of the FET was proposed in 1925. In their effort to create it, Bell Laboratories scientists unintentionally made the first point-contact bipolar transistor on a piece of polycrystalline germanium. Lack of single-crystal semiconductor materials hampered early efforts to develop the FET. The inventions of single-crystal germanium in 1950 and especially single-crystal silicon in 1952 changed that situation. Eventually a group led by M. M. Atalla at Bell Laboratories made the first practical MOSFET in 1960. MOSFET technology developed rapidly after that.

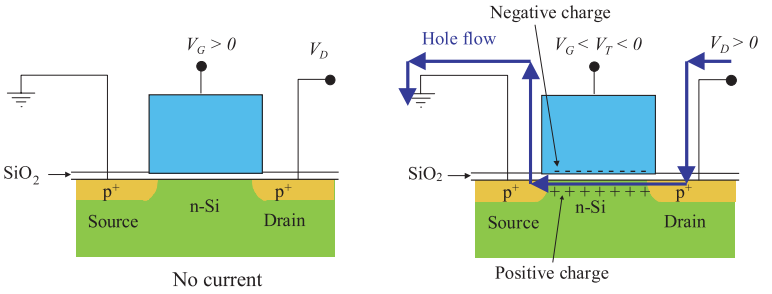


Figure 3.22 Structure and switching process of a pMOS.

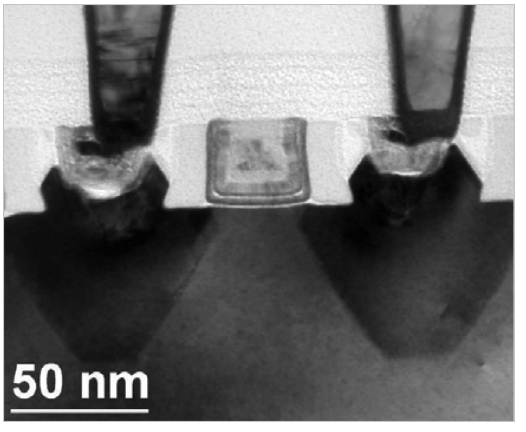


Figure 3.23 TEM image of a MOSFET manufactured with 32-nm technology (Chipworks).

Beginning in the 1980s, IC chips based on MOS transistors dominated the semiconductor industry. Since most IC chips are MOS-based, this text focuses mainly on the MOS IC processes.

In addition to silicon and germanium, there are many compound semiconductors. Gallium arsenate (GaAs) is commonly used to make high-frequency, high-speed IC devices for electronics instruments of defense and scientific applications. Other important applications of compound semiconductors are light-emitting diodes (LEDs), which are used in almost every electronic product as indicators and are widely used in traffic lights. Because of their highly efficient light emission and long lifetimes, it is possible that white LED lights will replace incandescent light bulbs and fluorescent lighting in homes and businesses in the future, when mass production brings down the cost of LEDs.

3.3 Integrated Circuit Chips

There are many different kinds of IC chips made in the world. They can be categorized into three groups: memory, microprocessor, and application-specific integrated circuits (ASICs).

3.3.1 Memory

Memory chips can memorize digital data by storing and discharging electrical charges. They are widely used in computers and other electronic devices for data storage. The largest portions of chips manufactured are memory chips, notably DRAM and NAND flash.

3.3.1.1 Dynamic random access memory

DRAM stands for dynamic random access memory. Random access indicates that each memory cell in the chip can be accessed to read or write in any order, in contrast to sequential memory devices where data have to be read or written in a certain order. For instance, hard disks and compact disks employ random access, while a cassette tape employs sequential access.

DRAM is the most commonly used memory chip, particularly inside computers for data storage. When people purchase a personal computer, one important factor is how many bytes of memory it has—that is, the data storage capability the DRAM chips have. The basic memory cell of a DRAM is formed by an nMOS and a capacitor, as shown in Fig. 3.24.

The n-type MOSFET serves as a switch. It allows electrons to flow and become stored in the capacitor, which holds the memory. The capacitor needs to be recharged periodically by the power supply V_{dd} to compensate for electron losses. That is why it is called dynamic RAM or DRAM. When power is removed from a DRAM, data are lost. While editing a file on a computer, all of the input materials, including words, graphs, and symbols, are stored in the DRAM main memory of the computer before the "save" command writes them permanently onto a hard disk or a universal serial bus (USB) flash drive.

The demand for more memory and faster memory chips in computer-related applications is one of the most important driving forces of technology development within the IC industry.

3.3.1.2 Static random access memory

SRAM stands for static random access memory and uses latching transistors to keep the instruction or data memory. It requires at least six components, either four transistors and two resistors or six transistors. Figure 3.25 shows the circuit of a six-transistor SRAM. We can see that it has four nMOS and two pMOS. SRAM

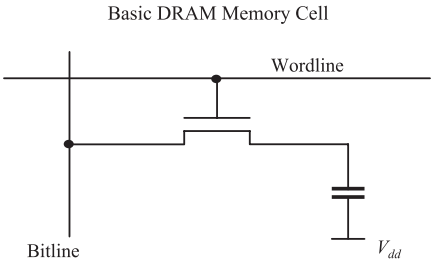


Figure 3.24 Circuit of a DRAM memory cell.

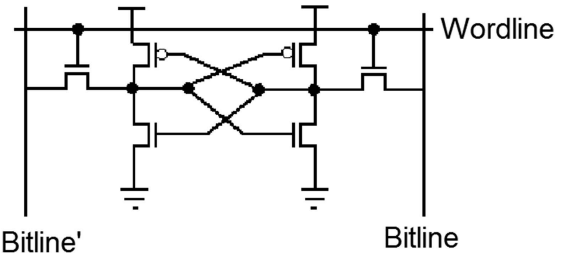


Figure 3.25 Circuit of a six-transistor SRAM.

is much faster than DRAM, since it does not need to recharge the capacitor to store data. However, for the same memory with the same processing technology, a SRAM chip is much larger and more expensive than a DRAM chip. SRAM is mainly used in computers as cache memory to store the most frequently used instructions. DRAM is used as the main memory to store less frequently used instructions and data. Flash drives and hard disk drives (HDDs) are nonvolatile memories, which are commonly used to permanently store the data files.

Most logic IC chips such as microprocessors, ASIC chips, and field programmable gate array (FPGA) chips have embedded cache memory, which integrates SRAM into their chip design. In these chips, SRAM arrays always have the highest device density. Almost all logic IC fabs use SRAM in their test vehicles for technology development and qualification.

3.3.1.3 Erasable programmable read-only memory, electric erasable programmable read-only memory, and flash

DRAM and SRAM need a power supply to keep the data. Without a power supply, the memory will be lost, thus it is called volatile memory. EPROM stands for erasable programmable read-only memory, and EEPROM stands for electric erasable programmable read-only memory. Both EPROM and EEPROM have nonvolatile memory. Their main applications are for permanent storage of data and instructions without power supply.

The structure of an EPROM cell is similar to an nMOS, as shown in Fig. 3.26. The fundamental difference is that it has a floating gate, which connects to nowhere and can permanently store electrons as digital data. By biasing the control gate with $V_G > V_T > 0$, electrons are attracted to the silicon–gate oxide interface and form an n-type channel underneath the floating gate. Electrons from the source and drain pass through the channel and conduct electric currents. While most electrons pass through the channel, some electrons are intentionally injected into the floating gate through the thin gate oxide by the electron tunneling effect, known as the hot electron effect. When electrons are injected into the floating gate, they can stay there for years because they have no way to escape. Thus, the gate can keep that memory, regardless of whether it is powered on or off. The memory writing process is shown in Fig. 3.27(a).

One can erase the memory from EPROM by shining UV light through the passivation dielectric. The UV light excites the electrons in the floating gate,

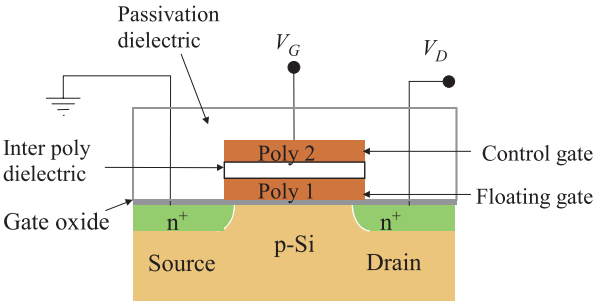


Figure 3.26 Cross section of an EPROM memory cell.

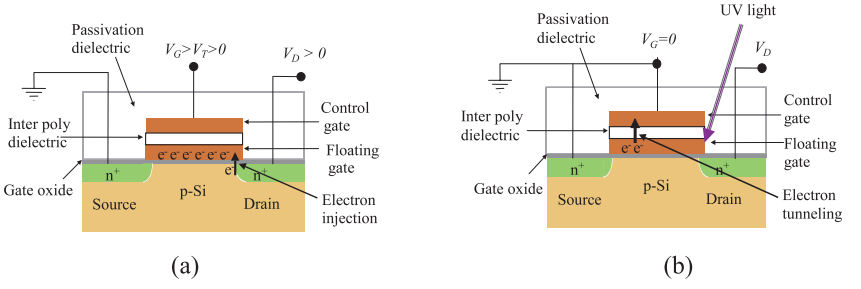


Figure 3.27 EPROM (a) programming and (b) erasing processes.

causing them to tunnel through the interpoly dielectric and drain to the ground at the control gate [see Fig. 3.27(b)].

Different from EPROM, which uses UV light to erase memory from the entire chip, EEPROM erases memory from individual cells electrically by biasing the control gate and causing electron tunneling from the floating gate to the control gate.

Flash memory is a specific EEPROM that costs significantly less to manufacture. Figure 3.28(a) shows a 64-bit NAND flash circuit. WL stands for word line, of which there are 64, numbered from 0 to 63. SG stands for select gate, which is an nMOS transistor. SG0 is the nMOS to select the string to the ground (source line), and SG1 is the nMOS to allow the string to select to the bit line. Figure 3.28(b) illustrates the cross section of a 64-bit NAND flash along the bit line direction. 64-bit NAND has been used in mass production in advanced NAND flash fabs with 19-nm technology.

NAND flash has been widely used in USB flash drives, solid state disks (SSDs), memory cards for mobile phones, digital cameras, digital camcorders, and other mobile devices. SSDs are faster and more reliable and also consume less power than HDDs. For example, a 256-Gb, 2.5-in. notebook SSD consumes about 20% of the power of a 250-Gb HDD of the same size. SSDs have been used in high-end mobile electronics such as notebook computers to replace HDDs for high volume data storage. The main advantage of HDDs is the cost, which is one of the most important factors for consumer electronics. For the same price as an SSD,

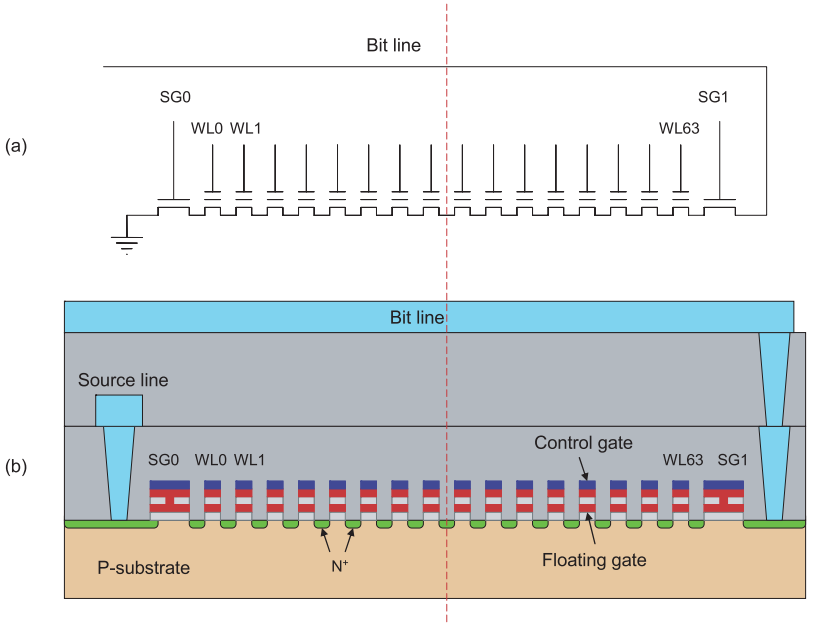


Figure 3.28 (a) Circuit of a 64-bit NAND flash and (b) the illustration of its cross section in the bit line direction.

a person can buy an HDD with more than ten times the storage. With the shrinking of feature size, it is possible that the cost per byte of flash memory could be reduced to a level comparable to a HDD. However, with the feature size shrinking, device standby leakage will increase. Also, as more devices are packed into one SSD with smaller feature size, power consumption of SSDs could increase rapidly. If flash memory can solve the leakage-induced power consumption issue, in the future, SSDs could replace HDDs for data storage in most computers and most other digital electronics, especially mobile devices. Otherwise, flash memory would only play a supporting role in a hybrid disk drive that integrates both SSD and HDD.

3.3.2 Microprocessor

The microprocessor is also called a central processing unit (CPU). It has two components: a control system and an arithmetic-logic unit (ALU). Most microprocessors also have a built-in cache memory. It is the brain of the computer and other control systems. Two kinds of microprocessor architectures are widely used: complete instruction set computers (CISCs) and reduced instruction set computer (RISCs).

A graphics processing unit (GPU) is a processor specially designed for graphics processing with high-speed floating point computation. It can free up the main CPU from computation-intensive 3D graphics processing and is widely used in embedded systems, personal computers (PCs), mobile phones, and game consoles.

3.3.3 Application-specific integrated circuits

Many chips belong in the ASIC category, including digital signal processing (DSP) chips, power devices, and chips for TV, radio, Internet, automobiles, wireless communications, telecommunications, etc. As the feature size shrinks, the cost of photomasks required for IC chip manufacturing increases dramatically. For high-volume products such as microprocessors and memory chips, the cost of the masks can be shared by thousands of wafers. However, ASIC chips with limited production volume have to face the tough challenge of cost effectiveness. FPGA chips, which allow users to configure system applications after chip manufacturing, become attractive alternatives.

3.4 Basis Integrated Circuit Processes

IC processing technology began with Kilby's germanium bar and then quickly switched to a single-crystal silicon chip with planar technology, first developed by Noyce. After 1960, the IC industry developed exponentially at a rate that matches well with Moore's law.

Bipolar-transistor-based IC chips dominated the semiconductor industry from the 1960s to the 1980s. Demands for electronic watches, calculators, computers, and other digital electronic products rapidly drove MOSFET-based IC process development and chip fabrication. The requirements for low-power-consumption circuits pushed the rapid development of CMOS IC chips. CMOS-based IC chips have dominated the IC industry since the 1980s. Many people predict that the semiconductor industry will keep the development pace described by Moore's law for the next ten years; then photolithography will become obsolete.

Basic IC processes can be divided into the following four categories:

- adding
- removing
- heating
- patterning.

Adding indicates that atoms are added into a (doped) wafer or a layer is added onto the wafer surface (thin-film growth or deposition). Ion implantation and diffusion processes are used to add dopants into the semiconductor substrate. In oxidation and nitridation processes, oxygen and nitrogen are added to chemically react with silicon to form silicon dioxide and silicon nitride. Thin-film depositions add thin layers of semiconductors, dielectric or metal via CVD, PVD, spin-on, or electrical plating processes.

Removing involves removing material from the wafer, either chemically, physically, or by a combination of both. Wafer cleaning, which uses a chemical solution to remove contaminants and particles from the wafer surface, is a removal process; patterned and blanket etching are as well. CMP, which removes materials from the wafer surface by both chemical reaction and mechanical grinding to achieve surface planarization, also belongs in this category.

In heating or thermal processes, wafers are heated to specific temperatures in a certain amount of time to achieve desirable physical or chemical results. Very little or no materials are added or removed from the wafer during heating processes, which include postimplantation annealing (high temperature > 1000 °C) and metal annealing (low temperature < 450 °C), alloying, and thermal reflow processes. Two kinds of thermal processes are commonly used in semiconductor manufacturing: conventional (furnace) process and rapid thermal process (RTP). The heating process is sometimes performed in situ, with oxidation and high-temperature deposition processes in a furnace or RTP chamber for oxide or thin-film annealing.

All adding, removing, and heating processes are employed in the patterning process, which transfers circuit-design layouts from the mask or reticle to photoresist on the wafer surface. The pattering process is the most frequently used process in IC fabrication, so it is very important. Most fabs use the photolithography process for pattern transferal in present-day IC chip manufacturing. Extreme ultraviolet (EUV) lithography, nanoimprint lithography (NIL), and electron beam direct write (EBDW) lithography could be used in the future when minimum feature size becomes too small to use photolithography.

The rest of this section briefly summarizes the basic processing steps involved in IC manufacturing. Since the majority of IC chips manufactured are MOSFET-based IC chips, this text focuses on MOS, especially CMOS processing steps.

3.4.1 Conventional bipolar transistor process

The main processing steps for bipolar-transistor-based IC chip manufacturing are: buried layer doping, epitaxial silicon growth, isolation and transistor doping, interconnection, and passivation.

Silicon-based bipolar-transistor ICs normally use a p-type wafer to make npn transistors. The basic processing steps originated with technology from the mid-1970s (ion implantation for the doping process). The process consists of seven mask steps, starting with a p-type doped wafer. The buried layer is formed to reduce the series collector resistance and to improve device speed. Its processing steps include wafer cleaning, oxidation, photolithography, ion implantation, photoresist stripping, and silicon dioxide stripping. Epitaxial silicon growth is a high-temperature CVD process. During epitaxial growth, the implanted buried layer anneals and diffuses slightly, as shown in Figs. 3.29(a) and 3.29(b). After a thin layer of oxide grows, repeated photolithography and ion implantation processes define isolations, as well as the emitter, base, and collector of the transistor [Figs. 3.28(c) and 3.29(d)]. Next, the screen oxide is stripped, and a thick layer of SiO_2 is grown. Photolithography and oxide etch define the contact holes. After photoresist stripping, an aluminum alloy metal layer is deposited to fill the contact holes and cover the entire wafer. Photolithography and metal etch form the metal line interconnection and make contacts with transistors [Fig. 3.29(e)]. After stripping the photoresist, a layer of CVD silicon oxide is deposited to protect the transistors and metal wires [Fig. 3.29(f)]. The final processes are photolithography, bonding pad etching, and photoresist stripping.

More-advanced process technologies such as dielectric trench fill isolation, deep collector, selective epitaxy growth, and self-aligned emitter base have been developed and applied in bipolar-transistor-based IC chip fabrications. Since the main focus of this text is MOSFET-based IC, these topics are not covered in detail.

The main applications of bipolar-transistor ICs are in analog systems such as chips for TVs, sensors, and power devices. They are also used with CMOS to form BiCMOS ICs for a variety of applications, including some microprocessors.

3.4.2 p-Channel metal-oxide-semiconductor process (1960s technology)

In the early stages of the IC industry, most fabs made bipolar-transistor-based IC chips. When they started to make MOSFET-based IC chips, they used p-channel MOSFET or pMOS because of the limitations of technology. Although it has exactly the same design (same gate material, geometry, substrate dopant concentration, and source/drain dopant concentration), nMOS is significantly faster than pMOS because electron mobility is two to three times higher than holes. However, difficulties associated with the nMOS process could not be solved

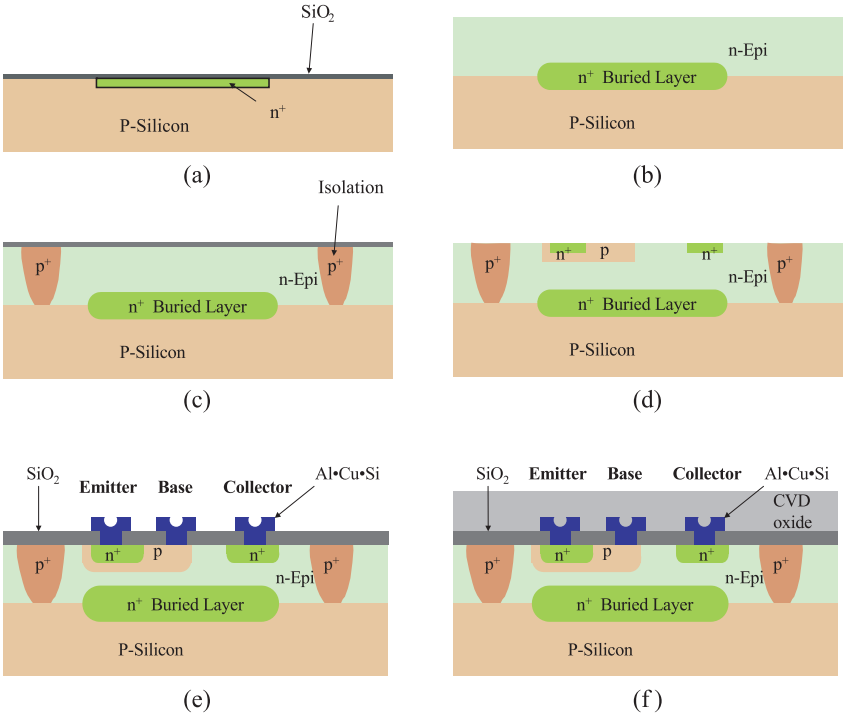


Figure 3.29 Bipolar-transistor process.

without ion implantation technology. It was much easier and more cost effective to make pMOS than nMOS using diffusion techniques for silicon doping.

A pMOS process flow with technology from the 1960s is given in Figs. 3.30 and 3.31 and Table 3.1. It uses blanket-field oxide for isolation, boron diffusion for source/drain doping, and aluminum-silicon alloy for gate and interconnection. About 1% of silicon was added into the aluminum to saturate the dissolving of the silicon, preventing junction spiking of the aluminum caused by the silicon dissolution. Minimum feature size was about 20 μm.

The entire pMOS process has five masking steps. Every masking step is a photolithography step, including wafer cleaning, prebake, primer application, photoresist coating, soft bake, alignment and exposure, development, pattern inspection, and hard bake. From the process flow it is evident that the IC process always repeats the removing, adding, pattering, and heating processes.

3.4.3 n-Channel metal-oxide-semiconductor process (1970s technology)

IC processing technology changed significantly after the introduction of ion implantation into mass production in the mid-1970s. Ion implantation processing replaced diffusion processing for silicon doping because it offered the advantages of independently controlled dopant concentration and junction depth, as well as an anisotropic doping profile. Polysilicon replaced aluminum as the gate material and local interconnection. It was used to form self-aligned sources and drains by taking advantage of the anisotropic profile of ion implantation and the high-temperature stability of polysilicon. Minimum feature size was about 7.5 μm. Figure 3.31 shows the self-aligned source/drain implantation process; this technique is still used in advanced MOSFET IC chip manufacturing.

Phosphorus silicate glass (PSG) was used as a (premetal dielectric) PMD material. PSG can trap mobile ions such as sodium and prevent them from diffusing to the gates and damaging the MOSFETs. Its use represents one of the major IC technology breakthroughs of the late 1960s. PSG can flow thermally at 1100 °C, at which temperature it can smooth and planarize the dielectric surface. A planarized surface is favorable for subsequent metallization and photolithography processes. A thin layer of undoped silicate glass (USG) was used as a barrier layer between the source/drain (S/D) and PSG, aluminum-copper-silicon alloy was used as a global interconnection, and CVD nitride was used for passivation dielectrics.

After the mid-1970s, IC manufacturers made nMOS-based IC chips because nMOS is faster than pMOS for the same geometry and dopant concentrations. An nMOS process flow is given in Table 3.2 and shown in Fig. 3.32.

3.5 Complementary Metal-Oxide Semiconductor

Starting in 1970, electronic watches and hand-held calculators developed very quickly, and LEDs were used for the displays. Because LEDs consume a great deal of power and consequently limit battery lifetime, great efforts were made to find their replacement for electronic watch and calculator applications. The liquid

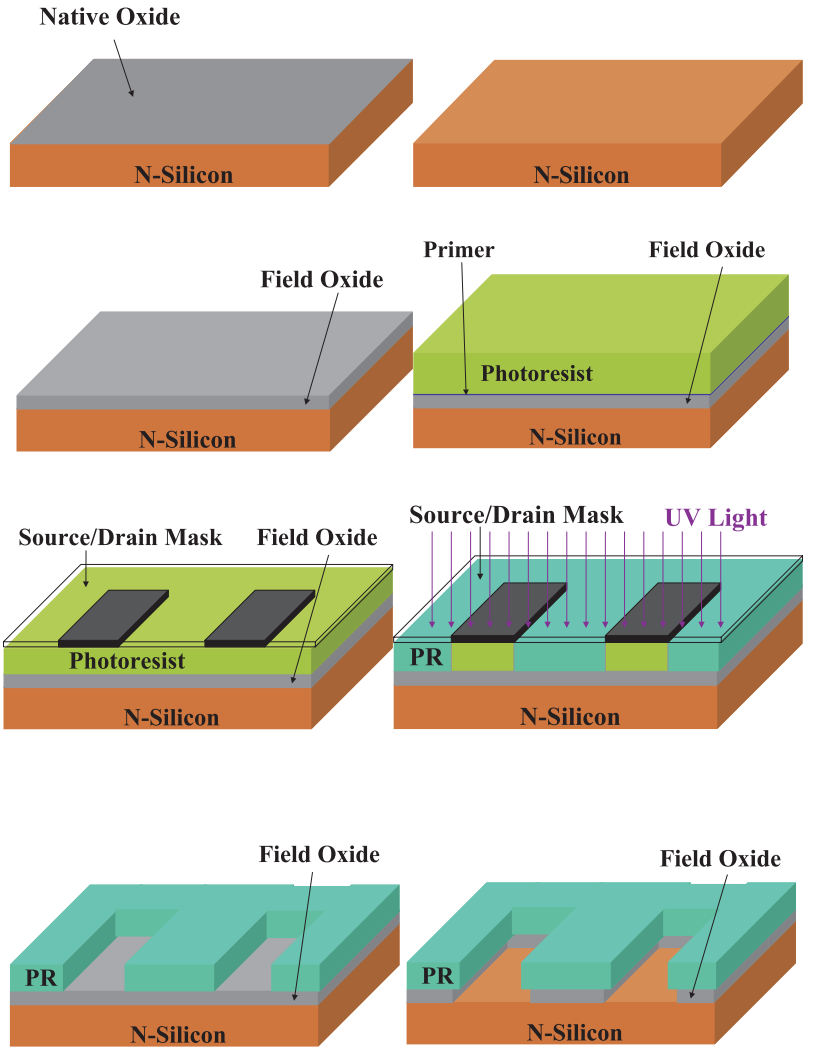


Figure 3.30 Processing steps of pMOS IC fabrication (*continued on next page*).

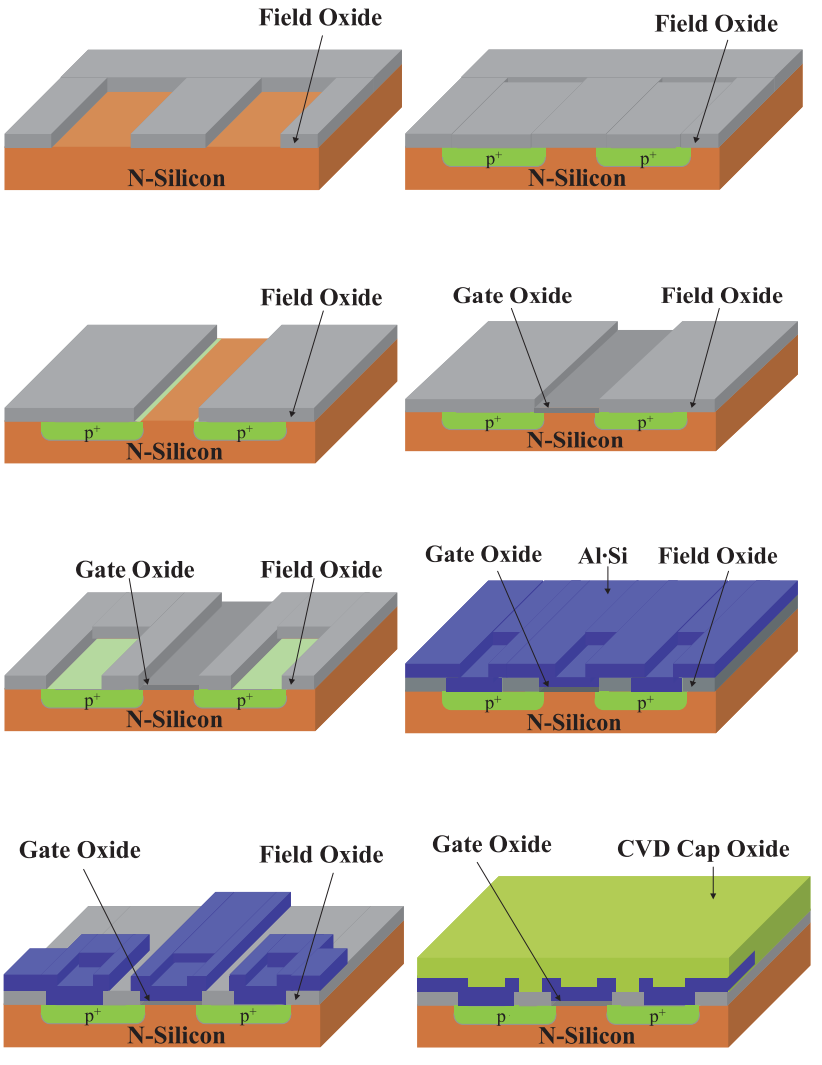


Figure 3.30 (*continued*)

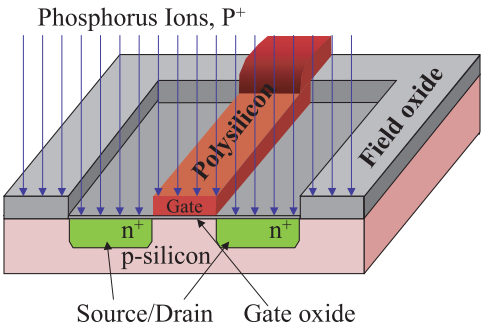


Figure 3.31 Self-aligned source/drain implantation.

Table 3.1 pMOS processing sequence (1960s); A stands for adding, H for heating, P for patterning, and R for removing.

Wafer cleaning	(R)	Etch oxide	(R)
Field oxidation	(A)	Strip photoresist	(R)
Mask 1: source/drain	(P)	Al deposition	(A)
Etch oxide	(R)	**Mask 4:** metal	(P)
Strip photoresist/clean	(R)	Etch aluminum	(R)
S/D diffusion (B)/oxidation	(A)	Strip photoresist	(R)
Mask 2: gate	(P)	Metal anneal	(H)
Etch oxide	(R)	CVD oxide	(A)
Strip photoresist/clean	(R)	**Mask 5:** bonding pad	(P)
Gate oxidation	(A)	Etch oxide	(R)
Mask 3: contact	(P)	Testing and packaging	

crystal display (LCD), which consumes much less power, quickly replaced LEDs for those applications after its introduction in early 1980.

The need to reduce power consumption of the circuits for calculators and electronic watches was one of the major driving forces for CMOS-based IC chip development. CMOSs are used in logic and memory chips, and have dominated the IC market since the 1980s.

3.5.1 Complementary metal-oxide-semiconductor circuit

Figure 3.33(a) shows a CMOS inverter circuit. Two MOSFETs can be seen: one is nMOS and the other is pMOS. If the input is high voltage, or logic 1, the nMOS is turned on and the pMOS is turned off. Therefore, the output voltage is ground voltage V_{ss}, and thus V_{out} is low voltage, or logic 0. Conversely, if the input is low voltage, or logic 0, the nMOS is switched off, and the pMOS is switched on. Output voltage is high voltage V_{dd}, thus V_{out} is high voltage, or logic 1. Because it inverts the input signal, it is called an inverter. This design is one of the basic logic gates used in logic circuits. Table 3.3 is the digital logic table of an inverter.

From Fig. 3.33(a), it is evident that when the nMOS is on, the pMOS is off, and vice versa. This is why the circuit is called the complementary MOS, or CMOS.

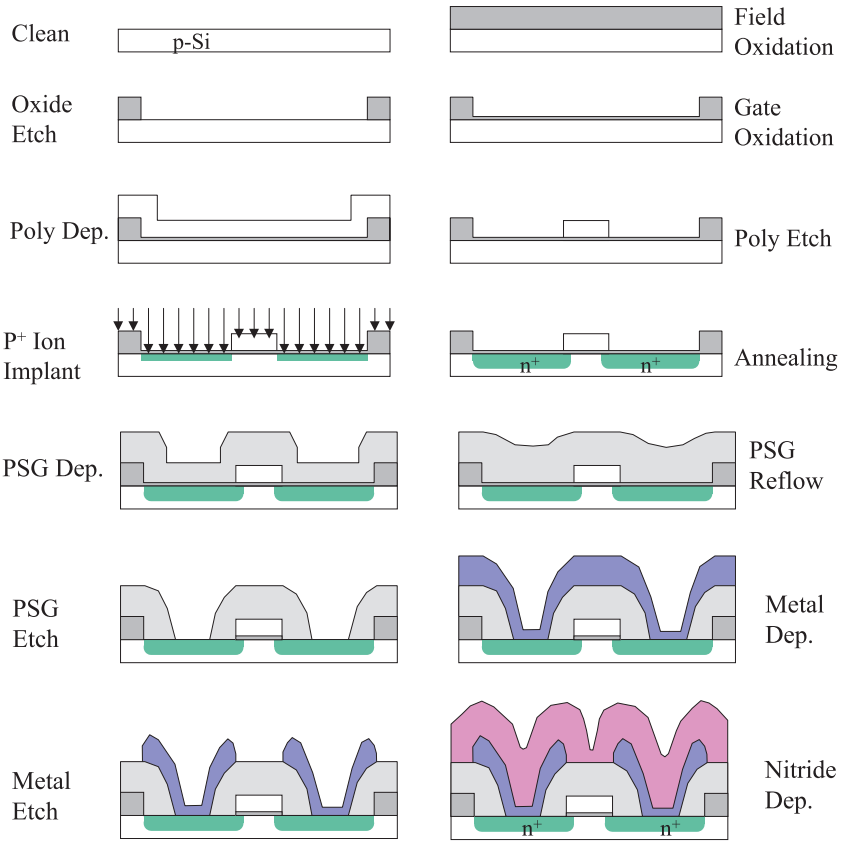


Figure 3.32 nMOS IC fabrication processing sequence.

Table 3.2 nMOS processing sequence (mid-1970s).

Wafer cleaning	PSG reflow
Grow field oxide	**Mask 3:** contact
Mask 1: active area	Etch PSG/USG
Etch oxide	Strip photoresist/clean
Strip photoresist/clean	Al deposition
Grow gate oxide	**Mask 4:** metal
Deposit polysilicon	Etch aluminum
Mask 2: gate	Strip photoresist
Etch polysilicon	Metal anneal
Strip photoresist/clean	CVD oxide
S/D and poly dope implant	**Mask 5:** bonding pad
Anneal and poly reoxidation	Etch oxide
CVD USG/PSG	Testing and packaging

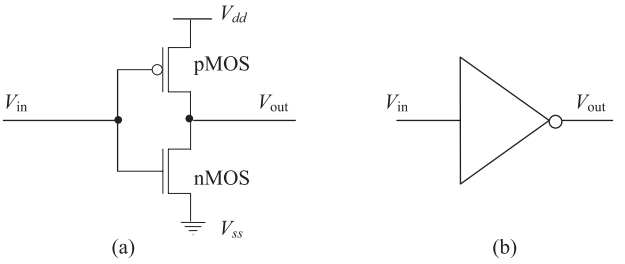


Figure 3.33 Circuit of (a) a CMOS inverter and (b) its logic symbol.

Table 3.3 Logic table of an inverter.

Input	Output
1	0
0	1

For the CMOS inverter, the circuit is always open between high-voltage-biased V_{dd} and grounded V_{ss}. Ideally, there is no current flow between V_{dd} and V_{ss}, so CMOS has very low inherent power consumption. The main power consumption of a CMOS inverter comes from leakage during switching, which has a very high frequency. Other advantages of CMOS over nMOS are that it has higher noise immunity, a lower chip temperature, wider operation temperature range, and less clocking complexity.

By combining both CMOS and bipolar technology, the BiCMOS IC developed rapidly in the 1990s. The CMOS circuit is used for the logic part, and the bipolar is used for input/output to increase device speed. Since the BiCMOS is not a mainstream product and is likely to lose its applications when IC power supply voltage drops below 1 V, this process is not covered in detail.

3.5.2 Complementary metal-oxide-semiconductor circuit process (1980s technology)

CMOS processing was developed from nMOS processing. It takes at least three more mask steps to form a CMOS than it does for an nMOS. One is for n-well formation, one is for pMOS source/drain implantation, and the third is for nMOS source/drain implantation.

Figure 3.34 and Table 3.4 show CMOS manufacture processing steps with technology from the 1980s. Local oxidation of silicon (LOCOS) replaced blanket-field oxidation for isolation between transistors. Borophosphosilicate glass (BPSG) was used for the PMD or interlayer dielectric (ILD0) to reduce the required reflow temperature. Because of the shrinkage of the critical dimension (CD), plasma etch (dry etch) replaced wet etch for most patterned etching. One layer of metal line was no longer enough to route all of the components on the IC chip with required conductance; therefore, a second metal layer was used. Deposition and planarization of dielectrics between metal lines [intermetal dielectric (IMD)],

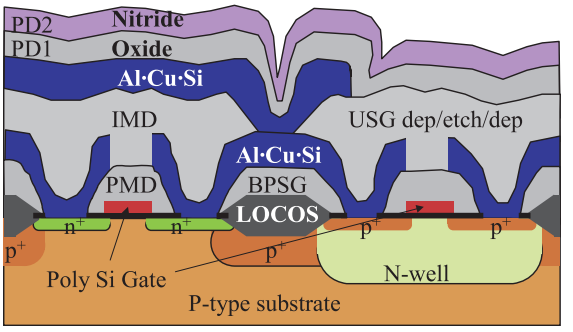


Figure 3.34 Cross section of a CMOS IC chip with two layers of metal.

Table 3.4 CMOS processing sequence (1980s).

Wafer cleaning	Strip photoresist/clean	USG deposition
Pad oxidation	Polysilicon annealing/oxidation	Etch etchback
Nitride deposition	**Mask 4:** p-channel S/D	USG deposition
Mask 1: LOCOS	Boron implant	**Mask 8:** via
Etch nitride	Strip photoresist	Etch IMD
Field V_T implant (boron)	**Mask 5:** n-channel S/D	Strip photoresist
Strip photoresist/clean	Phosphorous implant	Premetal clean
Field oxidation	Strip photoresist/clean	Sputter Al alloy
Strip nitride	Deposit USG (barrier layer)	**Mask 9:** metal 2
Mask 2: n-well	Deposit BPSG (ILD0)	Metal etch
n-well implant	BPSG reflow	Strip photoresist
Strip photoresist/clean	**Mask 6:** contact	Metal annealing
n-well drive-in	Etch oxide	CVD oxide/nitride
Strip pad oxide	Strip photoresist	**Mask 10:** bonding pad
Wafer cleaning	Premetal clean	Etch nitride/oxide
Gate oxidation	Sputter Al alloy	Strip photoresist
Polysilicon deposition	**Mask 7:** metal 1	
Poly dope implantation	Metal etch	
Mask 3: Gate	Strip photoresist/clean	*Testing and packaging*
Polysilicon etch	Metal anneal	*Final test*

posed a great challenge from the 1980s to the 1990s. The minimum feature size shrank from 3 to 0.8 μm in the 1980s.

Basic CMOS processing steps are wafer preparation, well formation, isolation formation, transistor making, interconnection, and passivation. Wafer preparation includes epitaxial silicon deposition, wafer cleaning, and alignment mark etch. Well formation defines the substrate for nMOS and pMOS transistors. Depending on the technology development, there are different techniques for well formations: single well, self-aligned twin well (also called a single-photo twin well), and double-photo twin well. The isolation process builds insulation blocks to electrically isolate the neighboring transistors. In the 1980s, LOCOS replaced blanket-field oxide as the main technique for device isolation. Transistor production involves gate oxide growth, polysilicon deposition, photolithography, polysilicon etch, ion implantation, and thermal annealing. These are the most crucial steps in

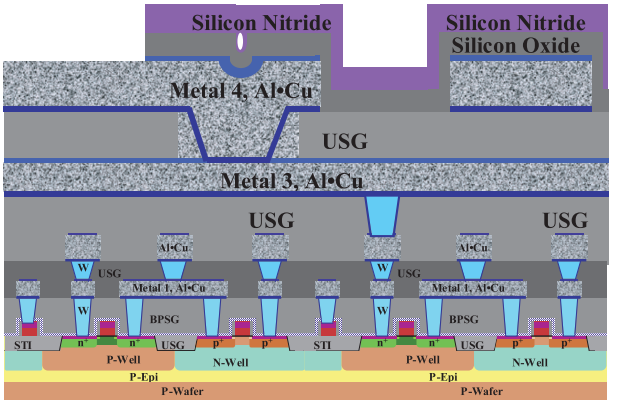


Figure 3.35 Cross section of a CMOS IC chip with four Al-Cu metallization layers.

IC processing. Interconnection processes use the combination of deposition, photolithography, and etch processes to define contact holes and metal wires to connect millions of transistors built on the silicon surface. Finally, passivation dielectric deposition, photolithography, and etch processing seal the IC chip from the outside world, leaving only the bonding pad open for testing and soldering.

3.5.3 Complementary metal-oxide-semiconductor process (1990s technology)

During the 1990s, the feature size of IC chips continued to shrink, from 0.8 μm to smaller than 0.18 μm, and several new technologies were adapted in IC manufacturing. Shallow trench isolation (STI) replaced LOCOS for isolation formation when feature size was smaller than 0.35 μm. Silicides were widely used to form gates and local interconnections, and tungsten (W) was widely used for metal interconnection as the plug between different metal layers. More fabs started using CMP to form STI and tungsten plugs, and they began to planarize the ILDs. High-density plasma etch and CVD processes became more popular, and copper metalization started ramping up in the production line. Figure 3.35 illustrates the cross section of a CMOS IC with four Al-Cu alloy metal interconnection layers. Figure 3.36 illustrates the cross section of a CMOS IC with four Cu metal interconnection layers and an Al-Cu alloy bond pad layer.

3.6 Technology Trends after 2000

CMOS IC technology entered nanometer technology nodes at the start of the new millennium. Technology nodes shrank from 130 to 32 nm. 193-nm became the dominant wavelength of optical photolithography. Immersion photolithography, which uses water as the medium between the objective lens of a scanner and photoresist on the wafer to further improve patterning resolution, has become widely used in IC manufacturing at the 45-nm node and beyond. The double-patterning technique has been used in IC manufacturing since the 45-nm

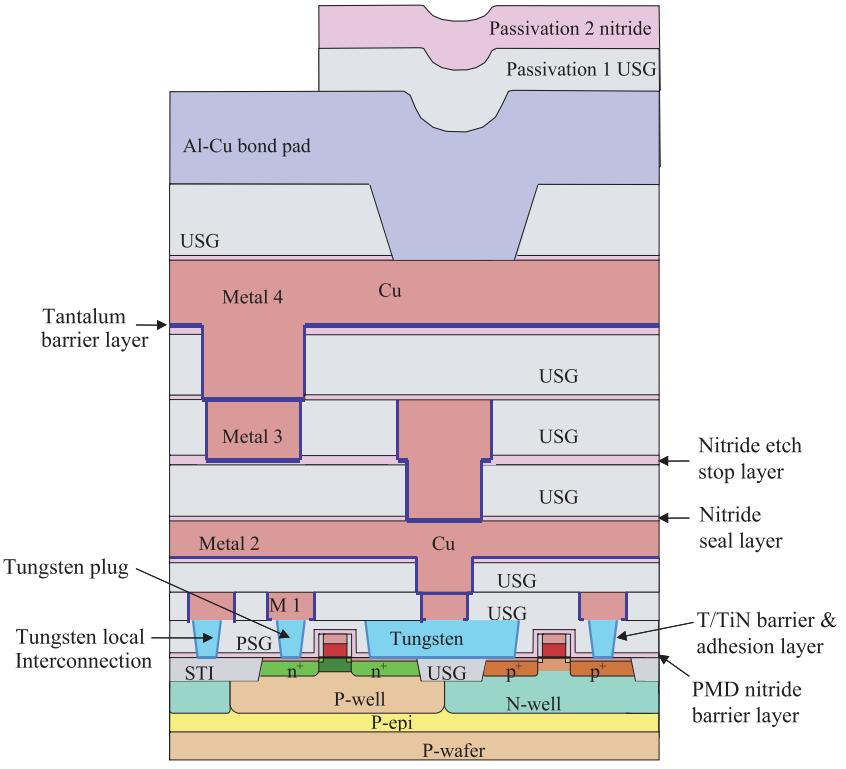


Figure 3.36 Cross section of a CMOS IC chip with four copper metallization layers.

technology node. The combination of immersion photolithography and double patterning has helped IC manufacturers to further shrink feature size without next-generation lithography. Starting from the 65-nm node, nickel silicide has replaced cobalt silicide as the material of choice for self-aligned silicide. High-κ and metal gates started to replace silicon dioxide and polysilicon as gate dielectric and gate electrode materials. Substrate engineering such as strain silicon is widely used to enhance device performance by improving carrier mobility. Techniques such as dual stress liners and selective epitaxial silicon germanium (eSiGe) growth are commonly used methods to strain the channel of the MOSFET to enhance carrier mobility and device speed. Figure 3.37 illustrates the cross section of a 32-nm CMOS with selective epitaxial SiGe and SiC, gate-last integration of a high-κ metal gate, nine layers of copper interconnection, and a lead-free bump.

3.7 Summary

- Semiconductors are materials with conductivity between the conductor and dielectric. Their conductivity can be controlled by dopant concentration and applied voltage.

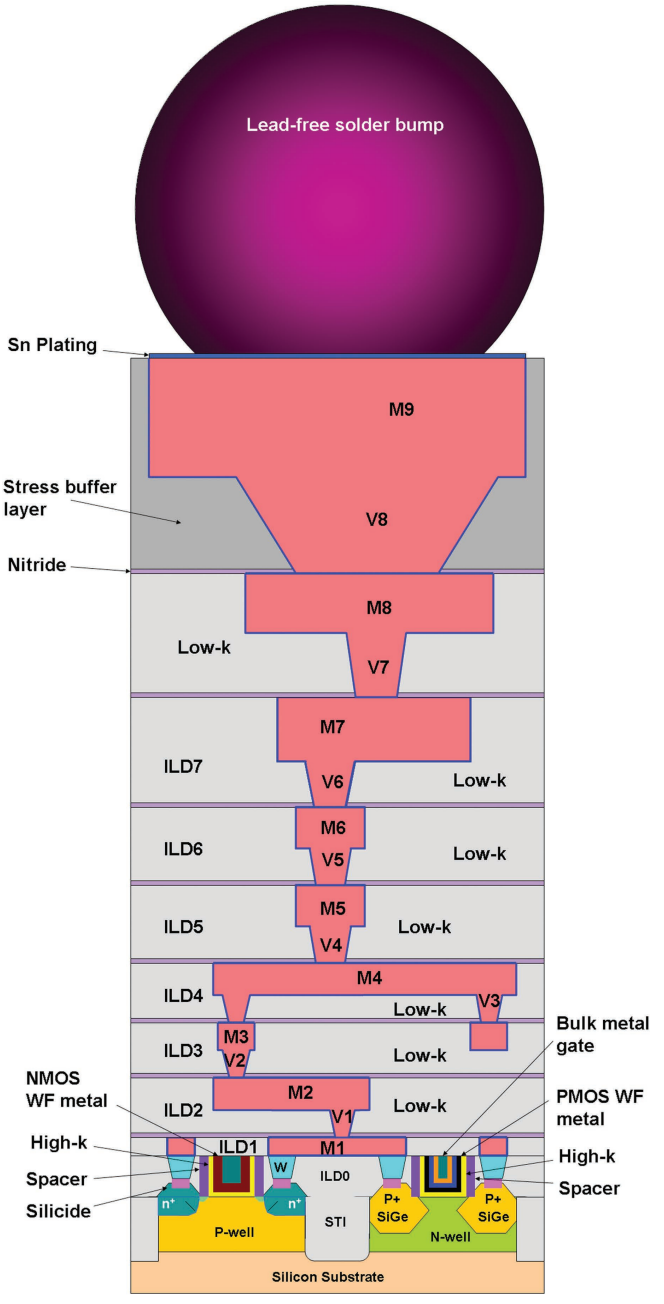


Figure 3.37 A 32-nm CMOS with selective epitaxial SiGe and SiC, gate-last high-κ metal gate, nine copper/low-κ layers, and a lead-free bump.

- Silicon, germanium, and gallium arsenide are the most commonly used semiconductor materials.
- p-type semiconductors are doped with atoms from column IIIA of the table of elements, mainly boron. The majority carriers are holes.
- n-type semiconductors are doped with atoms from column VA of the table of elements, principally P, As, and Sb. The majority carriers are electrons.
- The higher the dopant concentration, the lower the semiconductor resistivity.
- Electrons move faster than holes; therefore, n-type silicon has lower resistivity than p-type silicon for the same dopant concentration.
- Resistors are mainly made with polysilicon. Their resistance is determined by polysilicon line geometry and dopant concentration.
- Capacitors are used in DRAM to store charges and keep data memory.
- Bipolar transistors can amplify electric current and can also be used as switches.
- MOSFET can be turned on and off by biasing the gate at different voltages.
- MOSFET-based IC chips have dominated the semiconductor industry since 1980, and their market share is still increasing.
- Memory, microprocessor, and ASIC chips are the three most commonly made IC chips in the semiconductor industry.
- The advantages of CMOSs are low power consumption, low heat generation, noise immunity, and clocking simplicity.
- The basic CMOS process steps are wafer preparation, isolation formation, well formation, transistor making, interconnection, and passivation.
- The basic semiconductor processes are adding, removing, heating, and patterning (see Fig. 3.38).

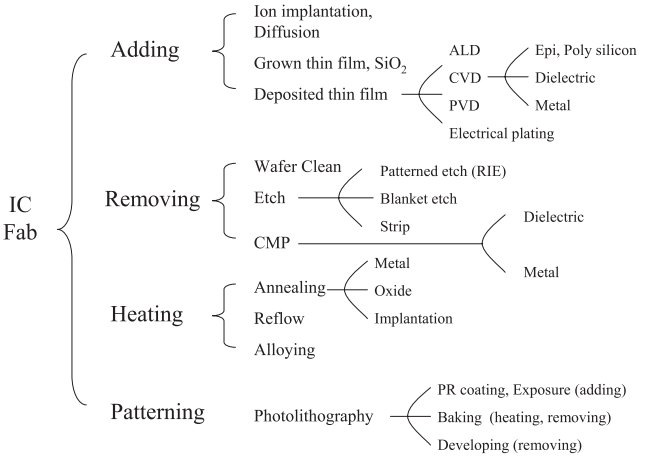


Figure 3.38 Summary of basic IC processes.

3.8 Bibliography

D. Manners, "50 Not Out," *Electronics Weekly*, Dec. 17, 1997.

P. Packan et al., "High Performance 32-nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors," *IEDM Tech. Dig.*, 659 (2009).

G. Stix, "Trends in Semiconductor Manufacturing: Toward 'Point One'," *Sci. Am.* **727**(2), 90–95 (1995).

S. M. Sze, *Physics of Semiconductor Devices, Second Edition*, John Wiley & Sons, New York (1981).

P. J. Zdebel, "Current Status of High Performance Silicon Bipolar Technology," *14th Ann. IEEE GaAs IC Symp. Tech. Dig.*, 15–18 (1992).

3.9 Review Questions

1. What is a semiconductor? List the three most commonly used semiconductor materials.
2. What is the majority carrier of a p-type semiconductor? What material is a p-type dopant?
3. List three n-type dopants. What does "n" stand for?
4. How does semiconductor resistivity change with dopant concentration?
5. At the same dopant concentration, which has higher electrical conductivity, phosphorus-doped silicon or boron-doped silicon?
6. What materials are commonly used to make resistors in IC chips? What factors determine the resistance of the resistor?
7. What kind of IC chip needs many capacitors? Why does it need them?
8. List the two kinds of memory chips that dominate memory IC manufacturing.
9. Describe how to turn an nMOS on and off.
10. List the four basic IC processes. Ion implantation and postimplant RTA (rapid thermal annealing) represent which ones? How about photolithography and etch?
11. What are the minimum mask steps required to make a functional pMOS? How about a bipolar transistor?
12. What was the major technological breakthrough in the mid-1970s in the IC industry?
13. Describe a self-aligned source/drain process. Why it is still used today?
14. Why are CMOS circuits widely used in semiconductor chips?
15. List the basic processes for CMOS IC chip manufacturing.
16. Why do advanced IC chips use multilayer metal interconnections?

Chapter 4
Wafer Manufacturing, Epitaxy, and Substrate Engineering

Objectives

After finishing this chapter, the reader will be able to:

- give two reasons why silicon is more commonly used than any other semiconductor material
- list at least two preferred orientations of single-crystal silicon
- list the basic steps for making silicon wafers from sand
- describe the Czochralski (CZ) and the floating zone methods
- explain the purpose of epitaxial silicon layer deposition
- describe the epitaxial silicon deposition process
- list two ways to make a silicon-on-insulator (SOI) wafer
- identify the advantage of strained silicon
- describe the selective epitaxial process and its application in generating strained silicon.

4.1 Introduction

Single-crystal silicon wafers are the most commonly used semiconductor material in IC manufacturing. This chapter explains why silicon wafers are used the most, and how to make them.

All materials are made from atoms. According to the arrangement of atoms inside solid materials, there are three different material structures: amorphous, polycrystalline, and single crystal. In an amorphous structure, there are no repeated patterns of atoms. A polycrystalline structure has some repeated patterns of atoms, which form grains. All atoms are arranged in the same repeated pattern in a single-crystal structure. Figure 4.1 shows cross sections of the three different structures.

In nature, most solid materials are either amorphous or polycrystalline in structure. Only a few solids have a single-crystal structure, and they are usually gem stones, such as quartz (single-crystal silicon dioxide), ruby (single-crystal aluminum oxide with the presence of the element chromium), sapphire (single-crystal aluminum oxide with different impurities other than chromium), and diamond (single-crystal carbon).

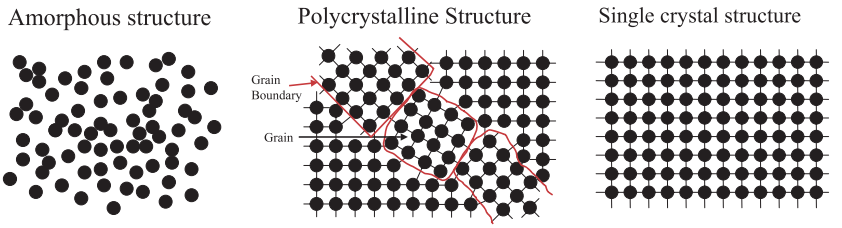


Figure 4.1 The three different structures of solid materials.

The first transistor was made from polycrystalline germanium. However, to make a miniature transistor, a single-crystal semiconductor substrate is required. Electron scattering from the grain boundary in a material can seriously affect p-n junction characteristics. Although they are not good for IC chip manufacturing, both polycrystalline silicon and amorphous silicon can be used to make solar panels, which directly convert photon energy from sunlight into electrical power.

4.2 Why Silicon?

In the early years of the semiconductor industry, germanium was the major semiconductor material used for making electronic devices, such as transistors and diodes. When Kilby made the first IC, it was built on a piece of single-crystal germanium substrate. However, starting in the 1960s, silicon quickly replaced germanium and dominated the IC industry.

About 26% of earth's crust is silicon, making it one of the most abundant elements in the earth's crust, second only to oxygen. It is not necessary to mine silicon. Quartz sand is mainly silicon dioxide, and it can be found in large quantities in many places. As silicon crystal technology developed, the cost of a single-crystal silicon wafer dropped, becoming much lower than single-crystal germanium wafers or any other single-crystal semiconductor material.

Another major advantage of a silicon substrate is that a layer of silicon dioxide can easily grow on its surface in a thermal oxidation process. Silicon dioxide is a strong and stable dielectric, whereas germanium dioxide is difficult to form, unstable at high temperatures (>800 °C), and worst of all, it is water soluble. In fact, in 1947 at Bell Laboratories, a technician mistakenly washed away the anodized germanium oxide layer from the germanium sample surface in the postanodization clean process. This paved the way for Bardeen and Brattain to make the first point-contact bipolar transistor. Otherwise, they might have made the first MOSFET, as they intended to do.

Compared with germanium, silicon has a larger bandgap. Therefore, it can tolerate higher operating temperatures and wider doping ranges. It also has a higher breakdown voltage than germanium. For dopants of interest such as phosphorus and boron, silicon dioxide can serve as a doping mask, since most dopants diffuse much more slowly in silicon dioxide than in silicon. For metal-insulator-semiconductor (MIS) transistors, the SiO_2–Si interface has superior electrical characteristics for MOS. Information on silicon is summarized in Table 4.1.

Table 4.1 Facts about silicon.

Symbol	Si
Atomic number	14
Atomic weight	28.0855
Discoverer	Jöns Jacob Berzelius
Discovery location	Sweden
Discovery date	1824
Origin of name	From the Latin word silicis, meaning flint
Bond length in single-crystal Si	2.352 Å
Density of solid	2.33 g/cm^3
Molar volume	12.06 cm^3
Velocity of sound in substance	2200 m/sec
Hardness	6.5
Electrical resistivity	100,000 $\mu\Omega \cdot$ cm
Reflectivity	28%
Melting point	1414 °C
Boiling point	2900 °C
Thermal conductivity	150 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	2.6×10^{-6} K^{-1}
Etchants (wet)	HNO_4 and HF, KOH, etc.
Etchants (dry)	HBr, Cl_2, NF_3, etc.
CVD precursor	SiH_4, SiH_2Cl_2, $SiHCl_3$, and $SiCl_4$

4.3 Crystal Structures and Defects

4.3.1 Crystal orientation

Figure 4.2 shows the atomic structure of a basic lattice cell of single-crystal silicon. It is a unit cell of the zincblende lattice structure, in which every silicon atom is bonded with another four silicon atoms. Diamond, which is single-crystal carbon, also has this type of crystal structure.

Crystal orientations are defined with Miller indexes, which show a cross section of the orientation plane to the x, y, and z axes. Figure 4.3 illustrates the ⟨100⟩ orientation plane and the ⟨111⟩ orientation plane of a cubic crystal. Note that the ⟨100⟩ plane has a square shape [see Fig. 4.3(a)], and the ⟨111⟩ plane is a triangle [see Fig. 4.3(b)].

The ⟨100⟩ and ⟨111⟩ planes are the most popular orientations used in single-crystal wafers for IC processing. The ⟨100⟩ wafer is commonly used for MOSFET IC chip manufacturing. The ⟨111⟩ orientation wafers are employed to make bipolar transistors and bipolar IC chips, because the ⟨111⟩ surface is stronger due to its higher atom surface density, which makes it ideal for higher-power devices. ⟨100⟩ and ⟨111⟩ orientation planes are illustrated in Fig. 4.4. When a ⟨100⟩ wafer breaks, the fragments normally have right angles (90 deg). If a ⟨111⟩ wafer breaks, the fragments are 60-deg triangles.

Crystal orientation can be determined by numerous methods. Visual identification depends on distinguishing features such as etch pits and growth facets, or x-ray diffraction. Single-crystal silicon can be wet etched. If there are some defects on

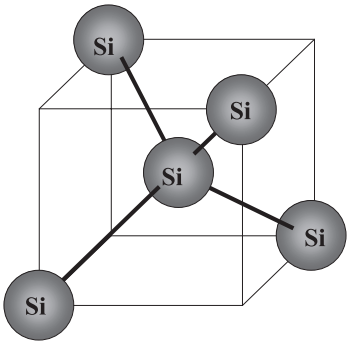


Figure 4.2 A unit cell of the single-crystal silicon lattice structure.

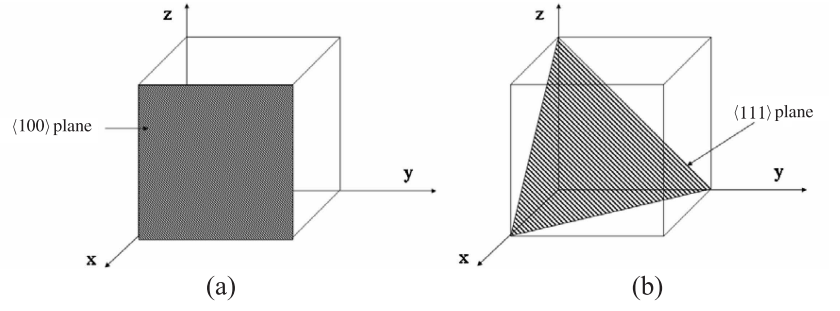


Figure 4.3 Definition of (a) ⟨100⟩ and (b) ⟨111⟩ orientation planes.

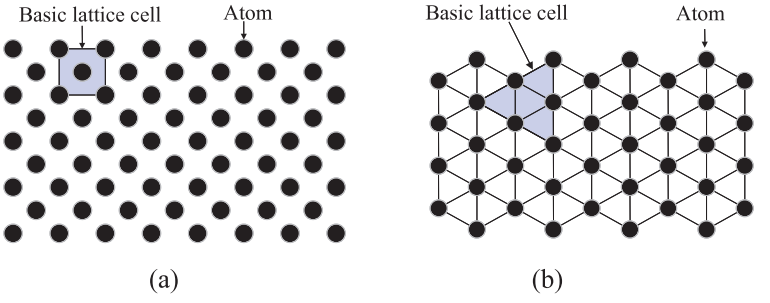


Figure 4.4 Lattice structure of ⟨100⟩ and ⟨111⟩ orientation planes.

the surface where the etch rate is higher, this can create etch pits. For the ⟨100⟩ wafer, an etch pit looks like an inverted four-sided pyramid when it is selectively etched with solutions such as potassium hydroxide (KOH), which etches much faster on the ⟨100⟩ plane than on the ⟨111⟩ plane. For the ⟨111⟩ wafer, an etch pit is a tetrahedron, or three-side inverse pyramid (see Fig. 4.5).

4.3.2 Crystal defects

There are numerous crystal defects that occur during the growth and subsequent processing of silicon crystals and wafers. The simplest point defect is vacancy, which is also called the Schottky defect, in which an atom is missing from the lattice (Fig. 4.6). Vacancy can affect doping processes, since the dopant diffusion rate in single-crystal silicon is a function of the vacancy number.

When an extra atom occupies the space between normal lattice sites, it is called an interstitial defect. If an interstitial defect and a vacancy occur in proximity, the pair is called a Frenkel defect (see Fig. 4.6).

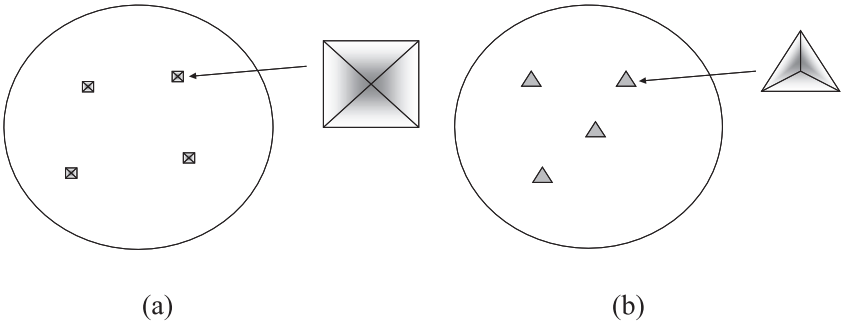


Figure 4.5 Etch pits on (a) ⟨100⟩ wafer and (b) ⟨111⟩ wafer.

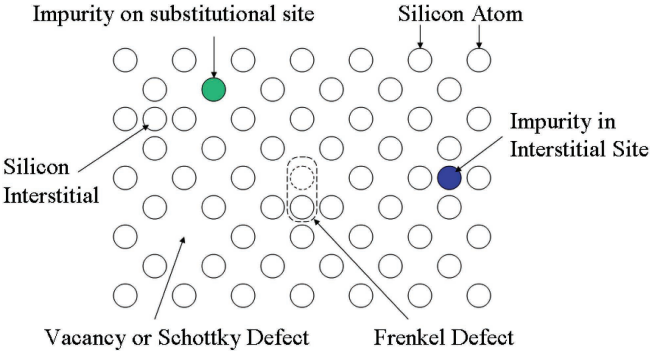


Figure 4.6 Potential silicon crystal defects.

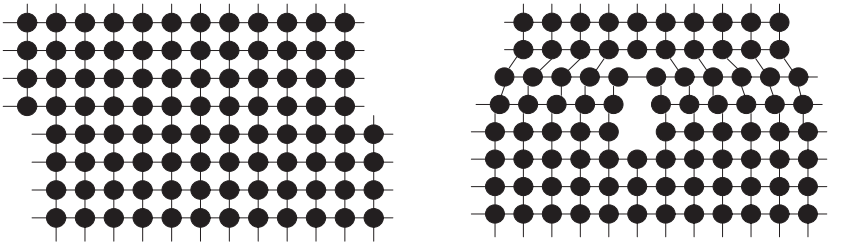


Figure 4.7 Dislocation defects.

Dislocations are geometric faults in the lattice, which can be caused by the crystal pulling process. During the wafer process, dislocation is mainly associated with excessive mechanical stress, often from uneven heating or cooling, dopant diffusion into the lattice, or film deposition, or from external forces such as tweezers. Figure 4.7 shows two examples of dislocation defects.

Dislocation and other defect densities must be minimized on the polished side of wafers, where transistors and other microelectronic devices are made. Defects on the silicon surface can scatter electrons, increasing resistivity and affecting device performance. Defects on a wafer surface can reduce the production yield of IC chips.

Each defect has some dangling silicon bonds, with which impurity atoms tend to bind and become immobilized. On the backside of the wafer, defects are intentionally created to trap unwanted mobile impurities inside the wafer. This prevents them from affecting the normal function of microelectronic devices on the front side.

4.4 Sand to Wafer

4.4.1 Crude silicon

Common quartz sand is mainly silicon dioxide, which can react with carbon at high temperatures. Carbon then replaces silicon to form silicon and carbon monoxide, or carbon dioxide. Because the silicon-oxygen bond is strong, silicon dioxide is very stable. Therefore, this carbon-reducing reaction requires very high temperatures. Manufacturers put quartz sand with carbon into a high-temperature furnace. Carbon used in this reaction does not need to be of very high purity; in fact, it could be in the form of coal, coke, or even pieces of wood. At high temperatures, carbon starts to react with silicon dioxide to form carbon monoxide. This process generates polycrystalline silicon with about 98 to 99% purity, called crude silicon or metallurgical-grade silicon (MGS). The chemical reaction to form MGS can be expressed as

$$\underset{\text{quartzite}}{SiO_2} + \underset{\text{coal}}{2C} \xrightarrow{\text{heat}} \underset{\text{MGS}}{Si} + \underset{\text{carbon monoxide}}{2\,CO}$$

Crude silicon has a high impurity concentration and must be purified for use in semiconductor device manufacturing.

4.4.2 Silicon purification

The purification of silicon has several steps. First, crude silicon is ground into a fine powder. Then the silicon powder is introduced into a reactor to react with hydrogen chloride (HCl) vapor to form a trichlorosilane (TCS, $SiHCl_3$) vapor at about 300 °C. The chemical reaction can be expressed as:

$$\underset{\text{MGS}}{\text{Si}} + \underset{\text{hydrochloride}}{3\,\text{HCl}} \xrightarrow{\text{heat (300 °C)}} \underset{\text{TCS}}{SiHCl_3} + \underset{\text{hydrogen}}{H_2}$$

TCS vapor then passes through a series of filters, condensers, and purifiers to reach ultrahigh-purity liquid TCS, with purity as high as 99.9999999% (nine 9s!), or less than one impurity per billion. Figure 4.8 gives the schematic of the high-purity TCS formation process.

High-purity TCS is one of the most commonly used silicon precursors for silicon deposition. It is widely applied to amorphous silicon, polycrystalline silicon (polysilicon for short), and epitaxial silicon depositions. At high temperatures, TCS can react with hydrogen and deposit high-purity polysilicon. The deposition reaction is

$$\underset{\text{TCS}}{SiHCl_3} + \underset{\text{hydrogen}}{H_2} \xrightarrow{\text{heat (1100 °C)}} \underset{\text{EGS}}{\text{Si}} + \underset{\text{hydrochloride}}{3\text{HCl}}$$

The high-purity polycrystalline silicon is called electronic-grade silicon (EGS). Figure 4.9 illustrates the deposition process of high-purity EGS, and Fig. 4.10 shows actual photographs of it. The EGS is now ready to be pulled into a single-crystal silicon ingot for creating wafers for IC processing.

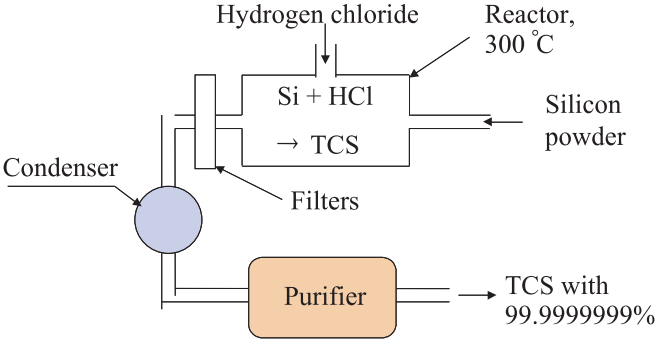


Figure 4.8 Schematic of the process for obtaining high-purity TCS from crude silicon.

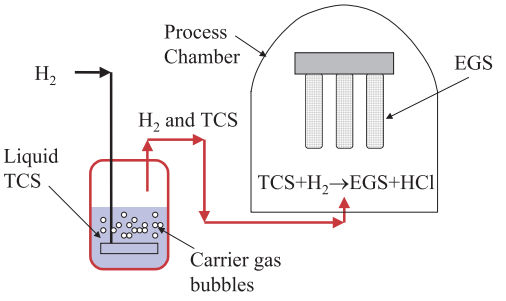


Figure 4.9 EGS deposition.





Figure 4.10 EGS photographs (courtesy of MEMC Electronic Materials).

4.4.3 Crystal pulling

To make a single-crystal silicon ingot, high temperatures are required to melt the EGS and a single-crystal silicon seed. The molten silicon then can recondense with the same crystal structure as the seed crystal. There are two methods commonly used in the semiconductor industry to generate single-crystal silicon: the Czochralski (CZ) method and the floating zone (FZ) method.

4.4.3.1 Czochralski method

The CZ method, which was invented by Polish chemist Jan Czochralski, is the more popular method in wafer manufacturing because it has many advantages over the only alternative, the floating zone method. Only the CZ method can make wafers with a diameter larger than 200 mm, and it has a relatively low cost, since pieces of single-crystal silicon and polycrystalline silicon can be used. Finally, it can create heavily doped single-crystal silicon by melting and recondensing the dopant with the silicon. A schematic of the CZ process is given in Fig. 4.11; the entire system exists in a sealed chamber with an argon ambient to control contamination.

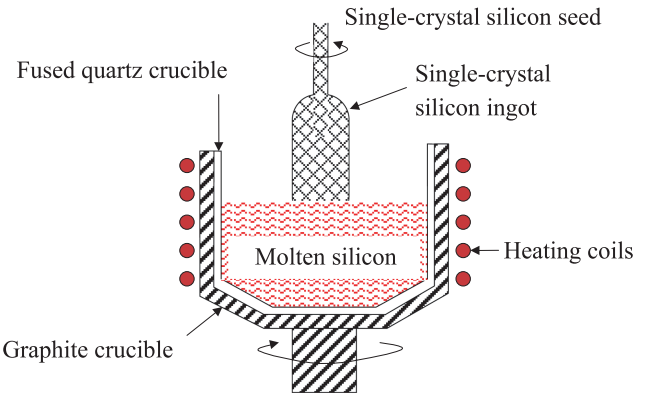


Figure 4.11 Schematic of the CZ crystal-pulling method.

In the CZ method, high-purity EGS is melted in a slowly rotating quartz crucible at 1415 °C, just above the silicon melting point of 1414 °C. It is heated by rf or resistive heating coils. A single-crystal silicon seed rod mounted on a slowly rotating chuck is gradually lowered into the molten silicon, and the surface of the seed crystal submerges into the molten silicon and starts to melt. Seed crystal temperature is precisely controlled at just below the silicon melting point. When the system reaches thermal stability, the seed crystal is withdrawn very slowly, dragging along some molten silicon to recondense around it with the same crystal orientation. The ingot, a whole piece of single-crystal silicon, is formed after up to 48 h of pulling. Rotation and melting of the silicon helps improve dopant uniformity across the crystal ingot, which is sawed into wafers in a later process. In some cases (40–100-$\Omega \cdot$ cm n-type wafers and 100–200-$\Omega \cdot$ cm p-type wafers), a magnetic field is also used to further improve radial uniformity of the dopant. Figures 4.12(a) through 4.12(e) illustrate the crystal pulling process, and Fig. 4.12(f) shows a photograph of a single-crystal silicon ingot pulled with the CZ method.

The diameter of crystals created using the CZ method can be controlled by the temperature and pulling rate. An automatic diameter control (ADC) system is commonly used to control crystal diameter. The ADC system uses infrared sensors to monitor the bright radiation ring at the crystal and molten silicon interface [see Fig. 4.12(c)], and the feedback signals control the pulling rate. If the ring is in the sensor range, it increases the pulling rate, reducing the crystal diameter; if the ring is out of the sensor range, it reduces the pulling rate, increasing the crystal diameter. The grooves on the side of the single-crystal ingot in Fig. 4.12(f) are caused by this feedback diameter control process.

Single-crystal silicon ingots pulled by the CZ method always have trace amounts of impurities of oxygen and carbon, which come from the crucible materials. Typically, the oxygen concentration in silicon crystal grown by the CZ method is approximately from $1.0 \times 10^{16}/\text{cm}^3$ to $1.5 \times 10^{18}/\text{cm}^3$. The carbon content varies from about $2.0 \times 10^{16}/\text{cm}^3$ to about $1.0 \times 10^{17}/\text{cm}^3$. Oxygen and

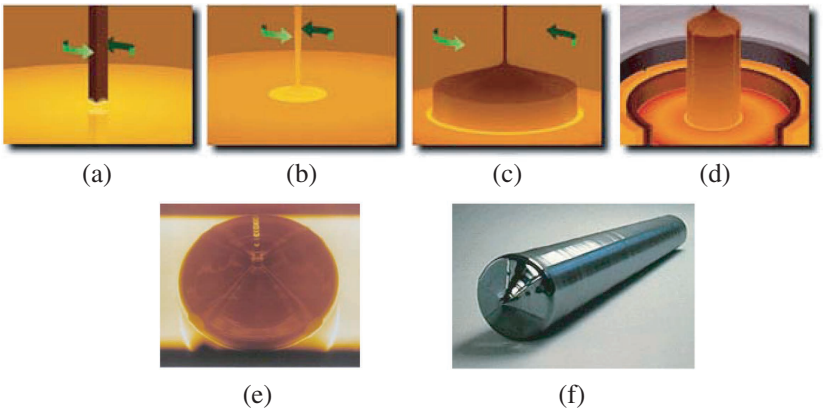


Figure 4.12 Single-crystal silicon ingot and its CZ pulling process (courtesy of MEMC Electronic Materials).

carbon concentrations in the silicon are the result of ambient pressure during crystal growth, pulling and rotation rates, and the ratio of the crystal diameter to its length.

4.4.3.2 Floating zone method

The floating zone method is the only other practical way to make single-crystal ingots. Figure 4.13 illustrates the floating zone crystal pulling method. As with the CZ method, the process takes place in a sealed furnace chamber with an Ar ambient.

The process starts with a polycrystalline silicon rod, about 50 to 100 cm in length, placed vertically in a furnace chamber. The heating coil heats and melts the lower end of the polysilicon bar, and the seed crystal fuses into the molten zone. The melt is suspended between the seed and polysilicon rod by the surface tension of the molten silicon. Then the heating coils move slowly upward, melting the polysilicon rod just above the molten silicon. The molten silicon at the seed end starts to freeze with the same crystal orientation as the seed crystal. After the heating coils move over the entire polysilicon rod, the rod converts to a single-crystal silicon ingot.

The diameter of the ingot is controlled by the relative rotation rates of the top and bottom sections. The largest diameter of a silicon wafer is 450 mm (18 in.) with the CZ method. With the floating zone method, the largest wafer diameter is 150 mm (6 in.). The CZ method has been used for mass producing the 300-mm (12-in.) wafer since the 1990s.

Since no crucible is used in this process, the primary advantage of the floating zone method is that it can achieve higher silicon purity due to the lower melt contamination, especially from oxygen and carbon. This method is mainly used to manufacture wafers needed for discrete power devices. These wafers require high-resistance starting materials.

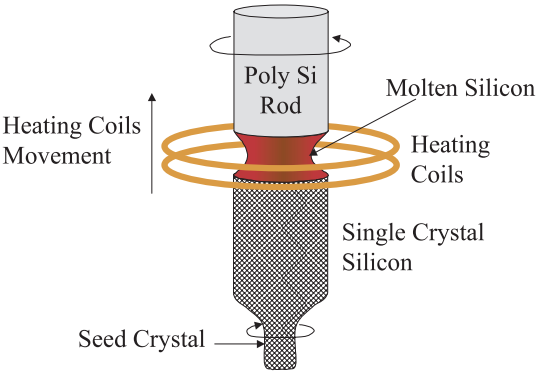


Figure 4.13 Schematic of the floating zone method.

There are two major disadvantages of the floating zone method. One is that the melt–crystal interface is very complex, thus it is difficult to achieve dislocation-free crystals. Another is that it is more expensive, since it requires high-purity polysilicon rods as the starting material, while for the CZ method, any kind of high-purity silicon, such as wafer saw powder, ingot end cuts, and the like, can be used as starting materials.

Because they are formed by freezing the rotating molten silicon, single-crystal silicon ingots—and the wafers sliced from them—are always round. It is possible to make square wafers by grinding the ingot into square columns before slicing wafers from them. However, a square wafer is much harder to handle mechanically. It can be very easy to chip the corners of a square wafer and break it down.

4.4.4 Wafering

After the crystal ingot cools down, machines cut both ends and polish the side to remove the grooves created by the automatic diameter control process. Then a flat (150 mm or smaller) or notch (200 mm or larger) is ground into the ingot to mark the crystal orientation (Fig. 4.14).

Now the ingot is ready to have wafers sliced from it. In a rapid-rotating inward-diameter diamond-coated saw, the ingot moves outward, and wafers are cut, slice by slice, from the ingot. Coolant pours constantly over both ingot and saw to control the considerable heat generated by the sawing process. Wafers are sawed as thin as possible; however, they must be thick enough to sustain the mechanical handling in wafer processing. Larger-diameter wafers require larger thickness. Thicknesses of the different wafer sizes are listed in Table 4.2.

During the sawing process, about one-third of a single-crystal ingot is reduced to sawdust. Fortunately, it can be reused in the CZ crucible as the starting silicon material. The sawing process is illustrated in Fig. 4.15. While sawing, the saw blade must be free of vibration, since any blade vibration will scratch the wafer surface and increase difficulties and costs of the subsequent tapping and polishing processes. Blade withdrawal must be well controlled to prevent retrace damage.

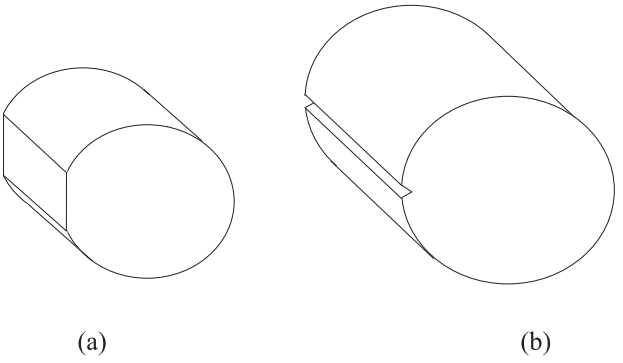


Figure 4.14 Single-crystal ingot after polishing, with (a) a flat or (b) notch.

Table 4.2 Wafer thickness for different wafer sizes. The asterisk indicates the preliminary standard set in October 2008.

Wafer size (mm)	Thickness (μm)	Area (cm^2)	Weight (g)
50.8 (2 in.)	279	20.26	1.32
76.2 (3 in.)	381	45.61	4.05
100	525	78.65	9.67
125	625	112.72	17.87
150	675	176.72	27.82
200	725	314.16	52.98
300	775	706.21	127.62
450	925 ± 25*	1590.43	342.77

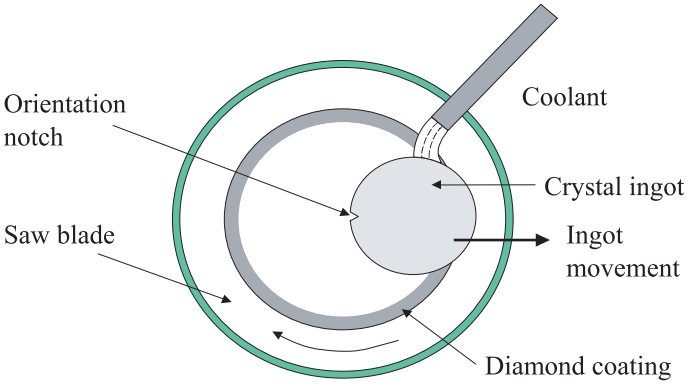


Figure 4.15 Wafer sawing process.

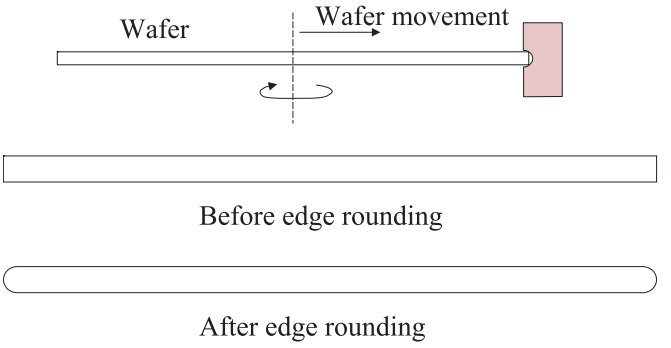


Figure 4.16 Edge-rounding process.

4.4.5 Wafer finishing

After wafer sawing, the wafer edge is ground in a mechanical process to round the sharp edges created in the slicing process. Round edges prevent edge chipping during mechanical handling of wafer processing. Figure 4.16 shows a schematic of the edge-rounding process.

The wafer is then rough polished by the conventional, abrasive, slurry-lapping process to remove the majority of surface damage caused by the wafer-slicing process, and to create the flat surface required for photolithography. The mechanical, double-sided lapping process is performed under pressure using glycerin slurry with fine alumina (Al_2O_3) particulate suspended in it. The lapping process can produce a wafer with a surface flatness within 2 μm and removes about 50 μm of silicon from both sides of a 200-mm wafer.

A wet etch process is used to remove particles and damage created during the sawing, edge-rounding, and lapping processes. Since some sawing damage can be 10 μm deep into the silicon, a wet etch needs to remove about 10 μm of silicon from both sides of a wafer. A commonly used etchant is a mixture of nitric acid (HNO_3), hydrofluoric acid (HF), and acetic acid (CH_3COOH). Nitric acid oxidizes the silicon to form silicon dioxide on the wafer surface. HF dissolves and removes the silicon dioxide. CH_3COOH helps to control the reaction rate. Due to the isotropic etch characteristic, the etch process further smoothes the wafer surface due to the isotropic etch characteristic with an HNO_3-rich solution. The usual formulation is a 4:1:3 mixture of HNO_3 (79 wt% in H_2O), HF (49 wt% in H_2O) and pure CH_3COOH. The combined chemical reaction can be expressed as

$$3\ Si + 4\ HNO_3 + 6\ HF \rightarrow 3\ H_2SiF_6 + 4\ NO + 8\ H_2O.$$

In the CMP process shown in Fig. 4.17, the wafer is held on a rotating wafer holder and pressed onto a rotating polishing pad, with slurry and water in between. The slurry is a colloidal suspension of fine silica (SiO_2) particles with diameters of about 100 Å in an aqueous solution of sodium hydroxide (NaOH). NaOH oxidizes the silicon surface (chemical process) with the help of heat generated by the friction

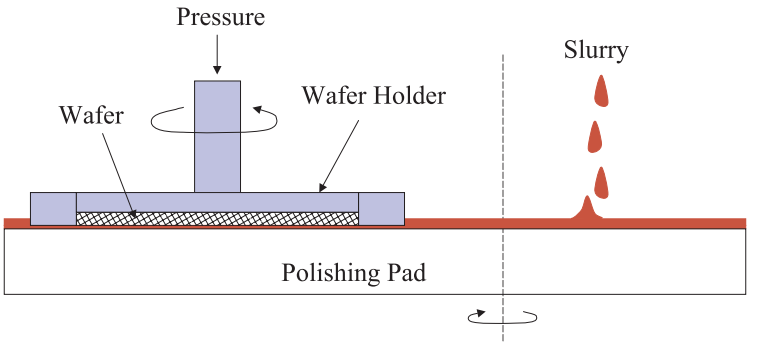


Figure 4.17 Schematic of CMP process.

between the wafer and polishing pad. Then the silica particles abrade the silicon oxide away from the surface (mechanical process). This combined chemical and mechanical process removes silicon from the wafer surface.

Post-CMP cleaning processes use mixtures of acid-oxidizer solutions to remove organic and inorganic contaminants and particles. This finishes the wafer-surfacing process and creates a defect-free surface ready for IC processing. Commonly used clean solutions are a mixture of ammonium hydroxide (NH_4OH) and hydrogen peroxide (H_2O_2), a mixture of hydrochloric acid (HCl) and hydrogen peroxide (H_2O_2), or a mixture of sulfuric acid (H_2SO_4) and H_2O_2.

Figure 4.18 illustrates 200-mm wafer surface roughness changes and wafer thickness changes during the wafer finishing processes. After post-CMP cleaning, inspection, and labeling, the wafer is ready to ship to the customer for semiconductor chip processing.

Manufacturers intentionally create defects and dislocations on the backside of wafers to trap heavy metals, mobile ions, oxygen, carbon, and other contaminants. Backside defects can also be created by argon ion implantation, polysilicon deposition, and heavily doped phosphorus. During IC processing, the backside of a wafer is always deposited with a CVD silicon oxide or silicon nitride layer to prevent any outward diffusion during thermal processes.

Since all transistors and circuits are made on one side of the wafer, most wafers need to be polished on only one side. The unpolished side is used for handling during wafer processing. Some processes, such as Fourier transform infrared (FTIR) measurement of thin-film moisture absorption, requires a double-side polished wafer. Otherwise, scattering of infrared light from the rough surface of the wafer backside will disrupt measurement signals and cause invalid results.

4.5 Epitaxial Silicon Deposition

The word epitaxy comes from two Greek words, *epi* meaning upon, and *taxis* meaning arranged or ordered. The epitaxial deposition process is a process used to grow a thin crystalline layer on a crystalline substrate. Unlike the CZ or floating

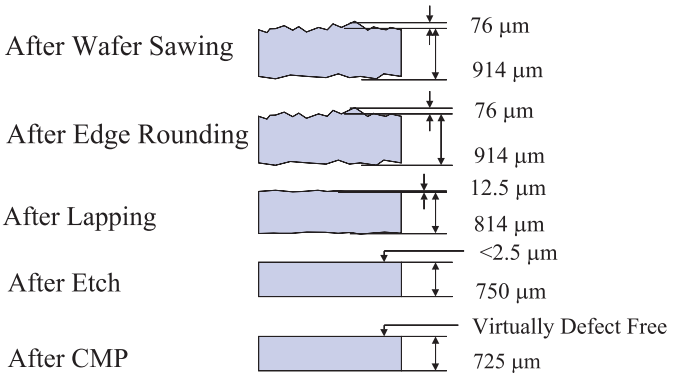


Figure 4.18 Changes of thickness and surface roughness of a 200-mm wafer during wafer manufacturing processes.

zone crystal growing processes, an epitaxy layer grows at a temperature much lower than the melting temperature of silicon.

Earlier development of epitaxial silicon deposition was driven by the requirement of bipolar transistors for high breakdown voltage of the collector. An epitaxial layer can provide a high-resistance layer over the low-resistance substrate, enhancing performance of the bipolar transistor. The epitaxial layer can also enhance the performance of DRAM and CMOS ICs. There are two fundamental advantages of using an epitaxial layer. One is that bipolar transistors require an epitaxial layer to form the heavily doped burial layer deep inside the silicon. This layer cannot be made by existing doping processes such as ion implantation or diffusion. Another advantage is that an epitaxial layer can allow for different physical properties in a substrate wafer. For example, a p-type epitaxial layer grown on an n-type wafer can provide a designer more freedom in the structure of a microelectronic device and circuits. The epitaxial layer is generally free of oxygen and carbon. This cannot be achieved in silicon wafers made with the CZ method because small amounts of oxygen in the quartz (SiO_2) crucible and carbon in the graphite liner can diffuse into the molten silicon during the CZ crystal-pulling process and end up in the silicon crystal.

Bipolar transistors always require an epitaxial layer in order to form a buried layer. For low-speed CMOS and DRAM chips, fabs try to avoid using an epitaxial layer. The epitaxial growth process is the most expensive processing step in IC fabrication, about $20 to $100 per wafer, compared with about $1 per wafer per step for other processes. For a higher-performance IC chip, an epitaxial layer is required because the oxygen impurity in the silicon wafer formed by the CZ method can reduce the carrier lifetime and slow down the device. Figure 4.19 illustrates the application of an epitaxial silicon layer in bipolar and CMOS IC chips.

Currently, epitaxial silicon layers are deposited onto wafers by wafer manufacturers, not IC manufacturers. This is why the epitaxy process is covered in this chapter.

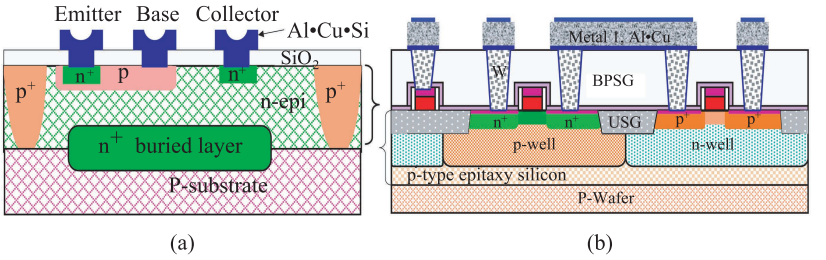


Figure 4.19 Epitaxial silicon layer applications for (a) bipolar and (b) CMOS IC chips.

There are two ways to grow epitaxy silicon layers on silicon wafers: a CVD epitaxy process and a molecular beam epitaxy process.

4.5.1 Gas phase epitaxy

High-temperature (~1000 °C) CVD epitaxial silicon layer growth is the most popular method in the semiconductor industry for growing a single-crystal silicon layer. Silicon source gases used are silane (SiH_4), dichlorosilane (DCS, SiH_2Cl_2), and trichlorosilane (TCS, $SiHCl_3$). The chemical reactions of epitaxial silicon growth are

$$\underset{\text{silane}}{SiH_4} \xrightarrow{\text{heat (1000 °C)}} \underset{\text{epi–Si}}{Si} + \underset{\text{hydrogen}}{2\,H_2}\,,$$

$$\underset{\text{DCS}}{SiH_2Cl_2} \xrightarrow{\text{heat (1100 °C)}} \underset{\text{epi–Si}}{Si} + \underset{\text{hydrogen chloride}}{2HCl}\,,$$

$$\underset{\text{TCS}}{SiHCl_3} + \underset{\text{hydrogen}}{H_2} \xrightarrow{\text{heat (1100 °C)}} \underset{\text{epi–Si}}{Si} + \underset{\text{hydrogen chloride}}{3HCl}\,.$$

As the film grows, epitaxial silicon can be doped by flow dopant gases, such as arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6), with silicon source gas in the reactor. At high temperatures, these dopant hydrides dissociate from the heat and release arsenic, phosphorus, and boron into the epitaxial silicon film. In this way, in-situ doping of epitaxial film can be achieved. Chemical reactions of the in-situ doping are

$$AsH_3 \xrightarrow{\text{heat (~1000 °C)}} As + 3/2\,H_2,$$

$$PH_3 \xrightarrow{\text{heat (~1000 °C)}} P + 3/2\,H_2,$$

$$B_2H_6 \xrightarrow{\text{heat (~1000 °C)}} 2\,B + 3\,H_2.$$

All three hydride dopant source gases mentioned are highly poisonous, flammable, and explosive. During epitaxial-layer growth, the dopant inside the substrate, driven by the high temperature, diffuses into the epitaxial layer. If the film growth is slower than the dopant diffusion, the entire epitaxial layer becomes

doped by the substrate dopant. This effect, called autodoping, is undesirable, since it can affect dopant concentration in the epitaxial layer. To avoid autodoping, the deposition rate of the epitaxial layer must outpace the dopant diffusion rate in the epitaxial layer so that the dopant concentration of the epitaxial film is determined by the dopant induced in the gas phase during deposition. High deposition temperatures can help meet this requirement.

4.5.2 Epitaxial growth process

An example of the epitaxial silicon growth and doping process is illustrated in Fig. 4.20. First, precursors such as DCS and AsH_3 are introduced into the reactor. Precursor molecules diffuse to the wafer surface, adsorb on the surface, dissociate, and react on the surface. Solid byproduct atoms called adatoms migrate onto the surface and bond with other surface atoms in the same crystal structure as the substrate crystal, while volatile byproducts desorb from the hot surface and diffuse out.

Figure 4.21 shows the relationship between temperature and epitaxial silicon growth rate for different silicon precursors. We can see that there are two deposition regimes: one at lower temperature, with growth rates highly sensitive to temperature, and another at higher temperatures with growth rates less sensitive to temperature. The first regime is called the surface-reaction-limited regime, and the second is called the mass-transport-limited regime. We discuss these in detail in Chapter 10.

The silane process is in the surface-reaction-limited regime when the temperature is below 900 °C. It changes to the mass-transport-limited regime when the temperature is higher than 900 °C.

Question: For the silane process, if the temperature continues to increase to 1300 °C, how will the growth rate change?

Answer: Silane is a very reactive gas. When the temperature is higher than 1200 °C, it starts to react in the gas phase (gas phase nucleation). This reduces the epitaxy growth rate and generates huge amounts of particles (a regime to be avoided at all costs). Dichlorosilane and trichlorosilane are less reactive than silane; they have higher reaction temperatures and can deposit at higher temperatures.

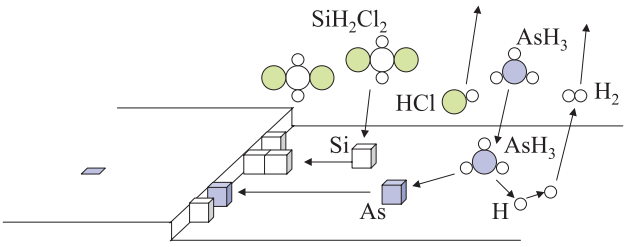


Figure 4.20 Schematic of epitaxial silicon layer growth and doping process.

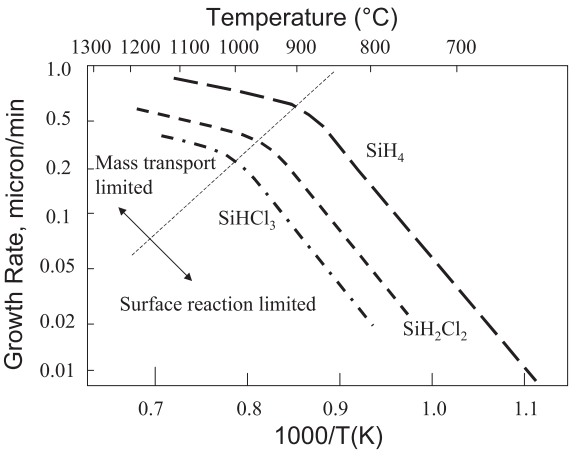


Figure 4.21 Epitaxial silicon film growth rate versus temperature [redrawn from F. C. Everstyn, *Phillips Research Reports* **29** (1974)].

At lower temperatures (550 to 650 °C) in a low-pressure reactor, a silane-based reaction can be used to deposit polycrystalline silicon on a single-crystal silicon wafer surface. This is because at lower temperatures, surface mobility of the adatoms is lower. Multiple nucleation sites form on the surface, causing different crystal grains to grow and form a polycrystalline silicon layer. At even lower temperatures (<550 °C), the silane-based process deposits amorphous silicon because of the extremely low surface mobility of the thermally dissociated radicals from silane molecules, SiH_3, SiH_2, and SiH.

4.5.3 Epitaxy hardware

There are two kinds of epitaxy systems, the batch system and the single-wafer system. The batch system can process multiple wafers at a time. As a result, it has a higher throughput. Three different batch reactors are widely used in the semiconductor industry; they are the barrel reactor, vertical reactor, and horizontal reactor, as illustrated in Fig. 4.22.

All three reactors have their own advantages and disadvantages. Barrel systems have very good uniformity but need extensive preventive maintenance when operating temperatures are greater than 1200 °C. This makes them unsuitable for the >1200 °C high-temperature processes. The horizontal system is simpler and less expensive. However, using this system, it is difficult to control wafer-to-wafer uniformity because it is hard to control the process over the entire susceptor. The pancake vertical reactor normally has good uniformity but has formidable mechanical complexity. All of the batch systems have problems with large-size wafers, especially within wafer uniformity and wafer-to-wafer uniformity on 300-mm wafers. Therefore, the single-wafer epitaxy reactors introduced in the 1990s are becoming more popular. Figure 4.23 gives a schematic of a single wafer epitaxy reactor.

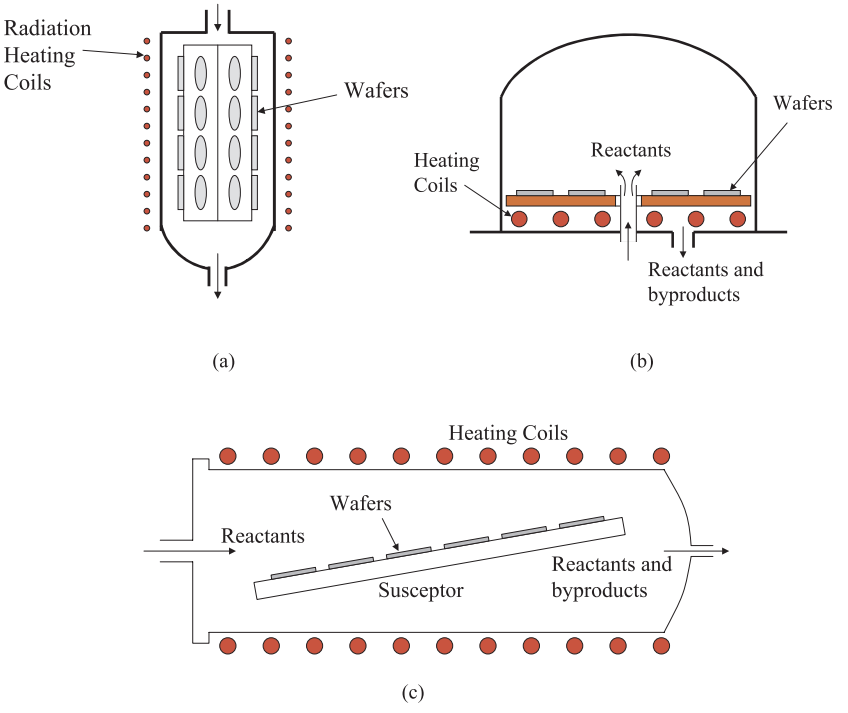


Figure 4.22 Three batch epitaxy reactors: (a) barrel, (b) vertical, and (c) horizontal.

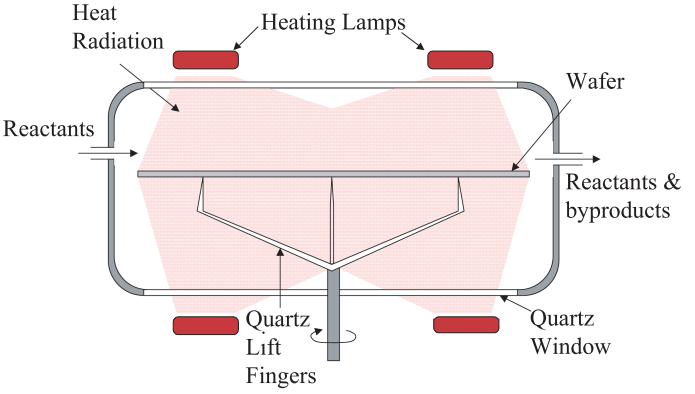


Figure 4.23 Schematic of single-wafer epitaxy system.

The single-wafer epitaxy reactor has a high epitaxial-layer growth rate and better reliability and repeatability than batch systems. It can deposit high-quality, low-cost film at both atmospheric and reduced pressures.

4.5.4 Epitaxy process

Epitaxy processes for these batch reactors are similar and have several processing steps. First, wafers are loaded into the chamber on the susceptor, then the chamber is closed and purged with hydrogen. After the hydrogen purge, the temperature is raised to 1150 to 1200 °C, and HCl then flows into the chamber for about three minutes to clean the chamber surface and etch the wafer surface to remove native oxide, particles, and surface defects. This helps minimize the mobile ions, especially those due to sodium contamination. Next, the chamber temperature is adjusted to the required process temperature. After the temperature is stabilized, silicon and dopant precursor gases flow into the chamber to grow the epitaxial silicon layer at a growth rate of 0.2 to 4.0 μm/min, depending on the process conditions (mainly pressure, gas flow rate, and temperature). After the epitaxial film is grown, process gases are eliminated, and the heater power is turned off. Hydrogen again flows into the chamber to purge the processing gases. When the temperature drops low enough, nitrogen purges the chamber to an ambient temperature, and the chamber is ready to open, unload, and reload. The whole process takes about an hour and can process 10 to 28 wafers at one time, depending on wafer size and reactor type.

For a single-wafer system, the process is similar. One difference is that the chamber does not need to cool down to an ambient temperature for loading and unloading. Because there is only one wafer, thermal capacity is small. When heated by an array of heating lamps, the wafer temperature can ramp-up very quickly in the single-wafer epitaxial system. After epitaxial-layer deposition, the wafer is removed from the deposition chamber by a transfer robot and sent to a cool-down chamber before it is placed into a plastic wafer cassette or a front-opening unified pod (FOUP).

Question: Nitrogen is the most commonly used purge gas in every type of processing equipment in semiconductor fabs. Why does epitaxy silicon processing use hydrogen instead of nitrogen as the main purge gas?

Answer: Nitrogen is very stable and abundant; 78% of the atmosphere is nitrogen. This makes it the most cost-effective gas to use for purging chambers and gas lines. However, at >1000 °C, nitrogen is no longer inert and can react with silicon to form silicon nitride, which can affect the epitaxial silicon deposition process. Therefore, hydrogen is used for the epitaxy chamber purge because it helps clean the wafer by forming gaseous hydride with contaminants on the wafer surface. Some facts of hydrogen are listed in Table 4.3.

Several possible defects are shown in Fig. 4.24. Dislocation of the epitaxial layer can be caused by dislocation of the substrate. This shows up as a slip or particulate

Table 4.3 Facts about hydrogen.

Symbol	H
Atomic number	1
Atomic weight	1.00794
Discoverer	Henry Cavendish
Discovery location	England
Discovery date	1766
Origin of name	From the Greek words hydro and genes, meaning water and generator.
Molar volume	11.42 cm^3
Velocity of sound	1270 m/sec
Refractive index	1.000132
Melting point	−258.99 °C
Boiling point	−252.72 °C
Thermal conductivity	0.1805 W m^{-1} K^{-1}
Main applications in IC processing	Epitaxial deposition, wet oxidation, premetal deposition reactive cleaning, and tungsten CVD
Main source	H_2

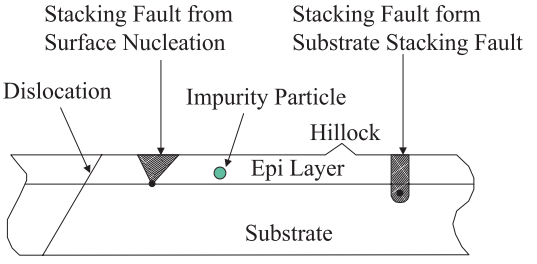


Figure 4.24 Different defects.

contamination on the wafer surface and can cause nucleation and stacking faults in the epitaxial layer. In film, particles can be generated during the epitaxy growth process due to spreading of stacking defects from the substrate surface to the epitaxial film.

4.5.5 Future trends of epitaxy

While device geometry continues to shrink and performance improves, higher quality, defect-free epitaxial layers with minimal thickness are required. Conventional epitaxial-layer growth requires a very high processing temperature, which can cause an autodoping effect and limit the minimum epitaxial-layer thickness. By reducing the epitaxy growth temperature, an abrupt transition between the epitaxial layer and substrate can be achieved. Therefore, low-temperature epitaxy (LTE) growth is of great importance. One way to reduce the epitaxy growth temperature is by reducing the processing pressure. Currently, reduced-pressure epitaxial growth processes operate at 40 to 100 torr and require about a 1000 °C processing temperature. Operating temperatures can be reduced to 750 to 800 °C when the processing pressure is further reduced to 0.01 to 0.02 torr.

Ultrahigh-vacuum CVD (UHV-CVD) epitaxial growth processes at 10^{-6} to 10^{-9} torr with low temperatures of 550 to 650 °C have been in the research and development phase for quite some time. It is likely to be the most promising LTE technology for the future of IC fabrication. However, atmospheric-pressure CVD (APCVD) epitaxy and molecule beam epitaxy (MBE) technologies could also be used for LTE applications in the future.

Another possible application of the UHV-CVD epitaxy process is growing a silicon-germanium (SiGe) epitaxial layer on a silicon substrate with SiH_4 and GeH_4. SiGe has higher hole mobility and therefore can be used to make faster CMOS IC devices. Silicon-carbine (SiC) and SiGeC epitaxy processes are also being researched and could be applicable for future IC devices.

4.5.6 Selective epitaxy

Selective epitaxy is another promising development. It uses silicon dioxide or silicon nitride as the epitaxy mask; thus, the epitaxial layer only grows on the area where silicon is exposed, a characteristic that can help to increase device packing density and reduce parasitic capacitance. Figure 4.25 illustrates the selective epitaxy process.

The selective epitaxial growth (SEG) process illustrated in Fig. 4.25 has been widely used to deposit SiGe in the pMOS S/D area to help create uniaxial compressive strain in the pMOS channel, increasing hole mobility and improving pMOS drive current and speed. It can also be used to deposit SiC in the nMOS S/D region to help create tensile strain in the nMOS channel, increasing electron mobility and improving nMOS drive current and speed. The SEG process can also be used in hybrid orientation technology, which allows two different silicon orientations to be on the same wafer surface, so that nMOS and pMOS can be built on different crystal orientations, maximizing carrier mobility and device performance. These are described in detail in the following section on substrate engineering.

4.6 Substrate Engineering

With the development of silicon IC technology, many methods have been employed to enhance MOSFET device performance. Global substrate engineering such as silicon-on-insulator (SOI), strained silicon, and strained silicon-on-insulator (SSOI) are some examples. In IC manufacturing, more fabs apply localized substrate engineering such as hybrid orientation technology (HOT), stress liners, selective epitaxial SiGe, and selective epitaxial SiC.

4.6.1 Silicon-on-insulator wafer

An SOI wafer allows semiconductor manufacturers to completely isolate IC devices from their neighbors, thus reducing interference and leakage, and in turn improving device speed and performance. There are two ways to make an SOI wafer: one method uses heavy oxygen implantation and high-temperature annealing, and the other uses hydrogen implantation and wafer bonding.

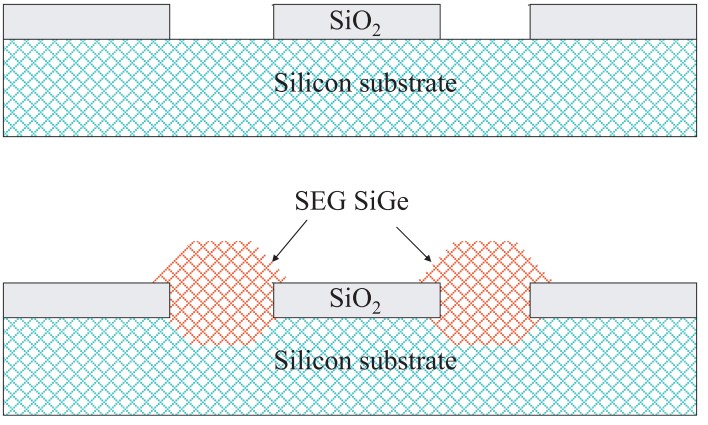


Figure 4.25 The selective epitaxy silicon process.

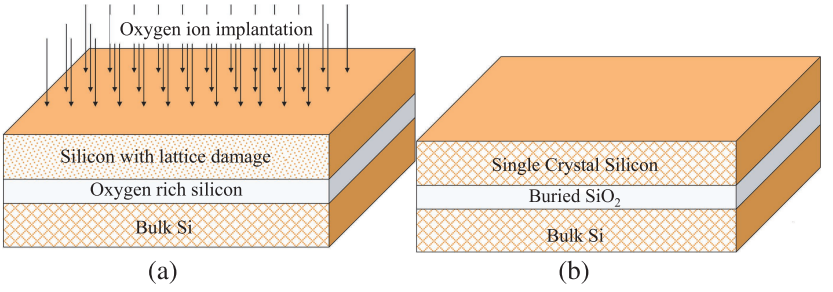


Figure 4.26 SOI wafer formation with (a) oxygen ion implantation and (b) high-temperature annealing.

Figure 4.26 illustrates the SOI formation of the first approach—separation by implantation of oxygen (SIMOX). First, a high-energy, high-current implanter (up to 10^{18} ions/cm^2) is used to implant oxygen ions deep into the silicon substrate to form an oxygen-rich silicon layer. A thermal annealing process at a high-temperature (~1400 °C) causes a chemical reaction between silicon and oxygen atoms and forms the buried silicon dioxide layer, while silicon near the surface relaxes and recovers the single-crystal structure. The thickness of the silicon on top of the oxide is determined by the oxygen implantation energy, and the thickness of the buried oxide is determined by the number of oxygen atoms.

The other method is called bonded SOI, which uses hydrogen implantation and wafer bonding (Fig. 4.27). First, wafer A is cleaned, and a silicon dioxide layer is grown with the thickness required for the buried oxide. Wafer A is then is implanted with hydrogen to form a hydrogen-rich layer [see Figs. 4.27(a) and 4.27(b)]. The depth of the hydrogen-rich layer is determined by the thickness of the silicon on top of the buried oxide. Wafer A is then flipped and bonded with wafer B, the handler wafer [Fig. 4.27(c)]. In a thermal process, silicon dioxide on the surface of wafer

A forms chemical bonds with the silicon surface of wafer B, while the hydrogen in wafer A bubbles up inside the silicon, causing wafer A to split [Figs. 4.27(d) and 4.27(e)]. After CMP and cleaning, the SOI wafer is ready, as shown in Fig. 4.27(f).

After processing, wafer A can be polished, cleaned, and recycled to become a new wafer A or B. The major advantage of bonded SOI is the cost. Because the hydrogen dosage needed to split a silicon wafer is significantly lower than the oxygen dosage needed to form buried oxide, bonded SOI has higher throughput than SIMOX SOI. Most SOI wafers used in IC manufacturing are made with bonded SOI technology.

4.6.2 Hybrid orientation technology

By using a ⟨110⟩ orientation wafer as wafer A in Fig. 4.27 and a ⟨100⟩ orientation wafer as wafer B, a hybrid orientation SOI wafer can be made, as shown in Fig. 4.28(a). Using the selective epitaxial growth technique illustrated in Fig. 4.25, hybrid ⟨110⟩ and ⟨100⟩ orientations can be achieved on the wafer surface [see Fig. 4.28(c)]. Figure 4.28(d) shows CMOS devices in which pMOS is made on a ⟨110⟩ substrate, while nMOS is made on a ⟨100⟩ substrate.

Because hole mobility is higher in the ⟨110⟩ orientation than in the ⟨100⟩ orientation of single-crystal silicon, a CMOS IC made with HOT could have a higher driving current and faster speed than an IC made with a regular ⟨100⟩ substrate. The tradeoff is cost: the CMOS IC would need an SOI wafer and an additional photolithography mask layer. If other techniques such as strain silicon

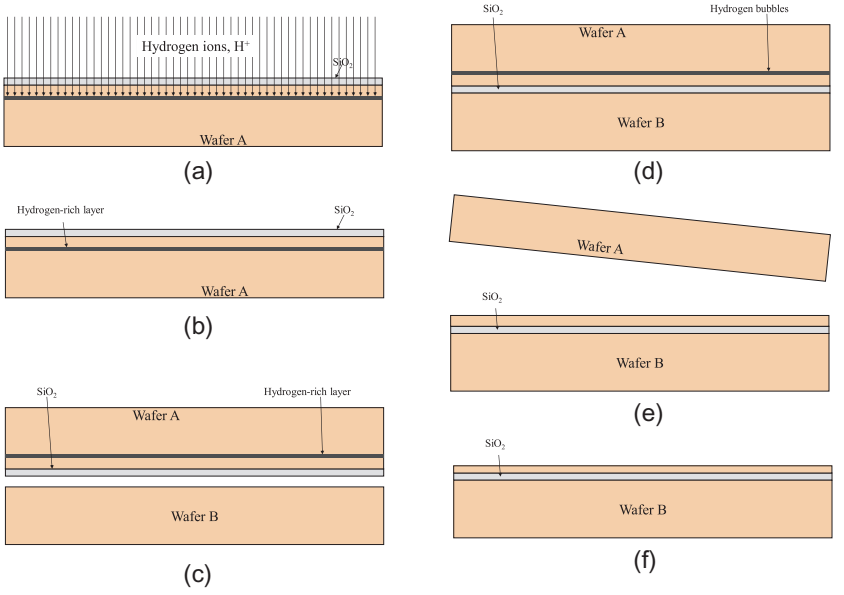


Figure 4.27 Bonded SOI formation process.

can achieve the same performance gain at a lower cost, it would be difficult for HOT to become mainstream in CMOS technology.

4.6.3 Strained silicon wafer

By heavily stressing single-crystal silicon, silicon lattices are stretched or compressed beyond their normally interatomic distance, forming what is called strained silicon. Carrier mobility in strained silicon can be significantly higher than that in strain-free silicon. Strained silicon can be achieved by depositing silicon onto a SiGe surface. Gradient SiGe with gradually increasing Ge concentration allows growth on a silicon surface without much lattice mismatch. After relaxed-SiGe layer deposition, an epitaxial silicon layer is deposited onto the SiGe surface, with a lattice structure that matches the SiGe underneath and the strained silicon layer above (see Fig. 4.29).

4.6.4 Strained silicon-on-insulator wafer

Using the strained silicon wafer described in Fig. 4.29(a) as wafer A of the bonded SOI process (shown in Fig. 4.27), an SSOI wafer can be made. These types of wafers have the high carrier mobility of strained silicon and the high device packing density of SOI (see Fig. 4.30).

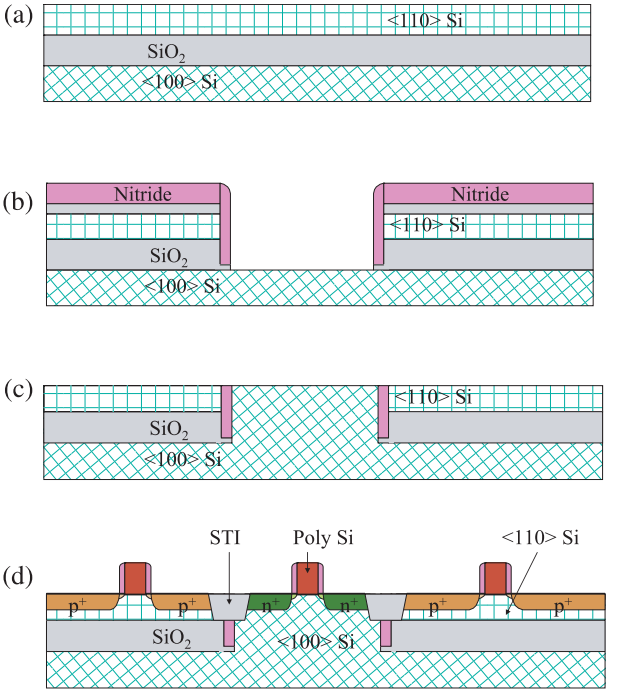


Figure 4.28 Hybrid orientation technology.

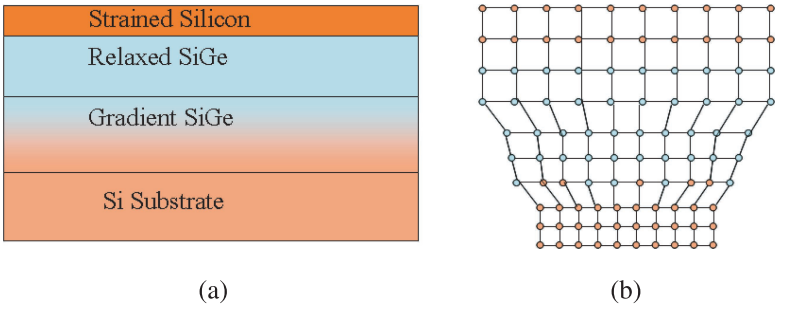


Figure 4.29 (a) Formation and (b) lattice of strained silicon.

The strained silicon and SSOI wafers described in this section are wafers with strained silicon on their top surfaces that is used for making microelectronic and nanoelectronic devices. However, scientists and engineers have found that by using existing process technologies, they can create strained silicon locally on a normal silicon wafer. Because only the channel underneath the gate oxide of a MOSFET needs to be strained, it is not necessary to strain the entire surface of a silicon wafer. Also, pMOS and nMOS channels require different types of strain; pMOS prefers compressive strain to enhance hole mobility, while nMOS needs tensile strain to enhance electron mobility. A strained silicon wafer cannot satisfy both needs.

4.6.5 Strained silicon in integrated circuit manufacturing

IC manufacturers started using strained silicon in MOSFET devices when the feature size shrunk into nanometer technology nodes. Dual stress liners applied compressive stress liners on pMOS and tensile stress liners on nMOS to boost the mobility of holes in the p-channel and electrons in the n-channel, respectively.

Because stress liners place stress mainly on the gate electrode and not directly on the channel, their effectiveness on channel strain is limited. As devices continue to shrink, fabs have found that by using selective epitaxial SiGe, they can generate higher compressive strain on the pMOS channel. Figure 4.31 shows a CMOS cross section with hybrid strain techniques, in which a pMOS channel compressive strain is achieved by SEG SiGe, and an nMOS channel tensile strain is achieved by a tensile stress liner. Recessed nMOS S/G allows for more effective channel strain from the stress liner.

Figure 4.32 illustrates the cross section of a state-of-the-art 22-nm high-κ metal gate CMOS with an SEG SiGe pMOS S/G that strains the p-channel compressively, and an SEG SiC nMOS S/G that strains the n-channel with tensile stress.

4.7 Summary

- Silicon is an inexpensive semiconductor material, and silicon dioxide is a strong and stable dielectric that can be easily grown on silicon surfaces in a thermal process.

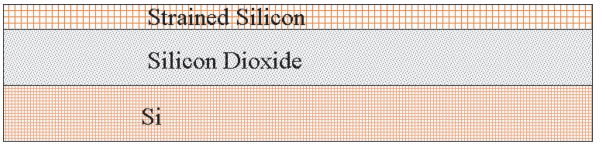


Figure 4.30 Strained SSOI.

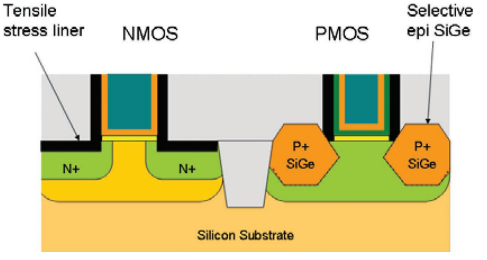


Figure 4.31 Cross section of a CMOS device with selective epitaxial SiGe for pMOS and tensile stress liner for nMOS.

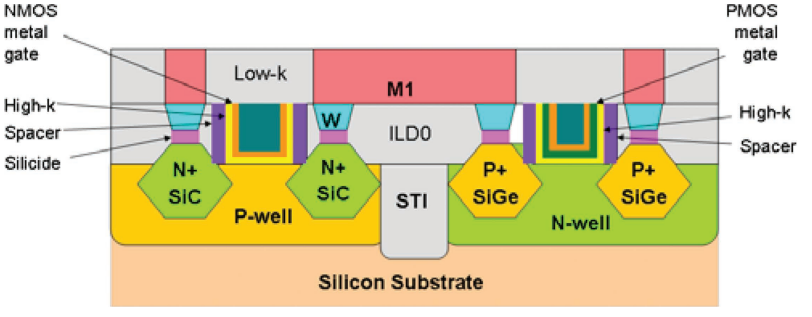


Figure 4.32 Cross section of a CMOS device with SEG SiGe for pMOS and SEG SiC for nMOS.

- The most commonly used silicon crystal orientations are ⟨100⟩ and ⟨111⟩.
- The wafer fabrication process converts sand (silicon dioxide) to MGS, MGS to TCS, TCS to EGS, EGS to single-crystal silicon ingots, and ingots to wafers.
- Both the CZ and floating zone methods are used for wafer manufacturing, with the CZ method being more common.
- The CZ method is cheaper and can produce larger-sized wafers.
- The floating zone method can produce wafers with higher purity.
- The epitaxial silicon layer is needed for bipolar devices and improves performance for CMOS and DRAM.
- The IC industry uses high-temperature CVD processes to grow epitaxial silicon layers.

- Most SOI wafers are made by bonded SOI technology.
- HOT uses SOI wafers and selective epitaxial growth. It allows IC manufacturers to make pMOS on ⟨110⟩ substrate and nMOS on ⟨100⟩ substrate to improve CMOS IC performance.
- Strained silicon techniques are widely applied in IC devices in the nanometer technology node.

4.8 Bibliography

M. S. Bawa, E. F. Petro, and H. M. Grimes, "Fracture strength of large-diameter silicon wafers," *SI Mag.* **18**(11), 115–118 (1995).

C. Y. Chang and S. M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

F. C. Eversteyn, "Chemical-reaction engineering in the semiconductor industry," *Philips Res. Rep.* **29**, 45–66 (1974).

L. Shon-Roy, A. Wiesnoski, and R. Zorich, *Advanced Semiconductor Fabrication Handbook*, Integrated Circuit Engineering Corp., Scottsdale, AZ (1998).

S. M. Sze, *VLSI Technology, Second Edition*, McGraw-Hill, New York (1988).

M. Yang et al., "High-performance CMOS fabricated on hybrid substrate with different crystal orientations," *IEDM Tech. Dig.*, 453–456 (2003).

See http://www.webelements.com for further information on the elements in this chapter.

4.9 Review Questions

1. Why are single-crystal materials needed for IC chip manufacturing?
2. Draw the ⟨100⟩ and ⟨111⟩ planes from a cube.
3. What are the main reasons silicon wafers are far more popular than any other semiconductor wafers in the IC industry?
4. What is the chemical used to purify MSG to ESG? List its safety hazards.
5. How does the CZ puller work? Why do wafers made from the CZ puller have higher oxygen concentration than wafers manufactured with the floating zone method?
6. Explain the purpose of the epitaxy process.
7. What is the autodoping effect? How can it be avoided?
8. List three epitaxial silicon precursors.
9. List the three dopants commonly used for epitaxy doping. What are the safety hazards of these dopant gases?
10. What are the advantages of the single-wafer epitaxy reactor over batch epitaxy systems?
11. What kind of ion implantation is required to make a bonded SOI wafer? What kind of ion implantation is required to make a SIMOX SOI wafer?

12. Explain why most IC manufacturers use local strain instead of strained silicon wafers to make MOSFETs.
13. Explain why more IC manufacturers use bulk silicon wafers and local strain techniques to make advanced IC chips instead of using HOT.

Chapter 5
Thermal Processes

One of the advantages of the silicon substrate over other semiconductor substrates is the high-temperature processing capability of silicon. Silicon wafer processing involves many high-temperature (700 to 1200 °C) procedures. This chapter covers these high-temperature thermal processes, including diffusion, oxidation, deposition, and annealing, in both standard furnace and rapid thermal processing.

Objectives

After finishing this chapter, the reader will be able to:

- list at least three important thermal processes
- describe the basic systems of vertical and horizontal furnaces, and list the advantages of a vertical furnace
- describe the oxidation process
- explain the importance of preoxidation cleaning
- identify the difference between dry and wet oxidation processes and applications
- describe the diffusion process
- explain why ion implantation replaced diffusion for silicon doping
- name at least three high-temperature deposition processes
- explain the reasons for postimplantation annealing
- describe the advantages of rapid thermal processing.

5.1 Introduction

The natural oxide of silicon, silicon dioxide, is a very stable and strong dielectric material that can be easily formed in a high-temperature process. This is one of the important reasons why silicon dominates the IC industry as a semiconductor substrate material. IC fabrication processing starts with an oxidation process, which grows a layer of silicon dioxide to protect the silicon surface. Silicon wafers then go through thermal processing in high-temperature furnaces and using rapid thermal processing (RTP) tools many times during the IC processing flow. Figure 5.1 shows the IC fabrication process flow.

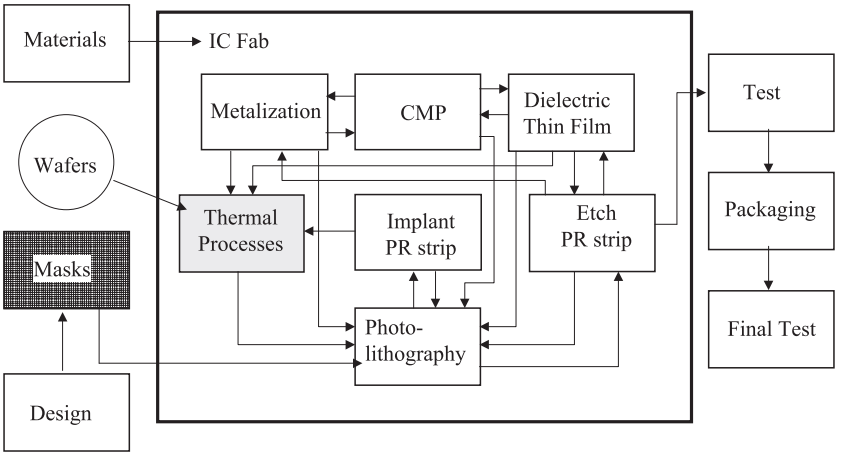


Figure 5.1 IC fabrication process flow.

5.2 Thermal Process Hardware

5.2.1 Introduction

Thermal processes occur in a high-temperature furnace called a diffusion furnace, because it was extensively used in the diffusion doping process in the early years of the semiconductor industry. There are two kinds of furnaces, horizontal and vertical, determined by the way the quartz tubes and heating elements are placed in the system. Typical requirements for a furnace include uniformity, accurate temperature control, low particulate contamination, high productivity, high reliability, and low cost.

A furnace consists of five basic components: a control system, process tube, gas delivery system, exhaust system, and loading system. The furnace used for low-pressure CVD (LPCVD) processing also requires a vacuum system. Figure 5.2 illustrates a schematic for a horizontal furnace.

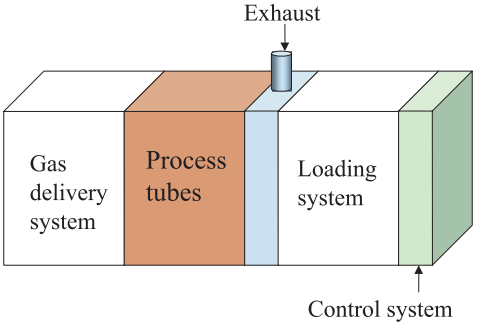


Figure 5.2 Layout of a horizontal furnace.

A vertical furnace has many advantages over a horizontal furnace, such as lower particle contamination, the ability to handle a large number of heavy wafers, better uniformity, lower maintenance cost, and a smaller footprint. It is very important for a processing tool to have a small footprint, since high-grade cleanroom space is very expensive in advanced fabs. Also, because wafers are placed in vertical stacks, large particles fall only onto the top wafer and do not reach the wafers underneath. Vertical furnaces are more commonly used in advanced semiconductor fabs.

Most of the parts in a thermal process furnace, such as the wafer boat, paddle, and wafer tower, are made of fused quartz. Quartz is a single-crystal silicon dioxide, which is a very stable material even at high temperatures. Its drawbacks are fragility and metallic impurities. Because quartz is not a sodium barrier, a trace amount of sodium can always penetrate the tube and cause damage to devices on the wafer. When the temperature is higher than 1200 °C, small flakes can occur and cause particle contamination.

SiC is another material used for high-temperature furnaces. Compared with quartz, SiC has a higher thermal stability and is a better mobile-ion barrier. The disadvantages of SiC are that it is heavier and more expensive. As device dimensions are further reduced, more SiC parts will be used in furnaces to meet processing requirements.

5.2.2 Control system

The controller consists of a computer connected to several microcontrollers. Each microcontroller connects to another system control interface board, which controls processing sequences such as wafer loading and unloading, processing times for each step, processing temperatures and temperature ramp rates, processing gas flow rates, exhaust, etc. It also collects and analyzes processing data, programs the processing sequence, and tracks the lot numbers. Figure 5.3 shows a block diagram of a furnace control system.

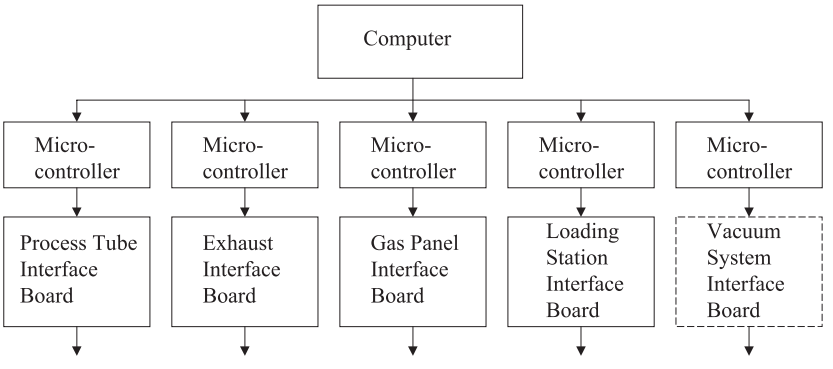


Figure 5.3 Block diagram of the control system of a furnace.

5.2.3 Gas delivery system

The gas delivery system handles processing gases and delivers them to the processing tube as required. A gas panel consists of regulators, valves, mass flow controllers (MFCs), and filters; it distributes processing and purge gases to the processing tube as needed. Processing gases are usually stored in high-pressure (more than 100 psi) bottles in remote gas cabinets. Processing gases are throttled to the gas panel by dozens of pounds per square inch (too high to be directly used in the process). Regulators and valves monitor the processing gas pressure and control pressure. Gas flow rate is precisely controlled by the MFC, which controls the rate by adjusting an internal control valve so that the measured flow rate equals the set value. Filters in the gas panel help minimize particulate contamination by preventing particles from flowing into the processing tube with the gases. Figure 5.4 shows a schematic of a gas delivery system.

5.2.4 Loading system

The loading station is used for wafer loading, unloading, and temporary storage. It loads wafers from cassettes into quartz boats, which sit on a quartz paddle; then the paddle is gently pushed into the processing tube by a loading mechanism, which is controlled by the microcontroller. After thermal processing, the paddle is gradually pulled out of the processing tube.

Several paddle types have been used in the loading systems of horizontal furnaces. Paddles with wheels are no longer used due to the particle generation caused by direct contact between the paddle and quartz tube. Soft-landing paddles, which gradually land on the quartz tube after the paddle is pushed into the desired position inside the processing tube, were developed next. Due to direct contact between the paddle and quartz tube, soft-landing paddles caused particulate contamination from friction between the paddle and tube wall during paddle landing and raising. To avoid this direct contact, suspended or cantilevered paddles are currently used in most systems. They allow the paddle to move in and out of the processing tube without any direct contact with the quartz surface, minimizing particle contamination. However, wafer loading can affect the paddle's suspension.

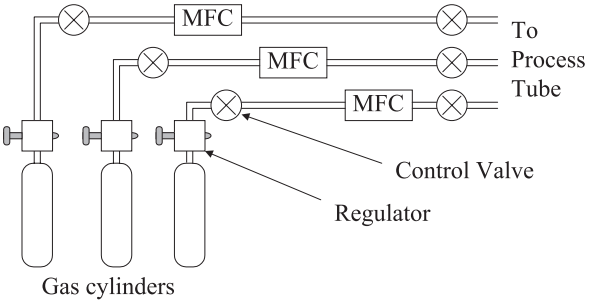


Figure 5.4 Schematic of the gas delivery system.

For a vertical furnace, wafers are loaded from the cassette into the tower by a robot. Then a lift mechanism (controlled by the microcontroller) raises the tower into the processing chamber, or the chamber elevator lowers the chamber until the entire tower is inside the chamber. This eliminates contact between the wafer holder and the quartz chamber wall, and there is no suspension problem for the wafer holder.

5.2.5 Exhaust system

Process byproducts and unused source gases are removed from the processing tube or chamber by an exhaust system. Exhaust gas is evacuated from the processing tube symmetrically, and purge gas is drawn in from the exhaust manifold to prevent backstreaming of the exhaust gas. For furnace processes involving pyrophoric or flammable gases, such as silane (SiH_4) and hydrogen (H_2), an additional chamber called a burn box is required. In the burn box, exhaust gases pass through a controlled combusting process in the presence of oxygen and are reduced to less harmful, less reactive oxide compounds. Filters are used to remove particles generated in the burning process, such as silicon dioxide, after burning silane. A scrubber with water or an aqueous solution absorbs most of the toxic and corrosive gases before the exhaust gas is vented to the atmosphere.

5.2.6 Processing tube

Wafers undergo a high-temperature procedure in the processing tube, which consists of a quartz tube chamber body and several heaters. Thermocouples touch the chamber wall and monitor chamber temperature. Each heater is independently supplied by a high-current power source. The power of each heater is controlled by feedback from the thermocouple data through the interface board and microcontroller and becomes stable when the setting temperature is reached. Temperature in the flat zone at the center of the tube is precisely controlled within 0.5 °C at 1000 °C. Figure 5.5 shows schematics of horizontal and vertical furnaces.

In a horizontal furnace, wafers are placed on quartz boats, which sit on a paddle made of SiC. The paddle (with wafer boats) is then slowly pushed into the quartz tube where it places the wafers in the flat zone of the furnace for thermal processing. After processing, the wafers must be pulled out very slowly to avoid warping caused by the overwhelming thermal stress of sudden temperature change.

In a vertical furnace, wafers are loaded into a wafer tower made of quartz or SiC. The wafers are placed in the tower face up, then the tower is slowly raised into the quartz tube to heat. Afterward, the tower is lowered slowly to avoid wafer warping.

5.3 Oxidation

The oxidation process is one of the most important thermal processes. It is an adding process that adds oxygen to a silicon wafer to form silicon dioxide on the wafer surface. Silicon is very reactive to oxygen; thus in nature, most silicon

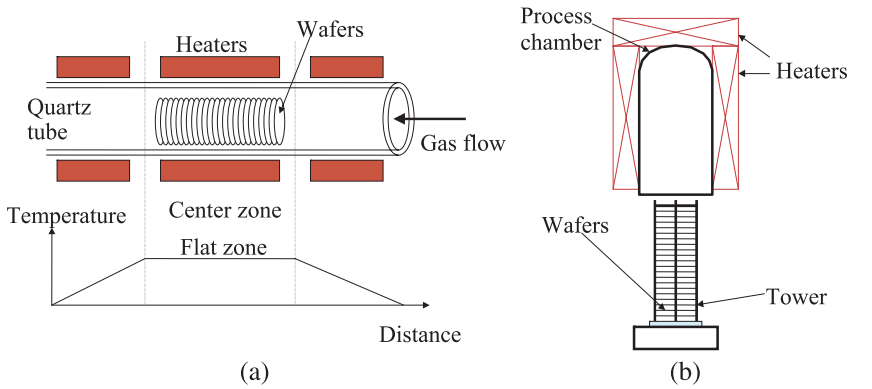


Figure 5.5 Schematics of (a) horizontal and (b) vertical furnaces.

exists in the form of silicon dioxide, such as quartz sand. It reacts with oxygen very quickly, and forms silicon dioxide on the silicon surface. The reaction can be expressed as

$$Si + O_2 \rightarrow SiO_2.$$

Silicon dioxide is a dense material that fully covers the silicon surface. To continue the oxidization of silicon, oxygen molecules have to diffuse across the oxide layer to reach the silicon atoms underneath and react with them. The growing silicon dioxide layer increasingly blocks and slows the oxygen. When bare silicon is exposed to the atmosphere, it reacts almost immediately with oxygen or moisture in the air and forms a thin layer (about 10 to 20 Å) of silicon dioxide, called native oxide. The layer of the native oxide is thick enough to stop further oxidation of the silicon at room temperature. Figure 5.6 illustrates the oxidation process.

In the oxidation process, oxygen comes from the gas phase, and silicon comes from the solid substrate. Therefore, while silicon dioxide is growing, it consumes the substrate silicon, and the film grows into the remaining silicon substrate, as shown in Fig. 5.6. Oxygen is used as an oxidizer in processes such as thermal oxidation, CVD, and reactive sputtering deposition. It is also commonly used in etch and photoresist stripping processes. Oxygen is the most abundant element in the earth's crust and the second most abundant element in the earth's atmosphere, after nitrogen. Some facts about oxygen are listed in Table 5.1.

At high temperatures, thermal energy causes oxygen molecules to move much faster, which can drive them to diffuse across an existing oxide layer and react with silicon to form more silicon dioxide. The higher the temperature, the faster the oxygen molecules, and the quicker the oxide film growth. The oxide film quality is also better than that grown at lower temperatures. Therefore, to get high-quality oxide film and fast growth rate, oxidation processes are always performed in a high-temperature environment, normally in a quartz furnace. Oxidation is a slow

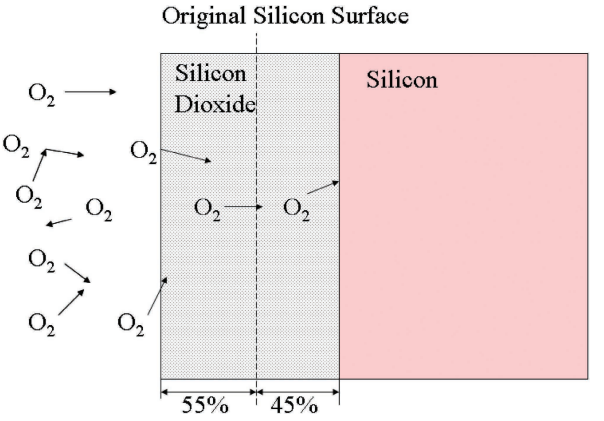


Figure 5.6 Silicon oxidation process.

Table 5.1 Facts about oxygen.

Symbol	O
Atomic number	8
Atomic weight	15.9994
Discoverer	Joseph Priestley, Carl Scheele
Discovery location	England, Sweden
Discovery date	1774
Origin of name	From the Greek words oxy genes, meaning acid (sharp) and forming (acid former)
Molar volume	17.36 cm^3
Velocity of sound	317.5 m/sec
Refractivity	1.000271
Melting point	54.8 K = −218.35°C
Boiling point	90.2 K = −182.95°C
Thermal conductivity	0.02658 W m^{-1} K^{-1}
Applications	Thermal oxidation, oxide CVD, reactive sputtering, and photoresist stripping
Main sources	O_2, N_2O, O_3

process; even in furnaces hotter than 1000 °C, a thick oxide (>5000 Å) still takes several hours to grow. Therefore, oxidation processes usually are batch processes, with a large number (100 to 200) of wafers processing at the same time to achieve reasonable throughput.

5.3.1 Applications

Oxidation of silicon is one of the basic processes throughout the IC process. There are many applications for silicon dioxide; one of them is as diffusion mask. Most dopant atoms used in the semiconductor industry such as boron and phosphorus have much lower diffusion rates in silicon dioxide than they do in single-crystal silicon. Therefore, by etching windows on the masking oxide layer, a silicon

substrate can be doped at a designated area by a dopant diffusion process, as shown in Fig. 5.7. The thickness of the masking oxide is about 5000 Å.

Screen oxides are also commonly used for ion implantation processes. They can help prevent silicon contamination by blocking the sputtered photoresist. They can also minimize the channeling effect by scattering incident ions before they enter a single-crystal silicon substrate. The thickness of a screen oxide is about 100 to 200 Å. Figure 5.8 illustrates a screen oxide for application of ion implantation.

Thermally grown silicon dioxide is used as a pad layer for silicon nitride in both local oxidation of silicon (LOCOS) and shallow trench isolation (STI) formation. Without the stress buffer from this pad oxide, LPCVD silicon nitride film could crack and in some cases even break a silicon wafer due to the high tensile stress (up to 10^{10} dynes/cm^2). The thickness of the pad oxide is about 150 Å.

Silicon dioxide is also used as a barrier layer to prevent contamination of the silicon substrate before trench fill in the STI process. Trench fill is a dielectric CVD process in which undoped silicate glass (USG) is deposited to fill the trench for electrical isolation of neighboring transistors. Since the CVD process always brings small amounts of impurities, a dense, thermally grown silicon dioxide barrier layer is necessary to block possible contamination. Figure 5.9 shows the pad and barrier oxides in the STI process.

One of most important applications of thermally grown silicon dioxide in the past was forming isolation blocks to electrically isolate neighboring transistors in an IC chip. Blanket field oxide and LOCOS were two kinds of blocks used

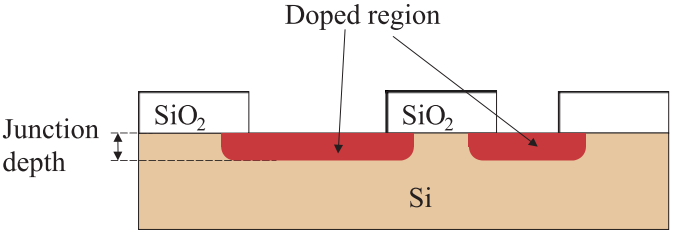


Figure 5.7 Diffusion masking oxide.

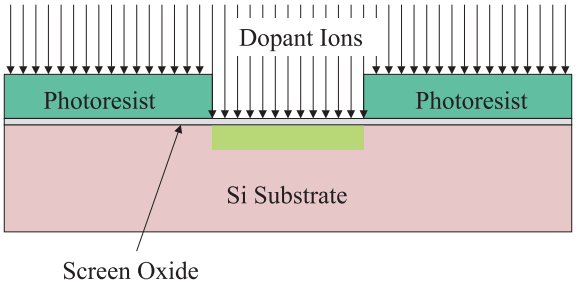


Figure 5.8 Screen oxide for application of ion implantation.

to isolate neighboring devices and prevent them from crosstalking. Blanket field oxide was the simplest isolation process and widely used in the early years of silicon IC manufacturing. By thermally growing a thick layer of silicon dioxide (5,000 to 10,000 Å), patterning it with photolithography, and etching the oxide with hydrofluoric acid (HF), the activation areas could be opened for transistor making, as shown in Fig. 5.10.

LOCOS has better isolation properties than blanket-field oxide. The LOCOS process uses a thin layer of oxide (200 to 500 Å) as the pad layer to buffer the strong tensile stress of LPCVD nitride. After the nitride etch, photoresist strip, and wafer cleaning, a thick layer of oxide (3000 to 5000 Å) is grown on the area not covered by silicon nitride. Silicon nitride is a much better barrier layer than silicon dioxide. Oxygen molecules cannot diffuse across the nitride layer; therefore, the silicon underneath the nitride layer does not oxidize. In the area not covered by the nitride, oxygen molecules continuously diffuse across the silicon dioxide layer,

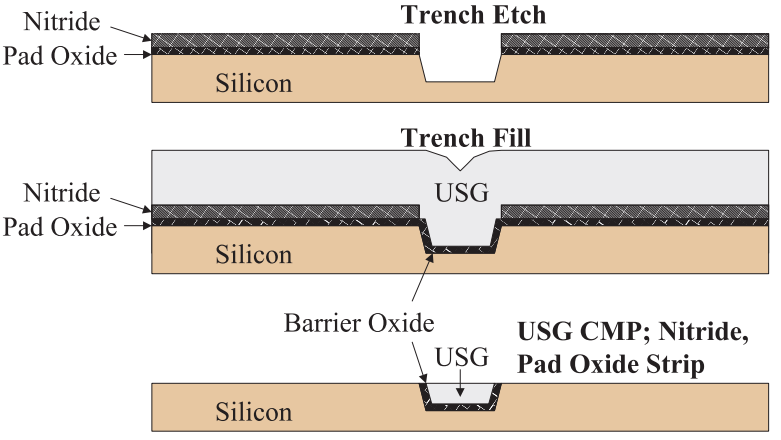


Figure 5.9 Pad oxide and barrier oxide in the STI process.

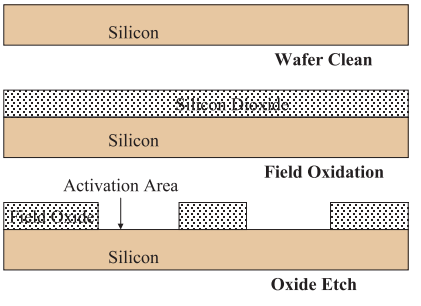


Figure 5.10 Blanket field oxide process.

where they react with silicon underneath to form more silicon dioxide. The LOCOS formation process is illustrated in Fig. 5.11.

Because oxygen diffusion inside silicon dioxide is an isotropic process, oxygen also reaches silicon at the side. This causes the oxide to grow underneath the nitride layer near the etched oxidation window, forming the so-called bird's beak (see Fig. 5.11). An undesirable consequence is that this bird's beak uses up a good deal of space on the wafer surface. Another disadvantage of LOCOS is the surface planarization problem due to the oxide-to-silicon surface step, caused by the oxide's growing characteristics, illustrated in Fig. 5.6.

Many different methods have been used to suppress the bird's beak. The most commonly used method is poly-buffered LOCOS (PBL). A thicker pad oxide allows a longer bird's beak to grow because of the wide oxide path for oxygen diffusion. By using a polysilicon layer (~ 500 Å) to buffer the high tensile stress of the LPCVD silicon nitride, the thickness of the pad oxide layer can be reduced from 50 to 100 Å, significantly reducing oxide encroachment. But, no matter how many different methods were tried, there was still an approximately 0.1- to 0.2-μm bird's beak at both sides of the LOCOS, a thickness that is intolerable when the minimum feature size is smaller than 0.35 μm. The STI process has been developed to avoid the bird's beak problem and has a more planarized surface topology as well. STI gradually replaced LOCOS isolation starting in the mid-1990s when device feature size was smaller than 0.35 μm.

Sacrificial oxide is a thin layer (< 1000 Å) of silicon dioxide grown on activation areas of a silicon surface and stripped in HF solution right after its growth. It is applied before the gate oxidation process to remove damages and defects on the silicon surface. This oxide growth and removal processing sequence helps to create a defect-free silicon substrate surface for growing a high-quality gate oxide layer.

The thinnest and most important silicon dioxide layer in a MOSFET-based IC chip is the gate oxide. While device dimensions have been shrinking, the thickness

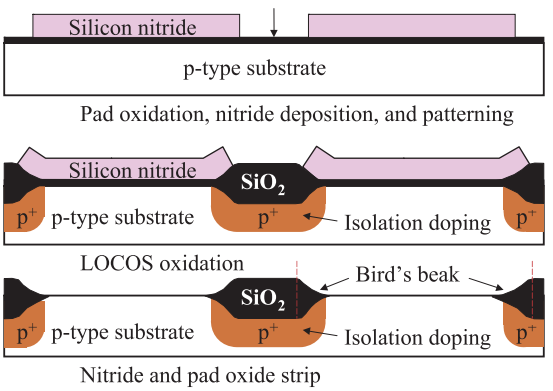


Figure 5.11 LOCOS process.

of gate oxides has been reduced from more than 1000 Å in the 1960s to about 15 Å in high-end chips in the mid-2000s; operation voltage of IC chips has been reduced from 12 to 1.0 V. The quality of gate oxide is vital for proper device function. Any defect, impurity, or particle contamination in a gate oxide can affect device performance and significantly reduce chip yield. Figure 5.12 illustrates the sacrificial oxidation and gate oxidation processes. Applications of thermally grown silicon dioxide in IC fabrication processes are summarized in Table 5.2. For nanometer technology node IC chips, the gate oxide is nitrided to increase its dielectric constant, thus it can increase in thickness with a higher breakdown voltage, while still possessing larger gate capacitance to maintain proper MOSFET switching properties. Also, a nitrogen-rich layer can block the diffusion of boron in heavily p-type doped polysilicon gate electrons into the n-type doped channel of pMOS.

5.3.2 Preoxidation cleaning

Thermally grown silicon dioxide is an amorphous material. It is unstable, and its molecules tend to cross-link to form crystalline structures. This is the main reason silicon dioxide exists in the form of quartz and quartz sand in nature. Since the crystallization of amorphous silicon dioxide takes millions of years at room temperature, amorphous silicon dioxide in IC chips is very stable in their lifetimes. However, the crystallization process is dramatically accelerated at the high temperatures (>1000 °C) required during silicon dioxide growth. If the silicon surface is not free of contaminants, defects and particles can serve as nucleation sites for crystallization during the oxidation process, and silicon dioxide will grow into a polycrystalline structure similar to ice crystals that forms on glass in the winter. Crystallization of silicon dioxide is very undesirable, since it is not uniform and the crystal boundaries provide easy paths for impurities and moisture.

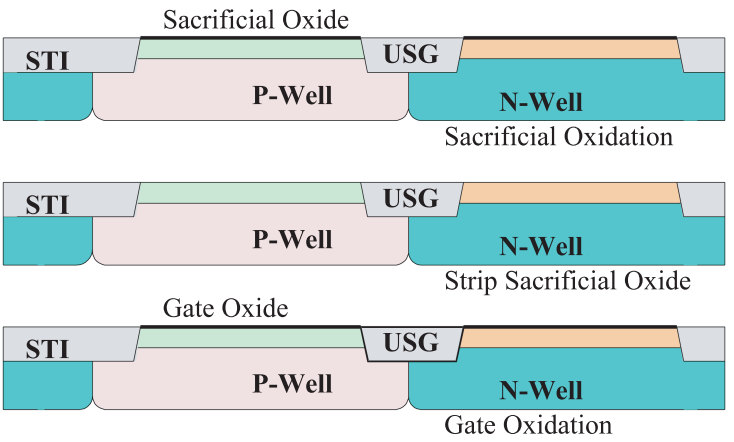


Figure 5.12 Sacrificial and gate oxidation processes.

Table 5.2 Summary of oxide applications.

Oxide name	Thickness	Application	Time of application
Native	15–20 Å	Undesirable	-
Screen	~200 Å	Implantation	Mid-1970s to present
Masking	~5000 Å	Diffusion	1960s to mid-1970s
Field and LOCOS	3000–5000 Å	Isolation	1960s to 1990s
Pad	100–200 Å	Nitride stress buffer	1960s to present
Sacrificial	<1000 Å	Defect removal	1970s to present
Gate	15–120 Å	Gate dielectric	1960s to present
Barrier	100–200 Å	STI	1980s to present

Therefore, proper preoxidation wafer cleaning—removing the particles, organic and inorganic contaminants, native oxide, and surface defects—becomes very important to eliminate crystallization. Figure 5.13 shows a crystalline structure of silicon dioxide grown on a poorly prepared silicon surface.

Wet clean processing is the most commonly used cleaning process in advanced semiconductor fabs. Strong oxidizers, such as the solutions $H_2SO_4 : H_2O_2 : H_2O$ or $NH_4OH : H_2O_2 : H_2O$, can remove particles and organic contaminants. When a wafer submerges into these solutions, particles and organic contaminants oxidize,

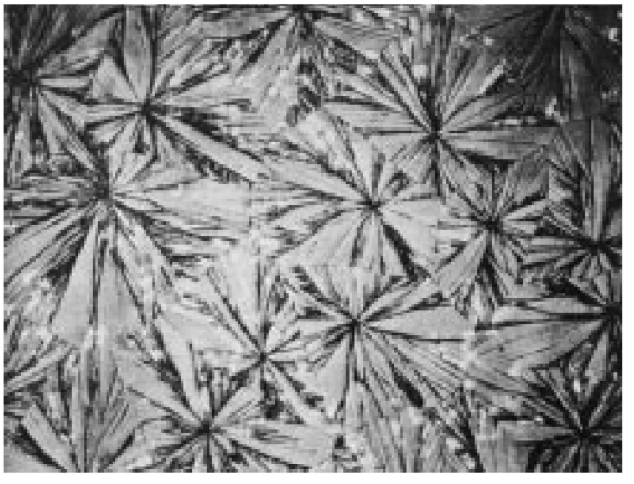


Figure 5.13 Oxide structure on a poorly prepared surface (Integrated Circuit Engineering Corporation).

and oxidation byproducts are either gaseous (CO, etc.) or soluble in the solution (such as H_2O). In most IC fabs, the composition of $NH_4OH : H_2O_2 : H_2O$ with a 1:1:5 to 1:2:7 ratio at 70 to 80 °C is widely used. This cleaning process is known as standard cleaning 1 (SC-1) of the Radio Corporation of America (RCA) clean, which was first developed by Kern and Puotinen (from RCA) in 1960. After SC-1, wafers are rinsed by DI water in a dunk tank and dried in a spin dryer.

After SC-1 and DI water rinse and dry, wafers are submerged in a solution with a 1:1:6 to 1:2:8 composition ratio of HCl: H_2O_2 : H_2O at 70 to 80 °C. This is called SC-2 of the RCA clean, which is commonly used to remove inorganic contaminants by forming soluble byproducts in a low-pH solution. In the SC-2 process, H_2O_2 oxidizes the inorganic contaminants, and HCl reacts with the oxides to form soluble chlorides, allowing desorption of contaminants from the wafer surface. SC-2 is also followed by a DI water rinse and spin dry processes.

Native oxides occurring in silicon have poor quality and need to be stripped, especially for gate oxides, which require the highest quality. HF is a chemical that is commonly used to dissolve native silicon dioxide. A native oxide strip can be performed either on a wet bench with $HF : H_2O$ solution, or in a HF vapor etcher, which uses HF vapor to react with silicon dioxide and vaporize the byproduct. After native oxide stripping, some fluorine atoms bind with silicon atoms and form Si-F bonds on the silicon surface.

5.3.3 Oxidation rate

When oxygen starts to react with silicon, it forms a silicon dioxide layer, which prevents silicon atoms from reacting with oxygen molecules. When the oxide first begins to grow and the oxide layer is very thin (< 500 Å), oxygen molecules can penetrate the oxide with very few collisions in the oxide layer and reach the silicon, where they react and continue to grow the silicon dioxide film. This is called a linear growth regime because the oxide thickness linearly increases with the growth time. When the oxide film thickness becomes thicker, oxygen molecules cannot pass through the oxide layer without numerous collisions with the atoms inside the oxide film. They must diffuse across the growing oxide to reach the silicon and react with it to form silicon dioxide. This is called the diffusion-limited regime because the oxide growth rate is slower than the linear growth regime. Figure 5.14 illustrates the two oxide growth regimes.

A and B in the equations in Fig. 5.14 are the two coefficients related to oxide growth rates; they are determined by many factors, such as oxidation temperature, oxygen source (O_2 or H_2O), silicon crystal orientation, dopant type, concentration, and pressure.

Oxide growth rates are very sensitive to temperature because the oxygen diffusion rate in silicon dioxide is exponentially related to temperature: $D \propto \exp(-E_a/k_BT)$. Here D is the diffusion coefficient, E_a is the activation energy, $k_B = 2.38 \times 10^{-23}$ J/K is the Boltzmann constant, and T is the temperature. Increasing the temperature can significantly increase both B and B/A, as well as the oxide growth rate.

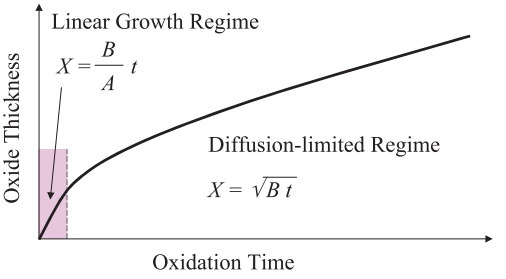


Figure 5.14 The two oxidation rate regimes.

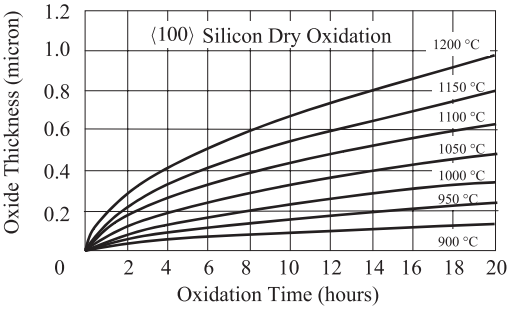


Figure 5.15 Dry oxide growth on a ⟨100⟩ surface.

Oxide growth rate is also related to the oxygen source: dry oxidation with O_2 has a lower oxide growth rate than wet oxidation with H_2O. This is because the diffusion rate of the oxygen molecule O_2 in silicon dioxide is lower than that of the hydroxide HO generated from the dissociation of H_2O molecules at high temperatures. Oxide growth rates of dry and wet oxidation processes are illustrated in Figs. 5.15 and 5.16.

The figures show that wet oxidation is significantly faster than dry oxidation. For example, for ⟨100⟩ silicon at 1000 °C, the wet oxide layer grows to ~ 2.2 μm after 20 h, while the dry oxide layer only grows to 0.34 μm. This is why the wet oxidation process is preferred for growing thick oxide layers such as masking and field oxides.

Oxide growth rates are also related to single-crystal silicon orientation. Normally, ⟨111⟩ orientation silicon has a higher oxide growth rate than ⟨100⟩ orientation silicon. This is because a ⟨111⟩ silicon surface has higher silicon atom density than that of a ⟨100⟩ silicon surface and can provide more silicon atoms to react with oxygen and form thicker silicon dioxide layers. By comparing Fig. 5.17, which shows the wet oxidation rate of ⟨111⟩ silicon, with Fig. 5.16, which shows the wet oxidation rate of ⟨100⟩ silicon, it is evident that ⟨111⟩ silicon has a higher growth rate than ⟨100⟩ silicon.

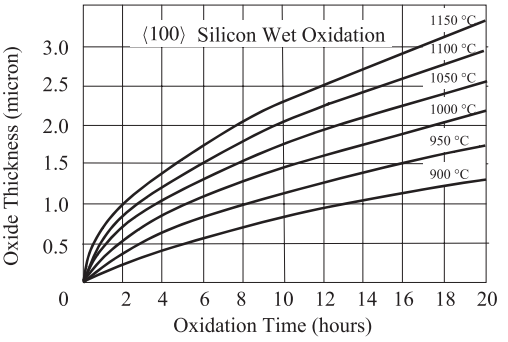


Figure 5.16 Wet oxide growth on a ⟨100⟩ surface.

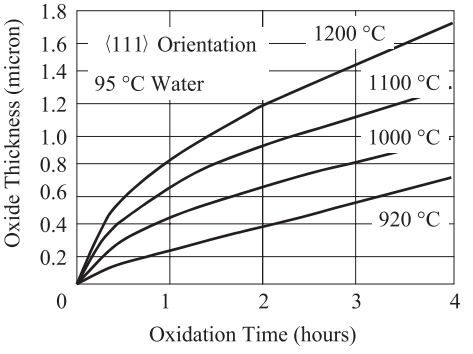


Figure 5.17 Wet oxide growth on a ⟨111⟩ surface.

Oxidation rate is related to dopant and dopant concentrations. Heavily doped silicon generally oxidizes faster than lightly doped silicon. During oxidation, boron in the silicon tends to be drawn up into the silicon dioxide, which causes depletion of the boron concentration at the Si–SiO_2 interface. n-type dopants such as phosphorus, arsenic, and antimony have the opposite effect. When oxidation occurs, these dopants are driven deeper into the silicon. Similar to snow piling up from a snowplow, the n-type dopant concentration in the Si–SiO_2 interface can become significantly higher than its original value. Figure 5.18 illustrates the pile-up effect of an n-type dopant and the depletion effect of a p-type dopant.

Oxidation rate is also related to the addition of gas, such as HCl, that is commonly used in the gate oxidation process to suppress mobile ions. With the appearance of HCl, the oxidation rate can increase by about 10%.

5.3.4 Dry oxidation

Dry oxidation has a lower growth rate than wet oxidation, yet it has better oxide film quality. Therefore, thin oxides such as screen, pad, and especially gate oxides, use a dry oxidation process. Figure 5.19 shows a dry oxidation system.

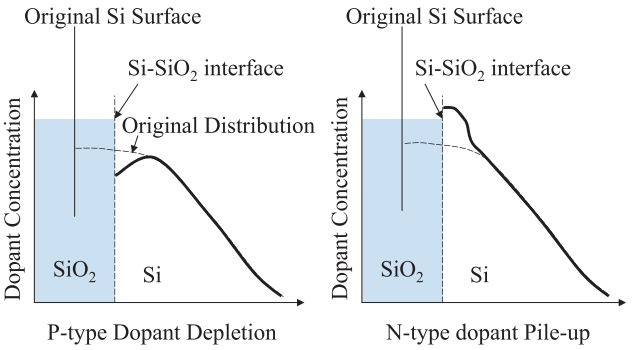


Figure 5.18 Dopant depletion and pile-up effects caused by oxidation.

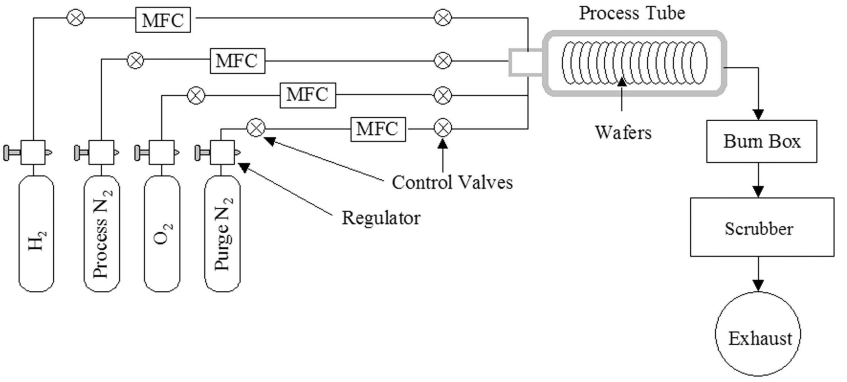


Figure 5.19 Schematic of a dry oxidation system.

There are two nitrogen (N_2) sources in most oxidation systems: one for the process application (with higher purity), and the other for the chamber purge (with lower purity and also lower cost). Because nitrogen is a very stable gas, it is used as an inert gas in the oxidation process during system idle time, wafer loading, temperature ramp, temperature stabilization, and wafer unloading steps. For dry oxidation, high-purity oxygen gas (O_2) is used as the oxygen source during the oxidization step to oxidize silicon. HCl is also used during the oxidation step to reduce mobile ions in the oxide and minimize interface state charge.

The quartz tube starts to sag with temperatures above 1150 °C, so oxidation processes cannot operate at those temperatures for very long. Dry oxidation processes operate at about 1000 °C. In dry oxidation, HCl is commonly used to remove mobile metallic ions, especially sodium, by forming immobile chloride compounds. This is very important, since a trace amount of sodium can cause MOS transistor malfunction and affect IC chip performance and reliability.

When silicon dioxide is forming on a single-crystal silicon surface, there is an abrupt change at the silicon–silicon-dioxide interface. There are always some dangling bonds at the interface due to crystal structure mismatch, as shown in

Fig. 5.20. The dangling bonds induce an interface state charge, which is a positive charge that strongly affects IC chip performance and reliability. This is because hydrogen or other atoms can diffuse to the silicon–silicon-dioxide interface during IC chip application and attach to the dangling bonds. This changes the interface state charge, the threshold voltage V_T of the MOS transistor, and the performance of the IC device. Although dangling bonds are always present at the silicon–silicon-dioxide interface, it is very important to minimize them as much as possible to achieve highly stable and reliable device performance.

Introducing HCl into the oxidation reaction means that some chlorine atoms can integrate into the silicon dioxide film and bind with silicon at the silicon-silicon-dioxide interface. This helps minimize the number of dangling bonds and improve IC device reliability. However, if chlorine concentration in the oxide is too high, device stability can be affected because of the response of the extra chlorine ions to the bias voltage.

Question: Fluorine has one unpaired electron and can attach to the dangling bonds at the Si-SiO_2 interface. Why don't fabs use HF in the oxidation process and integrate fluorine into the oxide to minimize the interface state charge?

Answer: HF can aggressively etch the silicon dioxide layer and the quartz tube, so it is undesirable to use HF in the thermal oxidation process. However, after BF_2 ion implantation, the integration of small amounts of fluorine atoms into the silicon substrate can help reduce the interface state charge.

A typical dry oxidation process sequence for gate oxide growth is as follows:

- idle with purge N_2 flow
- idle with process N_2 flow
- wafer boats pushed in with process N_2 flow
- temperature ramp-up with process N_2 flow

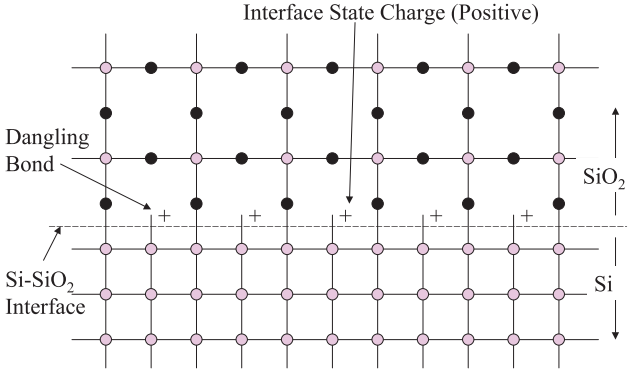


Figure 5.20 Example of dangling bonds inducing interface state charge.

- temperature stabilization with process N_2 flow
- oxidation with O_2, HCl; stop N_2 flow
- oxide annealing, stop O_2, start process flow N_2
- temperature cool down with process N_2 flow
- wafer boats pulled out with process N_2 flow
- idle with process N_2 flow
- bring in next boats and repeat process
- idle with purge N_2 flow.

During system idle time, the furnace is always kept at a high temperature, such as 850 °C, so that temperature ramp-up for processing does not require much time. N_2 purge gas is always used when the system is idle for long periods of time. Before wafer loading, process N_2 starts to flow into the processing tube in order to fill it with higher-purity nitrogen. It needs to slowly push the wafer boats into the processing tube to avoid wafer warping due to the high thermal stress induced by sudden temperature change. When the wafer boats are placed into the flat zone of the processing tube, temperature ramps up at a rate of 10 °C/min. Due to its high thermal capacity, the temperature of a furnace system cannot ramp up too quickly, as oscillations can occur due to over- or undershooting the temperature around the setting point. After the furnace reaches the set temperature required by the process (typically 1000 °C), a temperature-stabilization step with N_2 flow is required to allow for temperature oscillation to settle down and for the furnace to finally reach a steady state with the set temperature.

Now the system is ready for oxidation. By turning on O_2 and HCl flows and turning off the N_2 flow, oxygen begins to react with silicon and forms a thin layer of silicon dioxide on the silicon wafer surface. After the required oxide thickness is reached, the O_2 and HCl flows are terminated, and N_2 flow is resumed. The wafers stay at the high temperature for a while to anneal the oxide. This step improves the quality of the silicon dioxide, makes it denser, reduces the interface state, and increases breakdown voltage. A thin gate oxide (thickness of about 50 Å) can grow in a furnace at a lower temperature such as 700 °C, so that oxidation time will be long enough to control the process. After the oxide film grows, the film is annealed at >1000 °C in a nitrogen ambient to improve oxide quality. After oxide annealing, the furnace temperature is gradually cooled down to its idle temperature, and wafer boats are slowly pulled out from the furnace with a constant nitrogen flow.

5.3.5 Wet oxidation

When using H_2O instead of O_2 as the oxygen source, the silicon oxidation process is called wet oxidation, and the oxide is called steam oxide. The chemical reaction can be expressed as

$$2\,H_2O + Si \rightarrow SiO_2 + 2\,H_2.$$

At high temperatures, H_2O can dissociate and form hydroxide (HO), which can diffuse in silicon dioxide faster than O_2 can. Therefore, a wet oxidation process

has significantly higher oxidation rates than a dry oxidation process. Wet oxidation is used to grow thick oxides such as masking, blanket field, and LOCOS oxides. Table 5.3 shows an example of the oxidation time difference to grow a 1000-Å oxide film at 1000 °C.

Several systems have been used to deliver water vapor into the processing tube of a furnace. The boiler system vaporizes water at temperatures higher than 100 °C, and the water vapor flows to the processing tube by means of a heated gas line. In a bubbler system, nitrogen is bubbled through pure DI water, which carries the water vapor into the processing tube. Figure 5.21 gives the schematics of the boiler and bubbler systems.

In a flush system, small droplets of pure DI water are dropped onto a quartz hot plate to vaporize the water. Then an oxygen gas flow carries the steam into the processing tube. Figure 5.22 illustrates a flush system.

A major problem with boiler, bubbler, and flush systems is that they cannot precisely control the H_2O flow rate. The most commonly used wet oxidation system is the dryox system (or pyrogenic steam system), which burns hydrogen at the entrance of the processing tube, so that water vapor can be formed pyrogenically with the chemical reaction between H_2 and O_2:

$$2\,H_2 + O_2 \rightarrow 2\,H_2O.$$

The dryox system eliminates the liquid and vapor handling requirements, and can precisely control the flow rate. The tradeoff is the induction of flammable and explosive hydrogen gas. Figure 5.23 illustrates the schematic of the pyrogenic steam system.

Table 5.3 Time comparison of wet and dry oxidation processes.

Process	Temperature	Film thickness	Oxidation time
Dry oxidation	1000 °C	1000 Å	~ 2 h
Wet oxidation	1000 °C	1000 Å	~ 12 min

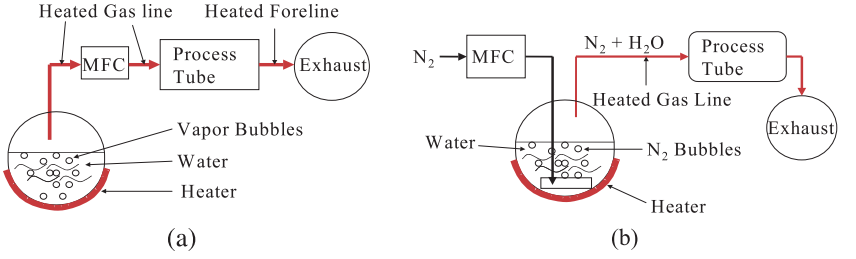


Figure 5.21 Schematics of (a) boiler and (b) bubbler system.

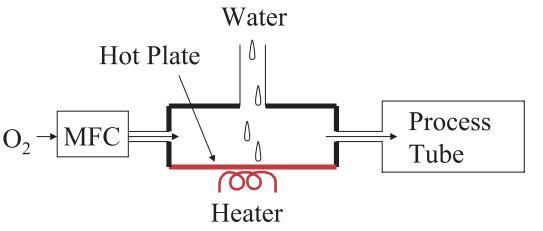


Figure 5.22 Schematics of a flush system.

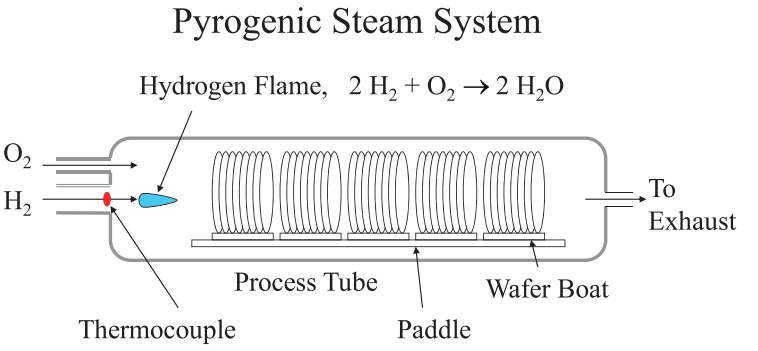


Figure 5.23 Schematics of a pyrogenic steam system.

A wet oxidation system with pyrogenic steam is illustrated in Fig. 5.24. A burn box is required in the exhaust system to burn out any residual hydrogen before releasing the exhaust gases into the atmosphere.

The autoignition temperature of hydrogen is about 400 °C. At a processing temperature with oxygen present, a flow of hydrogen gas into a processing tube will automatically react with the oxygen and form steam inside the tube. In a pyrogenic steam oxidation process, the ratio of the H_2 to O_2 flow rates is very important. The normal flow rate ratio of H_2:O_2 should be slightly lower than 2:1 to ensure the presence of excessive oxygen, which can fully oxidize hydrogen in the pyrogenic reaction. Otherwise, hydrogen can accumulate inside the tube, causing an explosion. The typical H_2:O_2 ratio is 1.8:1 to 1.9:1.

A normal pyrogenic wet oxidation process sequence is as follows:

- idle with purge N_2 flow
- idle with process N_2 flow
- ramp O_2 with process N_2 flow
- wafer boat pushed in with process N_2 and O_2 flows
- temperature ramp-up with process N_2 and O_2 flows
- temperature stabilization with process N_2 and O_2 flows
- ramp O_2, turn off N_2 flow
- stabilize O_2 flow
- turn on H_2 flow, ignition and H_2 flow stabilization

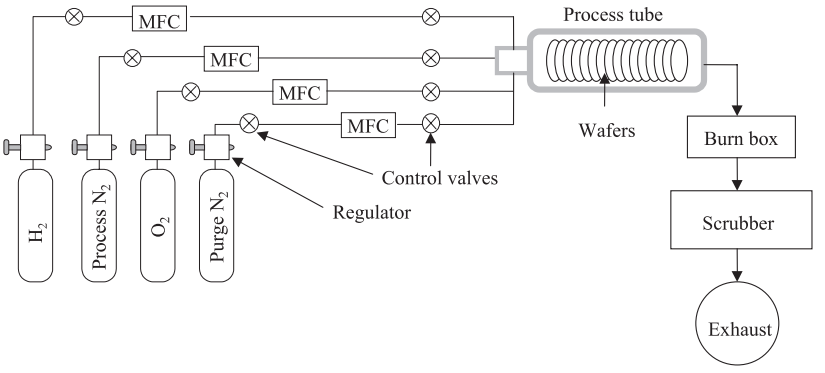


Figure 5.24 Schematics of a pyrogenic steam oxidation system.

- steam oxidation with O_2 and H_2 flow
- hydrogen termination, turn off H_2 while keeping flow
- oxygen termination, turn off O_2, start process N_2 flow
- temperature ramp-down with process N_2 flow
- wafer boat pulled out with process N_2 flow
- idle with process N_2 flow
- bring in next boats and repeat process
- idle with purge N_2 flow.

The appearance of oxygen during wafer push-in, temperature ramp-up, and temperature stabilization steps helps grow a thin layer (a few hundred angstroms) of better-quality dry oxide on the wafer surface. This serves as a barrier layer for the lesser-quality steam oxide and helps reduce defects at the silicon–silicon-dioxide surface. After steam oxidation, hydrogen flow is terminated, while oxygen continuously flows into the processing tube to scavenge any residual hydrogen. This oxygen flow also can help reduce hydrogen integration in the steam oxide.

5.3.6 High-pressure oxidation

Increasing pressure increases the oxygen or steam concentration inside the processing chamber as well as the diffusion rate inside the silicon dioxide, helping to increase the oxidation rate. High-pressure oxidation can reduce oxidation time at the same oxidation temperature, or reduce oxidation temperature at the same oxidation time. Generally, increasing pressure by one atmosphere can reduce oxidation temperature by 30 °C. Tables 5.4 and 5.5 show time and temperature reductions caused by using a high-pressure oxidation process to grow a 10,000 Å (1 μm) oxide film in wet oxidation processes.

The hardware required for high-pressure oxidation is different from that needed for other types of oxidation; Fig. 5.25 illustrates a high-pressure oxidation system. Due to hardware complexity and safety concerns, high-pressure oxidation processes are not very popular in advanced semiconductor fabs.

Table 5.4 Oxidation time to grow 10,000-Å wet oxide.

Temperature	Pressure	Time
1000°C	1 atm	5 h
	5 atm	1 h
	25 atm	12 min

Table 5.5 Oxidation temperature to grow 10,000-Å wet oxide in 5 h.

Time	Pressure	Temperature
5 h	1 atm	1000 °C
	10 atm	700 °C

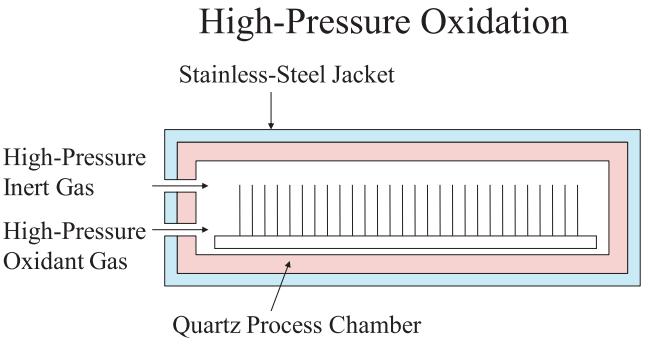


Figure 5.25 Schematics of a high-pressure oxidation system.

5.3.7 Oxide measurement

Monitoring the oxidation process means measuring oxide thickness and its uniformity. Ellipsometry is commonly used to measure dielectric thin-film refractivity and thickness. When a beam of light is reflected from a film's surface, the polarization state changes (see Fig. 5.26). By measuring this change, information about the film's refractive index and thickness can be acquired. An approximate value of film thickness is needed in advance, since the measured ellipsometric quantities are the periodic functions of the thickness. Because the refractive index of silicon dioxide is well known to be 1.46 for a light wavelength of 633 nm (red He-Ne laser), ellipsometry can also be used to measure the oxide film thickness.

After the oxide is grown, the color of the wafer surface changes. Color depends on film thickness, refractive index, and the angle of light. Reflected light from the oxide surface (light 1) and reflected light from the Si-SiO_2 interface (light 2) have the same frequencies, but with different phases, since light 2 travels a longer distance inside the oxide film, as shown in Fig. 5.27. The two reflected lights interfere with each other and cause both constructive and destructive interference at different wavelengths, since the refractive index is a function of wavelength. The color seen on the wafer surface is determined by constructive interference

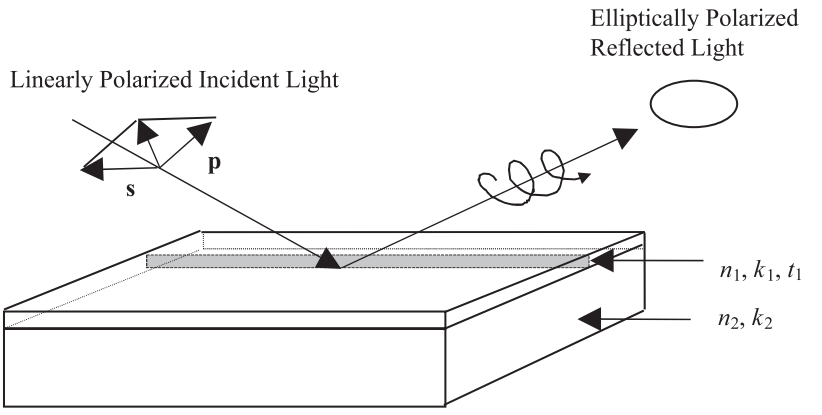


Figure 5.26 Ellipsometry system.

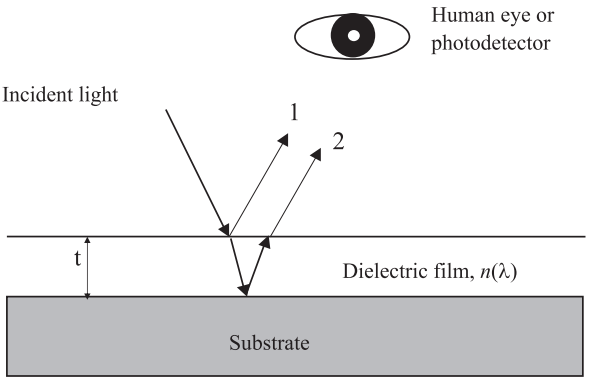


Figure 5.27 Reflected light and phase difference.

frequency,

$$\Delta\phi = 2tn(\lambda)/\cos\theta = 2N\pi,$$

where t is the thin-film thickness, $n(\lambda)$ is the thin-film refractive index, θ is the incident angle, and N is an integral number. When phase shift $\Delta\phi$ is larger than 2π, the color pattern repeats itself.

Color charts are a convenient tool to measure film thickness (see Table 5.6). Although these charts are no longer used for thickness measurements in advanced IC fabs, they are still a useful tool to estimate oxide thickness quickly and to detect obvious nonuniformity problems.

If a wafer with a thick oxide layer is placed into a HF solution, HF will etch silicon dioxide. By pulling back the wafer gradually, the oxide thickness changes gradually due to the different etch times, as does the color of the wafer. In this way,

Table 5.6 Color chart of silicon oxide.

Thickness (Å)	Color	Thickness (μm)	Color
500	Tan	1.0	Carnation pink
700	Brown	1.02	Violet-red
1000	Dark violet to red-violet	1.05	Red-violet
1200	Royal blue	1.06	Violet
1500	Light blue to metallic blue	1.07	Blue violet
1700	Metallic to very light yellow-green	1.10	Green
2000	Light gold or yellow, slightly metallic	1.11	Yellow-green
2200	Gold with slight yellow-orange	1.12	Green
2500	Orange to melon	1.18	Violet
2700	Red-violet	1.19	Red-violet
3000	Blue to violet-blue	1.21	Violet-red
3100	Blue	1.24	Carnation pink to salmon
3200	Blue to blue-green	1.25	Orange
3400	Light green	1.28	Yellowish
3500	Green to yellow-green	1.32	Sky blue to green-blue
3600	Yellow-green	1.40	Orange
3700	Green-yellow	1.45	Violet
3900	Yellow	1.46	Blue-violet
4100	Light orange	1.50	Blue
4200	Carnation pink	1.54	Dull yellow-green
4400	Violet-red		
4600	Red-violet		
4700	Violet		
4800	Blue-violet		
4900	Blue		
5000	Blue-green		
5200	Green		
5400	Yellow-green		
5600	Green-yellow		
5700	Yellow to yellowish (at times appears light gray or metallic)		
5800	Light orange or yellow to pink		
6000	Carnation pink		
6300	Violet-red		
6800	Bluish (appears between violet-red and blue-green, overall looks grayish)		
7200	Blue-green to green		
7700	Yellowish		
8000	Orange		
8200	Salmon		
8500	Dull, light red-violet		
8600	Violet		
8700	Blue-violet		
8900	Blue		
9200	Blue-green		
9500	Dull yellow-green		
9700	Yellow to yellowish		
9900	Orange		

a rainbow wafer can be created with color pattern changes occurring gradually and periodically.

For a precise measurement of silicon dioxide thickness, spectroreflectometry is used. It measures the reflected light intensity at different wavelengths, and the thin-film thickness can be calculated from the relation of reflected light intensity and wavelength of the light.

For gate oxides it is very important to measure breakdown voltage and fixed charge, measurements that can be achieved by depositing a patterned conductor layer onto the oxide layer to form a metal-oxide-semiconductor capacitor. By applying bias voltage, the response of the capacitance to the applied voltage, or the C–V curve, gives the information of the fixed change at the Si–SiO_2 interface. By increasing bias voltage until the silicon dioxide breaks down, the breakdown voltage can also be measured. By using a higher test temperature, such as 250 °C, the accelerated device failure time caused by thermal stress can help predict the device's lifetime. Figure 5.28 illustrates the C–V test system.

5.3.8 Recent oxidation trends

While feature size has continued to reduce, STI gradually has replaced LOCOS for isolation of neighboring transistors. There is no longer a need to grow a thick oxide layer in IC chip manufacturing. Most oxidation processes in fabs grow thin oxide layers, such as pad, screen, barrier, and gate oxides. All of the layers are grown in dry oxidation processes. Wet oxidation is mainly used to produce test wafers needed for process control and development.

As gate oxide thicknesses continue to shrink, rapid thermal oxidation (RTO) will replace furnace oxidation processes. RTPs have better temperature control at high temperatures, and better wafer-to-wafer uniformity. An RTO system can be integrated with a HF vapor etcher in one cluster tool to perform in-situ native oxide stripping, gate oxidation, and gate oxide annealing. Nitridation of a gate oxide can increase the dielectric constant of a film and help reduce effective oxide thickness (EOT), while keeping the physical thickness of the gate oxide thick enough to prevent electrical breakdown. Nitridation can be performed in a thermal annealing process using nitric oxide (NO) as an ambient gas. It can also be accomplished

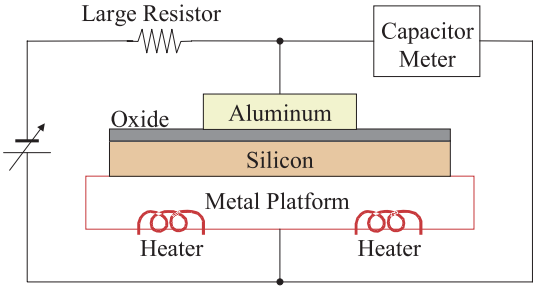


Figure 5.28 C–V test configuration.

with nitrogen plasma followed by annealing in ambient nitrogen. One advantage of this integrated process is that the silicon surface does not need to be exposed to the atmosphere before the oxidation process.

5.4 Diffusion

In the early years of the IC industry, diffusion was widely used for doping semiconductors. Because the most commonly used tool for silicon doping was the high-temperature quartz tube furnace, it was (and still is) called a diffusion furnace. For the same reason, in an IC fab, the area in which high-temperature furnaces are located is called a diffusion bay, although very few diffusion processes are actually performed in an advanced IC fab.

Diffusion is a basic physical phenomenon in which material moves from high-concentration regions to low-concentration regions, driven by the thermal motion of the molecules. The diffusion process happens anywhere and anytime. Perfume is a good example of diffusion in air; sugar, salt, and ink are examples of diffusion in liquid; and wood soaking in water or oil is an example of diffusion in a solid.

Diffusion doping processes dominated the early IC industry. By introducing a high concentration of dopant onto a silicon surface with high temperature, the dopant can diffuse into the silicon substrate, changing the conductivity of the semiconductor. Figure 5.29 illustrates a silicon diffusion doping process. Junction depth is reached when the diffused dopant concentration equals the substrate dopant concentration. Figure 5.30 shows the definition of junction depth.

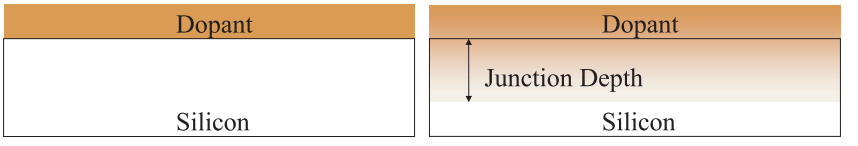


Figure 5.29 Diffusion doping process.

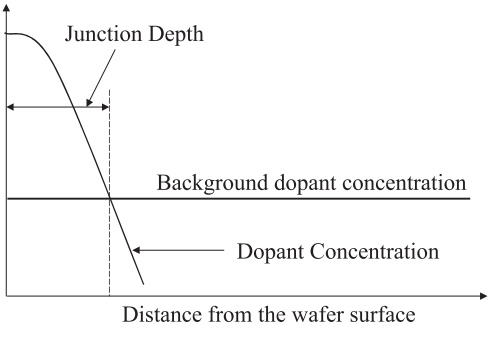


Figure 5.30 Definition of junction depth.

Because the diffusion rate in solid materials is exponentially related to temperature [$D \propto \exp(-E_a/k_BT)$], the diffusion process can be accelerated significantly at a high temperature (E_a is the activation energy, k_B is the Boltzmann constant, and T is the temperature).

For most dopants of interest in semiconductor processes, such as boron and phosphorus, their activation energy in silicon dioxide is higher than in single-crystal silicon, and their diffusion rate in silicon dioxide is much lower than in silicon. Therefore, silicon dioxide can be used as a diffusion mask to dope designated areas of the silicon surface, as shown in Fig. 5.31.

Compared with the ion implantation doping process, the diffusion doping process has several disadvantages. For instance, it cannot independently control dopant concentration and junction depth. Diffusion is an isotropic process; thus, dopants always diffuse underneath the masking oxide, as shown in Fig. 5.31. When used for small feature sizes, diffusion can cause neighboring junctions to short. Therefore, when ion implantation was introduced in semiconductor fabrication in the mid-1970s, it quickly replaced diffusion as a silicon doping process.

One important concept in the IC industry related to diffusion processes is thermal budget. After self-aligned S/D implantation, heavily doped S/Ds are formed separately by the dimension of the gate, as shown in Fig. 5.32. Any high-temperature process after S/D implantation can cause diffusion of the dopants in the S/D and expand the S/D junctions. If they expand too much, device performance can be dramatically affected. The combination of time and temperature that the wafer can spend in thermal processing after S/D implantation is called the thermal budget.

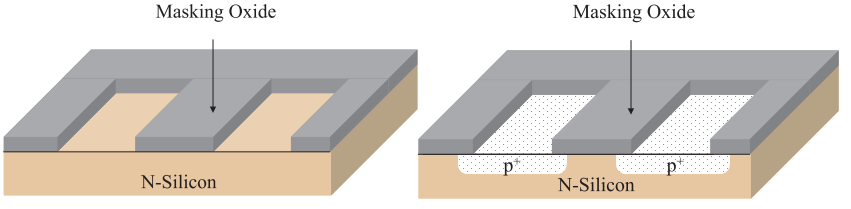


Figure 5.31 Patterned diffusion doping process.

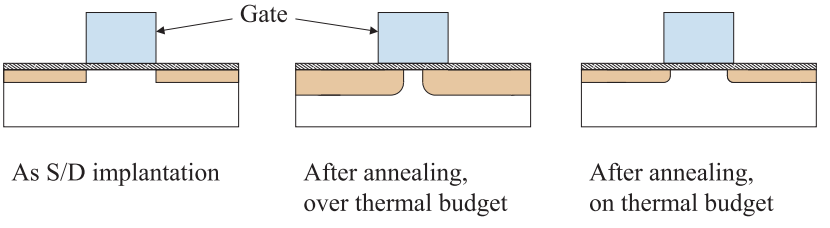


Figure 5.32 S/D diffusion in postimplantation thermal processes.

The thermal budget is determined by gate dimension, which usually is the minimum feature size of the IC chip. A device with smaller gate width has less room for the S/D junction to diffuse and therefore has a smaller thermal budget. As minimum feature size shrinks, a wafer requires less time in high-temperature (> 1000 °C) processes and needs a tighter thermal budget control. Figure 5.33 shows thermal budgets (time at temperature) for devices with different feature sizes at different temperatures. The figure assumes that surface dopant concentration is 10^{20} atoms/cm^3. A smaller device has a smaller thermal budget. For example, 0.25-μm devices only require 24 sec at 1000 °C after S/D implantation, while 2-μm devices require 1000 sec. By reducing temperature, a thermal budget can be significantly increased. For instance, 0.25-μm devices need about 200 sec at 900 °C after S/D extension (SDE) implantation. In comparison, 16-nm devices need only 1 sec at 900 °C after SDE, explaining why 16-nm devices need millisecond annealing, such as spike, laser, and flash annealing.

5.4.1 Deposition and drive-in

The commonly used diffusion doping process sequence is called predeposition and drive-in. First, a layer of dopant oxide, such as B_2O_3 or P_2O_5, is deposited onto the wafer surface at 1050 °C, followed by a thermal oxidation process that consumes the residue dopant gas and grows a layer of silicon dioxide, which caps the dopant and prevents out-diffusion. Chemical reaction of the predeposition and cap oxidation for the most commonly used boron and phosphorus sources [diborane (B_2H_6) and phosphorus oxychloride ($POCl_3$, commonly called POCL)]

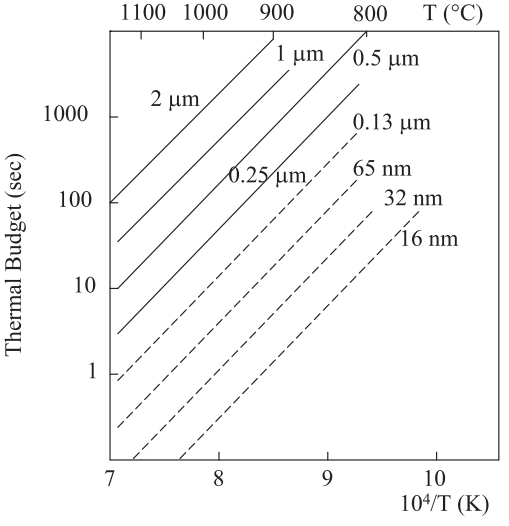


Figure 5.33 Available time-at-temperature (thermal budget) for devices with different feature sizes (after R. B. Fair).

can be expressed as

boron: predeposition $B_2H_6 + 2\ O_2 \rightarrow B_2O_3 + 3\ H_2O$
cap oxidation $2\ B_2O_3 + 3\ Si \rightarrow 3\ SiO_2 + 4\ B$
$2\ H_2O + Si \rightarrow SiO_2 + 2\ H_2$,

phosphorus: predeposition $4\ POCl_3 + 3\ O_2 \rightarrow 2\ P_2O_5 + 6\ Cl_2$
cap oxidation $2\ P_2O_5 + 5\ Si \rightarrow 5\ SiO_2 + 4\ P$.

Diborane (B_2H_6) is a poisonous gas with a burnt chocolate, sickly sweet smell. It can be fatal if absorbed through the skin or inhaled. Diborane is flammable, with an autoignition temperature of 56 °C, and becomes explosive when the concentration in air is higher than 0.8%. POCL is a corrosive liquid, which can cause burns to the skin or eyes. POCL vapor is irritating to the skin, eyes, and lungs, and can cause dizziness, headache, loss of appetite, nausea, and lung damage. Other commonly used n-type dopant chemicals are arsine (AsH_3) and phosphine (PH_3), both being poisonous, flammable, and explosive. Their predeposition and oxidation reactions are very similar to those of diborane (B_2H_6).

Figure 5.34 illustrates a furnace system used for boron predeposition and cap oxidation processes. A dedicated furnace processing tube for each dopant prevents cross-contamination.

After predeposition and cap oxidation, the temperature of the furnace is ramped up to 1200 °C in ambient oxygen, providing enough thermal energy for the dopant to diffuse into the silicon substrate quickly. Drive-in time is determined by the required junction depth and can be easily calculated from the existing theory for each dopant. Figure 5.35 illustrates the predeposition, cap oxidation, and drive-in steps for a diffusion doping process.

The diffusion process cannot independently control dopant concentration and dopant junction depth, since both are strongly related to the processing

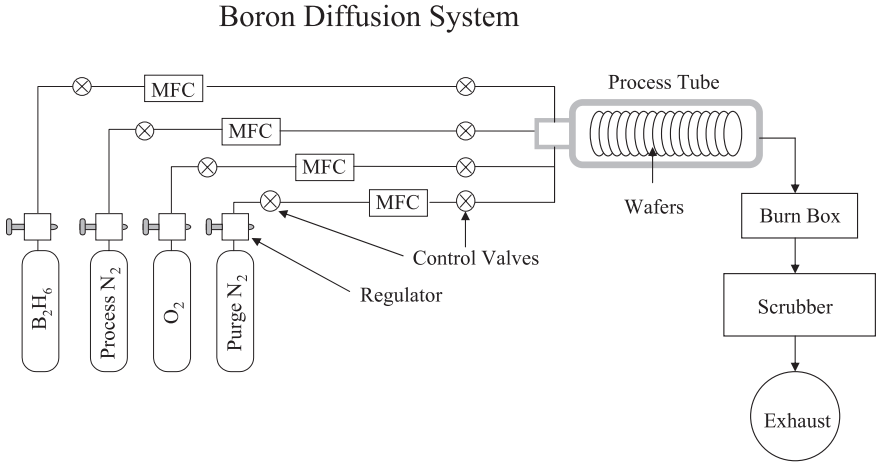


Figure 5.34 Schematic of a furnace system for the boron doping process.

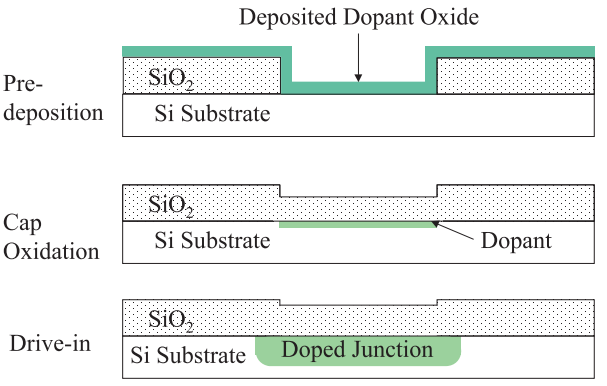


Figure 5.35 Diffusion doping process sequence.

temperature. The diffusion process is an isotropic process, and dopant atoms always diffuse underneath the masking silicon dioxide. In advanced IC processing, almost all semiconductor doping is done with an ion implantation process instead of a diffusion process, because ion implantation has much better control of dopant concentration and profile. The main application of the diffusion process in advanced IC processing is dopant drive-in during the well implantation annealing process.

The diffusion process was revisited by fab research and development departments for ultrashallow-junction (USJ) formation. First, a layer of borosilicate glass (BSG) with very high boron concentration is deposited onto the wafer surface by a CVD process. Then, by using an RTP, boron is driven out of the BSG and diffuses into silicon to form a shallow junction. Figure 5.36 illustrates the USJ formation by deposition, diffusion, and stripping processes.

5.4.2 Doping measurement

The four-point probe is one of the most commonly used measurement tools for monitoring doping processes. Resistivity of silicon is related to dopant concentration; therefore, measuring sheet resistance on the silicon surface can provide some useful information of the dopant concentration. Sheet resistance R_s is a defined parameter. The resistance of a conducting line can be expressed as

$$R = \rho \frac{L}{A},$$

where R is the resistance, ρ is the resistivity of the conductor, L is the length of the conducting line, and A is the area of the line cross section. If the wire is a band with a rectangular cross section, the area of the cross section simply changes to the product of the width W and thickness t. Line resistance can be expressed as

$$R = \rho \frac{L}{Wt}.$$

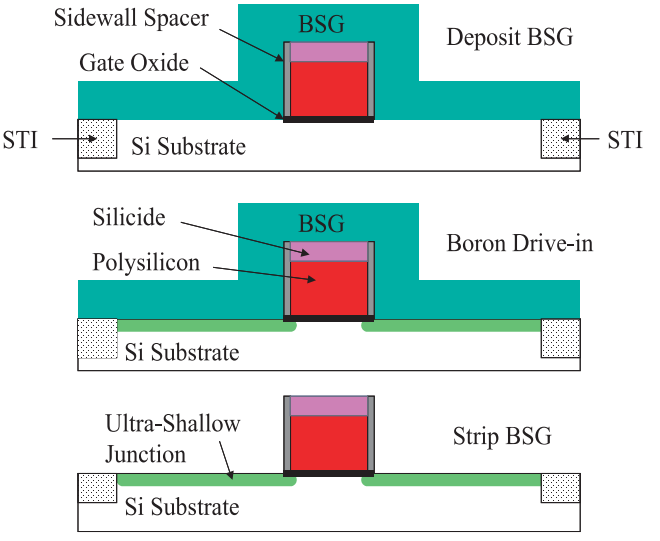


Figure 5.36 Diffusion process for USJ applications.

For a square sheet, the length is equal to the width, $L = W$, thus they cancel each other. Therefore, the resistance of a square conducting sheet, defined as sheet resistance, can be expressed as

$$R_s = \rho/t.$$

Resistivity ρ of doped silicon is mainly determined by dopant concentration, and thickness t is mainly determined by dopant junction depth. Therefore, a sheet resistance measurement can provide information about dopant concentration, since junction depth can be estimated by knowing the ion energy, ion species, and substrate materials.

By applying a certain amount of current between two of the pins and measuring voltage difference between the other two, sheet resistance can be calculated. Typically, the distances between the probes are: $S_1 = S_2 = S_3 = 1$ mm. If a current I is applied between P_1 and P_4, as shown in Fig. 5.37, sheet resistance $R_s = 4.53\ V/I$, where V is the measured voltage between P_2 and P_3. If the current is applied between P_1 and P_3, $R_s = 5.75\ V/I$, where V is measured between R_2 and R_4. These two equations are derived with the assumption of an infinite film area, which is not the case for thin-film measurements on wafers. During the measurement, an advanced tool normally runs four sets of measurements, programmed in a sequence with both configurations and reverse current for each configuration, to minimize the edge effect and achieve an accurate measurement.

Since the four-point probe makes direct contact and creates defects on the wafer surface, it is only used on test wafers for process development, qualification, and control. During measurement, it is important that probes touch the silicon surface

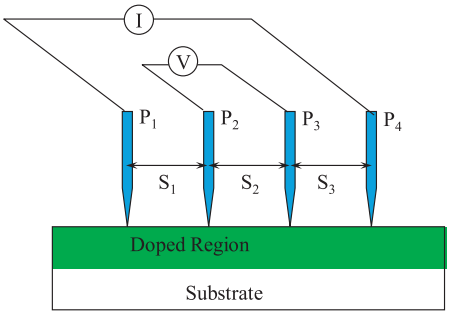


Figure 5.37 Four-point probe.

with enough force that the pins break through the thin (~10 to 20 Å) native oxide to make good contact with the silicon substrate.

5.5 Annealing

Annealing is a heating process in which a wafer is heated to achieve a desired physical or chemical change, with minimal material being added to or removed from the wafer surface.

5.5.1 Postimplantation annealing

In an ion implantation process, energetic dopant ions cause extensive damage to the silicon crystal structure near the wafer surface. To achieve device requirements, the lattice damage must be repaired in a thermal annealing process to restore the single-crystal structure and activate the dopant. Only when dopant atoms are in the form of a single-crystal lattice can they effectively provide electrons or holes as majority carriers for conducting electric current. In a high-temperature process, atoms move quickly, driven by thermal energy, and they find and rest at sites that have the lowest free energy, located at single-crystal lattices; this process restores the single-crystal structure. Figure 5.38(a) illustrates crystal damage after ion implantation, and Fig. 5.38(b) shows the crystal recovery and dopant activation after thermal annealing.

Before the 1990s, furnaces were widely used for postimplantation annealing. A furnace annealing process is a batch process operated from 850 to 1000 °C for about 30 min in ambient nitrogen and oxygen. A small amount of oxygen is used to prevent the formation of silicon nitride on the wafer surface where silicon is exposed. During idle time, a furnace is always kept at a high temperature, between 650 and 850 °C, and wafers must be pushed into and pulled out of the furnace very slowly to avoid wafer warping. Therefore, wafers at opposite ends of the paddle or tower will have different annealing times due to the slow entry and exit; this can cause wafer-to-wafer (WTW) nonuniformity.

Another concern for furnace annealing is the thermal budget, or the dopant diffusion problem during the annealing process. The furnace annealing process

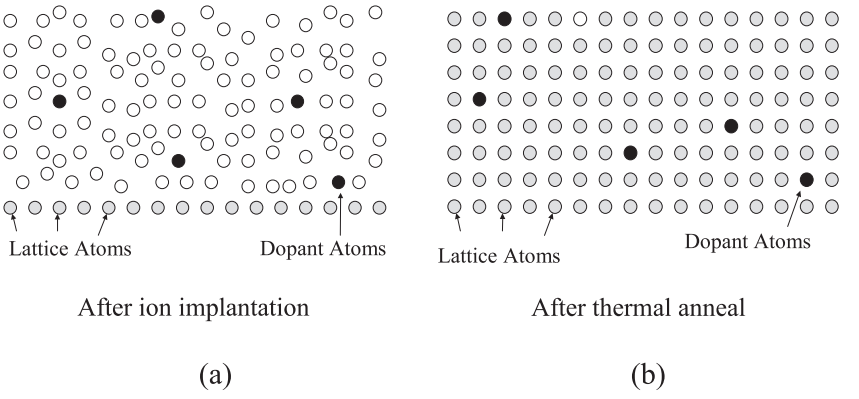


Figure 5.38 (a) Crystal damage and (b) annealing process.

takes a long time and can cause too much dopant diffusion, which is intolerable for a transistor with a small geometry. Therefore, the RTA process is preferred for most postimplantation annealing in an advanced IC fab. An RTA system can ramp up wafer temperature from room temperature to about 1100 °C in a very short time, normally within 10 sec. RTPs can precisely control wafer temperature and within-wafer (WIW) temperature uniformity. At approximately 1100 °C, a single-crystal structure can be recovered in about 10 sec, with minimum dopant diffusion. The RTA process also has much better WTW temperature uniformity control. Thus, only some noncritical postimplantation annealing processes, such as well implantation annealing and drive-in, still use furnaces for advanced IC processing.

5.5.2 Alloy annealing

Alloy annealing is a heating process in which thermal energy helps different atoms chemically bond with each other to form a metal alloy. In semiconductor processing, several alloy processes have been used. Formation of cobalt silicide ($CoSi_2$) in a self-aligned silicide (salicide) process is one of the most common, as shown in Fig. 5.39.

The first anneal at 450 °C forms CoSi, and the second anneal at 700 °C forms $CoSi_2$. With the RTP, annealing occurs at temperatures of 700 to 750 °C to directly form $CoSi_2$ in one step. $CoSi_2$ was widely used in logic IC with technology nodes from 0.25 μm to 90 nm. It is still used for some memory chips.

After the 65-nm node, nickel silicide (NiSi) was introduced as the silicide material for high-speed logic ICs. NiSi can be formed at lower temperatures, around 450°C, with resistance comparable to $CoSi_2$ and $TiSi_2$. This is a very important advantage because it can significantly reduce the thermal budget to meet requirements of multiple logic IC technology nodes, all the way to 10 nm and beyond.

Both furnace and RTP systems are used for titanium silicide and cobalt silicide alloy processes, but the RTP system clearly offers better control of thermal budget and WTW uniformity. For nickel silicide, only RTP systems are used.

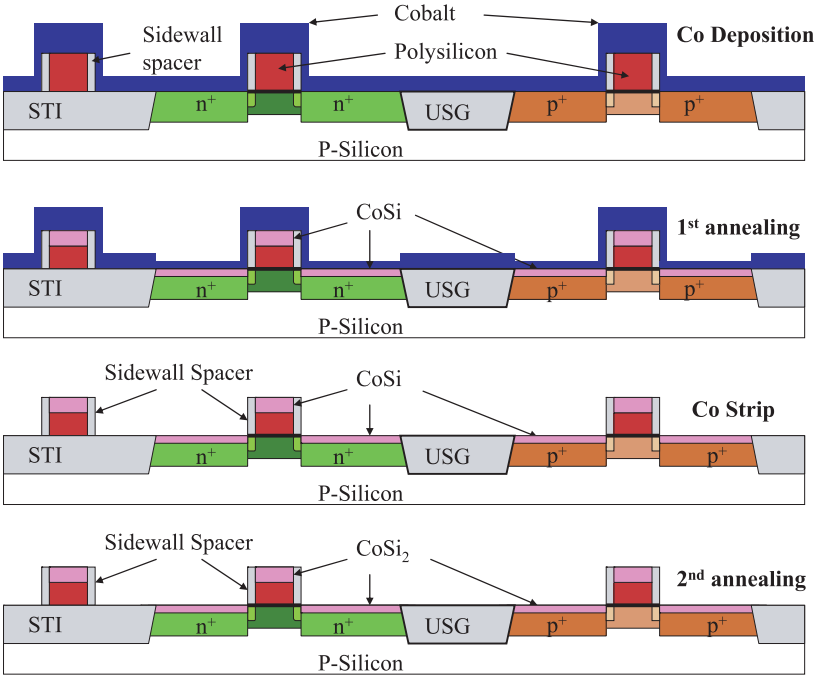


Figure 5.39 Self-aligned cobalt silicide process.

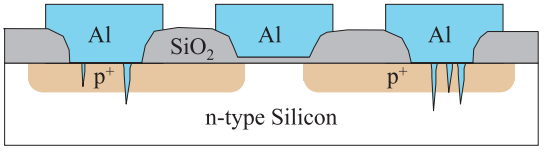


Figure 5.40 Junction spiking.

Furnaces have been used to form an aluminum-silicon alloy at 400 °C in nitrogen and ambient hydrogen to prevent silicon-aluminum interdiffusion, which can cause a junction spike, as shown in Fig. 5.40.

5.5.3 Reflow

When the temperature exceeds the glass transition temperature T_g of silicate glass, the glass becomes soft and starts to flow. This property is widely used in the glass industry to shape glass into different kinds of glassware. It is also applied in wafer processing to smooth silicate glass surfaces in a thermal process commonly called reflow. At 1100 °C, phosphosilicate glass (PSG) becomes soft and starts to flow.

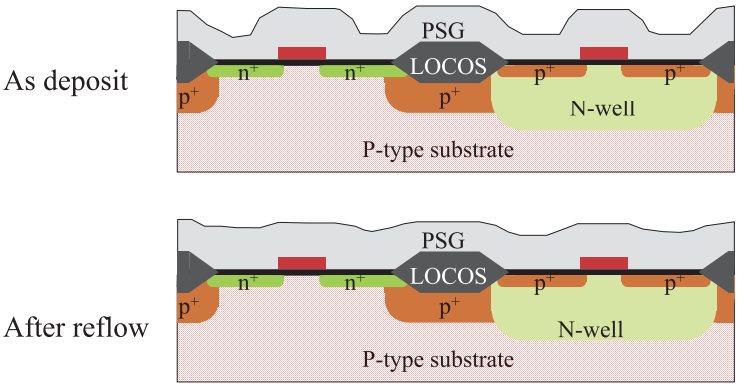


Figure 5.41 PSG reflow process.

The softened PSG follows surface tension, which smoothes and planarizes the dielectric surface. This improves photolithography resolution and makes it easier for metallization processes to succeed. PSG deposition and reflow is shown in Fig. 5.41.

As minimum feature size shrinks, the thermal budget becomes tighter. Borophosphosilicate glass (BPSG) reduces reflow temperature to about 900 °C, which significantly increases thermal budget. The reflow process takes approximately 30 min in a furnace with ambient nitrogen (after the wafer boats are pushed in, and the temperature ramps up and stabilizes).

When the minimum feature size is smaller than 0.25 μm, reflow alone can no longer meet the surface planarization requirement for high photolithography resolution, and the tight thermal budget limits the application of the reflow process. The CMP process then replaces reflow for dielectric surface planarization.

5.6 High-Temperature Chemical Vapor Deposition

CVD is an adding process in which a layer is deposited onto a wafer surface. High-temperature CVD processes include epitaxial silicon deposition, polysilicon deposition, and LPCVD nitride deposition.

5.6.1 Epitaxial silicon deposition

Epitaxial silicon is a single-crystal silicon layer deposited on a single-crystal silicon wafer surface in a high-temperature process. An epitaxial silicon layer is needed for bipolar transistors, BiCMOS IC chips, and high-speed, advanced CMOS IC chips.

Silane (SiH_4), dichloride silane (DCS, SiH_2Cl_2), and trichloride silane (TCS, $SiHCl_3$) are the three most commonly used source gases for silicon-epitaxial

growth processes. Chemical reactions of the epitaxial silicon growth process are:

$$\underset{\text{silane}}{SiH_4} \xrightarrow{\text{heat } (1000^\circ C)} \underset{\text{epi-Si}}{Si} + \underset{\text{hydrogen}}{H_2},$$

$$\underset{\text{DCS}}{SiH_2Cl_2} \xrightarrow{\text{heat } (1100^\circ C)} \underset{\text{epi-Si}}{Si} + \underset{\text{hydrochloride}}{2HCl},$$

$$\underset{\text{TCS}}{SiHCl_3} + \underset{\text{hydrogen}}{H_2} \xrightarrow{\text{heat } (1100^\circ C)} \underset{\text{epi-Si}}{Si} + \underset{\text{hydrochloride}}{3HCl}.$$

Epitaxial silicon can be doped as the film grows by flowing dopant gas such as arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6) with the silicon source gas into the reactor. All three dopant gases are very poisonous, flammable, and explosive. Blanket epitaxial silicon deposition (growth) is normally done outside the IC fab by wafer manufacturers. The epitaxial process is discussed in detail in Chapter 4.

5.6.2 Selective epitaxial growth processes

By patterning silicon oxide or silicon nitride masking film, epitaxial layers can be grown at locations where the masking film is removed and silicon is exposed. This process is called selective epitaxial growth (SEG). Figure 5.42 illustrates the application of SEG SiC to form tensile-strained channels of nMOS and SEG SiGe to form compressive-strained channels of pMOS.

5.6.3 Polycrystalline silicon deposition

Polysilicon has been widely used as a gate material since the introduction of ion implantation as a silicon doping process in the IC industry in the mid-1970s. It is also widely used as contact plugs and capacitor electrodes in DRAM chips. Figure 5.43 illustrates the applications of polysilicon in an advanced DRAM chip.

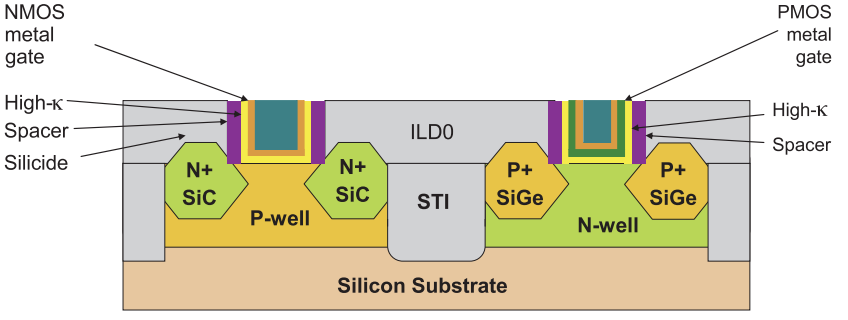


Figure 5.42 SEG applications in advanced CMOS ICs.

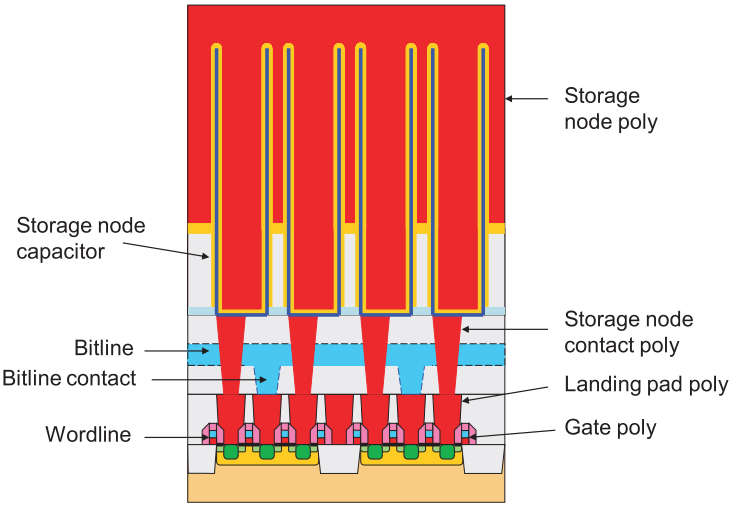


Figure 5.43 Applications of polysilicon in a DRAM chip.

The first poly forms the gate electrodes and wordline and has tungsten on top to reduce resistance. The second poly forms the plugs, which provide landing pads for contact between the S/D and the bitline, and contact between the S/D and the storage capacitor. The third poly forms the contact plugs between the storage capacitors and the landing pad poly. The fourth poly forms the grounding electrode of the storage capacitor of the DRAM, and titanium nitride forms another electrode with high-κ dielectric sandwiches between the two electrodes. By using a high-κ dielectric, the size of the capacitor can be reduced while keeping the required capacitance.

Polysilicon deposition is a LPCVD process that takes place in a furnace with a vacuum system, as illustrated in Fig. 5.44. Polysilicon deposition uses silane (SiH_4) chemistry. At high temperatures, silane will dissociate, and silicon can be deposited on the heated surface. The chemical reaction can be expressed as

$$SiH_4 \rightarrow Si + 2\,H_2.$$

Polysilicon can also use dichlorosilane (DCS, SiH_2Cl_2) chemistry for deposit. At high temperatures, DCS will react with hydrogen and deposit silicon on a heated surface. The DCS process requires a higher deposition temperature than the silane process. The DCS chemical reaction is:

$$SiH_2Cl_2 + H_2 \rightarrow Si + 2\,HCl.$$

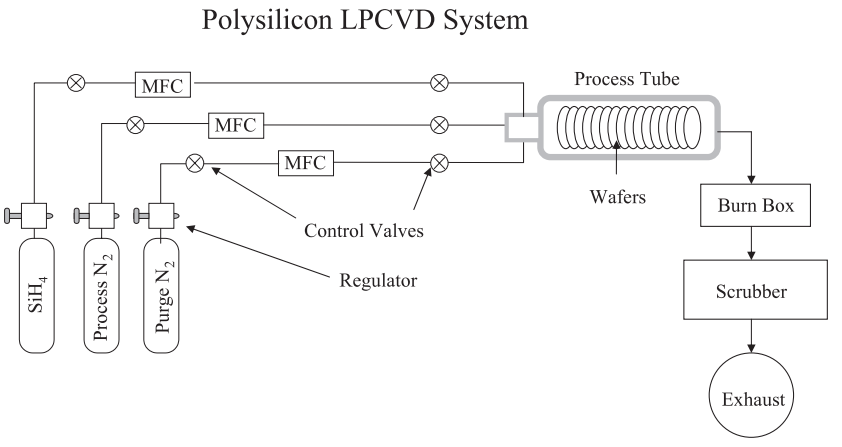


Figure 5.44 Schematic of a LPCVD system for polysilicon deposition.

LPCVD polysilicon can be doped in situ by flowing dopant gas such as arsine (AH_3), phosphine (PH_3), or diborane (B_2H_6) with silane or DCS into the processing chamber (tube).

Polysilicon is deposited at low pressures of 0.2 to 1.0 torr, with deposition temperatures between 600 and 650 °C, using either pure silane or 20 to 30% silane diluted in nitrogen. Deposition rates for both processes are from 100 to 200 Å/min, mainly determined by deposition temperature. Nonuniformity of film thickness within the wafer is usually less than 4%.

The polysilicon deposition processing sequence is as following:

- idle with purge N_2 flow
- idle with process N_2 flow
- wafer loading with process N_2 flow
- process tube (bell jar) lowers with process N_2 flow
- pump down chamber to base pressure (< 2 mtorr) by turning off N_2 flow
- stabilize wafer temperature with N_2 flow and leak check
- pump-to-base pressure (< 2 mtorr) by turning off N_2 flow
- set up process pressure (~250 mtorr) with N_2 flow
- turn on SiH_4 flow and turn off N_2, start deposition
- turn off silane and open gate valve, pump to base pressure
- close gate valve, fill N_2, and ramp-up pressure to atmospheric pressure
- raise bell jar and cool down wafer temperature with process N_2 flow
- unload wafer with process N_2 flow
- idle with purge N_2 flow.

The LPCVD polysilicon deposition process is mainly controlled by process temperature, total process pressure, silane partial pressure for the diluted process, and dopant concentration. Wafer spacing and load size have minimum effect on deposition rate, but they are important for uniformity.

The resistivity of polysilicon film is highly dependent on deposition temperature, dopant concentration, and annealing temperature. Increasing deposition temperature causes resistivity to decrease. Increasing dopant concentration also reduces resistivity, and higher annealing temperatures form large grain sizes, which reduce resistivity as well. Larger polysilicon grain size can cause difficulties in the etch process because it can induce rough sidewalls; therefore, polysilicon is deposited at lower temperatures with smaller grain size and higher resistivity. After polysilicon etch and photoresist stripping, the film is annealed at a high temperature to form large grain size and lower resistivity. In some cases, amorphous silicon (α-Si) is deposited at about 450 °C, then patterned, etched, and annealed to form polysilicon with larger and more uniformly sized grains.

Polysilicon can also be deposited in a single-wafer system. One of the advantages of such deposition is that it can deposit tungsten silicide in situ with the polysilicon. In some DRAM chips, polysilicon-tungsten silicide-staked films are applied to form the gate/wordline and the bitline. The in-situ polysilicon/silicide deposition process eliminates requirements of deglazing and surface cleaning of the polysilicon layer before tungsten silicide deposition, processes that are required in traditional furnace polysilicon deposition and the CVD tungsten silicide processing sequence. By using a polysilicon-tungsten silicide integrated system, throughput can be improved significantly. The single-wafer polysilicon deposition chamber is similar to the single-wafer epitaxial silicon deposition chamber illustrated in Fig. 4.23. A schematic of an integrated polysilicon plus tungsten silicide (WSi_x) deposition system, or polycide system, is shown in Fig. 5.45.

In an integrated polycide system, a wafer is first loaded from the loading station and then sent by a robot to the polysilicon chamber via the transfer chamber. After

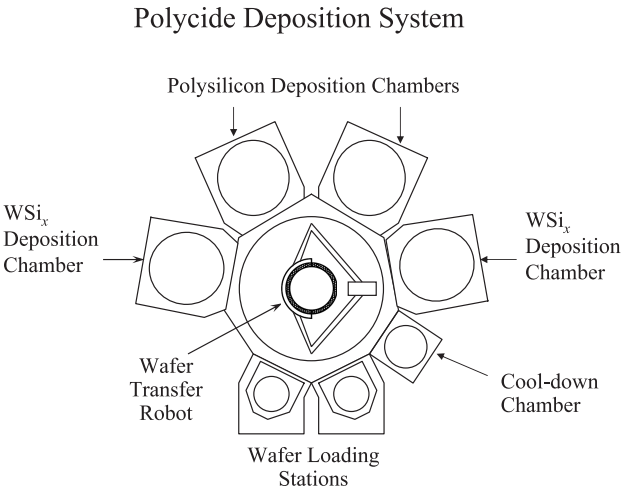


Figure 5.45 Multichamber polycide system.

polysilicon deposition, the wafer is taken out of the polysilicon chamber and sent to the WSi_x chamber to deposit tungsten silicide (Si_x $2 < x < 3$) via the transfer chamber, which is under vacuum conditions. When polycide deposition finishes, the robot picks up the wafer and sends it to the cool-down chamber. Nitrogen gas flow then removes heat from the wafer before the robot places it into a plastic wafer cassette in the loading station for unloading.

For advanced DRAM chips, polysilicon, tungsten silicide, tungsten nitride, and stacked tungsten (poly/WSi_x/WN/W) are commonly used for the gate/wordline, and tungsten nitride and stacked tungsten (WN/W) are used for the bitline. One of the most advanced DRAM chips uses buried wordline (bWL) technology, which uses stacked TiN/W for array transistor gates and wordline, and poly/WSi_x/WN/W for bitline and periphery transistor gate electrodes. A cluster system (illustrated in Fig. 5.45) can be used to deposit stacked poly/WSi_x/WN/W, with four deposition chambers to deposit each film in one processing sequence.

Single-wafer polysilicon deposition uses silane chemistry at a reduced pressure from 10 to 200 torr. Deposition temperature is from 550 to 750 °C, with a deposition rate of up to 2000 Å/min. Hydrogen chloride is used in the chamber dry-clean process to remove polysilicon film deposited on the chamber wall, helping reduce particle generation.

5.6.4 Silicon nitride deposition

Silicon nitride is a dense material widely used as a diffusion barrier layer in IC chips. It is used as an oxidation mask to block oxygen diffusion in the LOCOS formation process, as shown in Fig. 5.11. It is also used in the STI formation process as a CMP stop layer because it has a significantly lower polish rate than USG (see Fig. 5.9).

LPCVD silicon nitride is also used for sidewall spacer formation, either as an etch-stop layer for oxide sidewall spacers or as the spacer itself. Silicon nitride is deposited as a dopant diffusion barrier layer before interlay dielectric (ILD0) PSG or BPSG deposition, to prevent boron and phosphorus from diffusing across the ultrathin gate oxide and penetrating the silicon substrate, possibly causing device malfunction. The nitride barrier layer can also serve as an etch-stop layer for self-aligned contact application, as shown in Fig. 5.46.

These front-end-of-line (FEoL) nitrides can be deposited with LPCVD processes. For the nitride diffusion barrier, some advanced IC chips use plasma-enhanced CVD (PECVD) processes due to thermal budget concerns because PECVD requires significantly lower temperatures than LPCVD. Some advanced CMOS IC chips use nitride stress liners to create different channel strains for pMOS and nMOS. The dual stress liner technique employs PECVD nitride for compressive stress liners to form compressive strain channels of pMOS and uses LPCVD nitride for tensile stress liners to create tensile strain channels of nMOS.

A thin layer of silicon nitride is also commonly used for copper metallization as an ILD etch-stop layer (ESL). A thick layer of nitride is always used as the

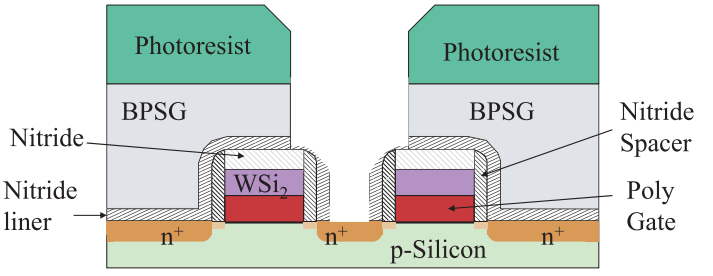


Figure 5.46 Silicon nitride sidewall spacer and self-aligned contact etch stop.

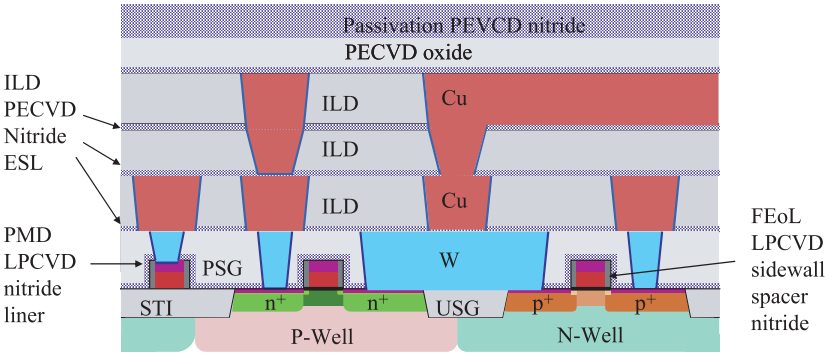


Figure 5.47 Application of silicon nitride in a CMOS IC chip.

passivation dielectric (PD) for IC chips. The ILD0 and passivation applications of silicon nitride in a copper chip are illustrated in Fig. 5.47.

Since wafers cannot tolerate any processing with temperatures higher than 450 °C after the first aluminum-alloy metal layer deposition, most ILD and PD nitride deposition processes are PECVD processes operated around 400 °C. PECVD can achieve high deposition rates at relatively low temperatures because the plasma-induced free radicals dramatically enhance the chemical reaction rate. PECVD processing is discussed in detail in Chapter 10.

LPCVD processes are widely used for LOCOS nitride, STI nitride, spacer nitride, and ILD0 barrier nitride deposition because LPCVD nitride has better film quality and less hydrogen integration than PECVD nitride. In addition, LPCVD nitride is not prone to plasma-induced device damage, which is unavoidable with the PECVD process. LPCVD processing uses a furnace with a vacuum system, as shown in Fig. 5.48. Silicon nitride is deposited with dichlorosilane (SiH_2Cl_2) and ammonia (NH_3) chemistry at 700 to 800 °C. The chemical reaction can be expressed as

$$3\ SiH_2Cl_2 + 4\ NH_3 \rightarrow Si_3N_4 + 6\ HCl + 6\ H_2.$$

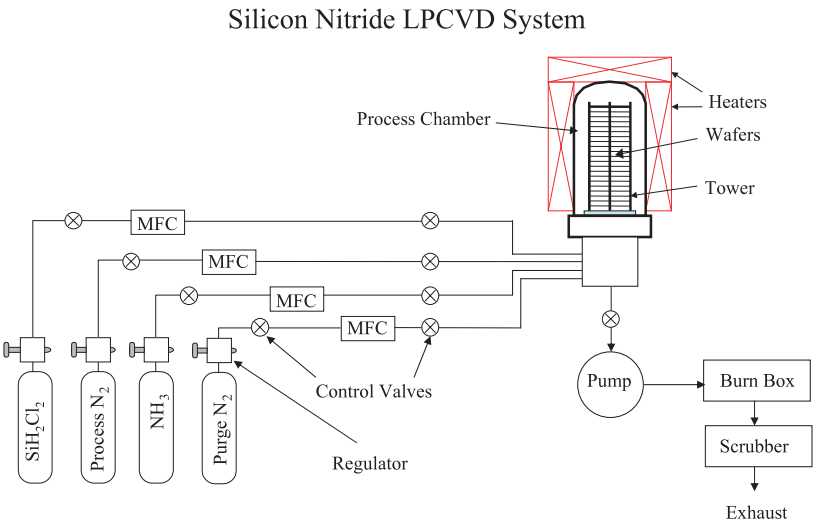


Figure 5.48 Schematic of silicon nitride LPCVD system.

Silicon nitride can also be deposited with silane (SiH_4) and ammonia (NH_3) chemistry at about 900 °C. The chemical reaction is expressed as

$$3\ SiH_4 + 4\ NH_3 \rightarrow Si_3N_4 + 12\ H_2.$$

The process sequence is:

- load wafers
- raise wafer tower with constant nitrogen flow and processing chamber temperature
- pump the chamber down to base pressure without nitrogen flow
- backflow nitrogen to stabilize wafer temperature
- turn off nitrogen and pump down to base pressure
- flow nitrogen and ammonia to set up and stabilize processing pressure
- turn off nitrogen and turn on dichlorosilane to deposit silicon nitride
- pump chamber down to base pressure with all gas flows turned off
- start nitrogen flow into the chamber to raise pressure to atmospheric pressure
- lower wafer tower and unload wafers with nitrogen flow.

The process flow diagram is illustrated in Fig. 5.49.

One of the possible byproducts of a DCS-based nitride LPCVD process is ammonium chloride (NH_4Cl), which is a solid and can cause particle contamination and pump damage. Therefore, researchers are seeking other precursors for the LPCVD nitride process. A possibility is bis(tertiary-butylamino)silane—$SiH_2[NH(C_4H_9)]_2$ or BTBAS, which is a liquid with a boiling point of 164 °C. At 550 to 600 °C, it can react with ammonia to deposit uniform silicon nitride film with high film quality and good step coverage, and free of ammonium chloride.

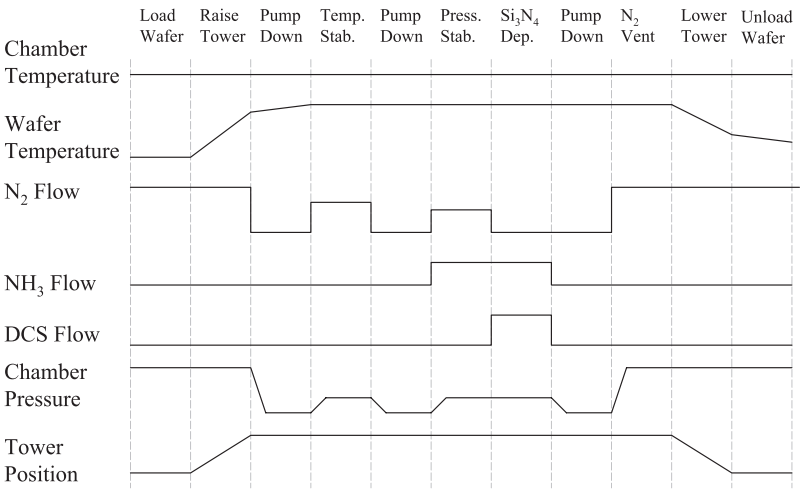


Figure 5.49 Silicon nitride LPCVD process flow diagram.

5.7 Rapid Thermal Processing

Rapid thermal processing is a batch processing tool that processes hundreds of wafers at the same time. Due to the large thermal capacity, the temperature of processing tubes or chambers must ramp up and down slowly. AN RTP system is used for single-wafer processing and can ramp up temperature at rates between 75 to 200 °C/sec. AN RTP system is the favored system for postimplantation annealing, silicide annealing, and ultrathin silicon dioxide layer growth.

RTP systems include rapid thermal annealing (RTA), rapid thermal oxidation (RTO), and rapid thermal CVD (RTCVD) processes. There are several different RTA processes, including postimplantation annealing (>1000 °C), alloy annealing (~700 °C), dielectric annealing (700 °C to 1000 °C), and metal annealing (<500 °C).

RTP uses a quartz chamber and many quartz parts. The heating element is a tungsten-halogen lamp, which can generate intense heat in the form of infrared (IR) radiation. The wafer temperature is precisely measured by an infrared pyrometer. Figure 5.50 illustrates one type of RTP system.

Lamps on the top and bottom are positioned in a vertical array to allow wafers to be heated uniformly by the IR radiation. Wafer temperature is monitored by a pyrometer and feedback is controlled by the lamp power. Figure 5.51 shows the lamp arrangement in an RTP system.

Another type of RTP system places heating lamps in gold-plated lamp houses with a honeycomb structure, as shown in Fig. 5.52. Gold-plated lamp housing improves the power transfer efficiency. During the process, a wafer rotates to improve heating uniformity. Several pyrometers monitor the temperature across wafer and feedback signals to control the heating power of the lamps in different heating zones and achieve very precise, very uniform wafer heating.

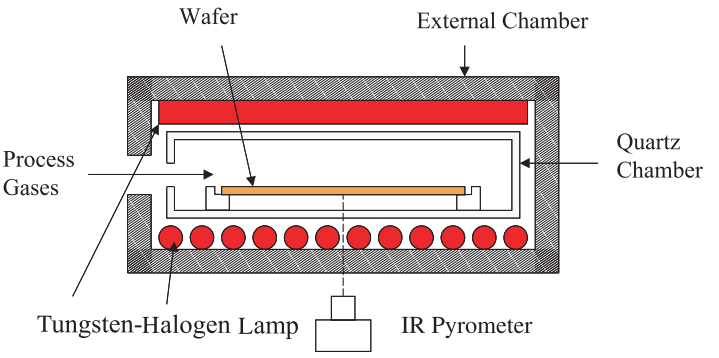


Figure 5.50 Schematic of an RTP chamber.

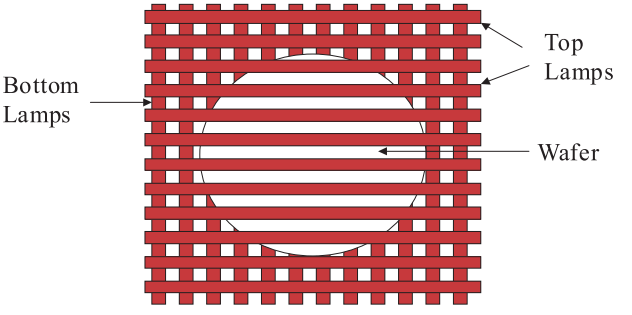


Figure 5.51 Heating lamp array in an RTP chamber.

5.7.1 Rapid thermal annealing

The most common application of RTP is the postimplantation RTA process. After ion implantation, the crystal structure of silicon near the surface is heavily damaged by energetic ion bombardment. A high-temperature process is required to anneal the damage, restore the single-crystal structure, and activate the dopant. During the high-temperature annealing process, dopant atoms diffuse fast, driven by the thermal energy. It is very important to minimize the dopant diffusion during thermal annealing. As device dimensions shrink into nanometer ranges, there is much less room for dopant atoms to diffuse. Therefore, precise control of the thermal budget is critical.

In amorphous silicon, thermal motion of the dopant atoms is unrestricted, whereas the dopant atoms in single-crystal lattices are severely restricted by binding energy. Therefore, dopant atoms can diffuse much faster than those bonded in a single-crystal lattice. At lower temperatures, the diffusion process outpaces the annealing process; at high temperatures (>1000 °C), the annealing process is faster. This is because the annealing activation energy (~5 eV) is higher than the diffusion activation energy (3 to 4 eV). Because furnace processing requires more time and the annealing temperature is relatively low, dopant diffusion cannot be minimized.

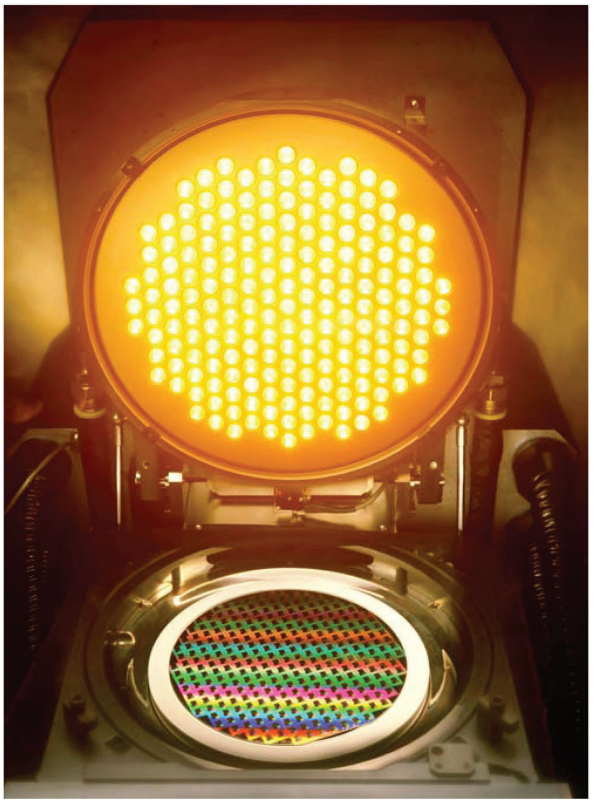


Figure 5.52 RTP system with honeycomb heating source (courtesy of Applied Materials, Incorporated).

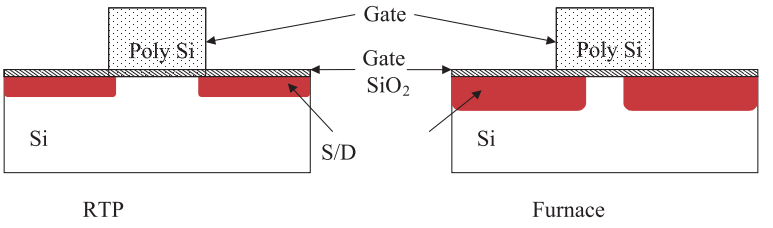


Figure 5.53 Dopant diffusion in RTP and furnace annealing processes.

Diffusion of dopant atoms can be quite significant and becomes intolerable for small devices. For some noncritical implantation processes, such as well implantation, furnace annealing could still be used; but for most postimplantation annealing, the RTA process is required. Both processes are shown in Fig. 5.53.

A RTP system can ramp up and cool down wafer temperature very quickly. Normally, it takes less than 10 sec to reach the required annealing temperature,

from 1000 to 1150 °C. Annealing takes about 10 sec, then the wafer is cooled down very quickly by turning the lamps off and flowing nitrogen gas. The faster the temperature ramp, the less the diffusion of the dopant atoms. For sub-0.1-μm device applications, a ramp-up rate of 250 °C/sec can be needed to achieve junction annealing with minimized dopant diffusion.

A postimplantation RTA process sequence is as follows:

- wafer in
- temperature ramp-up, temperature stabilization
- annealing
- wafer cool-down
- wafer out.

Temperature ramp-up rate is normally 75 to 150 °C/sec, and the annealing temperature is approximately 1100 °C. The entire processing sequence takes less than two minutes in ambient nitrogen with a constant nitrogen flow. Figure 5.54 shows the temperature change of the RTP system in a postimplantation annealing process.

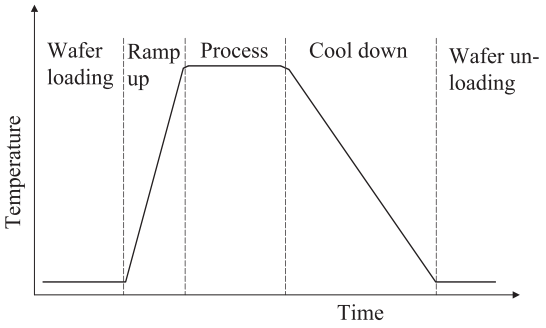


Figure 5.54 Temperature change in an RTA process.

Question: If the pyrometer is not well calibrated, what could be the consequence in the RTA process?

Answer: A poorly calibrated pyrometer can fail to measure wafer temperature accurately; thus, the stabilized processing temperature will not be the set value. If it is too low, sheet resistance will be high due to incomplete dopant activation and damage repair. If it is too high, in extreme cases, it can melt a silicon wafer. Therefore, routine calibration of the pyrometer is an important part of PM procedures.

Other RTP annealing processes include alloy annealing, especially silicide processes such as those for titanium silicide, cobalt silicide, and nickel silicide. $TiSi_2$ and $CoSi_2$ undergo RTA around 700 °C in ambient nitrogen, and NiSi undergoes RTA between 400 to 450 °C. It takes about one minute to anneal and form silicide.

RTP is also used for thermal nitridization processes, in which ammonia reacts with titanium to form titanium nitride on the surface as a barrier and adhesion layer for aluminum metallization. The chemical reaction can be expressed as

$$NH_3 + Ti \rightarrow TiN + 3/2\ H_2.$$

After titanium deposition, the wafer goes into the RTA chamber. After temperature ramp up and stabilization in the ambient nitrogen, the nitrogen is turned off and ammonia is turned on. Soon after the nitridization finishes, ammonia is turned off and nitrogen is turned on. The transfer robot removes the wafer from the processing chamber and sends it to a cool-down chamber until it is ready to be placed into a plastic wafer cassette. The processing temperature is about 650 °C. Figure 5.55 illustrates the titanium nitridization process.

5.7.2 Rapid thermal oxidation

When transistor size decreases, the thickness of the gate oxide also decreases. The thickness of gate oxides has been thinned to about 15 Å, a thickness that can no longer be reduced because of the leakage issue. With such a fine thickness, it is difficult to precisely control oxide thickness and WTW thickness uniformity within a multiwafer batch system, such as an oxidation furnace. There are several advantages to using single-wafer RTP systems to grow this ultrathin, high-quality oxide. Because RTP systems can precisely control wafer temperature and temperature uniformity across a wafer, the rapid thermal oxidation (RTO) process can grow a very thin and uniform oxide layer. Because it is a single-wafer system, the RTO process has better WTW uniformity control than furnace processing, especially for an ultrathin oxide layer. Another advantage is that the RTO chamber can be integrated with a HF vapor etch chamber in one mainframe. After the HF vapor etch process removes native oxide on the silicon wafer surface, the wafer is sent to the RTO/RTA chamber through a transfer chamber, which is always kept under high vacuum. The possibility of reoxidation of the silicon surface is eliminated, since the wafer does not become exposed to the atmosphere

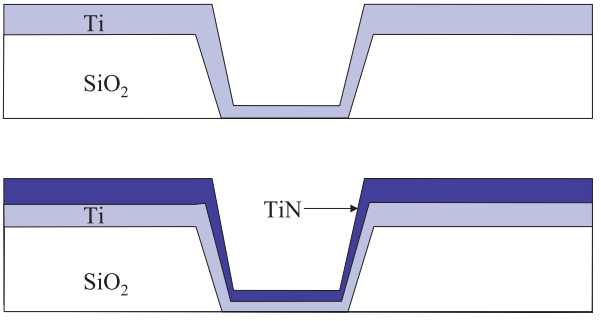


Figure 5.55 The titanium nitridization process.

or moisture before it is sent to the RTO chamber for HCl cleaning, oxidation, and annealing processes.

An RTO process diagram is illustrated in Fig. 5.56. After loading the wafer into the processing chamber, the heating lamp is turned on and the temperature ramps up in two steps, first to about 800 °C, at a higher ramp rate, then to the required oxidation temperature, such as 1150 °C, at a lower ramp rate. This two-step temperature ramp up minimizes ramping time by reducing the required stabilization time at a lower ramp rate to reach the processing temperature. After temperature stabilization, oxygen starts to flow into the processing chamber to react with the silicon and grow silicon dioxide on the wafer surface. HCl is used in the oxidation process to minimize mobile ion contamination and reduce interface state charge. After the oxide is grown, O_2 and HCl flows are turned off and N_2 flows into the chamber. Wafer temperature is ramped up to about 1100 °C to anneal the oxide layer, helping improve the oxide film quality and further reducing the interface state charge. Thermal nitridation with nitric oxide (NO) can be performed in this annealing step. If plasma nitridation is required, the wafer is sent to another chamber designed for that process, before going through the annealing step. When the annealing step is finished, the heat lamp is turned off and the wafer starts to cool down. A robot in the transfer chamber picks up the still-hot wafer and sends it to the cool-down chamber before sending it to the cassette.

The gate dielectric has changed from commonly used silicon dioxide ($\kappa = 3.9$) to nitrated silicon oxide (SiON), and finally to a dielectric with high dielectric constant (high-κ), which allows a thicker gate dielectric layer for preventing gate leakage and gate dielectric breakdown. Atomic layer deposition (ALD) is the most commonly used method for depositing a high-κ dielectric, and RTA processing is used to improve film quality and reduce interface state charge.

RTO processing has been widely used for gate oxidation processes due to better process control, especially WTW uniformity control. Except for wet oxidation,

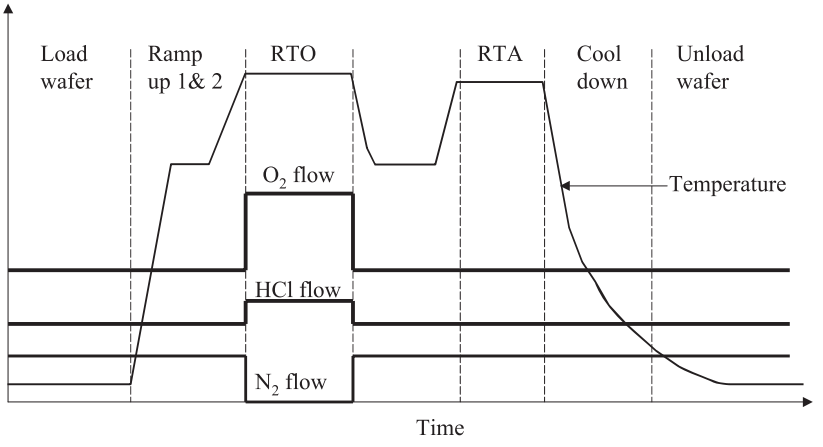


Figure 5.56 RTO processing diagram.

most oxidation processes for advanced IC chips are most likely to be performed in the RTO processing chamber due to better thermal budget control.

5.7.3 Rapid thermal chemical vapor deposition

Rapid thermal CVD (RTCVD) processing is a thermal CVD process performed in a single-wafer, cold-wall processing chamber with the capability of rapid temperature change and precise temperature control (see Fig. 5.57). Because it is a single-wafer system, the deposition rate has to be high enough that thin-film deposition can be finished within 1 to 2 min to achieve a reasonable throughput of 30 to 60 wafers per hour.

Figure 5.58 shows temperature changes of RTCVD and LPCVD processes. Compared with furnace LPCVD processing, RTCVD processing has better control of thermal budget and WTW uniformity. With device dimensions shrinking, the thickness of deposited thin films in the FEoL processes also decreases, typically from 100 to 2000 Å. With deposition rates from 100 to 1000 Å/min, single-wafer RTCVD processing becomes more attractive for FEoL thin-film deposition.

RTCVD processing can be used to deposit polysilicon, silicon nitride, and silicon dioxide, such as the CVD oxide used for trench-fill in shallow trench isolation processes. At high temperatures, tetraethoxysilane (TEOS) or $Si(OC_2H_5)_4$ can be used to deposit USG with high film quality and gap-fill capabilities.

In-situ process monitoring will become more important for better process control and higher throughput. The precise measurement of temperature, ±2 °C at 1000 °C, is the key to success for RTP. In-situ, real-time thickness measurement is required for oxidation and RTCVD processing endpoint detection.

A cluster tool consists of a mainframe and several processing chambers, and can integrate different processes into one system. It can improve throughput due to less interval time between different processes and achieve higher yield, because wafer transfer between different processes occurs in a vacuum environment, which reduces the chance of contamination. Figure 5.59 illustrates a cluster system for complete gate oxidation, nitridation, annealing, polysilicon deposition, and polysilicon annealing processes.

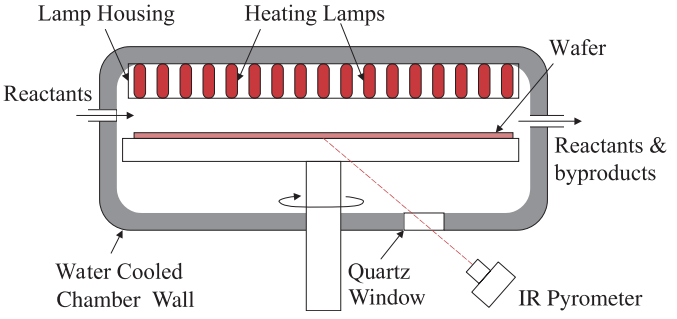


Figure 5.57 Schematic of RTCVD chamber.

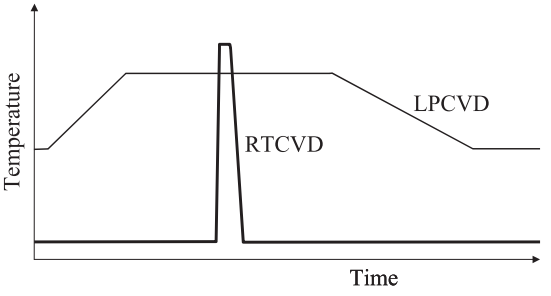


Figure 5.58 Temperature changes in RTCVD and LPCVD.

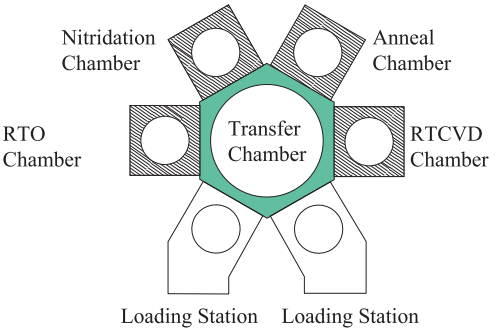


Figure 5.59 A cluster tool for gate oxide/polysilicon processes.

5.8 Recent Developments

Recently, there has been more focus on the development and application of rapid thermal processing (RTP), in-situ process monitoring, and cluster tools, while furnace processes are still used in noncritical thermal processes.

Junction depth becomes extremely shallow as devices grow smaller (less than 200 Å), making postimplantation annealing very challenging. Annealing ion implantation damage requires high temperatures; however, thermal budget requirements limit annealing time to the millisecond range. Spike, laser, and flush annealing techniques have been developed to meet the requirements of ultra-shallow junctions.

Spike annealing is an RTA process with a very short peak time, usually far less than 1 sec. It uses high peak temperatures to maximize dopant activation, a high temperature ramp rate to minimize dopant diffusion, and an aggressive temperature ramp-down rate to minimize dopant deactivation. Figure 5.60 illustrates the temperature variations of a spike annealing process. From the temperature curve, it is easy to see why this process is called spike annealing. Figure 5.61(a) is the simulation result of a dopant profile of a 0.13-μm pMOS with standard 1025 °C and 15-sec RTA, and Fig. 5.61(b) is the same device with 1113 °C spike annealing. It is evident that spike annealing significantly reduces dopant diffusion and becomes

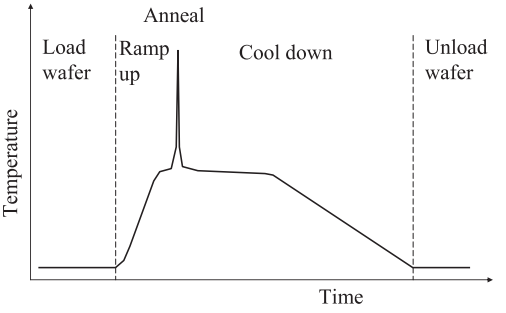


Figure 5.60 Temperature variations in a spike annealing process.

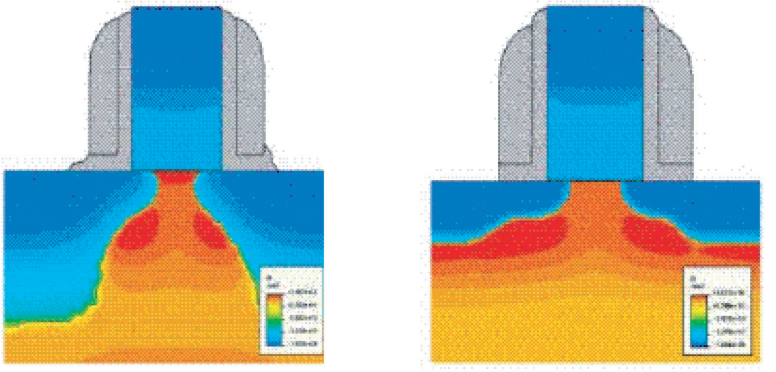


Figure 5.61 Simulated dopant profile of a 0.13-μm pMOS with a standard RTA at 1025 °C and 15 sec, and (b) spike annealing at 1113 °C. (E. Josse, et al., 2002)

more important for devices in nanometer technology nodes because the required junction depths become much thinner, and dopant diffusion during the annealing process must be minimized.

A laser annealing system uses energy from laser light to very quickly heat the wafer surface to the submelting point. Because of the high thermal conductivity of silicon, the wafer surface can be cooled down very quickly, in about a tenth of millisecond. Laser annealing systems can activate a dopant after ion implantation with minimum dopant diffusion and have been used in sub-45-nm process development. Laser annealing systems can be used with spike annealing systems to achieve optimized processing results.

Other postimplantation annealing technologies, such as flush annealing, have also been developed to achieve dopant activation with minimum diffusion for application of ultrathin junctions. Another technique is low-temperature microwave annealing, which uses microwaves to heat only the implantation-damaged area while the rest of the wafer stays in a low-temperature environment, thus achieving dopant activation without dopant diffusion.

5.9 Summary

- Thermal processing is a high-temperature process. In thermal processing, either a layer is being added onto the wafer surface (oxidation, deposition, and doping), or the chemical condition (alloying) or physical condition (annealing, diffusion, and reflow) of wafer materials is changed.
- Oxidation, annealing, and deposition are the three most important thermal processes.
- In the oxidation process, oxygen or water vapor reacts with silicon and forms silicon dioxide.
- It is very important to clean silicon wafer surfaces prior to the oxidation process. Contaminated silicon surfaces can provide nucleation sites that form a layer of silicon dioxide polycrystalline, which has very poor quality.
- Dry oxidation has a lower growth rate than wet oxidation. Dry oxide film quality is better than that of wet oxide. Thick oxide films, such as field oxide, usually use the wet oxidation process, while most thin oxide films use dry oxidation processes to grow.
- Diffusion processing was commonly used in doping processes of the IC industry. It uses silicon dioxide as a diffusion mask because most dopant atoms diffuse much slower in silicon dioxide than in single-crystal silicon.
- Diffusion processing consists of three steps: dopant oxide deposition, oxidation, and drive-in.
- Diffusion doping processes cannot independently control dopant concentration and dopant junction depth. Diffusion processing is an isotropic process; therefore, it always diffuses underneath the masking oxide. It had been replaced by ion implantation, which was introduced in the mid-1970s.
- Polysilicon and front-end silicon nitride depositions are LPCVD processes, which use a furnace with a vacuum system.
- After ion implantation, energetic ions cause extensive crystal structure damage; therefore wafers need postimplantation annealing to recover the single-crystal structure and activate dopants.
- RTP can ramp up temperatures at 50 to 250 °C/sec, compared to 5 to 10 °C/min in a furnace process. RTP has more control over thermal budget than does furnace processing.
- Postimplantation RTA is the most commonly used RTP. It is fast, can minimize dopant diffusion during the annealing process, and has excellent thermal budget control.
- Other RTP applications are RTA for dielectric annealing and silicide alloying, RTO, and RTCVD processes.
- RTP chambers possessing multiple controllable heating zones, in-situ process monitoring, and cluster tools are important trends for future IC fabrication thermal processes.
- Furnaces will continue to be used in future fabs for noncritical thermal processes due to their high throughput and low cost.

- To meet the requirements of diminishing device geometry, millisecond annealing techniques such as spike, laser, and flush annealing are being developed.

5.10 Bibliography

D. G. Baldwin, M. E. Williams, and P. L. Murphy, *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 2nd edition, OME Press, Beverly, MA (1996).

A. E. Braun, "Thermal Processing Options, Focus and Specialize," *SI* **22**(5), 56 (1999).

C. Y. Chang and S. M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

R. B. Fair, "Challenges in manufacturing submicron, ultra-large scale integrated circuits," *Proc. IEEE* **78**, 1687 (1990).

E. Josse, F. Arnaud, F. Wacquant, D. Lenoble, O. Menut, and E. Robilliart, "Spike anneal optimization for digital and analog high performance 0.13 μm CMOS platform," *Proc. Eur. Solid-State Dev. Res. Conf.*, 207–210 (2002).

J. M. Kowalski, J. E. Kowalski, and B. Lojek, "Microwave annealing for low-temperature activation of As in Si," *Proc. IEEE*, 51–56 (2007).

R. K. Laxman, A. K. Hochberg, D. A. Roberts, and F. D. W. Kaminsky, "Low temperature LPCVD silicon nitride using a chlorine-free organosilicon precursor," VLSI Multilevel Interconnection Conf., 1998, pp. 568.

SEMATECH, *Furnace Processes and Related Topics, Participant Guide* (1994).

R. Sharangpani, R. P. S. Thakur, N. Shah, and S. P. Tay, "Steam-based RTP for advanced processes," *Solid State Technol.* **41**(10), 91 (1998).

L. Shon-Roy, A. Wiesnoski, and R. Zorich, *Advanced Semiconductor Fabrication Handbook*, Integrated Circuit Engineering Corp., Scottsdale, AZ (1998).

See http://www.webelements.com for further information on the elements in this chapter.

5.11 Review Questions

1. List at least three thermal processes.
2. Describe a thermal oxidation process. Why does the oxide film grow into the silicon substrate during LOCOS formation?
3. For field oxide formation, which oxidation process is preferred, wet or dry? Explain your answer.
4. What are the advantages and disadvantages of a pyrogenic wet oxidation system compared with other wet oxidation systems?
5. In a pyrogenic wet oxidation process, why is the flow ratio H_2:O_2 slightly lower than 2:1?

6. List all gases used in a gate oxidation process, and explain the roles of each.
7. When temperature increases, how does the oxide growth rate change? What is the effect on the oxide growth rate when pressure increases?
8. Pad, barrier, gate, screen, and field oxides have been used in IC chip fabrication. Which is the thinnest and which is the thickest?
9. Why are furnaces commonly called diffusion furnaces even though they are not used for diffusion processes?
10. What are the advantages of a vertical furnace versus a horizontal furnace?
11. List three steps of the diffusion doping process.
12. Why can silicon dioxide be used as a diffusion mask?
13. What is junction depth?
14. Describe a titanium silicide process.
15. Why does a wafer need to be annealed at a high temperature after implantation? What are the advantages of using RTA processing in this application?
16. Describe a PSG reflow process. Can USG reflow be used? Explain the answer.
17. List processing gases that can be used for p-type-doped polysilicon LPCVD processing.
18. In LPCVD nitride deposition processes, why is ammonia used as a nitrogen source instead of nitrogen gas?
19. What is the temperature ramp rate of an RTP system? What is the rate of a furnace? Why can an RTP system ramp up temperature faster than a furnace?
20. Compared with the RTP system, what are the advantages of furnace systems?
21. What is the difference between spike annealing and RTA?
22. What is the driving force to develop millisecond annealing techniques?

Chapter 6
Photolithography

Objectives

After finishing this chapter, the reader will be able to:

- list the four components of photoresist
- describe the difference between positive and negative photoresists
- describe a photolithography processing sequence
- list four alignment and exposure systems
- identify the most commonly used alignment and exposure system in IC production
- describe the wafer movement in a track-stepper integrated system
- explain the relationship between resolution and depth of focus with wavelength and numerical aperture
- list at least three candidates for next-generation lithography.

6.1 Introduction

Photolithography is a patterning process that transfers a designed pattern from a mask or reticle to the photoresist on a wafer's surface. It was first used in the printing industry and has long been used to make printed circuit boards (PCBs). Photolithography has been a part of the semiconductor industry for transistor and integrated circuit manufacture since the 1950s. It is the most crucial processing step in IC fabrication because device and circuit designs are transferred to wafers by either etch or ion implantation by means of a pattern; that pattern is defined on a photoresist on a wafer's surface by the photolithography process.

Photolithography is the core of the IC manufacturing process flow, as shown in Fig. 6.1. From a bare wafer to bonding pad etch and photoresist strip, the simplest MOS-based IC chip needs five photolithography processing or masking steps; an advanced IC chip can take more than 30 masking steps. IC processing is very time consuming: even with a 24-7 nonstop working schedule, it takes six to eight weeks to go from bare wafer to finished wafer. The photolithography process takes about 40 to 50% of total wafer processing time.

The requirements of photolithography are high resolution, high sensitivity, precise alignment, and low defect density. The development of IC processing

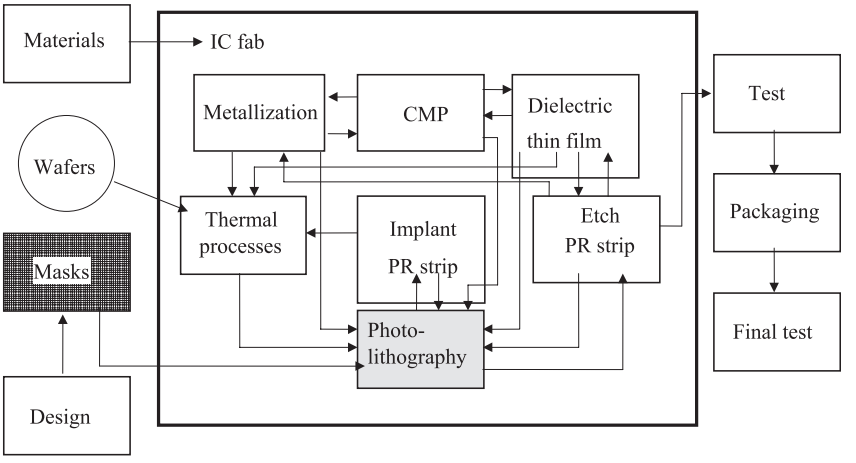


Figure 6.1 IC fabrication process flow.

technology is measured by the half-pitch (HP) of the minimum feature size on the production wafer, which was 25 nm (in 2010), as shown in Fig. 6.2. The smaller the HP, the more devices one can make on a wafer. 25-nm technology in 2010 was limited by photolithography resolution and patterning technology. By improving photolithography resolution and patterning technology, the HP can be further reduced. A photoresist must be very sensitive to exposure light to achieve reasonable throughput because the higher the photoresist sensitivity, the shorter the exposure time, and the higher the throughput. However, if the sensitivity is too high, other photoresist characteristics can be affected, such as line edge roughness (LER) and line width roughness (LWR), and those factors in turn affect the resolution. Thus, a careful balance between resolution and sensitivity is required. An advanced IC chip needs more than 30 patterning process steps; each one must be precisely aligned with the previous one to achieve a successful pattern transfer for the entire chip design. There is very little room for alignment error, since maximum tolerance is 10 to 20% of the CD. For 25-nm technology, misalignment needs to be controlled to within 2.5 to 5 nm, which is why automatic alignment systems are required. Advanced photolithography systems are very challenging systems, since everything must be precisely controlled. For example, a 1 °C difference on a 300-mm wafer can cause a 0.75-μm difference in wafer diameter due to silicon thermal expansion (or contraction when the temperature is lower) at a 2.5×10^{-6}/°C expansion (or contraction) rate. The photolithography process has to control defect density, since defects introduced in this process are transferred to devices and circuits via succeeding etch or ion implantation processes, affecting the yield and reliability of the products.

The photolithography process can be subdivided into three main operations: photoresist coating, alignment and exposure, and photoresist development. First, the wafer is coated with a thin layer of photosensitive material, called photoresist, which is exposed by ultraviolet (UV) light through a mask or reticle, with the

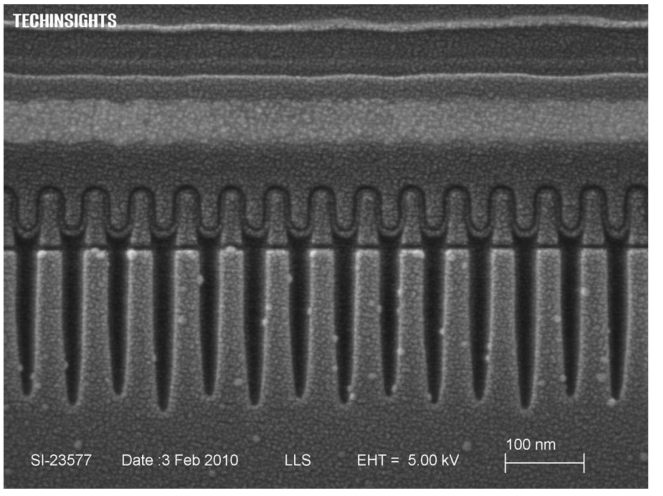


Figure 6.2 Cross-section SEM image of the IM Flash Technologies 25-nm flash memory array along the wordline direction. (Image from www.semiconductorblog.com)

pattern of transparent and opaque areas generated by a mask writer. The chemistry of the exposed photoresist is changed by photochemical reactions under the transparent areas where UV light passes through. For a positive photoresist, which is most commonly used in advanced semiconductor fabs, the exposed area is dissolved in developer solution, leaving the unexposed photoresist on the wafer surface, which reproduces a dark pattern on the mask or reticle.

6.2 Photoresist

Photoresist is a photosensitive material used to temporarily coat a wafer and then transfer an optical image of a chip design onto a mask or reticle to the wafer surface. It is very similar to the photosensitive material that is coated onto the plastic of photographic films, which can transfer optical images from a camera lens onto the film's surface. But unlike these photosensitive coatings on films, photoresists are not very sensitive to visible light, nor are they required to be sensitive to changes in color. Since it is mainly sensitive to UV light and is insensitive to visible light, photolithography processing does not need a darkroom. Because photoresists are insensitive to yellow light, all semiconductor fabs use yellow lights to illuminate photolithography areas, commonly called photo bays.

There are two kinds of photoresists, positive and negative. For the negative photoresist, the exposed parts become cross-linked and polymerized due to a photochemical reaction, then harden and remain on the wafer surface after development; the unexposed parts are dissolved by the developer solution. Positive photoresists normally are cross-linked and polymerized before exposure. After exposure, the exposed photoresist changes from an aqueous-base-insoluble material to an aqueous-base-soluble material in a photochemical reaction called

photosolubilization and is dissolved by the developer; the unexposed parts remain on the wafer surface.

Figure 6.3 illustrates the two different kinds of photoresists and their pattern transfer processes. The image of the positive photoresist is the same as the image on the mask or reticle, and the image of the negative photoresist is the reversed image on the mask or reticle. Before digital cameras, a roll of photographic film had to be purchased and loaded into a camera; that film was a negative film. The images left on negative film after development are the reversed images of those images the camera has taken; they require exposure and development again on negative photographic papers to print normal images. For a higher price, a roll of positive film can be purchased that gives the same images taken by the camera after development. Positive films are normally used to make the slides used for slide shows.

Presently, most advanced semiconductor fabs use positive photoresist because it can achieve the high resolution required for nanometer feature sizes. Photoresist has four basic components: polymer, sensitizer, solvent, and additives.

Polymer is the solid organic material that adheres to the wafer surface and withstands etch and ion implantation processes for the masking of the pattern transfer process. It is formed by organic compounds, which are carbon-hydrogen molecules (C_xH_y) with complicated chain and ring structures. A commonly used positive photoresist polymer for 248- and 193-nm photoresists is polymethylmethacrylate (PMMA). Phenol-formaldehyde or novolac resin are commonly used polymers for photoresist of i-line (365-nm) and g-line (436-nm) lithography. The most commonly used negative photoresist polymer is polyisoprene rubber.

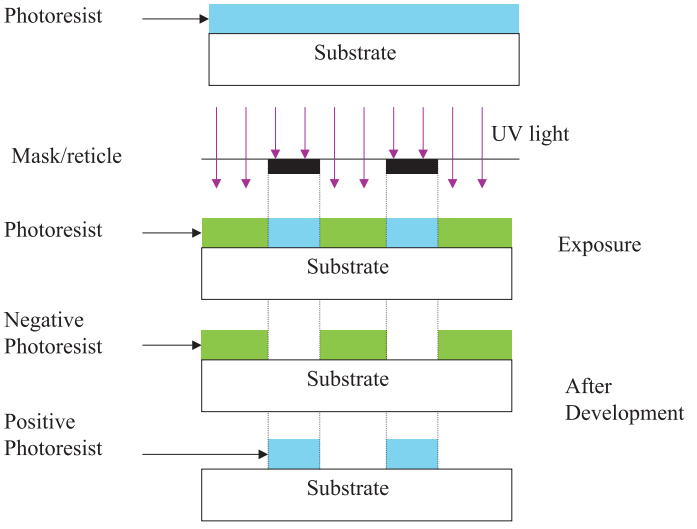


Figure 6.3 Patterning process with negative and positive photoresists.

A sensitizer is an organic compound with very high photoactivity that controls and modifies the photochemical reaction of a photoresist during exposure. The sensitizer for a positive photoresist is a dissolution inhibitor, which is cross-linked within the resin. During the exposure process, energy from the light dissociates the sensitizer and breaks down the cross-links causing the exposed resin to become soluble in an aqueous developer solution. The sensitizer for a negative photoresist is an organic molecule containing the N_3 group. Exposure to UV light liberates N_2 gas, forming free radicals that help cross-link the rubber molecules. The cross-linking chain reaction polymerizes the exposed areas, which have greater bonding strength and higher chemical resistance.

Solvent is the liquid that dissolves the polymer and sensitizer. It makes photoresist easy to apply on a wafer surface as a thin film of 0.5 to 3 μm thickness. The solvent thins the photoresist to allow application of thin layers by spinning, analogous to painting. Before the spin-coating process, about 75% of the photoresist is solvent. Positive photoresists use acetate-type solvents, and negative photoresists use xylene (C_8H_{10}).

Additives control and modify the photochemical reaction of photoresists during exposure to achieve optimized lithography resolution. Dye is a commonly used additive for both positive and negative photoresists.

For a negative photoresist, the developer solution is mainly xylene. The developer solution dissolves the unexposed photoresist, but often some of the developer solvent is absorbed in the exposed, cross-linked photoresist. This causes the photoresist swelling effect, which distorts pattern features and limits resolution to about two to three times that of the thickness of the photoresist. Negative photoresists were widely used in the semiconductor industry before the 1980s, when minimum feature size was larger than 3 μm. They are no long commonly used in advanced semiconductor fabs due to their poor resolution. Positive photoresists do not absorb developer solvents. They can achieve much higher resolution and are widely used for photolithography processes. Figure 6.4 illustrates the resolution of negative and positive photoresists.

Question: Positive photoresists can achieve much higher resolution than negative photoresists. Why were they not used before the 1980s?

Answer: Positive photoresist is more expensive than negative photoresist; therefore, manufacturers used negative photoresist until it had to be

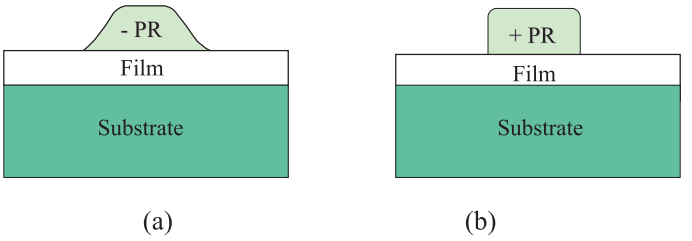


Figure 6.4 (a) Negative and (b) positive photoresists (PRs).

replaced when the minimum feature size shrank to smaller than 3 μm. Photoresist is one of the most important materials used in the photolithography process. Unless it is absolutely necessary, engineers are extremely reluctant to switch from one photoresist to another.

To pattern a small feature, shorter-wavelength exposure light is required. Photolithography processes using DUV (248 nm) or ArF (193 nm) require photoresist that is different from photolithography processes that use mercury g-line (436 nm) and i-line (365 nm). This is because the intensity of DUV light sources, usually excimer lasers, is much lower than that of mercury lamps. Chemically amplified photoresists are developed for DUV photolithography processes in 0.25-μm or smaller feature patterning applications. A catalysis effect is used to increase the sensitivity of this kind of photoresist. A photoacid is created in the photoresist when it is exposed to DUV light. In the postexposure bake (PEB) process a wafer is heated, and this heat drives acid diffusion and amplification in a catalytic reaction, as shown in Fig. 6.5.

To achieve completed pattern transfer, the photoresist needs to have good resolution, high etch resistance, and good adhesion. High resolution is the key to achieving successful pattern transfer. Without high etch resistance and good adhesion of the photoresist, the next etch or ion implantation process will most likely fail to meet the processing requirements and cause intolerable error. The thinner the photoresist film, the higher the resolution will be. However, it is also true that the thinner the photoresist film is, the lower the etching and ion implantation resistance. There is always a tradeoff between these two conflicting requirements.

The latitude of the photoresist process is tolerance of the photoresist to varying spin rates baking temperature, and exposure flux. The wider the process latitude, the more stable the process. This is an important factor when selecting a photoresist for a given process technology.

6.3 Photolithography Process

The photolithography process includes three major steps: photoresist coating, exposure, and development. To achieve high resolution, photolithography also has many baking and chilling steps. For older, all-manual processing technologies,

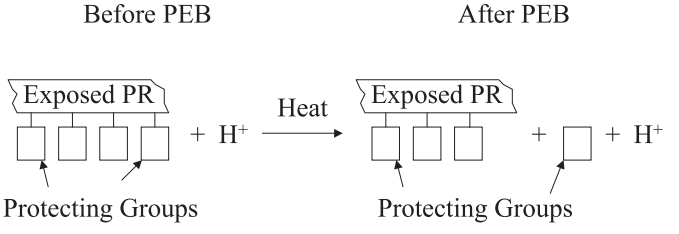


Figure 6.5 Chemically amplified photoresist.

the entire photolithography processing flow had eight steps: wafer cleaning, prebake, spin coating primer and photoresist, soft bake, alignment and exposure, development, pattern inspection, and hard bake. If the wafer failed to pass inspection, it would bypass the hard-bake step, the photoresist would be stripped, and the entire process would be repeated until the wafer passed inspection.

For advanced photolithography processes, three basic steps are the same. Detailed steps are added to achieve higher photolithography resolution. Integrated track-scanner systems are widely used to improve process yield and throughput. Because all of the coating, baking/cooling, exposure, and development processing steps are performed in a track-scanner system, and pattern inspection is performed after the hard bake. Figure 6.6 shows a flow chart of the photolithography process, and Fig. 6.7 gives an advanced photolithography processing sequence on the wafer surface.

6.3.1 Wafer cleaning

Before wafers reach the photolithography process, they have undergone some previous processes such as etch, ion implantation and annealing, oxidation, CVD, PVD, and CMP. The wafer might have some organic contaminants (from the photoresist, etch byproducts, bacteria, and debris of human skin) and inorganic contaminants (such as particles and residues from storage containers, improper wafer handling, and nonorganic materials in the environment, e.g., dust and mobile ions). The wafers need to be cleared of these contaminants before they are ready for the photolithography process. Even if the wafers are contamination free, it has been found that further cleaning is desirable for better adhesion of the photoresist to the wafer surface.

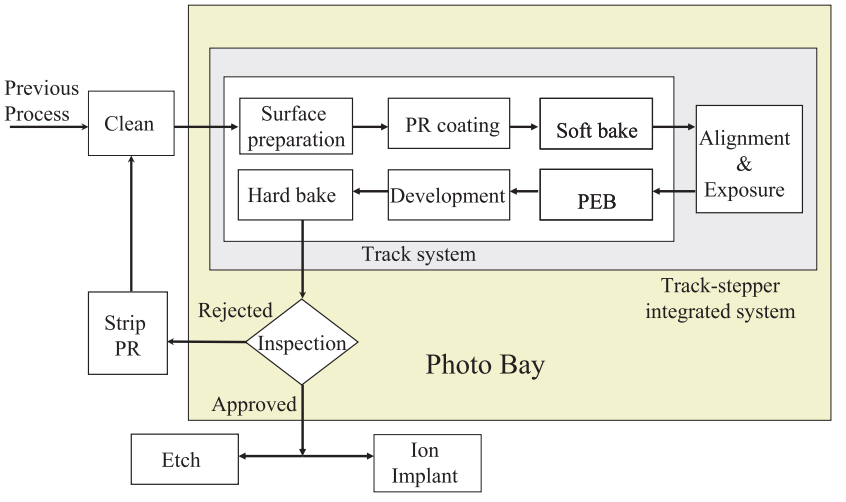


Figure 6.6 Flow chart of photolithography process.

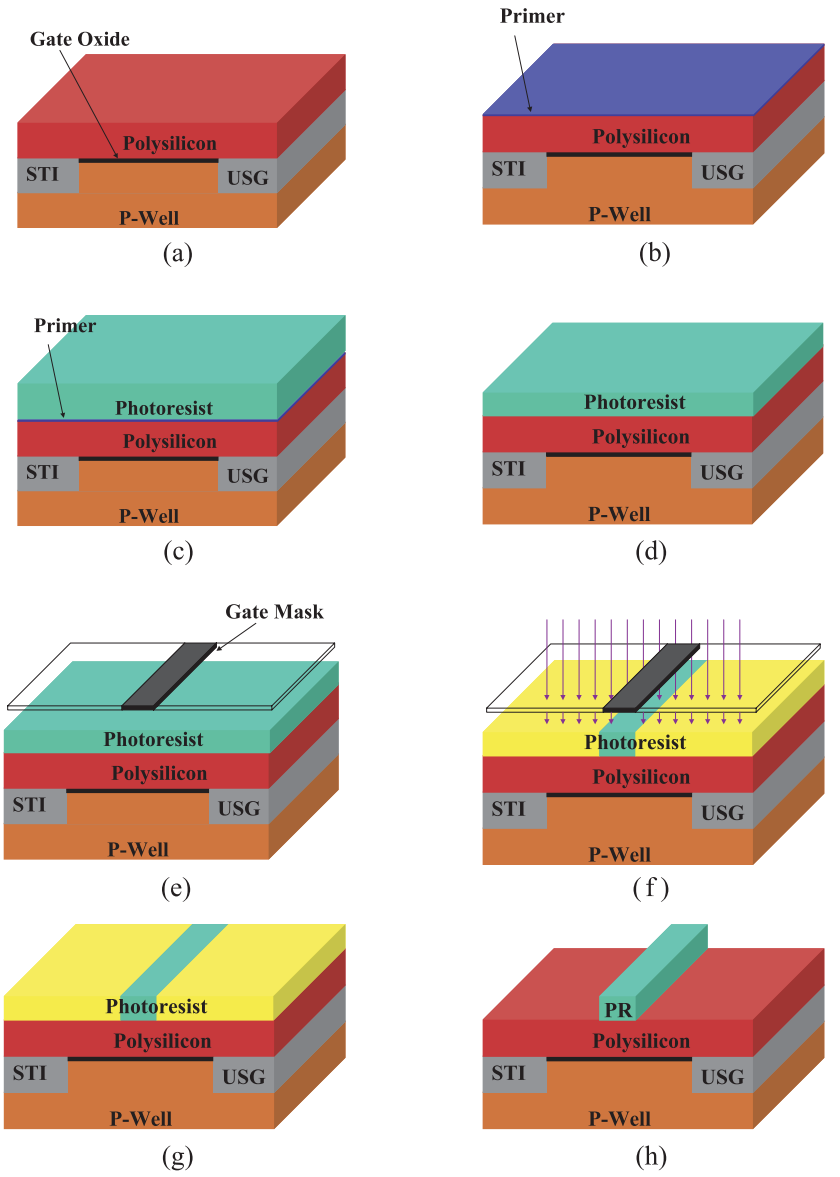


Figure 6.7 Photolithography process: (a) wafer cleaning, (b) prebake and primer vapor coating, (c) photoresist spin coating, (d) soft bake, (e) alignment, (f) exposure, (g) postexposure bake, and (h) develop, hard bake, and pattern inspection.

Chemical cleaning is most commonly used for wafer cleansing, which uses solvents and acids to remove organic and inorganic contaminant residues, respectively. It is usually followed by DI water rinse and spin dry processes, as illustrated in Fig. 6.8.

Other methods such as dry air or nitrogen blow, high-pressure steam blow, oxygen plasma ashing, and mechanical brushing have been used before and could still be used in some fabs. When feature size shrinks, killer particle size also shrinks. These methods worked well for the removal of larger particles, but they failed to remove smaller particles on smaller wafers, even adding more killer particles than they were able to remove.

Particles on a wafer's surface can cause pinholes in the photoresist; organic and inorganic contamination can cause photoresist adhesion problems and device and circuit defects. It is very important to minimize or eliminate these contaminants before the photolithography process begins to ensure process yield.

6.3.2 Preparation

Preparation is a two-step process performed in a sealed chamber commonly called a prep chamber. The first step is a heating process for the removal of moisture adsorbed on the wafer surface, called dehydration bake or prebake. A clean, dehydrated wafer surface is essential for adhesion of the photoresist on the wafer surface. Poor adhesion can lead to failure of photoresist patterning and can cause undercuts during subsequent etch processes. In most cases, the wafer is baked for about one or two minutes on a hot plate, with temperatures from 150 to 200 °C. Baking temperature and time are critical to achieving optimized processing results. If the baking temperature is too low or the baking time too short, insufficient surface dehydration can cause photoresist adhesion problems. If the baking temperature is too high, it can cause primer dissociation, which can cause contamination and affect photoresist adhesion.

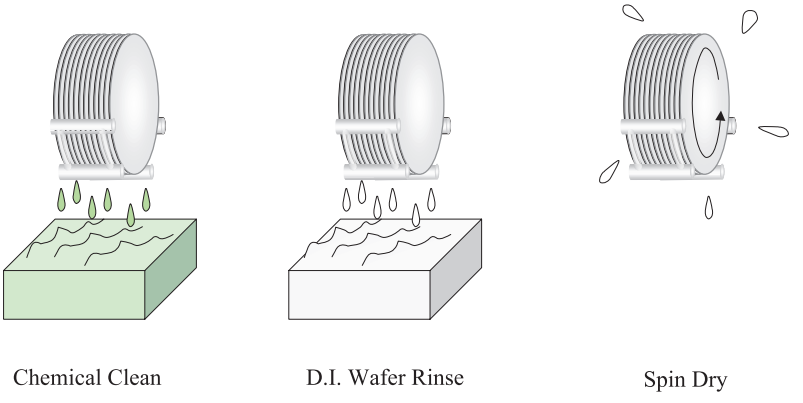


Figure 6.8 Wafer cleaning process.

The second step is a deposition process called priming. In this process, a thin layer of primer, which can wet the wafer surface and promote adhesion between the organic photoresist and the inorganic silicon or silicon compound wafer surface, is coated onto the wafer surface before the photoresist coating. Hexamethyldisilazane [HMDS, $(CH_3)_3SiNHSi(CH_3)_3$] is the most commonly used primer in IC photolithography processes. In advanced photolithography processes, HMDS is vaporized, introduced into the prep chamber, and deposited onto the wafer surface in situ with the prebake process. It is very important to coat the surface with photoresist immediately after the primer coating to prevent surface rehydration. Thus, the prep chamber is always placed in-line with the photoresist coater on a track system. Primer can also be coated by a spin-on process in situ with the photoresist coating, but this is not as popular as vapor priming in advanced IC fabs. Primer vapor coating is preferred over primer spin coating because it uses less HMDS and minimizes the potential for surface contamination by particulate carried in liquid chemicals. Figure 6.9 illustrates the in-situ prebake and vapor priming processes.

If the photoresist is still hot during photoresist spin coating, the solvents in the photoresist evaporate rapidly and cool down the wafer at the same time. This makes a very undesirable processing condition, since both solvent loss and temperature change can affect photoresist viscosity, which affect the thickness and thickness uniformity during the photoresist spin coating. Therefore, after the preparation process, the wafer needs to be cooled to the ambient temperature before applying the photoresist. Then the wafer is cooled on a chill plate, which is water cooled by a heat exchanger, on the same track system.

6.3.3 Photoresist coating

Photoresist coating is a deposition process in which a thin layer of photoresist is applied to the wafer surface. The wafer is placed on a spindle with a vacuum chuck that can hold the wafer during the high-speed rotation. Liquid photoresist

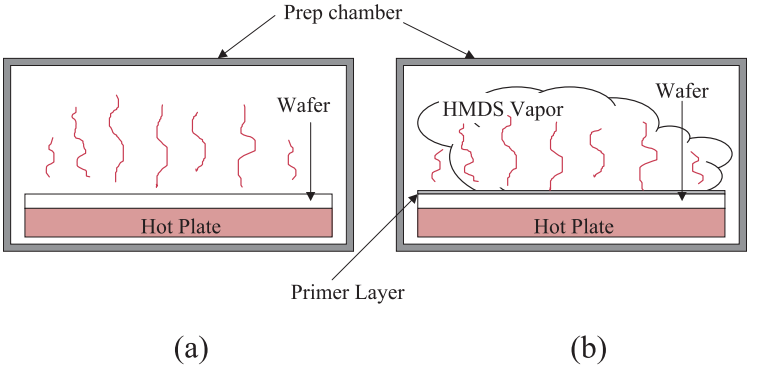


Figure 6.9 (a) In-situ prebake and (b) primer vapor coating.

is applied to the wafer surface, and the centrifugal force from the wafer rotation spreads the liquid across the entire wafer. After the solvents in the photoresist have evaporated, the wafer is left coated with a thin layer of photoresist. Photoresist thickness is related to both photoresist viscosity and wafer spin rate, as illustrated in Fig. 6.10. The higher the spin rate, the thinner the photoresist layer, and the better the thickness uniformity. The photoresist thickness is inversely proportional to the square root of the spin rate. Since photoresists have high viscosity and very high surface tension, a very high spin rate is required for uniform photoresist spin coating. The higher the viscosity, the thicker the photoresist films at the fixed spin rate. Viscosity of the photoresist can be controlled by the solid content of the photoresist solution. A typical photoresist thickness in photolithography processing is between 3,000 to 30,000 Å.

Photoresist can be dispensed with either a static or dynamic dispensing method. For static dispense, the photoresist is distributed onto a stationary wafer surface and allowed to spread across part of the wafer surface. When the photoresist puddle spreads to a certain diameter, the wafer is spun rapidly at a spin rate of up to 7000 rpm to distribute the photoresist evenly across the entire wafer surface. The thickness of the photoresist is related to photoresist viscosity, surface tension, photoresist drying characteristics, the spin rate, the acceleration rate, and the spin time. Photoresist thickness and thickness uniformity is especially sensitive to the acceleration rate.

For dynamic dispense, the photoresist is applied to the center of the wafer while the wafer is rotating at a low spin rate (around 500 rpm). After the photoresist is dispensed, the wafer is accelerated to a high spin rate of up to 7000 rpm to spread the photoresist uniformly across the wafer surface. The dynamic dispense method uses less photoresist; however, static dispense can achieve better

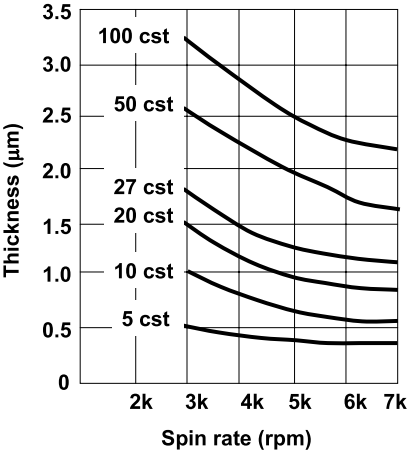


Figure 6.10 Relationship of photoresist thickness and spin rate at different viscosities (Integrated Circuit Engineering Corporation).

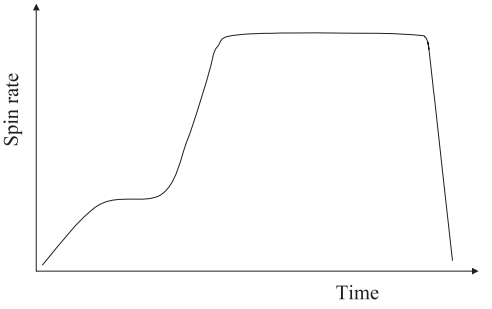


Figure 6.11 Spin rate change for spin coating processes.

photoresist coating uniformity. Figure 6.11 shows the change of spin rate in a dynamic dispense spin coating process.

Solvents in the photoresist evaporate rapidly during spin spreading processes, changing the photoresist viscosity. Therefore, it is very important to increase the spin rate as quickly as possible after the photoresist is applied to the wafer surface to reduce the effects of photoresist viscosity change due to solvent evaporation. In some processes, a thin layer of solvent is spin coated onto the wafer surface before the photoresist coating to improve photoresist adhesion and uniformity.

The photoresist spin coating process is illustrated in Fig. 6.12. The photoresist "suck back" feature is designed to prevent unwanted photoresist droplets from depositing onto the wafer after photoresist dispenses. If not removed, dry photoresist droplets could form at the tip of the dispense nozzle and fall onto photoresist-coated wafers, causing defects in subsequent processing steps.

Spin coating is also used for primer (HMDS) deposition just prior to the photoresist coating. First, liquid HMDS is applied to the wafer surface at a low spin rate to coat the wafer, then the spin rate is ramped up to 3000 to 6000 rpm to dry the HMDS in 20 to 30 sec. The advantage of the primer spin coating is that the primer spin coating occurs *in situ* as part of the photoresist coating process. This can effectively prevent rehydration of the wafer surface before photoresist coating. Currently, most fabs prefer primer vapor coating due to less primer usage (HMDS is very expensive), better coating uniformity, less risk of particulate contamination, and photoresist dissolving by the wet HMDS associated with liquid applications.

A schematic of a photoresist spin coater is illustrated in Fig. 6.13. The photoresist is drawn into the dispenser nozzle by means of a tube with a water sleeve in which water from a heat exchanger is used to maintain the photoresist at a constant temperature, since viscosity is related to temperature. Spin rate and spin rate ramp are precisely controlled; air flow temperature and air flow rate are also controlled in the coater, since they both affect the drying characteristics of the photoresist. The spindle is either nitrogen or water cooled to avoid wafer temperature nonuniformity, since the spindle at the center can become very hot during high-speed spins if it is not properly cooled. Excess photoresist and edge-bead removal (EBR) solution are collected and drained out of the station

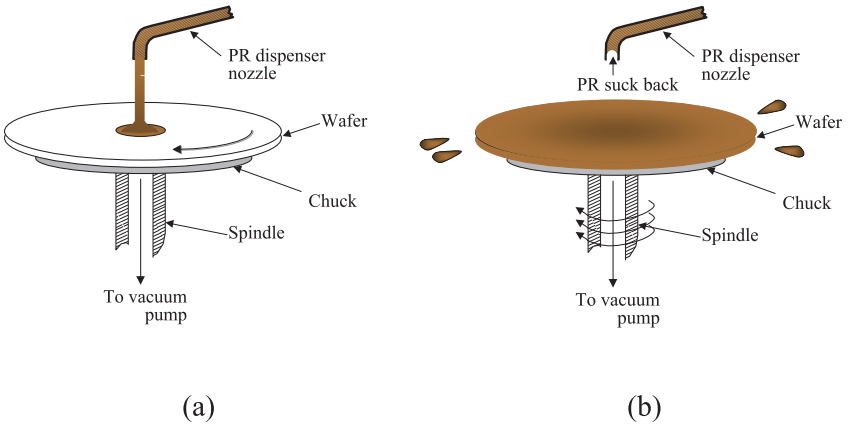


Figure 6.12 (a) Application of photoresist and (b) spin coating.

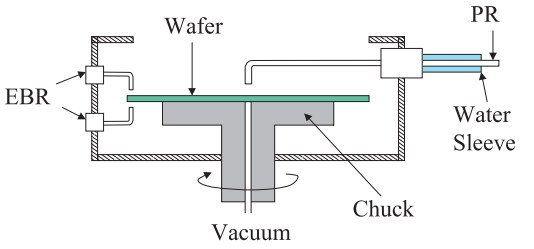


Figure 6.13 Schematic of a photoresist spin coater.

from the bottom, and vaporized solvent is evacuated from the exhaust. Photoresist thickness and uniformity are also related to exhaust gas temperature and gas flow rate, since they can affect a photoresist's drying characteristics. In fact, the ideal thickness uniformity of a photoresist is achieved without exhaust flow. However, without exhaust, the accumulation of solvent vapor fumes can be dangerous. Increasing the exhaust flow rate causes edge-thick photoresist profiles—faster drying near the edge of the wafer increases the viscosity of the photoresist, increasing thickness.

After photoresist spin coating, both sides of the wafer near the edge are covered by photoresist. It is absolutely necessary to perform EBR to eliminate this photoresist buildup. As the wafers move into the next process, e.g., etch or ion implantation, mechanical handlers such as robot fingers or wafer clamps can crack the photoresist buildup left on the wafer's edge and cause particulate contamination. A thick edge bead can also cause focusing problems during the exposure process. Both chemical and optical methods are used in EBR.

Chemical EBR is performed in the spin coater, after photoresist spin coating. In this process, solvents are injected onto the edge of both sides of the wafer while the

wafer is rotating; they dissolve the photoresist only at the edge and wash it away, as shown in Fig. 6.14.

Optical EBR is performed after the exposure process, in a specially designed station located on a track system, before the wafer is sent into developer. It uses a light source such as an LED to expose the top side of the wafer edge while the wafer is rotating. The exposed photoresist is removed during the developing process to ensure that there is no particle contamination during wafer clamping (this is especially important for some etch processes). Optical EBR processes are illustrated in Figs. 6.15(a) and 6.15(b).

Other photoresist coating methods, such as moving arm dispensers and roller coaters, could still be used in some less-advanced IC fabs; however, they are not as popular as spin coating.

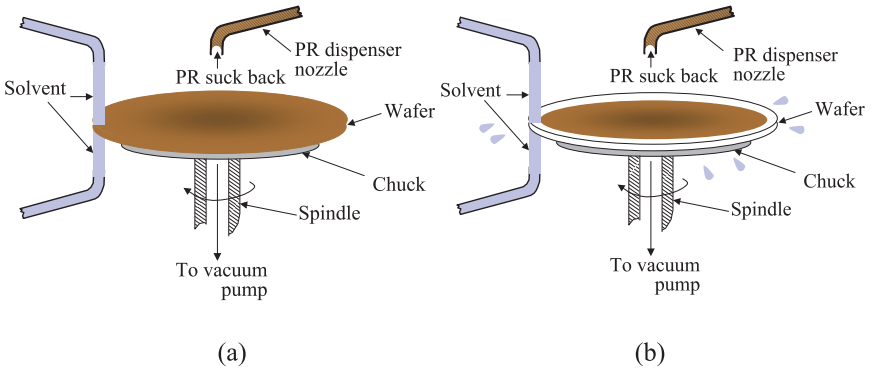


Figure 6.14 Chemical removal of an edge bead.

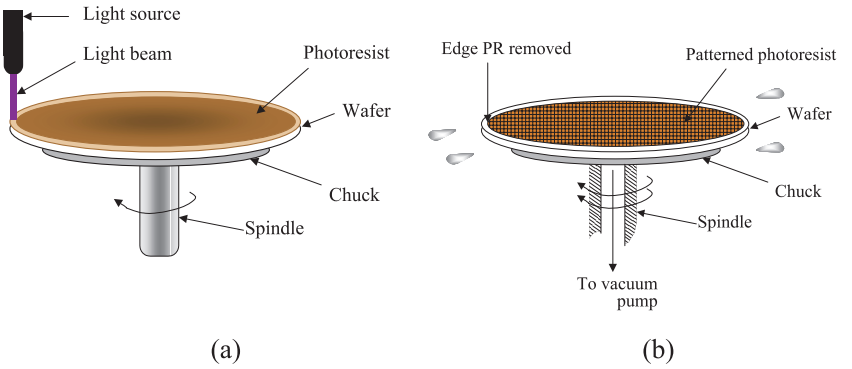


Figure 6.15 Optical EBR (a) after edge exposure and (b) after developer.

6.3.4 Soft bake

After photoresist coating, the wafer is again put through a thermal process to drive out the majority of solvent inside the photoresist and transform the photoresist from a liquid to a solid state. The soft bake also improves photoresist adhesion on the wafer surface; in some fabs it is called a pre-exposure bake. After soft bake, photoresist thickness shrinks about 10 to 20%, and the photoresist consists of about 5 to 20% solvent residue.

Soft-bake temperature and time depend on the type of photoresist (negative or positive) and vary with specific processes. Each photoresist has its own optimized baking time and baking temperature. If the photoresist is underbaked, either because the baking temperature was too low or because baking time was too short, the photoresist could peel off from the wafer surface during subsequent processes due to adhesion failure. Underbaked photoresist also can affect pattern resolution, first because excessive solvents in the photoresist cause insensitivity to light exposure, and second because of the microscopic shaking of the photoresist. Underbaking causes insufficient hardening, and a jelly-like photoresist can shake in a microscopic scale during wafer stepping, enough to create a blurred image on the photoresist (think of shooting a picture with a shaky camera).

Overbaking in the soft-bake step can cause premature polymerization of the photoresist, making it insensitive to light exposure. For chemically amplified photoresists used in DUV lithography processes, some residue solvent in the photoresist is needed for acid diffusion and amplification during the PEB step. Overbake can cause an insufficient catalysis chemical reaction, which could cause image underdevelopment.

There are several methods used for soft-baking processes: convection oven, infrared oven, microwave (MW) oven, and hot plate (see Fig. 6.16). A convection oven uses convection flow of heated nitrogen gas in an oven to heat a wafer at the required temperature, from 90 to 120 °C for about 30 min. An infrared oven can bake a wafer in a shorter time period than convection or MW ovens but can also burn the wafer bottom, since infrared radiation can pass through the photoresist layer and heat the wafer first, before the photoresist. A microwave oven is another method used for the baking process.

A hot plate can heat and dry photoresist on a wafer from the bottom up, since it heats the wafer first through heat transfer between the heated plate and the wafer. Thus, using a hot plate can avoid the photoresist crust problem due to surface heating and drying associated with convection oven baking processes. Unlike convection oven batch systems, a hot plate is a single-wafer system, which can achieve uniform heating within the wafer and, more importantly, stabilize processing results from wafer to wafer. Hot plates can be easily integrated into a track system for in-line coating, baking, and developing. Although it is possible that some less advanced IC fabs still use other methods, the hot plate is the most common method for all baking processes. It can be found in almost all photo-track systems in advanced IC fabs.

After the soft bake, the wafer is put on a chill plate to cool down to ambient temperature. It is very important to keep the wafer temperature constant during

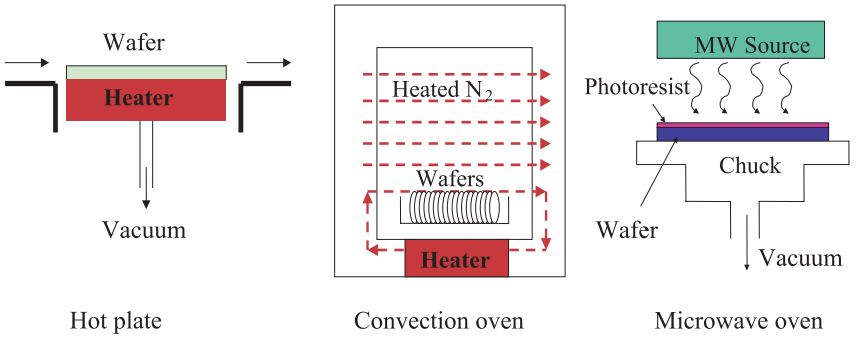


Figure 6.16 Different baking methods.

alignment and exposure, because 1 °C difference in temperature can cause 0.75-μm difference for a 300-mm silicon wafer due to thermal expansion effects.

6.3.5 Alignment and exposure

Alignment and exposure are the most critical steps of the photolithography process and for IC processing as a whole. Alignment and exposure determine the transfer success of the IC design pattern on the mask or reticle to the photoresist on the wafer surface.

The exposure process is very similar to taking a photograph with a camera: the patterned image on the mask or reticle is exposed to the photoresist on the wafer, just as an image is exposed on the film or image sensor inside a camera. Resolution of an IC photolithographic exposure system is much higher than a camera's, thus the IC photo exposure tool, called a stepper or scanner, is much more expensive than the most expensive photographic camera. In addition to the resolution requirement, precise alignment is also vital. Advanced IC chips have more than 30 mask steps, and each mask or reticle needs to be precisely aligned with the original alignment mark. Otherwise, the mask or reticle cannot successfully transfer the designed pattern to the wafer surface, causing device and circuit failures. Other requirements are high repeatability and reliability, high throughput, low defectivity, and low cost of ownership.

6.3.5.1 Contact and proximity printers

In the early years of the semiconductor industry, contact and proximity printers were widely used for the alignment and exposure process. A contact printer was the earliest and simplest tool. In the contact printing process, the mask makes direct contact with the photoresist on the wafer, and UV light passes through the clear pattern from the mask, exposing the photoresist underneath. A contact printer can achieve very good resolution, close to what can be achieved on the mask. However, due to the different curvatures of the mask and wafer, only a few points on the wafer actually make direct contact with the mask, and most of the area on the wafer's surface has about a 1- to 2-μm air gap between the mask and

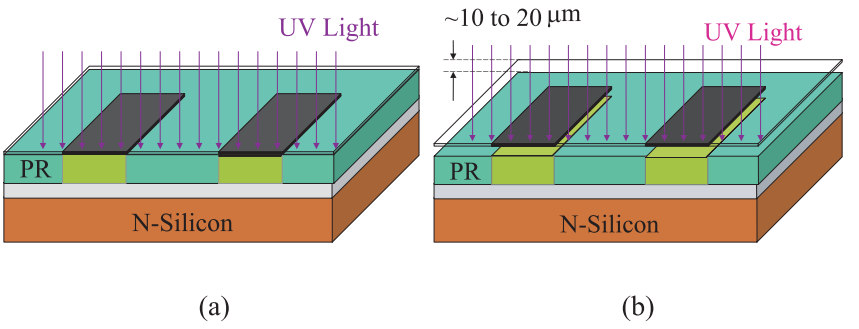


Figure 6.17 (a) Contact printing and (b) proximity printing.

photoresist. Despite this, the highest resolution of a contact printer can still be in the submicron range.

For every contact printing alignment and exposure, the touch and detach between the mask and photoresist can generate particles on both wafer and mask surfaces. Particles quickly accumulate on the mask surface and cause defects on the wafer due to both the particulate contamination and particle image transfer. A mask lifetime is seriously limited by particle contamination. To solve the particle generation problem, engineers adapted another approach that places the mask about 10 to 20 μm away from the photoresist. They called it a proximity printer. Since there is no direct contact, it has much less particle contamination and a much longer mask lifetime than that of a contact printer. The tradeoff is worse resolution, due to more light diffraction in a larger gap. The highest resolution a proximity printer can achieve is about 2 μm. Neither contact nor proximity printers are used in very large-scale integration (VLSI) or ultralarge-scale integration (ULSI) chip fabrication. Figure 6.17 illustrates exposure processes with contact and proximity printing systems.

6.3.5.2 Projection printer

To further improve exposure resolution while maintaining a low particle contamination level, the projection exposure system was developed and became widely used in VLSI semiconductor fabs, as shown in Fig. 6.18.

A projection system works in a manner very similar to an overhead projector. The mask is like a transparency foil, and the image is refocused on the wafer surface in a 1:1 ratio. In comparison, an overhead projector refocuses the image on a transparency foil to the screen in about a 1:10 ratio. In a projection exposure system, mask and wafer are far apart, which eliminates the possibility of particle generation from mask–wafer contact. By taking advantage of the optical characteristics of lenses and mirrors, projection exposure systems can achieve 1-μm minimum feature sizes and were widely used for VLSI device fabrication.

The most commonly used projection system is the scanning projection system (see Fig. 6.19), which uses a slit to block partial light from the light source to reduce light scattering and improve exposure resolution. The light is focused on

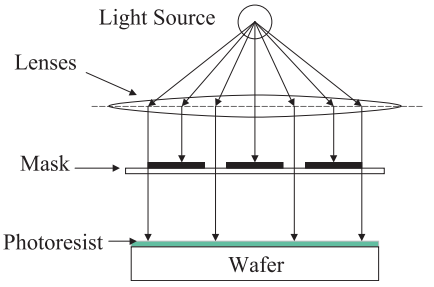


Figure 6.18 Schematic of a projection exposure system.

the mask by a lens and is refocused on the wafer surface by a projection lens as a slit. The mask and wafer move synchronously, allowing UV light to scan across the mask, refocus on the wafer surface, and expose the photoresist across the wafer.

6.3.5.3 Stepper/scanner

When feature size approaches the submicron level, a projection system can no longer meet the resolution requirements. Thus, step-and-repeat systems have been developed for VLSI and ULSI chip manufacturing.

In a projection system, image transfer from mask to wafer is 1:1, and a wafer requires only one exposure for the pattern transfer process. By shrinking and refocusing the image of the mask/reticle to the wafer surface at a ratio of 4:1 or 10:1, patterning transfer resolution can be improved. However, the exposure system must be completely redesigned because it would be impractical to make a mask and high-precision optical system with dimensions more than four to ten times the wafer diameter (which is 800 to 2000 mm for a 200-mm wafer) in order to expose the entire wafer in one exposure. It is also very difficult to find a UV light source

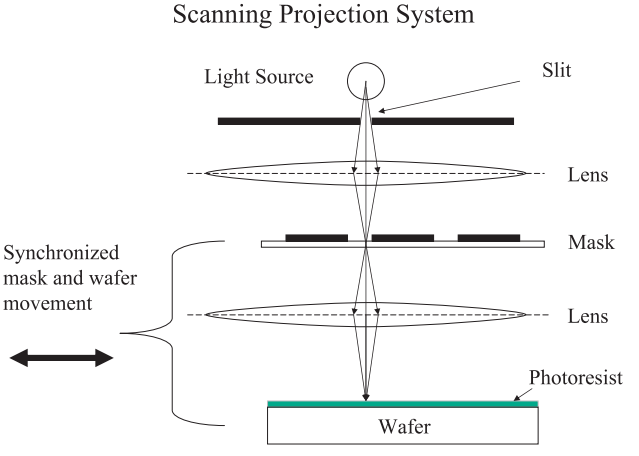


Figure 6.19 A scanning projection exposure system.

intense enough to do such an exposure. Therefore, the design pattern is made from a piece of chromium glass, called a reticle, which is used to expose only part of a wafer. In comparison, masks used in projection, contact, and proximity printers expose an entire wafer. Since the stepper shrinks the image from the reticle, the feature size on the reticle is much larger than that on the wafer surface. For instance, if the minimum feature size is 25 nm on the wafer surface, the minimum feature size on the reticle is 100 nm for a 4:1 shrink ratio. This is much easier to create than a 1:1 ratio photomask.

Question: Why is a 4:1 shrink ratio more popular in the semiconductor industry than a 10:1 shrink ratio?

Answer: Choosing between a 4:1 shrink ratio and a 10:1 shrink ratio involves a tradeoff between resolution and throughput. Obviously, 10:1 image shrinking will have better photolithography resolution than 4:1 image shrinking; however, it only exposes 16% of the area that is exposed by a 4:1 shrink ratio for a reticle. This means that total exposure time will be 6.25 times longer. Keep in mind that area changes as the power of the square of the dimension change, $A \propto d^2$.

Because a stepper only exposes a small portion of a wafer at a time, the exposure step must be repeated multiple times to expose the entire wafer. Figure 6.20 shows the basic structure of a stepper system and two steps of the exposure process. The stepper system is much more complicated than other optical exposure systems. For instance, the stepper needs to be aligned for every step, and each wafer needs 20 to 100 steps (determined by wafer size and product specification) to be entirely covered. In comparison, projection and contact/proximity printers only need one alignment per wafer. In submicron photolithography processes, there is very little room for alignment error. To meet throughput requirements, less than one second for each alignment and exposure step is allowed. Therefore, an automatic alignment system is required for stepper systems. A schematic of a stepper system is illustrated in Fig. 6.21.

After the wafer finishes photoresist coating and soft baking, it is sent into a stepper and put onto the wafer stage. It is first aligned with a previous alignment mark by a computer-controlled reticle and lens mechanism. For the very first photolithography step of the process, alignment is accomplished by using a notch or flat on the wafer, which is designed to indicate wafer crystal orientation and also serves as an alignment mark. The stepper usually adjusts the alignment optically with an automatic laser interferometer positioning system.

To further improve resolution of the image transfer, engineers combined scanning projection printer and stepper technology and developed step-and-scan systems, commonly called scanners, which are widely used in nanometer technology nodes of IC fabrication.

Because of high-precision, high-resolution, and high-throughput requirements for optical, mechanical, and electric systems, a scanner normally is the most expensive processing tool in a semiconductor fab. For example, an advanced 300-mm high-NA 193-nm immersion scanner can reach $30 to $40 million per system.

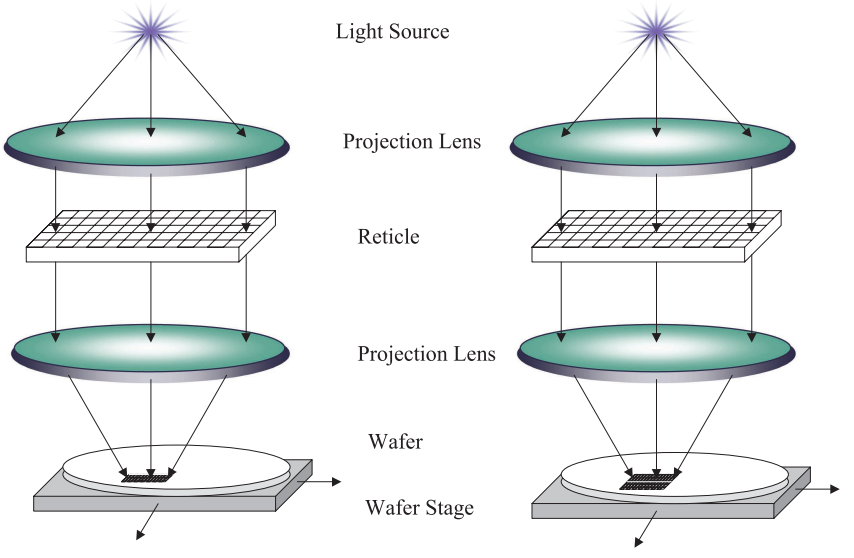


Figure 6.20 Step-and-repeat exposure system.

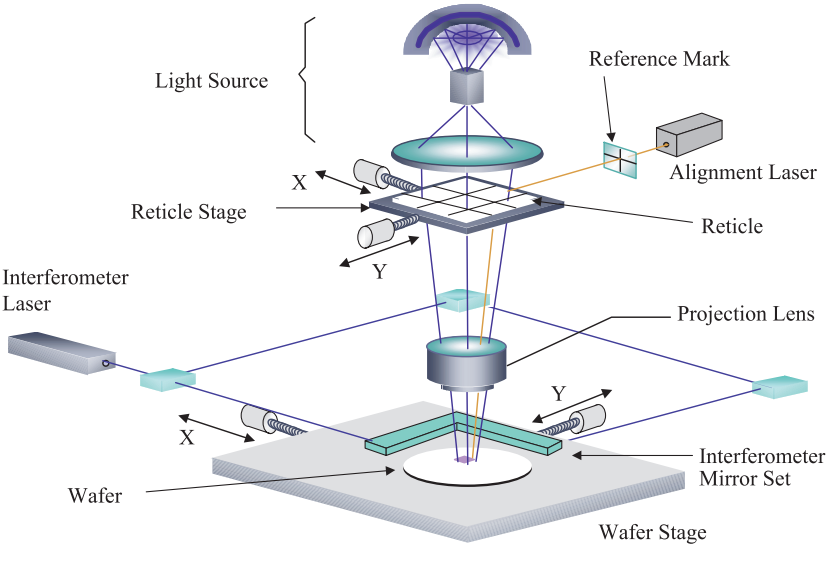


Figure 6.21 Step-and-repeat alignment and exposure system.

Many photoresists used for submicron semiconductor fabrication need to be exposed as soon as they have finished the soft bake; otherwise, the exposure resolution could be affected due to sensitizer decay in the photoresists. Therefore, in most fabs, scanners are integrated with coater-developer track systems.

6.3.5.4 Exposure light sources

The exposure process of photolithography is similar to the exposure of camera film. A picture taken in bright sunlight requires less exposure time and can achieve higher resolution than a picture taken in candlelight. Therefore, high-intensity light sources are important for achieving high resolution and high throughput. The wavelength of the UV light source used to expose the photoresist is a key factor in photolithography. Since certain photoresists are only sensitive to a narrow range of UV wavelengths, fabs choose the exposure light wavelength according to photoresist sensitivity and feature size of the circuit. The shorter the wavelength, the better the patterning resolution. When feature size shrinks, the wavelength of the exposure light also needs to shorten to match the patterning resolution requirement. There are two kinds of light sources widely used for photolithography processes: mercury lamps and excimer lasers. The exposure light source must be stable, reliable, of high intensity, and long in life.

For feature sizes larger than 2 μm, a broadband (multiwavelength) mercury lamp is used as the light source for contact/proximity and projection printers. When the feature size shrinks, single-wavelength light sources are required to achieve the desired resolution. High-pressure mercury lamps were the most commonly used UV light sources for submicron photolithography processes in projection systems and steppers in the 1980s and 1990s. The mercury UV lamp wavelength spectrum is given in Fig. 6.22. For photolithography exposure processes of IC chips with 0.50- and 0.35-μm feature sizes, g- and i-lines are most commonly used, respectively. These systems are still used in back-end processes of advanced IC fabs for layers so that their resolution can meet the requirements.

For photolithography processes with 0.25- and 0.18-μm minimum feature sizes, light sources with even shorter wavelengths are needed. A KrF excimer laser with a DUV wavelength of 248 nm is most commonly employed as a light source for steppers used in 0.25-μm processes, and it is capable of patterning feature sizes as small as 0.13 μm. Steppers using ArF excimer lasers with wavelengths of 193 nm have been used in IC production to pattern features since the 0.18-μm technology node and have extended to the 22-nm technology node, thanks to techniques such as immersion and double patterning. Research and development efforts on 157-nm scanners with F_2 excimer lasers were discontinued after the successful development of immersion technology, which extended applications of ArF 193-nm scanners for multiple technology nodes. Light sources commonly used for semiconductor photolithography are summarized in Table 6.1.

6.3.5.5 Exposure control

Exposure is controlled by exposure light intensity and exposure time. The total exposure light flux is the product of the intensity and exposure time and is very similar to the exposure of a camera.

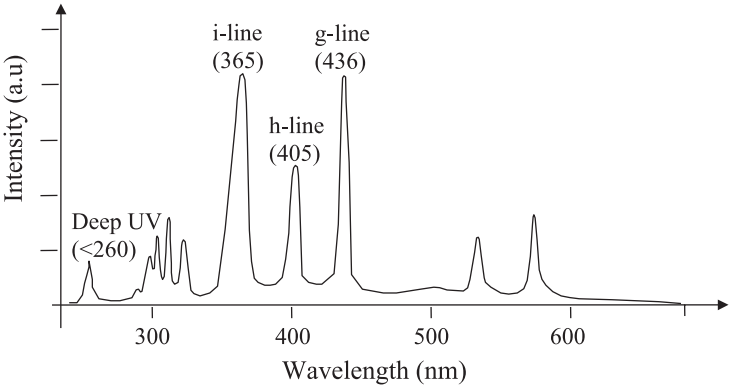


Figure 6.22 Light spectrum of the mercury lamp.

Table 6.1 Light sources for photolithography in fabrication, and research and development.

	Name	Wavelength (nm)	Application technology node (nm)
Mercury lamp	g-line	436	500
	h-line	405	
	i-line	365	350 to 250
Excimer laser	XeF	351	
	XeCl	308	
	KrF (DUV)	248	250 to 130
	ArF	193	180 to 14
Fluorine laser	F_2	157	No longer required
Laser-produced plasma (LPP) or discharge-produced plasma (DPP)	Extreme ultraviolet (EUV)	13.5	14 or less

Light intensity is mainly controlled by the electric power of a lamp or laser. By increasing applied power, the output light intensity can be increased; however, this increase can affect the reliability and lifetime of the lamp or laser.

A scanner must expose wafers with different photoresists and different reticles, so fab technicians need to be able to adjust a scanner's light intensity with precision. Even after a scanner runs the same process for a certain time period, light intensity can shift and cause patterning problems. Frequent light-intensity calibration is required to maintain a stable photolithography process. Illuminator intensity I is normally measured in mW/cm^2 by a photodetector. The amount of exposure is light intensity multiplied by exposure time, which is measured in mJ/cm^2.

Question: A routine illuminator intensity calibration was performed with a reticle still on the reticle stage. What kind of problem will it induce?

Answer: Because the reticle blocks some light from the illuminator, the photodetector on the wafer stage will receive fewer photons. Therefore, it will have a lower intensity than it should. To calibrate the light source to the required level, the applied power will be increased, and the light intensity will be too high. During the proceeding exposure process, the wrongly calibrated illuminator will cause overexposure of the photoresist. Similar to overexposed film, this will cause bad resolution. For a positive photoresist, overexposure can induce critical dimension (CD) loss. Often it is difficult to find the source of this kind of problem; however, a correct recalibration procedure will solve the problem.

6.3.6 Postexposure bake

When exposure light reflects off of the photoresist–substrate interface, it can interfere with incoming exposure light and cause a standing wave effect, due to constructive interference and destructive interference at the different depths. The standing wave pattern is illustrated in Fig. 6.23.

The standing wave effect causes striations of overexposed and underexposed areas throughout the photoresist, as shown in Fig. 6.24. The distance between two peaks is the wavelength of the exposure light λ divided by two times the refractive index of the photoresist $2n_{PR}$.

When feature size is large, the standing wave effect is not a big concern. As minimum feature size has continued to shrink, several methods have been used to reduce the standing wave effect. Dye has been added in the photoresist to reduce reflection intensity. Metallic and dielectric layers have been deposited on the wafer surface as antireflective coatings (ARCs) to reduce and minimize reflection. Another approach is using a spin coater to apply an organic antireflective layer prior to photoresist spin-on. Because the layer is at the bottom of the spin-on photoresist

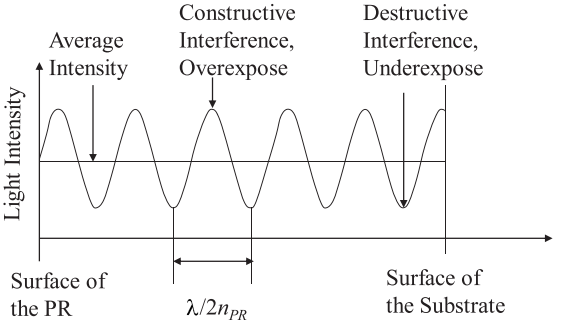


Figure 6.23 The standing wave effect induces light intensity changes.

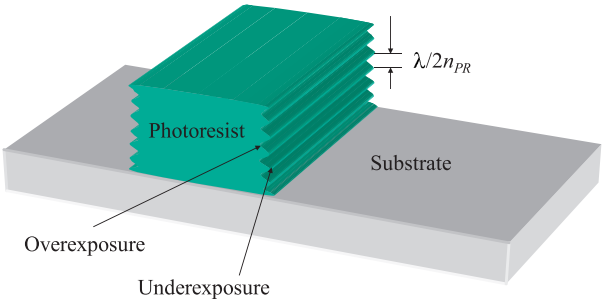


Figure 6.24 The standing wave effect on a photoresist.

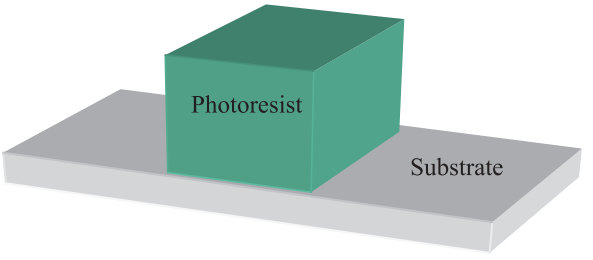


Figure 6.25 PEB minimizes the standing wave effect.

stack, it is commonly called bottom antireflective coating (BARC). PEB between exposure and develop processes also helps minimize the standing wave effect (see Fig. 6.25).

Photoresists have a special property called glass transition. When the temperature is above glass transition temperature T_g, photoresist molecules become more mobile. A baking process with temperatures higher than T_g provides the heat necessary for sufficient thermal movement of molecules in the photoresist. Thermal motion rearranges the over- and underexposed molecules, averaging and smoothing the standing wave effect, and improving photolithography resolution.

For chemically amplified photoresists used for DUV processes, PEB provides the heat necessary for acid diffusion and amplification. After the PEB process, images of the exposed areas appear on the photoresist due to the significant chemical change after acid amplification.

PEB processes use a hot plate at 110 to 130 °C for about 1 min. For the same type of photoresist, a PEB requires a higher baking temperature than a soft bake. Insufficient PEB will not completely eliminate a standing wave pattern, which affects resolution. On the other hand, overbaking causes polymerization of the photoresist and affects the developing process, which can result in failure of the pattern transfer. After PEB, the wafer is placed on a chill plate to cool down to the ambient temperature before continuing to the development process.

6.3.7 Development

After the photoresist-coated wafer proceeds through exposure, PEB, and for some processes, optical EBR, it is sent to a developer station for development. The development process removes unwanted photoresist and forms the desired pattern defined by the mask or reticle. For commonly used positive photoresist, the exposed portions are dissolved in the developer solution.

There are three steps in the development process: develop, rinse, and dry, as illustrated in Fig. 6.26. In the develop step, unwanted photoresist is dissolved by developer solution. A rinse dilutes the developer solution and prevents overdevelopment, and the drying process prepares the wafer for the next process.

Positive photoresists use a weak base solution as the developer solution. Alkaline water-based solutions such as sodium hydroxide (NaOH) and potassium hydroxide (KOH) can be used. However, these solutions can introduce mobile ions such as sodium and potassium, which are very undesirable since they can cause device damage. Therefore, most semiconductor fabs use nonionic base solution for positive photoresist developments. The most commonly used is tetramethyl ammonium hydroxide [TMAH, $(CH_3)_4NOH$].

The most commonly used developer solution for negative photoresists is xylene. Usually, *n*-butylacetate is used for the rinse. Mixtures of alcohol and trichloroethylene (TCE) and milder-acting Stoddart solvent can also be used for a negative photoresist rinse.

Development was previously performed in batch immersion processes in wet sinks, as illustrated in Fig. 6.26. Currently, most fabs use a spinner developer for this process. The main advantage of the spinner is that it can run the developing, rinsing, and drying processes in situ and can be easily integrated into a track system with a coater and baker. A spin developer system is very similar to a spin coater.

Figure 6.27 shows a schematic of a spin developer system. The development process is a chemical process that is sensitive to temperature. Therefore, both

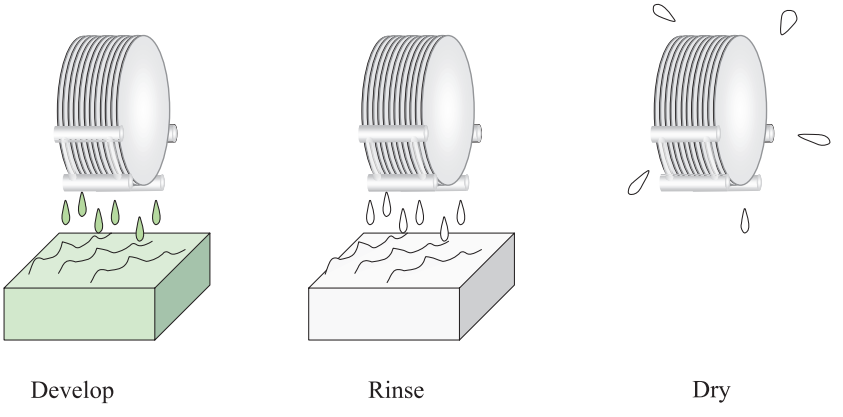


Figure 6.26 Three steps of the development process.

photoresist and wafer temperatures need to remain constant during the process. During the development process, higher temperatures can cause higher chemical reaction rates, which can induce photoresist overdevelopment and cause CD loss. Lower temperatures can cause lower chemical reaction rates, which can induce photoresist underdevelopment and cause CD gain. Both cases affect optimal photolithography resolution (see Fig. 6.28).

First, the developer solution is sprayed onto the wafer surface and spread across it by centrifugal force from the spinning. After development has finished, DI water is sprayed onto the wafer surface for rinsing, then the DI water is turned off and the spin rate increases to dry the wafer. The process is illustrated in Fig. 6.29.

Another method is called puddle development, which is very similar to spray development. Although it uses the same type of spin system, it starts with a certain amount of developer solution being sprayed onto a stationary wafer surface rather than a spinning one. The developer forms a puddle and covers the entire wafer due to surface tension. After required puddle time and the majority of the development process is finished, more developer solution is sprayed on the wafer as it starts to spin, washing away the dissolved photoresist; then the wafer is rinsed and dried at a high spin rate.

During the development process, both exposed and unexposed photoresist are dissolved in the developer solution. Selectivity between the two needs to be high enough to achieve good resolution. For the developing process, temperature

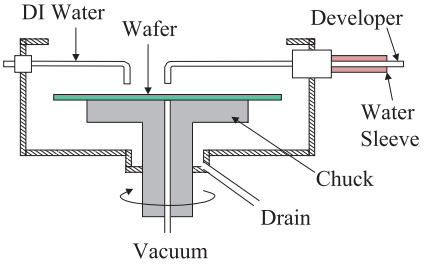


Figure 6.27 Schematic of a spin developer system.

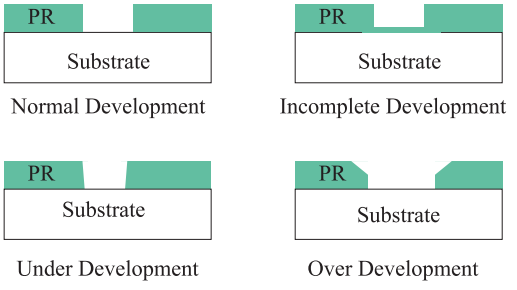


Figure 6.28 Photoresist profile for different developments.

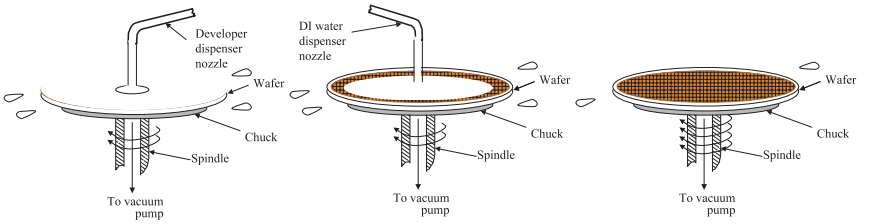


Figure 6.29 Developer spin-on, DI water rinse, and spin-dry processing sequence.

controls for the developer solution and wafer are very important. Different photoresists use different developer solution and require different development temperatures.

6.3.8 Hard bake

After development, the wafer is put through a bake process called hard bake. Hard bake drives out any remaining solvents and strengthens the photoresist, improves photoresist etch and implantation resistance due to further polymerization, and also improves adhesion of the photoresist to the wafer because of thermal dehydration. Several methods can be used for the hard bake process, as described earlier in the soft bake section. The most commonly used method is the hot plate, in which the temperature is usually from 100 to 130 °C and baking time is about 1 to 2 min, depending on the photoresist requirement. The hard bake temperature is higher than the soft bake temperature for the same kind of photoresist. For some applications, UV light is also used with a high-temperature (>100 °C) bake to harden photoresists.

Hard bake time and temperature need to be carefully controlled, since underbaking can cause high photoresist etch rates and affect wafer adhesion. Overbaking can cause bad resolution. If the baking temperature is too low, the photoresist will not achieve the required strength due to insufficient thermal polymerization and less photoresist thermal flow to fill the pinholes. It is important that the baking temperature is slightly higher than the photoresist transition temperature, so the photoresist flows a little bit to fill the pinholes and smooth the edges, as shown in Fig. 6.30. High temperature also can help to further dehydrate the photoresist and improve adhesion.

If the photoresist is overbaked (either the temperature is too high or the baking time is too long), it could flow too much, which will affect the resolution of the photolithography process, as shown in Fig. 6.31.

Question: When a photoresist has been used up or has been in use for too long, it needs to be replaced. If someone uses the wrong bottle, what could be the consequence?

Answer: Different photoresists have different sensitivities to certain wavelengths. They require different spin rates, spin ramp rates, and spin times. They also need different baking times and temperatures, and demand different

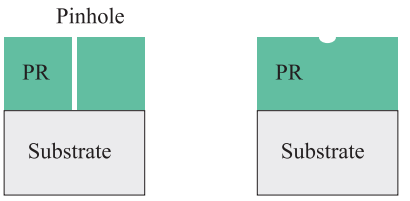


Figure 6.30 Pinhole filling by photoresist thermal flow.

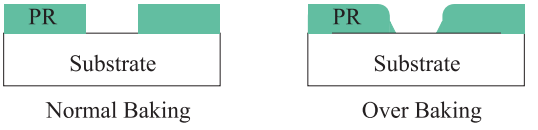


Figure 6.31 Overbake causes photoresist overflow.

exposure intensities and exposure times. They need different developer solutions and development conditions. Therefore, the pattern transfer will fail if the wrong type of photoresist is used. Using the wrong photoresist would significantly affect the throughput due to the required wafer rework and photoresist line clean and purge.

6.3.9 Metrology and defect inspection

In a production line of an IC fab, large numbers of wafers are processed with the highest throughput possible. Therefore, it is critical to capture processing shifts or defects on a wafer as soon as possible to quickly identify the problematic processing equipment causing the shift or defects. Then other wafers can be diverted from that equipment to avoid further loss.

After photoresist patterning, a wafer is processed through metrological and inspection steps to make sure that the controlled parameters of photoresist patterns, such as overlay, CD, and defect density, are within the tolerance limit. If one of them is out of control, the photoresist needs to be stripped and the wafers could be sent back for rework in the photolithography process. It is very important to measure and inspect a wafer before etch and ion implantation processes, since the pattern on the photoresist is only temporary, whereas after etch or ion implantation, it becomes permanent. It is impossible to rework a wafer after etch or ion implantation; if errors are found at that point, the wafer must be scrapped.

More than 30 masks are used to fabricate advanced CMOS IC chips. Overlay measurement is required for every photolithography step to ensure that each mask is accurately aligned. Figure 6.32 illustrates several kinds of misalignment cases, including run-out, run-in, reticle rotation, wafer rotation, misplacement in the x direction, and misplacement in the y direction. If misalignment is over the tolerance limit, the device will fail. For example, if the contact is misaligned with the metal

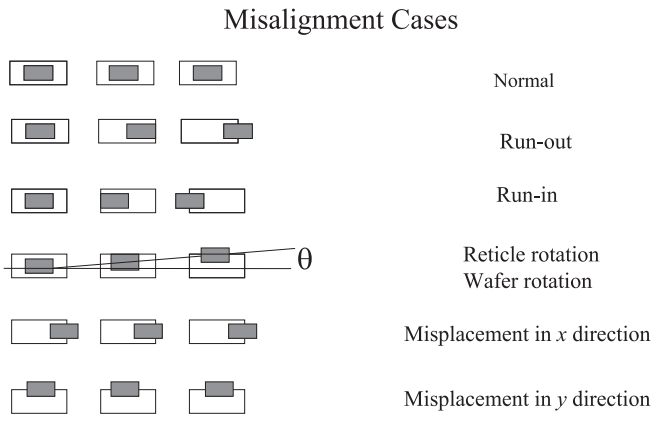


Figure 6.32 Examples of misalignment cases.

pad underneath, contact resistance could be too high and cause failure of the IC chip.

Overlay metrology uses optical measurements on designed alignment marks. The box-in-box alignment mark illustrated in Fig. 6.33 is one of the overlay patterns used in IC fabrication. The overlay shift in the x and y directions can be measured by:

$$\Delta X = X_1 - X_2 \text{ and } \Delta Y = Y_1 - Y_2.$$

Recently, grating-over-grating marks (see Fig. 6.34) have been developed for scatterometry-based overlay measurements in nanometer technology node IC manufacturing. Scatterometry is a technique that uses the spectral change of polarized light to measure pattern properties, which are described in detail later.

With the advancement of technology nodes, overlay metrology on test patterns within scribe lines might not be enough for proper yield management of the photolithography process. Overlay measurements of real devices after pattern etch using an SEM-based metrology system could be necessary in the future. For CD measurement, two systems are commonly used: an SEM-based metrology system called CD-SEM, and an optical system called scatterometry.

A wafer is inspected to determine whether the photolithography produced a desired and usable pattern on the photoresist. This is usually accomplished with microscopes (optical or electron) or in automatic inspection systems. An optical microscope can be used for visual inspection of larger-dimension features, while half-micron or submicron feature inspection requires an SEM.

Question: Can an optical microscope resolve a 0.25-μm feature?
Answer: No, because the feature size (0.25 μm = 2500 Å) is smaller than the wavelength of the visible light, which is from 3900 Å (violet) to 7500 Å (red).

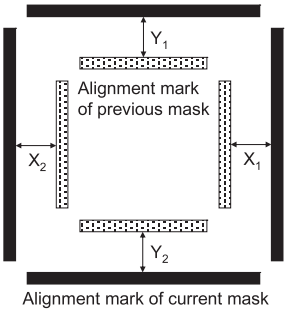


Figure 6.33 Box-in-box alignment mark.

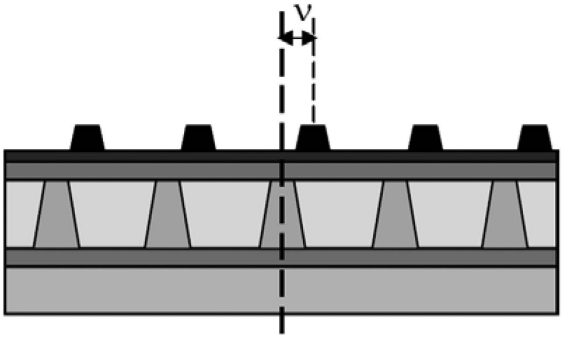


Figure 6.34 Grating-over-grating alignment mark for scatterometry overlay measurement. (B. Dinu et al., *Proc. SPIE*).

In his 1924 doctoral thesis, "Recherches sur la théorie des quanta (research on the quantum theory)," French physicist Louis de Broglie introduced a novel idea of the wave–particle duality of an electron; an electron is a tiny particle, and it is also a wave. His idea was immediately endorsed by Albert Einstein, who discovered the wave–particle duality of light; light is an electromagnetic wave and a particle called a photon. In 1927, experiments of electron diffraction of crystal lattices confirmed the particle–wave duality of an electron. In 1929, de Broglie became the youngest solo winner of the Noble Prize in physics. A material particle wave is called de Broglie's wave. The wavelength of an electron in de Broglie's wave is determined by the electron's momentum, which is related to its energy; the higher the electron's energy, the shorter its wavelength. Therefore, energetic electron beams can be used to detect very small features. When an electron beam hits material, it can excite secondary electrons. By mapping the detected electrical signal from these secondary electrons, an image of the tiny feature can be observed. Sharp edges on the photoresist, such as feature corners, have higher secondary electron yield rates and therefore higher intensity of the detected signal and a brighter image (see Fig. 6.35).

Scatterometry measures the optical signals of reflectometry or ellipsometry on measurement targets that consist of line/space arrays patterned in transparent or nonopaque films. Incident light strikes the target and causes an interaction of light reflection, diffraction, and refraction (see Fig. 6.36). Reflected light from the target contains information about phase and intensity that can be used (along with computational modeling) to reconstruct the film stack, pattern profile, and pattern CD of a feature on the measurement target. Because it measures the average CD value from an array, scatterometry CD measurements have very high repeatability.

Scatterometry CD measurements are performed on test structures on the scribe line, which are arrays of dense and isolated patterns, as shown in Figs. 6.37(a) and 6.37(b), respectively. CD-SEM measurements are performed on test structures called CD bars, as illustrated in Fig. 6.37(c). The CD value of the test structures usually are close to the CD value of the real device in that layer. There are

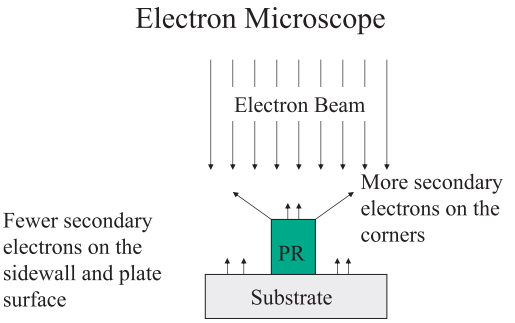


Figure 6.35 Electron beam and secondary electron distribution.

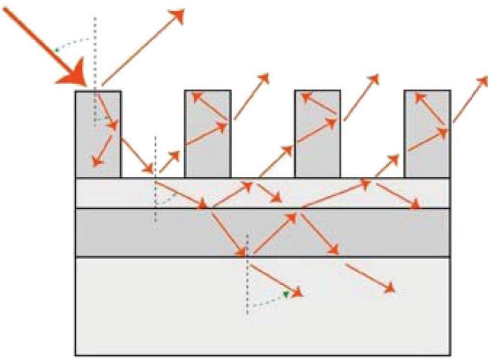


Figure 6.36 Light interaction of reflections, diffractions, and refractions. Scatterometry measurements based on this interaction can provide information of film stack, pattern CD, and profile information. (A. H. Shih)

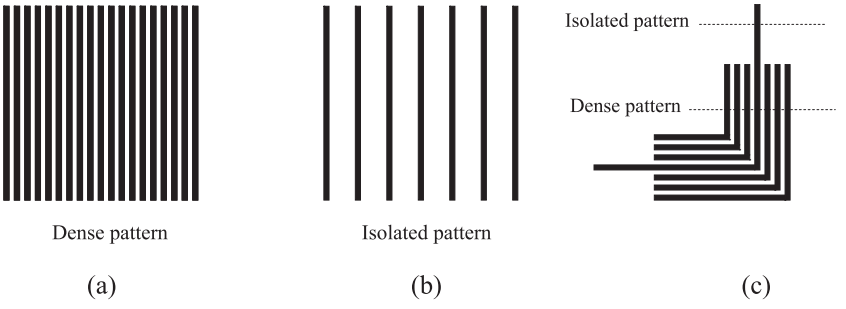


Figure 6.37 Test structures for (a) and (b) scatterometry, and (c) CD-SEM.

currently more requirements of design for manufacturing (DFM) due to smaller feature sizes. CD-SEM is widely used for design-based metrology applications, which directly measure the CDs on specific device patterns. Compared with CD-SEM, scatterometry is faster, more repeatable, and does less damage to the photoresist. On the other hand, CD-SEM shows pattern images and allows an engineer to see problems when they occur. CD-SEM can easily measure isolated devices with 2D structures that scatterometry cannot. Therefore, both scatterometry and CD-SEM coexist in advanced wafer fabs and will continue to coexist in the near future.

In advanced IC fabs, CD loss (or gain) causes many photolithography reworks. For critical layers such as the polysilicon gate of CMOS, less than 10% CD variation is allowed. Figure 6.38 shows some patterning issues in both cross-sectional views and top views that are similar to CD-SEM images. If after-development inspections (ADIs) find any density defects, overlays, or CD values outside of specifications, the photoresist can be stripped from the wafer surface for rework in the photolithography process. However, rework lowers throughput and affects yield, both of which always negatively affect a fab's profitability.

Bright-field optical inspection systems are used for defect inspection after photoresist development. A bright field inspection system projects high-intensity short-wavelength incident light onto the wafer surface and collects the reflected light on the image sensor to form an image of the patterns on the wafer. By comparing images of different dies at the same location, it can detect tiny defects by sensing the differences between each pixel in the images. This is called die-to-die (D2D) inspection. For chips with repeating patterns such as memory cell arrays, defects can be captured by comparing images of repeating cells, which is called cell-to-cell (C2C) inspection, or array mode inspection. Another method for detecting defects is comparing an inspection image to designed features extracted from design database files, which is commonly called die-to-database (D2DB) inspection. Figure 6.39 illustrates a bright-field defect inspection system.

If the wafers pass inspection, they move out of the photo bay and proceed to the next processing step, which is either etch or ion implantation.

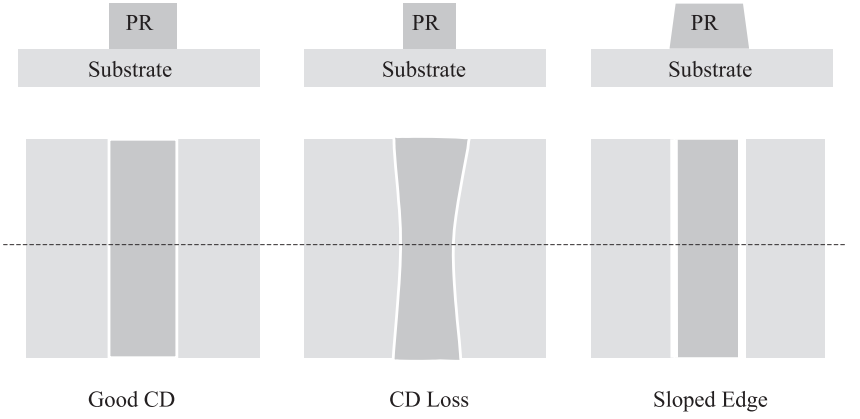


Figure 6.38 CD problems.

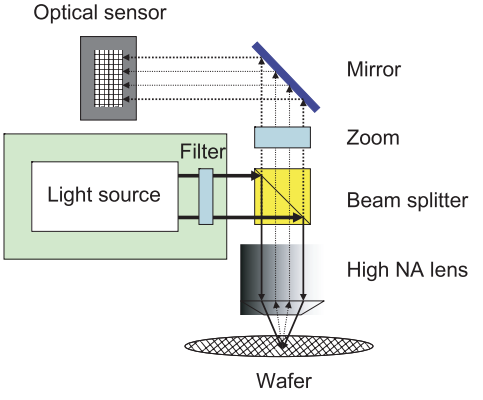


Figure 6.39 Bright-field defect inspection system (H. S. Kim, et al. *Proc. SPIE*).

6.3.10 Track-stepper integration system

All advanced semiconductor fabs use track-stepper integrated systems for photolithography processes. A track system has a wafer loading/unloading deck, prep chambers for dehydration bake and HMDS vapor primer coating, hot plates for the different baking processes, chill plates for cooling the wafer after baking, spin coaters for photoresist coating, and developers for photoresist development. Some systems also have an edge exposure system for optical EBR. A computer-controlled central robot transfers the wafers from one station to another. The in-line process improves throughput and process yield. Figure 6.40 gives a schematic of a track-stepper integrated system and wafer movement.

The track-stepper system improves throughput significantly by reducing wafer handling and increases yield by reducing the time interval between primer and spin coating, and soft bake and exposure. Improved throughput and increased yield are critical to photolithography resolution and photoresist adhesion.

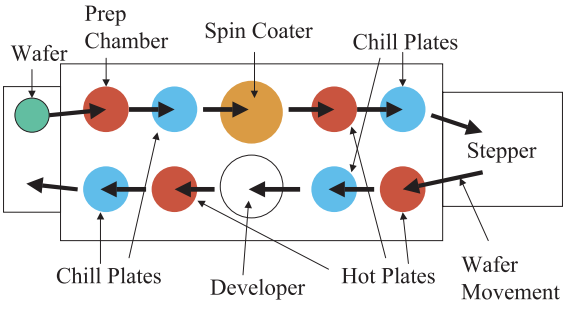


Figure 6.40 Track-stepper integrated system.

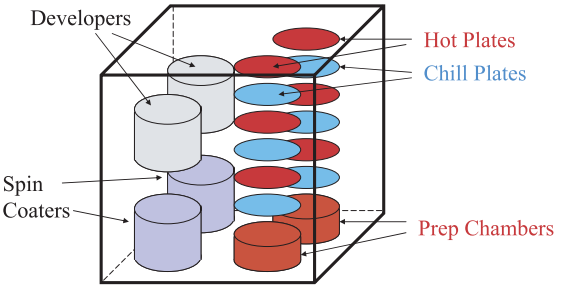


Figure 6.41 Schematic of a stacked track system.

Track systems in advanced semiconductor fabs no longer look like the 2D model in Fig. 6.40. Cleanroom grades have become higher and higher (as has the cost per square foot of the cleanroom) to keep up with shrinking feature sizes. Increasingly expensive cleanroom footage is the major driving force for stacked systems, e.g., hot plates and chill plates that are stacked to make a system more compact. In some advanced systems, coaters and developers are also stacked, saving even more space (see Fig. 6.41).

Some metrology systems, such as overlay and scatterometry CD metrology tools, can be integrated into track-scanner systems, commonly called litho cells, to further improve process throughput.

6.4 Lithographic Technology Trends

Optical photolithography systems have been used for pattern transfer in the semiconductor industry since its inception. In the near future, lithographic technology will change dramatically. This section speculates on possible changes.

6.4.1 Resolution and depth of focus

When a light wave passes through a gap or hole on a mask, it has diffraction, and the projected image will never be as sharp as that on the mask. Using a lens to focus